EMNLP 2024 Main Conference

Outcome-Constrained Large Language Models for Countering Hate Speech

Lingzi Hong (University of North Texas)
Pengcheng Luo (Peking University)
Eduardo Blanco (University of Arizona)
Xiaoying Song (University of North Texas)

Motivation

LLMs have shown superior capability in understanding and responding to human language

Ability to understand the complexities of human interactions?

Adapting to the complexities of human emotions, behavioral patterns, and even cultural contexts?

How to improve?

Image is generated by DALL-E

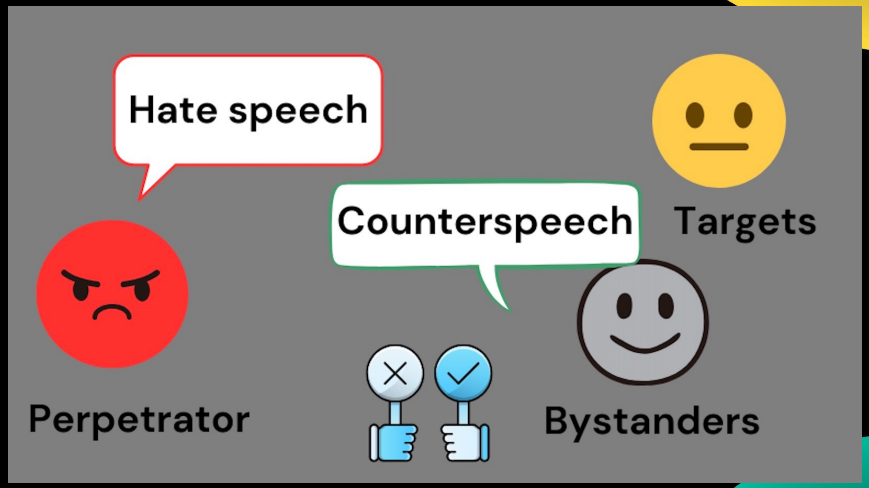
Background

Hate speech

aims to demean, or incite hostility and violence against specific groups

Counterspeech

uses constructive, positive, or factual responses to challenge or counteract hate speech



What has been done

- Human-written counterspeech
- Language models to generate counterspeech

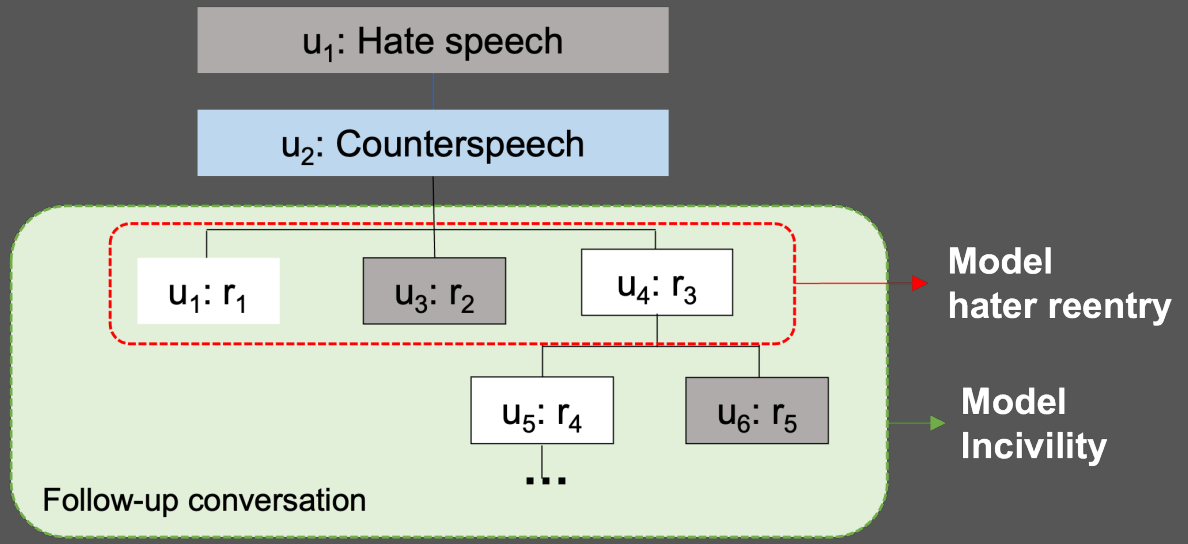
- Polite counterspeech
- informative counterspeech
- counterspeech with different intents

Research Questions

- How can constraints on conversation outcomes be incorporated into developing LLMs for generating counterspeech?
- How effective are these methods in generating outcome-oriented counterspeech?

