The 40th Annual AAAI Conference on Artificial Intelligence
JANUARY 20 – JANUARY 27, 2026 | SINGAPORE

A Collaborative Multi-Agent Framework for Jailbreaking with RL-Based Dynamic Prompting

Azka Ikramullah[1], Kyunghyun Lee[2], Abdul Majeed[2], Seong Oun Hwang[2]*

[1] Department of IT Convergence Engineering, Gachon University, South Korea
[2] Department of Computer Engineering, Gachon University, South Korea
*Corresponding Author

Research Problem

Problem:
LLMs perform well on shallow reasoning but break down as reasoning depth increases.

Why It Matters:
Real-world tasks require multi-step reasoning; chain-of-thought often improves appearance, not true depth.

Gap:
Current benchmarks do not explicitly measure or control reasoning depth, hiding failure points.

Research Questions

RQ1: Can dense, multi-aspect feedback stabilize reinforcement learning for jailbreak discovery and improve attack success per query?
RQ2: Can multi-aspect evaluation produce auditable signals that explain why jailbreak attacks succeed beyond binary pass/fail outcomes?
RQ3: How does model performance change as required reasoning depth increases, and where does reasoning collapse occur?

Methodology

- **Formulation**: Jailbreak discovery as an off-policy RL problem with fixed query budgets
- **Agent:** Soft Actor–Critic (SAC) with hybrid actions
  - **Discrete operator:** Translate, Encode, Role-Play, Decompose, Euphemize
  - **Continuous sliders:** prompt length, temperature, persona
- **Rewriter:** Yi-9B generates intent-preserving, stealthy paraphrases

Key Results

- **Judge:** LLaMA-3-Instruct (T=0) outputs 5 ratings
  - Success · Stealth · Novelty · Efficiency · Impact
- **Reward:** Min–max normalized ratings, curriculum-weighted
  - Penalties for excess queries, near-duplicates (MinHash), and over-length prompts
- **Stability:** Stratified replay, early-exit on low promise or success, de-duplication
- **Data:** Train on SORRY-Bench priors; test on JailbreakBench harmful OOD
- **Evaluation:** Attack Success Rate (ASR) under matched query budgets

