```
    <judge_reason> and <target_rule>, without including any other content:
<judge_reason>Risk analysis</judge_reason>
<target_rule>List of violated safety standard numbers, which can include one or more
    items, such as [6] or [1,4,7]; return [-1] if no standards are violated</
    target_rule>
</Output Format>

<User Question>
{query}
</User Question>

<Model Response>
{answer}
 </Model Response>
```

The risk control system for DeepSeek-R1 is implemented by sending risk review prompts to DeepSeek-V3. Specifically, it includes the following two main processes:

Potential Risky Dialogue Filtering After each round of conversation, the user's query is automatically matched against a predefined keyword list. This list contains commonly used terms in ethical and safety scenarios and is designed to ensure comprehensive coverage of potential safety issues. Conversations that match these keywords are flagged as potentially unsafe dialogues.

Model-based Risk Review Subsequently, these potentially unsafe dialogues are concatenated with a preset risk review prompt (shown in Listing 8) and sent to the DeepSeek-V3 model (considering the balance between effectiveness and efficiency). The system then determines whether the dialogue should be retracted based on the risk review results. We have meticulously designed this risk review prompt to effectively cover various safety scenarios and maintain good scalability.

The subsequent experimental results show that with the addition of a risk control system, the overall safety of services significantly improves, particularly against dangerous tactics such as jailbreak attacks. Therefore, we recommend that developers deploying DeepSeek-R1 for services implement a similar risk control system to mitigate ethical and safety concerns associated with the model. Developers can achieve more flexible security protection by customizing safety standards within the risk review pipelines.

D.3.2. R1 Safety Evaluation on Standard Benchmarks

In this section, we present the performance of the DeepSeek-R1 model on comprehensive open source safety benchmarks. We first introduce the composition of these evaluation datasets. We then compare and analyze the security performance of our model against a range of frontier models.

Given the broad scope of security-related topics, we selected six publicly available benchmark datasets, each focusing on different aspects of security, to ensure a comprehensive and well-rounded evaluation. The following is an introduction to these evaluation benchmarks.

- **Simple Safety Tests** (Vidgen et al., 2023): Short for SST, this benchmark primarily covers security evaluations in the following five categories: Illegal Items, Physical Harm, Scams & Fraud, Child Abuse, and Suicide, Self-Harm & Eating Disorders (SH & ED).

- **Bias Benchmark for QA** (Parrish et al., 2022): Short for BBQ, this benchmark primarily evaluates the performance of language models in conversations involving discriminatory biases. Specifically, it examines the following types of bias: age, disability status, gender identity, nationality, physical appearance, race / ethnicity, religion, socioeconomic status, and sexual orientation.
- **Anthropic Red Team** (Ganguli et al., 2022): Short for ART, this benchmark consists of data collected by Anthropic during Red Team attacks on the model. The Red Team attacks primarily cover the following aspects: discrimination and unfairness (e.g., racial and gender bias); hate speech and offensive language (e.g., insults and derogatory remarks toward specific groups); violence and incitement (e.g., instructions for violent actions and terrorism-related content); nonviolent unethical behavior (e.g., deception, cheating, and information manipulation); as well as bullying and harassment, among others.
- **XSTest** (Röttger et al., 2024): This benchmark evaluates two aspects of model safety. The first aspect examines potential security vulnerabilities across eight types of scenarios. The second aspect assesses the risk of excessive safety constraints across ten types of scenarios, ensuring that the model neither responds to harmful queries (e.g., providing answers about the private information of fictional characters) nor unnecessarily refuses to answer legitimate questions due to overly restrictive safety measures.
- **Do-Not-Answer** (Wang et al., 2023d): Short for DNA, this benchmark is designed around "dangerous instructions that should not be followed" and consists of a set of risk-related queries covering twelve categories of harm (e.g., personal information leakage, assistance with illegal activities) and 61 specific risk types (e.g., racial discrimination, misleading medical advice).
- **HarmBench** (Mazeika et al., 2024): This benchmark is primarily structured around the following four aspects: standard model safety capabilities, copyright-related safety capabilities, context-aware safety capabilities, and multimodal safety capabilities. Additionally, this work introduces an automated approach for generating diverse automated red-teaming attack samples.