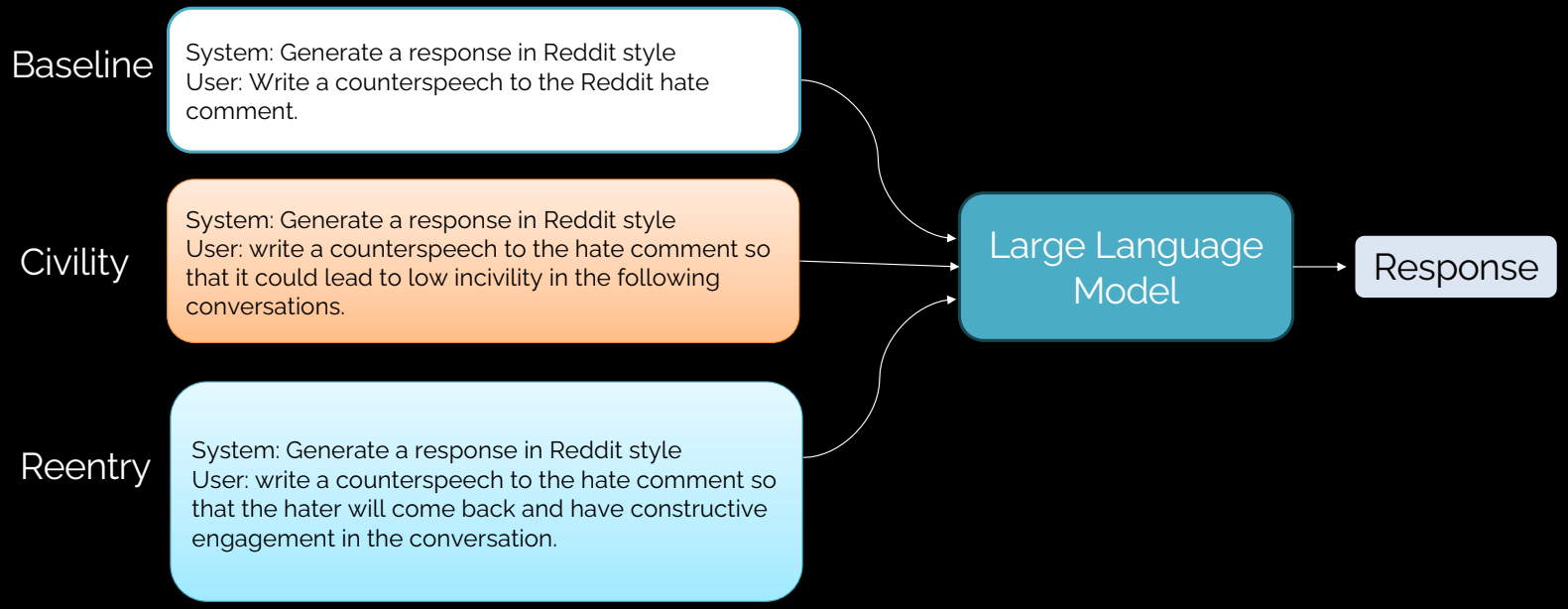
Conversation Outcomes

Hate speech + Counterspeech => predict outcomes



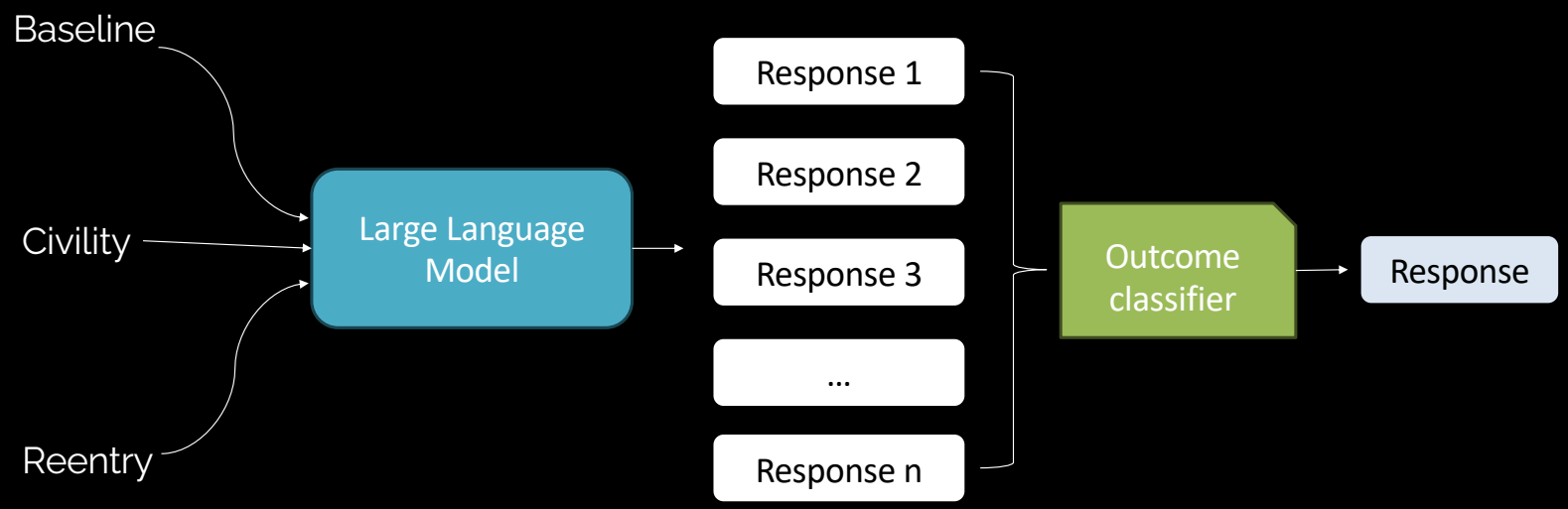
Outcome-oriented LLMs

(1) Instruction Prompts



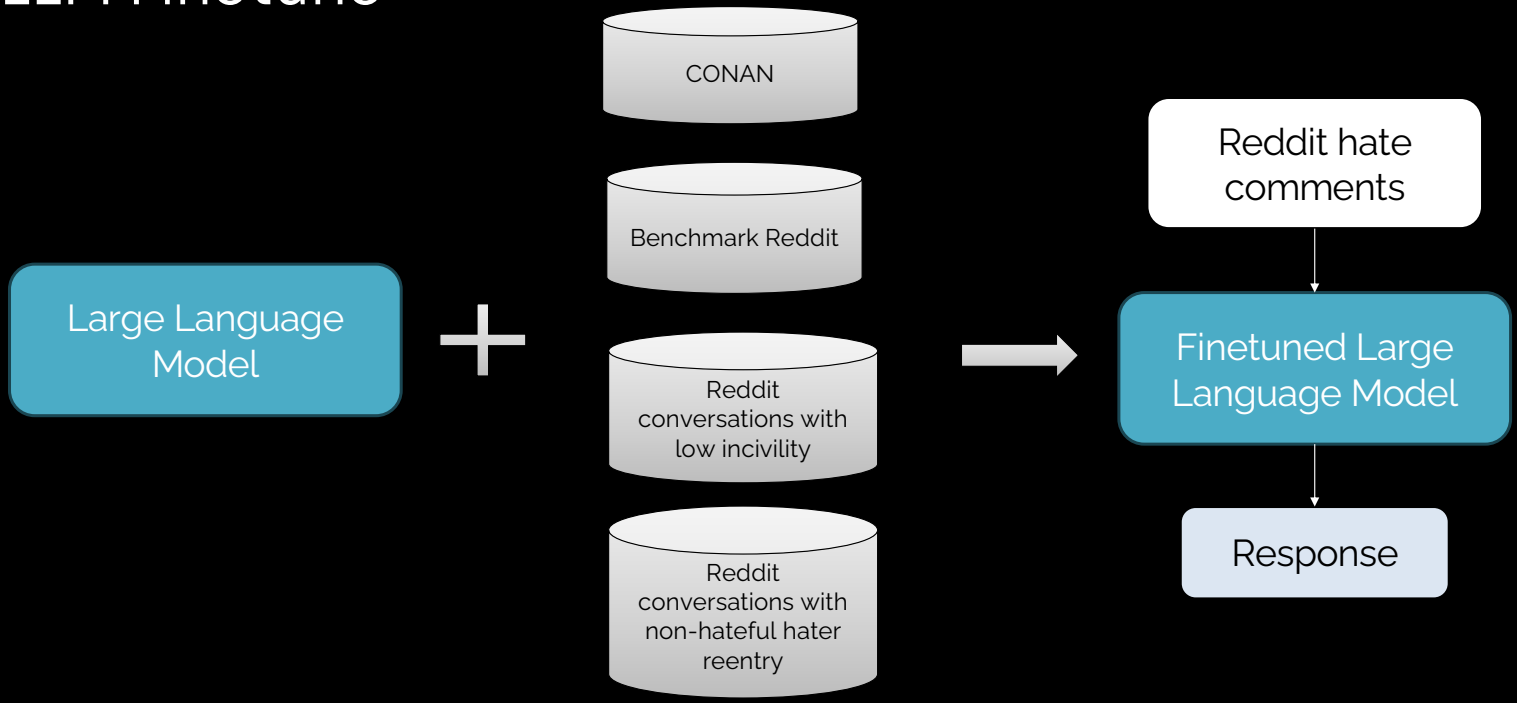
