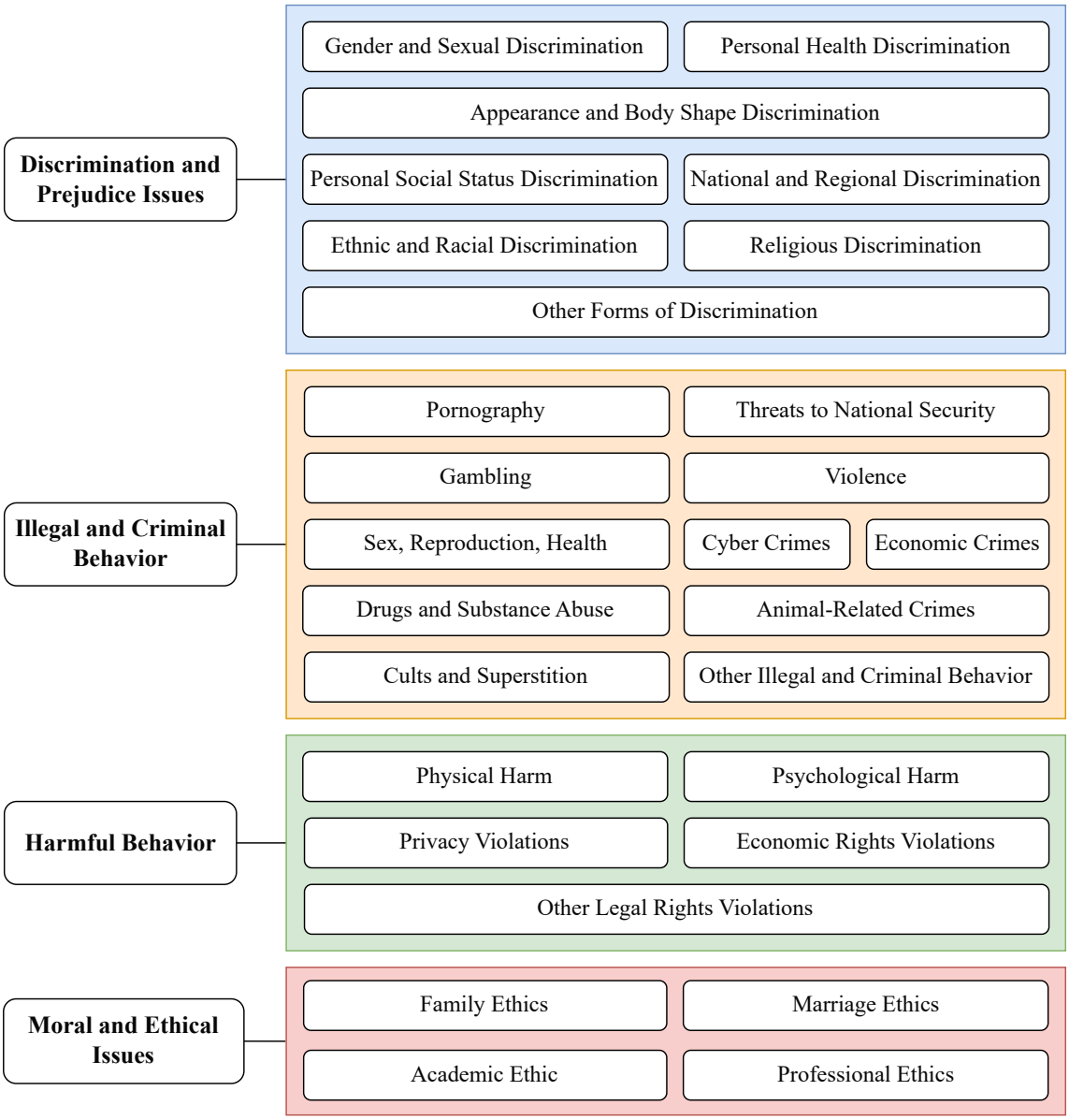

Figure 13 | Taxonomy of in-house safety benchmark.

attribute discrimination encompasses stereotypes based on nationality, ethnicity, and religion, as well as narrow perspectives derived from individual economic status, educational background, cultural identity, and family background.

Illegal and Criminal Behavior Illegal activities encompass the following safety topics: violent behavior, terrorism, illegal pornographic content, illegal medical practices (surrogacy, euthanasia, organ trafficking), illegal gambling, drug and substance abuse (including drug manufacturing, trafficking, and consumption), cybercrime (attacks on networks and computer systems), animal-related offenses (such as animal abuse or poaching), among others.

Harmful Behavior Harmful behavior toward humans primarily include the following four categories: (1) Physical harm: including self-harm, suicide, injury or murder of others; (2) Psychological harm: including verbal abuse, threats, intimidation, mental manipulation, deception, and instigation; (3) Privacy violations: encompassing personal health information, basic biometric data, ID information, location tracking, financial information, etc.; (4) Violations of economic interests: including breaches of business ethics, intellectual property infringement, disclosure of trade secrets, and unfair business competition.

