PAPER DIGEST

CERTIFICATE OF RECOGNITION

MOST INFLUENTIAL NEURIPS 2023 PAPER

This is to certify that the paper

Jailbroken: How Does LLM Safety Training Fail?

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt

presented at NEURIPS, 2023

is recognized among the Most Influential NEURIPS 2023 Papers





Paper Digest
New York, New York

Edition 2026-03
Issued 2026-03 - Impact Factor 8

Scan to verify



Verify this certificate at paperdigest.org

Certificate ID f4414a5e5e38ddf6 - Ranking constructed from citations in research papers and granted patents.