Outcome-oriented LLMs

(2) Prompt and Select



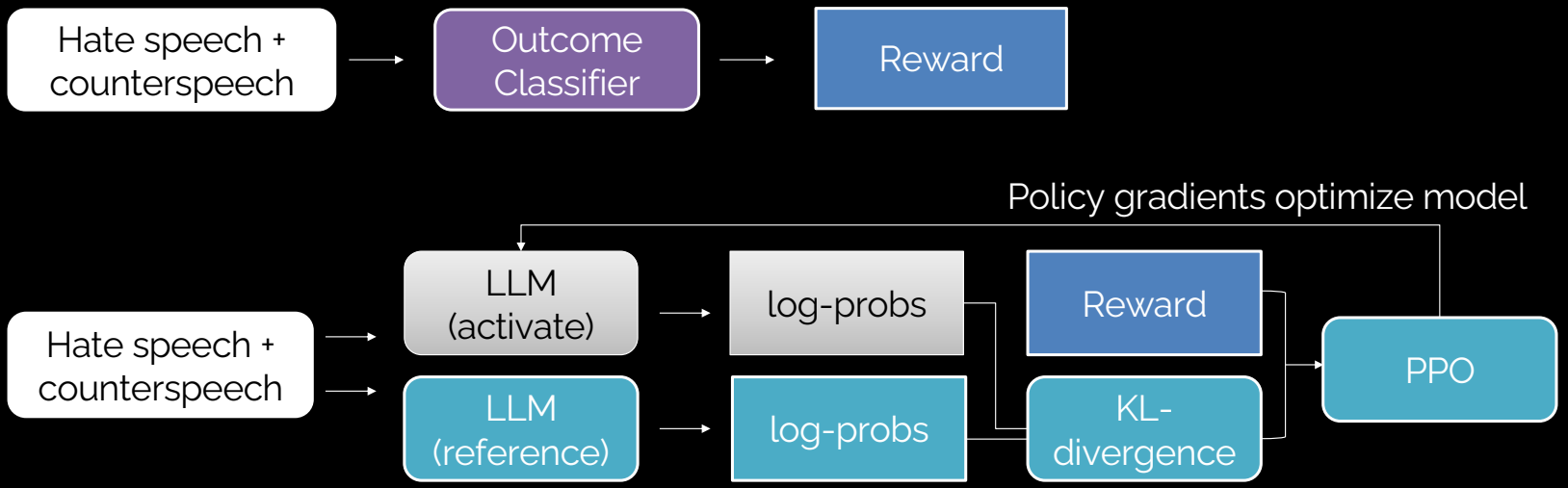
Outcome-oriented LLMs

(3) LLM Finetune



Outcome-oriented LLMs

(4) LLM Reinforcement Learning



Dataset and Experiment

Reddit

