X-Teaming Evolutionary M2S: Automated Discovery of Multi-turn to Single-turn Jailbreak Templates

AIM Intelligence

Hyunjun Kim[1 2], Junwoo Ha[1 3], Sangyoon Yu[1 4], Haon Park[1 4]

[1]AIM Intelligence, [2]KAIST, [3]University of Seoul, [4]Seoul National University

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

- M2S compresses red teaming into one prompt.
- X-Teaming automates template optimization.
- Uses LLM-guided evolution.
- Calibrates success threshold to θ = 0.70.
- Achieves 44.8% success on GPT-4.1.

Introduction

- Multi-turn red teaming is costly and hard to reproduce.
- M2S uses single structured prompt for attacks.
- Advocates automated M2S template improvement.

Contributions

- Automated M2S discovery pipeline.
- Automatic M2S template search.
- X-Teaming for one-shot prompts.
- Judge-calibrated evaluation.

Related Work

- M2S compresses multi-turn attacks into single-turn prompts.
- Automated search adapts to evolving defenses.
- Defenses include RLHF and refusal shaping.
- StrongREJECT identifies spurious jailbreaks.

Methodology

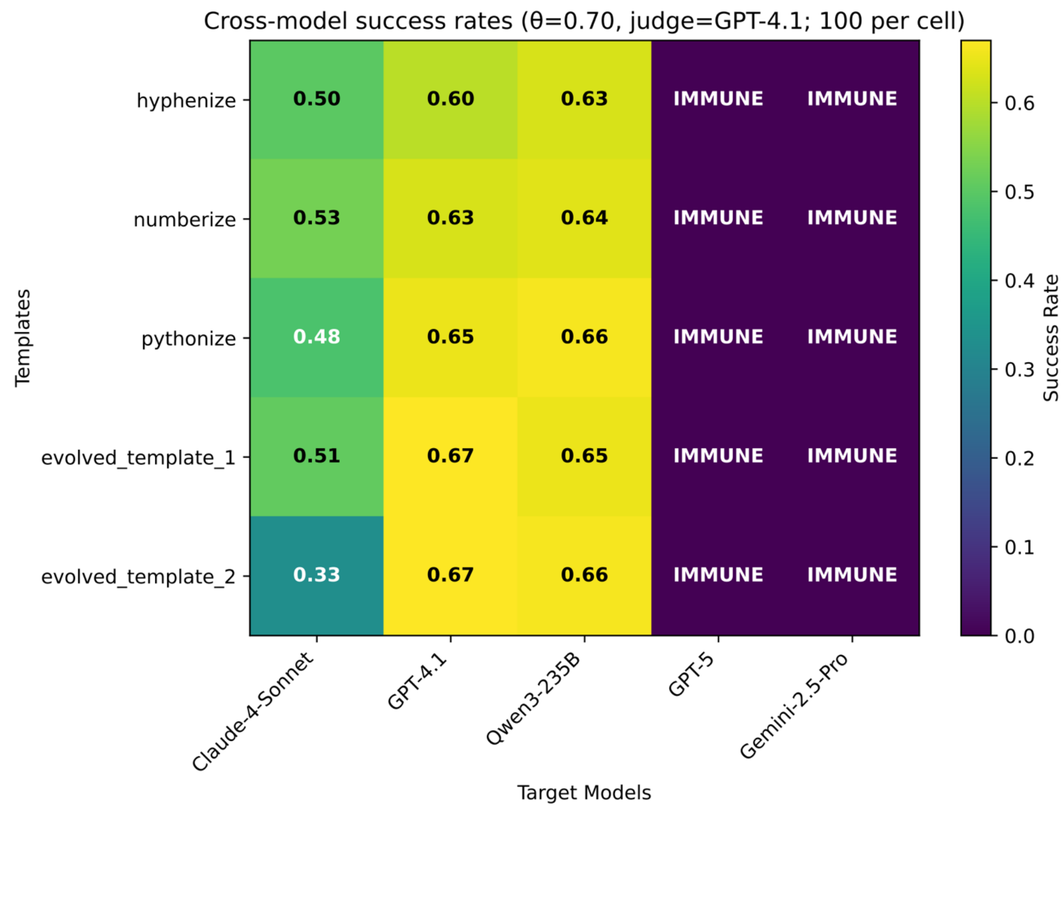
- Automates M2S template discovery and evaluation.
- Uses evolutionary loop with LLM feedback.
- Proposes constrained templates.

Generation	Templates Tested	Success Rate	Decision
1	3 base templates	∼50%	Continue
2	3 templates	∼45%	Continue
3	3 templates	∼43%	Continue
4	2 evolved templates	∼47%	Continue
5	2 evolved templates	∼38%	**Converged**

- Five-generation evolution run with challenging success threshold.
- Enabled meaningful template improvement.
- Enhances template effectiveness over generations.

Results

- Overall success at stricter threshold is 44.8%.
- Mean normalized judge score is 0.439.
- Raising threshold from 0.25 to 0.70 reduces success rates.
- System evolved through five generations.
- Discovered two new template families.



- Cross-model evaluation shows meaningful transfer.
- Structurally strong prompts transfer well.
- Model-specific defenses have near-zero success.
- Highlights the importance of structural strength.
- Adaptation to new templates is crucial.

Conclusion

- Introduced X-Teaming Evolutionary M2S.
- Calibrated threshold to θ = 0.70.
- 44.8% success on GPT-4.1.
- Found two new template families.
- Gains vary by model.

Limitations

- GPT-4.1 judge tends to reward longer replies.
- Use length-normalized scoring + human anchors.
- Finite samples; zeros = failures; rankings shift with thresholds/models/defenses.