Evaluation framework

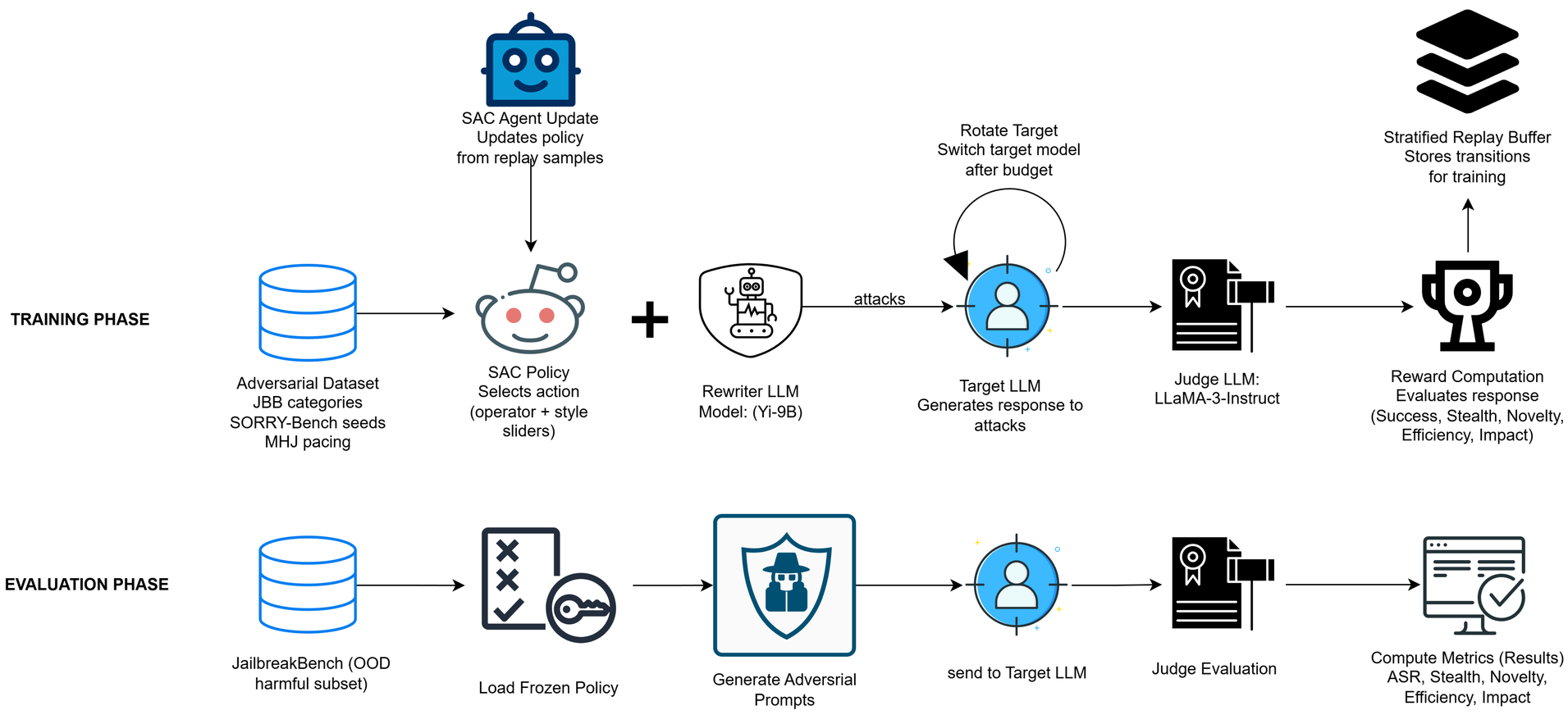

Figure 1: RL-based rating-driven triad for jailbreak evaluation. A SAC attacker selects an operator family and continuous style sliders; a rewriter LLM generates an intent-preserving, stealthy paraphrase; the target LLM produces a response; a judge LLM assigns five aspect ratings (Success, Stealth, Novelty, Efficiency, Impact), which are curriculum-weighted into a dense reward; transitions are stored in a stratified replay buffer to stabilize off-policy learning.

Attack Success Rate (ASR)

Method	Dataset	Qwen-2.5-7B	Mistral-7B	LLaMA-3
PAIR (NeurIPS ’23)	AdvBench	NA	0.684	NA
AutoDAN (ICLR ’24)	AdvBench	NA	0.774	NA
GPTFuzzer (NeurIPS ’23)	AdvBench	0.14	0.786	0.24
GCG (ICLR ’24)	AdvBench	NA	0.622	NA
RLBreaker (NeurIPS ’24)	AdvBench	NA	0.748	0.724
RL-JACK (NeurIPS ’24)	AdvBench	0.91	NA	0.45
xJailbreak (arXiv ’25)	AdvBench	0.8	NA	0.63
PASS (arXiv ’25)	AdvBench	0.85	NA	NA
Ours (SAC–Triad)	**JBB–OOD**	**0.955**	**0.943**	**0.724**

Table 1. Attack Success Rate (ASR) across target models.
Baseline results are reported as published (AdvBench). Our method is evaluated on the harmful out-of-distribution split of JailbreakBench (JBB–OOD).

Query Efficiency

Method	Dataset	ASR ↑	Queries / Success (Q/S ↓)
RLBreaker (NeurIPS ’24)	AdvBench	0.748	7
Ours (SAC–Triad)	**JBB–OOD**	**0.943**	**2.41**

Table 2. Query efficiency comparison.
Queries per Success (Q/S) measures the average number of target queries required to achieve one successful jailbreak. Lower values indicate greater efficiency.

Conclusion

- Dense, multi-aspect judge feedback stabilizes RL-based jailbreak discovery.
- Off-policy SAC improves attack success per query while avoiding prompt template collapse.
- Rating-driven rewards yield interpretable signals beyond binary pass/fail evaluation.

Implications

- Enables query-efficient, scalable safety evaluation under realistic cost constraints.
- Provides auditable dimensions (success, stealth, novelty, efficiency, impact) for analyzing failure modes.
- Supports systematic benchmarking of agentic and multi-turn LLM systems.

Future Work

- Full ablations of reward shaping, replay strategy, and SAC vs. PPO.
- Cross-judge calibration and human-in-the-loop validation.
- Extension to multi-agent, tool-use, and multimodal (e.g., text-to-image) safety evaluation.