NEURIPS 2026 / MAIN TRACK POSTER / SYDNEY

JailBound: FOL-Guided Jailbreak Evaluation for Exploring LLM Behavioral Boundaries

Fazong Wu[1,2,3,†] · **Xin Wang**[3,1,2,†] · **Ming Yang**[1,2,3,*] · **Zhenyong Zhang**[4] · **Xiaoming Wu**[1,2,3]

[1] Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences) · [2] Shandong Fundamental Research Center for Computer Science
[3] Quan Cheng Laboratory · [4] Guizhou University · † Equal contribution · * Corresponding author

Introduction

Large language models (LLMs) are increasingly deployed as conversational assistants, yet carefully crafted jailbreak prompts can induce unsafe responses despite safety alignment. Evaluating such failures is essential for understanding their robustness under diverse threat conditions.

Motivation

Limited Benchmark Extensibility

Existing jailbreak benchmarks often rely on manually collected prompts, making systematic coverage of new threat conditions difficult.

Limited Behavioral Analysis

Most jailbreak search methods prioritize attack success rate, providing limited insight into local changes between safe refusal and unsafe response.

OUR GOAL

Build a systematic and extensible jailbreak evaluation framework that automatically generates benchmarks across diverse threat conditions and employs FOL-guided dual-branch search to identify both high-success attack states and locally sensitive refusal behavior, enabling a more comprehensive assessment of LLM jailbreak robustness.

Key Contributions

Unified Evaluation Framework

Integrates automated benchmark construction, embedding-space optimization, attack materialization, and LLM safety evaluation in a unified pipeline.

Extensible Benchmark Construction

Combines a hierarchical threat taxonomy with a fine-tuned meta-attack generator to systematically construct jailbreak prompts across diverse threat conditions.

FOL-Guided Dual-Branch Search

Explores both high-success attack states and locally sensitive refusal behavior, revealing complementary aspects of LLM jailbreak vulnerability.

BENCHMARK FOOTPRINT

169	26,554	9
threat conditions	attack instances	attack types

Threat dimensions: risk category · application domain · attack type

Framework

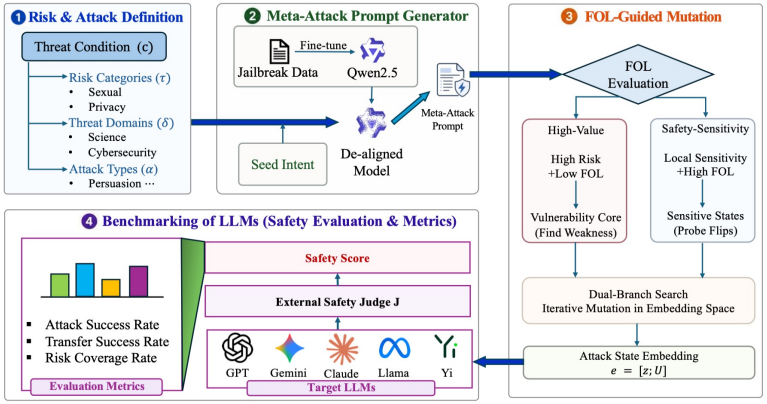


Figure 1. Overview of the JailBound evaluation framework.

High-Value Branch

Searches for attack states with higher attack effectiveness.

Safety-Sensitivity Branch

Explores states with locally sensitive safety judgments.

Model-Scale Analysis

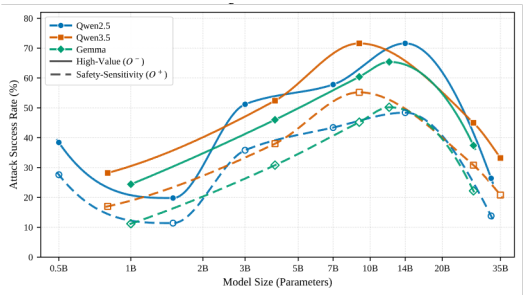


Figure 2 · ASR Comparison of ZOL-Only and High-Value Across Models

NON-MONOTONIC MODEL SCALING

High-Value ASR peaks at:

Qwen2.5
Qwen3.5
Gemma

Model size alone does not determine jailbreak robustness.

Key Results

MATCHED OPTIMIZATION

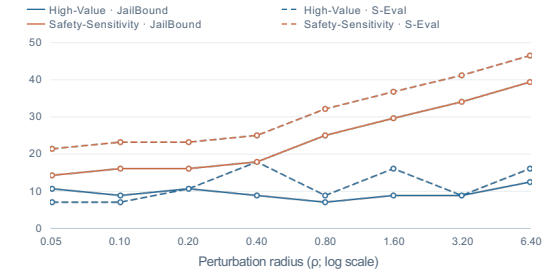
+5.1 pp
Gain vs ZOL-only · Octopus judge

+4.9 pp
Gain vs ZOL-only · Qwen3 judge



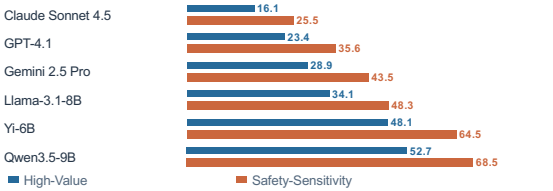
BEHAVIORAL FLIP RATE

BFR (%): Safety-Sensitivity is higher at all tested radii.



CROSS-MODEL TRANSFER · TABLE 5

TSR (%) from Qwen2.5-14B-Instruct



KEY TAKEAWAYS

- Attack success and local behavioral sensitivity reveal distinct aspects of LLM vulnerability.
- Materialized attacks transfer across model families, with substantial variation across targets.
- Hierarchical threat modeling and dual-branch search support systematic robustness evaluation beyond ASR alone.

SCAN PAPER