

# Differentially Private Adapters for Safe and Utility-Preserving NLP Alignment

Assignee Research

June 6, 2026

## Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Can differentially private adapter methods maintain alignment safety scores on ToxicChat while preserving utility on standard NLP benchmarks. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Differentially Private Fine-tuning of Language Models. Research question: Can differentially private adapter methods maintain alignment safety scores on ToxicChat while preserving utility on standard NLP benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

## 3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.5/10.

## 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.

## References

- <http://arxiv.org/abs/2305.06212v3>
- <http://arxiv.org/abs/2507.05660v3>
- <http://arxiv.org/abs/2110.06500v2>