Hate speech and counterspeech pairs and follow-ups from 39 subreddits

Build conversation outcome classifiers

Benchmark

Hateful comments from Reddit and counterspeech written by human

Counterspeech Generate and evaluate

Use Llama2-7b-Chat

for

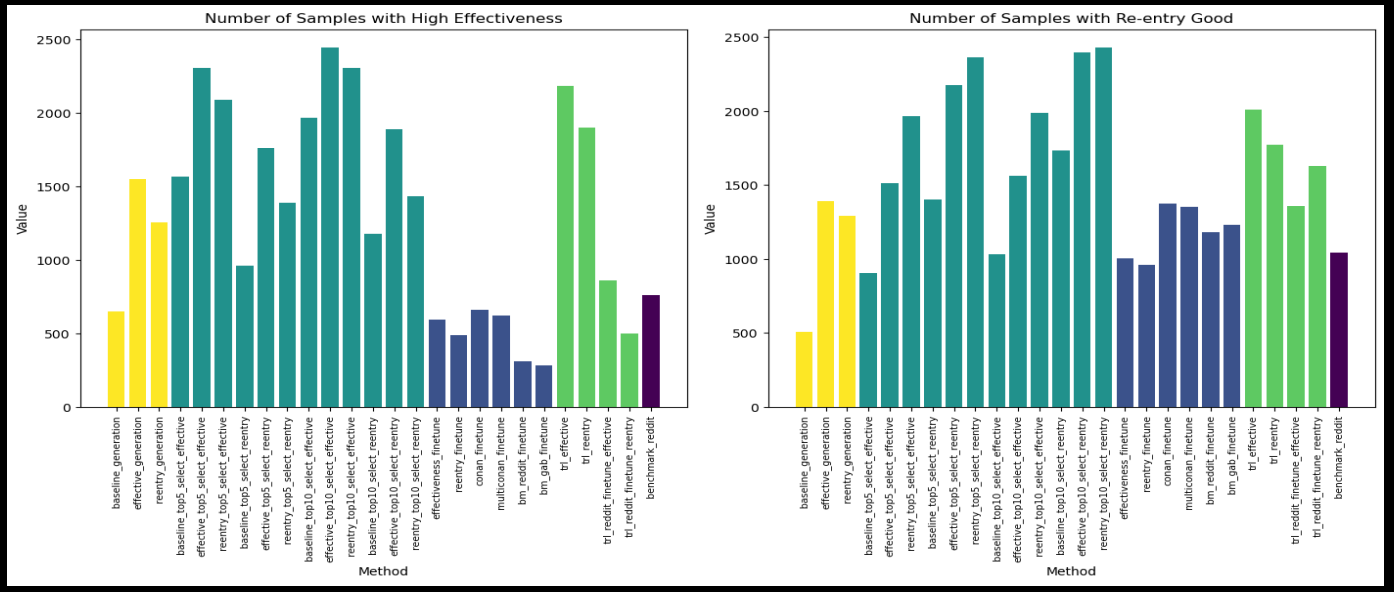
instruction prompts

finetuning

RL

Results

Conversation Outcomes



- *Human-written counterspeech often fail to result in desired outcomes*
- *Effective strategies: Instruction prompts, generate and select, and RL*

