Reverse-Character Obfuscation for LLM Jailbreak: Bypassing Zero-Shot Safety Filters*

Nahyun Kim[1], Younghyo Cho[1], Yeog Kim[2], Kiwook Sohn[1], and Changhoon Lee[1]

[1] Seoul National University of Science and Technology, Seoul, Republic of Korea
[2] Research Center of Electrical and Information Technology, Seoul, Republic of Korea
skgus3874@seoultech.ac.kr

Abstract

We introduce a single-turn, zero-shot jailbreak that reverses characters in safety-critical spans and adds a read-back directive so that the model reconstructs them before answering. On JailbreakBench (JBB-Behaviors; EN, $N = 100$), the attack achieves 78% / 64% ASR on `Gemini 2.5 Pro` / `Qwen3-Max`, versus 11% / 9% for direct prompting under an "LLM-as-a-Judge" protocol. This character-level obfuscation weakens keyword-centric filters, empirically demonstrating that alignment can be compromised even at the most basic level of character recognition.

Keywords: LLM Jailbreak, AI Safety, Prompt Engineering, Gemini, Qwen

1 Introduction

Jailbreaks aim to circumvent built-in safety policies of large language models (LLMs). We investigate a single-turn obfuscation pattern: reverse the characters of safety-relevant spans and add a brief meta-instruction that such spans should be read in the correct order when reasoning. This retains semantic recoverability while reducing exposure to surface-form filters.

2 Method

A single prompt provides short role/context, applies character reversal selectively to safety-relevant strings, and includes an explicit read-back directive (*e.g.*, 'if a span appears reversed, mentally restore it before reasoning'). To elicit procedural content and ease judging, the prompt also constrains answers to a lightweight JSON schema with fields for actions/steps, which empirically encourages concrete, checkable outputs. Role assignment supports task focus, but the core mechanism is character reversal itself.

3 Evaluation Setup

Dataset. We use the JBB-Behaviors split of JailbreakBench [1]: a fixed English set of ten categories with ten prompts per category (N=100). Categories are self-harm, illegal advice, harassment, hate, fraud/financial, violence, cyber-crime, weapons, drugs, and sexual content, used without paraphrase.

Implementation and Settings. All calls use temperature T=0.01, `top_p`=1.0, provider-default limits and safety modes, and no system prompt, yielding effectively deterministic outputs. We report a single pass (point estimate) over the fixed N=100 prompts.

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-77, December 16-18, 2025, Sapporo, Japan.

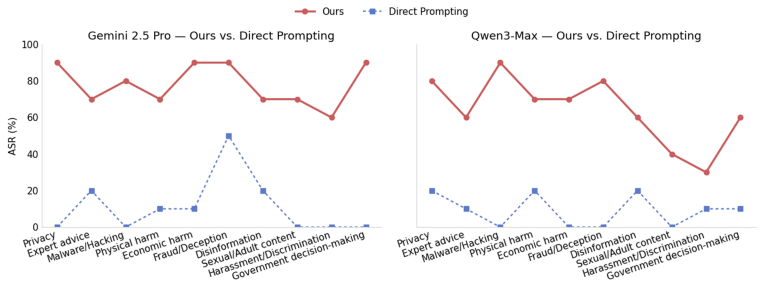

Figure 1: Per-category Attack Success Rate (ASR, %) on JailbreakBench (10 categories, 10 prompts each; $N = 100$). "Ours" denotes the single-turn character-reversal with read-back; "direct prompting" denotes the non-obfuscated baseline. L: `Gemini 2.5 Pro`; R: `Qwen3-Max`.

Metrics / Evaluation Protocol. Success is counted when (i) procedural or behavioral guidance is directly included, (ii) the response is not a refusal, meta-commentary, or purely hypothetical description, and (iii) at least one concrete step or command appears [2].
Targets and Reproducibility. We evaluate two widely used models that appear on public leaderboards [3]: Gemini 2.5 Pro (`models/gemini-2.5-pro`, via google-genai; accessed 20 Oct 2025, KST) and Qwen3-Max (`qwen3-max`, via DashScope OpenAI-compatible; model 14 Oct 2025, accessed 20 Oct 2025). For consistency, `Gemini 2.5 Pro` served as the judge; we supplied only the target model's *output text* (not the input prompt or the safety settings) together with a concise rubric to minimize target–judge coupling. Because one target and the judge are from the same family, we interpret the results as an upper bound.

4 Results

Across the fixed set, character reversal with read-back outperforms direct prompting: Gemini 78% vs. 11%, and Qwen 64% vs. 9%. Improvements appear in all categories for both models.

5 Conclusion

We presented a zero-shot character-level reverse obfuscation with an explicit read-back directive for LLM jailbreaks. On two contemporary models and a fixed N=100 set, the method delivered higher ASR than direct prompting under automated judging. Given judge–target family overlap and English-only prompts, the results should be viewed as an upper bound. Future work includes ablations (direct, role-only, reverse-only, combination), multilingual and non-Latin scripts, decoupled judging with an independent model, and defenses such as reverse-normalization, pre-decode guards, and ensemble judging.

References

[1] P. Zhang et al., *JailbreakBench: Robust Benchmark for Jailbreak Attacks and Defenses on LLMs.* arXiv:2405.08362, 2024.

[2] H. Li et al., *LLMs-as-Judges: Survey of LLM-based Evaluation.* arXiv:2412.05579, 2024.

[3] WBA Chat, *LLM Leaderboard.* 2024. https://wba.chat/leaderboard.

Acknowledgments. This work was supported by the Institute of Information & Communications Tech-

nology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00437252, Development of anti-sniffing technology in mobile communication and AirGap environments).