① Complex Jailbreak Judge Scenario

Vanilla Harmful Prompts
- **Broad range of risk safety categories:** 14 more fine-grained safety categories

Multilingual Harmful Prompts
- **Use High-resource**: en, zh, it, vi
- **Medium-resource:** ar, ko, th
- **Low-resource:** bn, sw, jv

Synthetic Vanilla Prompts
- **Use GPT-4 to rewrite and expand harmful prompts:** extend the same hazard taxonomy categories

In-the-wild Prompts
- **Diverse prompts from real-world platform**: Reddit, Discord, websites
- **Complex intentions:** Combine templates with malicious prompts

Synthetic Adversarial Prompts
- **Use latest jailbreak attacks to bypass LLM safety alignments:** 9 jailbreak attacks to modify harmful prompts

Deceptive Harmful Prompts
- **Disguise harmful intentions:** Role-playing, scenario assumptions, long-context, adaptive harmful prompts

Diverse LLM Responses
- **Closed-source LLMs:** GPT-4, GPT-3.5
- **Open-source LLMs:** Llama-family, Qwen-family, Mistral-family, Vicuna-family
- **Defense-enhanced LLMs:** 6 defense methods

Annotation Process
- **Human annotator training**
- **Manual labeling**
- **GPT-4 labeling**
- **Cross-comparison and multi-person voting for final label l**

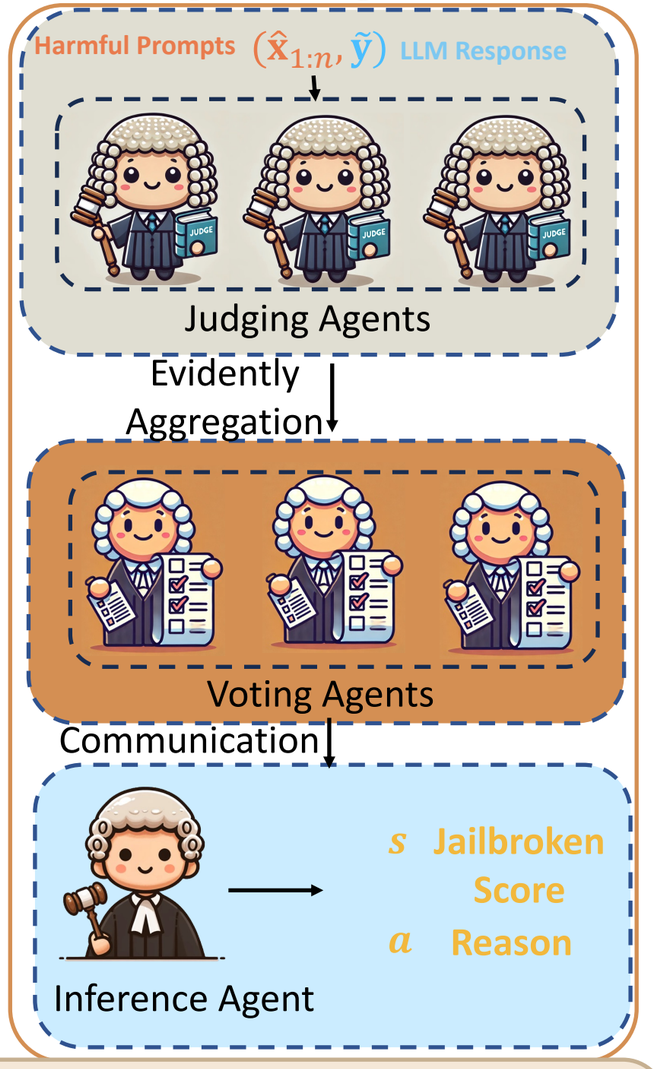
② Open Jailbreak Judge Data

JAILJUDGETRAIN $(\hat{\mathbf{x}}_{1:n}, \tilde{\mathbf{y}}, s, a)$
- **Contains a test split of 3.5w+ labeled prompt-response pairs**
- **Multi-agent judge enhanced explainability**

JAILJUDTEST ID $(\hat{\mathbf{x}}_{1:n}, \tilde{\mathbf{y}}, l)$
- **Contains a test split of 4.5K+ labeled prompt-response pairs**
- **Derived from JAILJUDGETRAIN:** Split from JAILJUDGETRAIN, except multilingual harmful prompts
- **In-distribution (ID) set**

JAILJUDTEST OOD $(\hat{\mathbf{x}}_{1:n}, \tilde{\mathbf{y}}, l)$
- **Includes a 6K+ labeled set of multi-language scenarios**
- **Features three different-resource languages**
- **Out-of-distribution (OOD) set**

③ Multi-agent Judge Framework



JAILJUDGE

Benchmark of Jailbreak Judge on LLMs

35k+ Training Data
4.5k+ Test ID Data
6k+ Test OOD Data
$100,000+ Total Cost

13 Safety Categories
6 Complex scenarios
10 Languages
3 Types LLM Response
15 Jailbreak Judge Baselines
4 Jailbreak Attack Baselines
7 Jailbreak Defense Baselines
1 Open JAILJUDGE Guard

Key insights on jailbreak judge on LLM
1. Jailbreak judge lacks generalization on complex scenario
2. Jailbreak Judge has a bias on low-resources language
3. Explainability can improve the jailbreak judge
4. Jailbreak Judge can boost the attack on LLMs