Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO (Professor))		
Student ID	2411432	Submission date	2026 / 7 / 22
Name	YU SHUYI		
Thesis title	Random Character-Level Perturbations Amplify LLM Jailbreak Attacks		

Abstract

Modern large language models (LLMs) rely on subword tokenization, which is vulnerable to small character-level perturbations that re-partition text into unfamiliar subwords and degrade performance. We show that this vulnerability also undermines safety: minimal word-internal perturbations increase the success rates of both simple and complex jailbreak attacks across multiple LLMs. Mechanistically, such perturbations cause over-fragmented tokenization and token-representation drift, substantially shifting the semantic representations of words. Through word-level semantic recovery and sentence-level spelling-error correction, we find that models cannot reliably reconstruct the original meaning, and layer-wise probe classifiers fail to detect the harmful intent of perturbed prompts. We further observe that perturbations can occasionally reduce attack success by inducing off-topic or incoherent responses. Together, our findings show that tokenization-induced vulnerabilities compromise safety mechanisms, underscoring the need for mitigation strategies.