



PAPER DIGEST

CERTIFICATE OF RECOGNITION

MOST INFLUENTIAL NEURIPS 2024 PAPER

This is to certify that the paper

Refusal in Language Models Is Mediated By A Single Direction

Andy Arditi, Oscar Obeso, Aaquib Syed, Nina Panickssery, Daniel Paleka, Wes Gurnee, and Neel Nanda

presented at NEURIPS, 2024

is recognized among the Most Influential NEURIPS 2024 Papers



Paper Digest
New York, New York



Edition 2026-03
Issued 2026-03 - Impact Factor 6

[Verify this certificate at paperdigest.org](https://paperdigest.org)

Certificate ID 9919bd15e433aece - Ranking constructed from citations in research papers and granted patents.

Scan to verify