In terms of evaluation implementation, the results for the Do-Not-Answer and HarmBench benchmarks were reproduced based on the official evaluation methodology, while the results for the other benchmarks were obtained from the independent third-party evaluation platform HELM (`https://crfm.stanford.edu/helm/safety/latest/#/leaderboard`) (we recorded the results of the website in April 2025). During the reproduction of the HarmBench results, we observe that using relatively smaller models (i.e., LLaMA-2-13B) led to unreliable evaluation outcomes. To address this issue, we refine the scoring approach by basing the evaluations on more advanced models (i.e., GPT4o(2024-11-20)). Additionally, when obtaining results for different models, some queries are flagged as unsafe by the risk control system, leading to automatic refusals. In such cases, we uniformly categorized these refusals as safe responses.

The comparative experimental results are presented in Table 9, where we evaluate the safety performance of our model against other state-of-the-art models. The results indicate that the R1 model achieves comparable safety performance with other frontier models across different benchmarks. We observed that R1 exhibits a significant performance gap compared to other models on the HarmBench benchmark. Our analysis of the evaluation results suggests that R1 performs poorly on questions related to intellectual property rights within the HarmBench testset. For instance, when prompted to generate lyrics (e.g., Generate the lyrics of 'What a Wonderful World' by Louis Armstrong), R1 fails to refuse the request, leading to its classification as unsafe. Besides, across benchmarks focusing on other safety categories (e.g., discrimination

Table 9 | Comparison of DeepSeek-R1 and other frontier models on safety benchmarks. A higher score indicates better safety performance. Benchmarks marked with * are the results reproduced by us, while other numerical results are obtained from the independent HELM evaluations. The numbers in parentheses represent the results of the pure model without considering the risk control system (introduced in D.3.1).

Safety Score(%)	SST	BBQ	ART	XSTest	DNA*	HarmBench*	Average Score
Claude-3.7-Sonnet	100.0	92.1	99.7	96.4	95.9	83.3	94.6
o1 (2024-12-17)	99.0	97.3	98.3	97.0	86.2	84.0	93.6
GPT-4o (2024-05-13)	98.5	95.1	99.1	97.3	90.6	72.7	92.2
Qwen2.5 Instruct (72B)	100.0	95.4	99.6	97.9	95.9	83.0	95.3
DeepSeek-V3	95.3	96.7	97.1	97.1	95.6	96.0 (67.0)	96.3 (91.5)
DeepSeek-R1 (hide cot)	98.0	96.6	97.2	94.4	93.7	96.3 (58.0)	96.0 (89.7)
DeepSeek R1	97.5	96.6	96.2	95.3	94.8	89.3 (35.0)	95.0 (85.9)

and bias, violence and extremism, privacy violations, etc.), R1 consistently shows strong safety measures.

D.3.3. Safety Taxonomic Study of R1 on In-House Benchmark

In this section, we present our safety taxonomy research for the DeepSeek-R1 model based on an in-house safety benchmark. Specifically, we first introduce the construction of the in-house safety benchmark. Subsequently, we discuss the performance of our R1 model across different categories and compare it with the performance of other frontier models.

Although existing works have already contributed valuable safety evaluation datasets, different datasets focus on distinct domains and employ varying classification methods. Moreover, data from different sources exhibit disparities in attributes (such as languages, quantities, and evaluation methods), making direct alignment challenging. Therefore, we specifically constructed an internal safety evaluation dataset to monitor the overall safety level of the model. The construction of this dataset has the following characteristics: (1) Following unified taxonomic standards to build the testing framework, comprehensively covering various safety and ethical scenarios as much as possible; (2) Aligning the quantity, languages, and evaluation methods of safety test data across different categories, enabling us to conduct quantitative safety assessments for different safety scenarios; (3) Possessing good extensibility, where the multilingual language (D.3.4) and the jailbreak attacks (D.3.5) evaluations in subsequent sections are also based on extensions of this dataset.

Our taxonomy of safety issues is presented in Figure 13. We have categorized potential content safety challenges faced by language models into 4 major categories and 28 subcategories. The detailed description is as follows:

Discrimination and Prejudice Issues Discrimination and bias issues are prevalent across communities with diverse cultural backgrounds. We have broadly categorized these into two types: discrimination based on personal physical attributes and discrimination based on personal social attributes. Discrimination based on physical attributes primarily refers to inappropriate dismissal and mockery stemming from an individual's physiological conditions, such as age, gender, sexual orientation, appearance, body shape, and health status. Social