Soft Prompt Attack Success Rates in Quantized vs Full-Precision Open-Source LLMs

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the success rate of soft prompt attacks on safety alignment compare between quantized and full-precision open-source LLMs across standard harmlessness benchmarks. Current research in adversarial robustness of LLMs focuses on discrete input manipulations in the natural language space, which can be directly transferred to closed-source models. However, this approach neglects the steady progression of open-source models. 4 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Soft Prompt Threats: Attacking Safety Alignment and Unlearning in Open-Source LLMs through the Embedding Space. Research question: How does the success rate of soft prompt attacks on safety alignment compare between quantized and full-precision open-source LLMs across standard harmlessness benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

4 papers retrieved. 4 claims extracted; 0 independently verified. Quality review score: 3.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Embedding space attacks increase the ROUGE score considerably to at least 0.49 for individual and universal attacks with	×	0.09
Universal embedding attacks generalize to unseen samples on the TOFU benchmark, achieving a ROUGE score of up to 0.52 wi	×	0.02
We created a dataset of sentence pairs from the first Harry Potter book, splitting it into 527 training pairs and 100 te	×	0.02
We optimized a universal embedding space that was attached to the instruction to improve the prediction of the target se	×	0.07

References

- http://arxiv.org/abs/2412.03235v2
- http://arxiv.org/abs/2312.12321v2
- http://arxiv.org/abs/2402.09063v2