Results

Relevance and Text Quality

Relevance

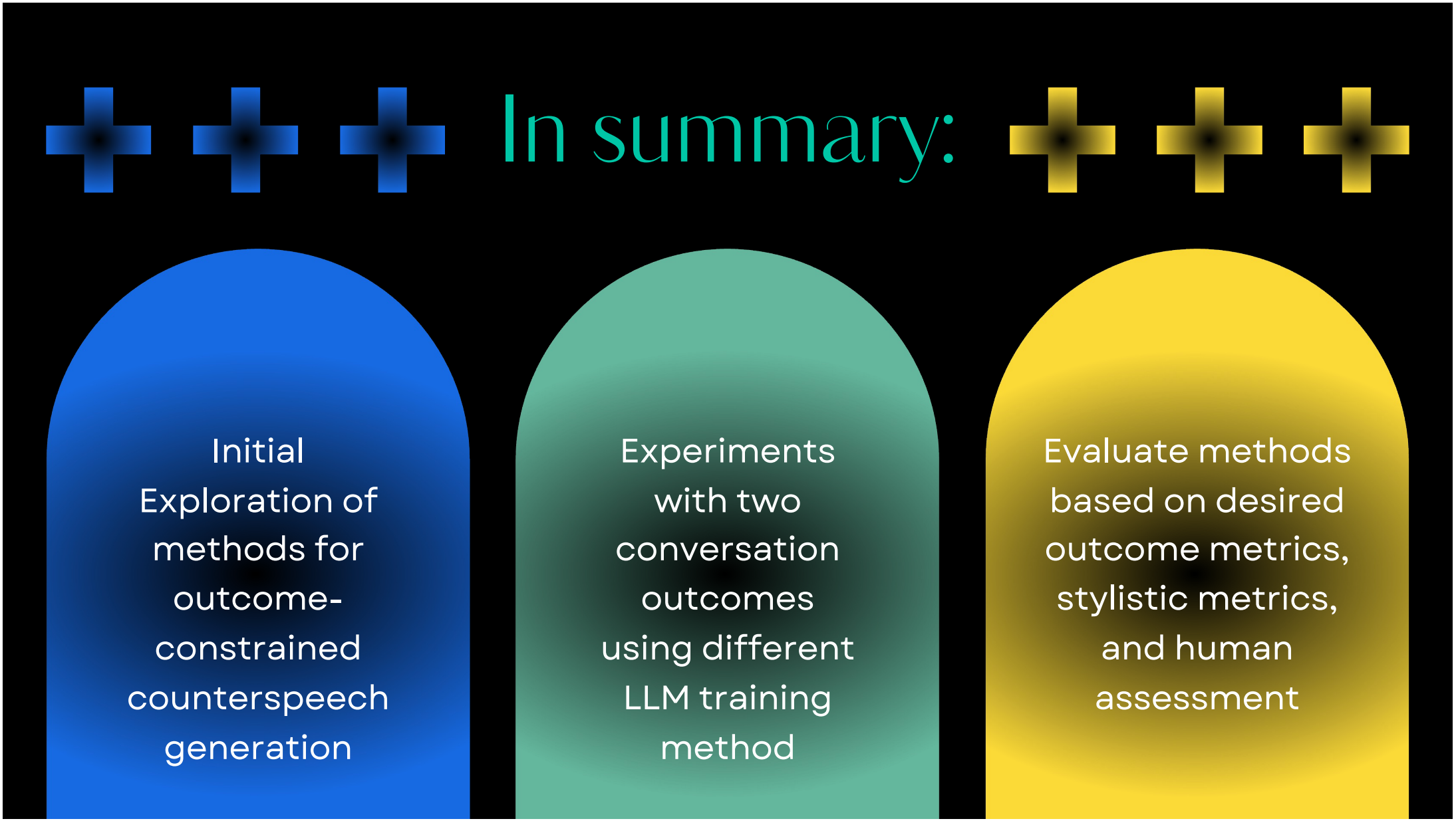
Exact matching measured by METEOR, BLEU, and ROUGE are all low

Semantic similarity measured by BERTScore are all high

Text quality

Counterspeech by instruction prompts are less focused, more redundant, and in low quality

Counterspeech by LLM finetune and RL is more focused and less redundant


In summary:
Initial Exploration of methods for outcome-constrained counterspeech generation
Experiments with two conversation outcomes using different LLM training method
Evaluate methods based on desired outcome metrics, stylistic metrics, and human assessment


Thank you for your time!
Questions?
contact: lingzi.hong@unt.edu