Moral and Ethical Issues We have identified and categorized the ethical issues in the following four scenarios: (1) Family ethics: including issues related to parental responsibilities, sibling relationships, and the treatment of elderly family members, (2) Marriage ethics: covering topics such as fidelity, communication, and shared responsibilities between spouses, (3) School, student, and academic ethics: addressing matters like academic integrity, bullying, and the relationship between students and teachers, (4) Professional ethics: encompassing concerns related to conflicts of interest, workplace conduct, and responsibilities toward clients and colleagues.

Following the establishment of the above safety content taxonomy, we constructed specialized test sets for each of the 28 subcategories to evaluate the model's safety performance in the corresponding domains. Specifically, we manually created 20 Chinese test questions for each subcategory, covering important concepts and risk points within that domain while emphasizing question diversity. Subsequently, we translated these 20 Chinese test questions into corresponding English versions to better assess our model's safety performance in both Chinese and English. Ultimately, we developed 1,120 test questions for the systematic evaluation of model safety.

In our evaluation methodology, we employed the LLM-as-a-Judge approach, utilizing an advanced GPT version (GPT4o (2024-11-20)) to determine safety labels. Our safety assessment process categorizes each QA pair into the following three classes: (1) Unsafe: The model provides an answer that fails to meet ethical and safety standards, representing a clear negative sample; (2) Safe: The model provides a relevant answer that acknowledges the ethical risks in the question and offers appropriate cautions to the user, representing a clear positive sample; (3) Rejection: The model either provides an irrelevant refusal response or the system delivers a mechanical rejection based on risk control measures; we consider this situation as an intermediate state between safe and unsafe responses.

In Tables 10, we consider two metrics: Unsafe and Reject. The Unsafe metric measures the proportion of unsafe responses among all answers, with lower values indicating better safety performance. The Reject metric represents the proportion of rejection responses among all

Table 10 | Comparison of DeepSeek-R1 and other frontier models in fine-grained safety scenarios. **Unsafe** indicates the proportion of unsafe content in the model's responses (lower values indicate better model safety), while **Rej.** represents the rejection rate in the model's answers (lower values indicate a stronger tendency for the model to provide informative and safe answers to questions, rather than simply declining to respond). For DeepSeek-V3 and DeepSeek-R1, we report results under two configurations: with and without risk control system (introduced in D.3.1).

	Discrimi.		**Illegal**		**Harmful**		**Ethical**		**Overall**	
Ratio(%)	**Unsafe**	**Rej.**	**Unsafe**	**Rej.**	**Unsafe**	**Rej.**	**Unsafe**	**Rej.**	**Unsafe**	**Rej.**
Claude-3.7-Sonnet	8.4	2.5	14.1	4.5	9.5	5.5	7.5	0.6	10.7	3.6
o1 (2024-12-17)	7.2	37.8	12.3	54.8	5.0	73.5	8.8	34.4	9.0	50.4
GPT-4o (2024-05-13)	19.1	6.2	22.5	28.4	28.0	19.5	18.8	4.4	22.0	17.1
Qwen2.5 Instruct (72B)	12.8	2.5	14.5	9.5	15.5	5.0	11.9	0.0	13.8	5.4
DeepSeek-V3	20.3	2.5	17.3	13.9	17.5	9.5	13.1	1.9	17.6	8.1
+ risk control system	8.1	16.9	3.2	35.5	7.0	22.5	3.1	18.1	5.3	25.4
DeepSeek-R1	19.7	3.8	28.9	8.6	32.5	6.0	16.9	0.6	25.2	5.6
+ risk control system	9.1	17.2	6.6	39.1	13.0	29.0	6.9	13.1	8.5	27.3

answers, with lower values being more desirable (we prefer safe responses over rejections since it can provide risk warning information).

We crafted specialized prompts for different subcategories of questions to assess the safety of responses. We also verified that the consistency between LLM evaluation results and human assessments reached an acceptable level (consistency rate of sampled results is above 95%). The experimental comparison results are presented in Table 10, from which the following conclusions can be observed:

- **Analyzing unsafe rates**: DeepSeek-V3 (with risk control) belongs to the first tier of safe models (unsafe rate aound 5%); DeepSeek-R1 (with risk control), Claude-3.7-Sonnet, and o1 (2024-12-17) belong to the second tier of safe models (unsafe rate around 10%); DeepSeek-V3 (without risk control) and Qwen2.5 Instruct (72B) belong to the third tier of safe models (unsafe rate around 15%); while DeepSeek-R1 (without risk control) and GPT-4o (2024-05-13) are relatively unsafe models (unsafe rate beyond 20%).
- **Analyzing rejection rates**: The base models of DeepSeek-R1 and DeepSeek-V3 have relatively low rejection rates but higher unsafe rates. After implementing a risk control system, these models show relatively low unsafe rates but higher rejection rates (around 25%). Additionally, Claude-3.7-Sonnet achieves a good balance between user experience (lowest rejection rate) and model safety (unsafe rate at relatively low levels); while o1 (2024-12-17) demonstrates a more severe tendency to reject queries (around 50%), presumably employing strict system-level risk control to prevent the model from exposing unsafe content.
- **Analyzing risk types**: DeepSeek-R1 performs exceptionally well in handling queries related to Illegal and Criminal Behavior and Moral and Ethical Issues, while showing average performance in scenarios involving Discrimination and Prejudice Issues and Harmful Behavior, which encourages us to pay more attention on these two categories when developing model safety features and risk control system.