Session 5: Safety Evaluation and Red-teaming (perspectives, prompting techniques, and frameworks)

Notes:

Presentation from Sandeep.

Question on infrastructure: response on Argo. We host the models on our system (in house). Can run models at scale.

Unlearning: did you look obliteration techniques? Censored, gard rails. Need to have more details on how we do obliteration.

Second speaker (Bhavya):

LLNL GUARD project.

Develop LLM for different applications: bio, security, chemistry, etc.

List of risks: Space security, Hallucination, Powergrid, privacy, data reveal, misuse (CBRNE), unsafe response.

Existing red-teaming for LLMS do not work. → methods are not scalable. No domain expert who can write comprehensive benchmark.

Static benchmark cannot handle

NNSA mission, need security guarantee. None available red teaming offer that.

→ developing a read teaming (risk detection), Blue team (automatic risk reduction), Assurance team (safety guarantees): safe AI.

Not a static pipeline: how is it dynamic?

Model compression can improve safety

Bo Li

LANL

Cyberfire (increasing puzzle challenges)?

# Evaluation of open responses

# Scribe: