multi-agent robust collaboration

Hengyu An

LINs Lab, Westlake University

November 4, 2024

WESTLAKE UNIVERSITY | SCHOOL OF ENGINEERING

Derail Yourself: MULTI-TURN LLM JAILBREAK ATTACK THROUGH SELF-DISCOVERED CLUES

hello hengyu