22, August 2023

CST, IIT KGP

MTECH SEMINAR(CS69045)

under the supervision of Prof : Mainack Mondal

Topic : Some recent Attack on LLMs.

Presented by:
Rajdeep Ghosh
23CS60R10

Table of Contents

Introduction & overview

1. Telling about the paper
2. some old work based on it
3. what LLMs are
4. clearing out the objective of this paper

Approach in this paper

1. Briefing about the 3 steps of work

Lets dive deep

Part.1:

Lets dive deep

Part.2:

Lets dive deep

Part.3:

Experimental results

discussing about the

setup....metrics.....results

Attacks on White-box Models

discussing about the setup

for these & results

Transfer attacks

Enhancing transferability.

CONCLUSION

ANY QS ?

Thank you
for listening!