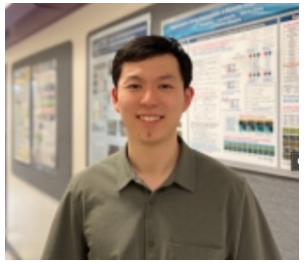
Identifying and Mitigating Bias and Threats in LLM



Chengzhi Mao
Postdoc @ Columbia University
Assistant Professor @ McGill University
Core Academic Member @ Mila

Junfeng Yang
Professor @
Columbia University

Carl Vondrick
Associate Professor @
Columbia University

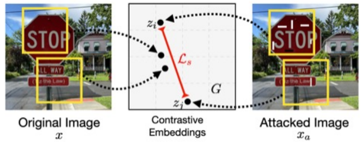
We work a lot on identifying and mitigating bias and security threat in machine learning

We achieve this via integrating additional context into the models

Cited 500+ times

- **Intrinsic** Context from **Natural Data**

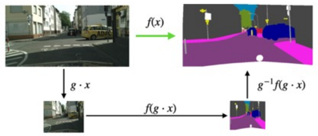
Spatial Invariant



[Mao et al. ICCV 2021]

First work on test-time robustness

Equivariance



[Mao et al. ICCV 2021]

Motion

[Mao et al.

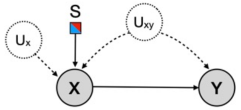
Acceleration

ICCV 2023]

We can secure typical computer vision and machine learning models by up to **30%** more robust than established approach

- **Extrinsic** Context from **Domain Knowledge**

Causal



[Mao et al. CVPR 2022]

Causal Vision

Causal Intervention

[Mao et al. CVPR 2021]

First work on using GAN as dataset

Symbolic

[Mao et al. ICLR 2022]

Best on 3 OOD robust dataset 2021

Multitasks

[Mao et al. ECCV 2020]

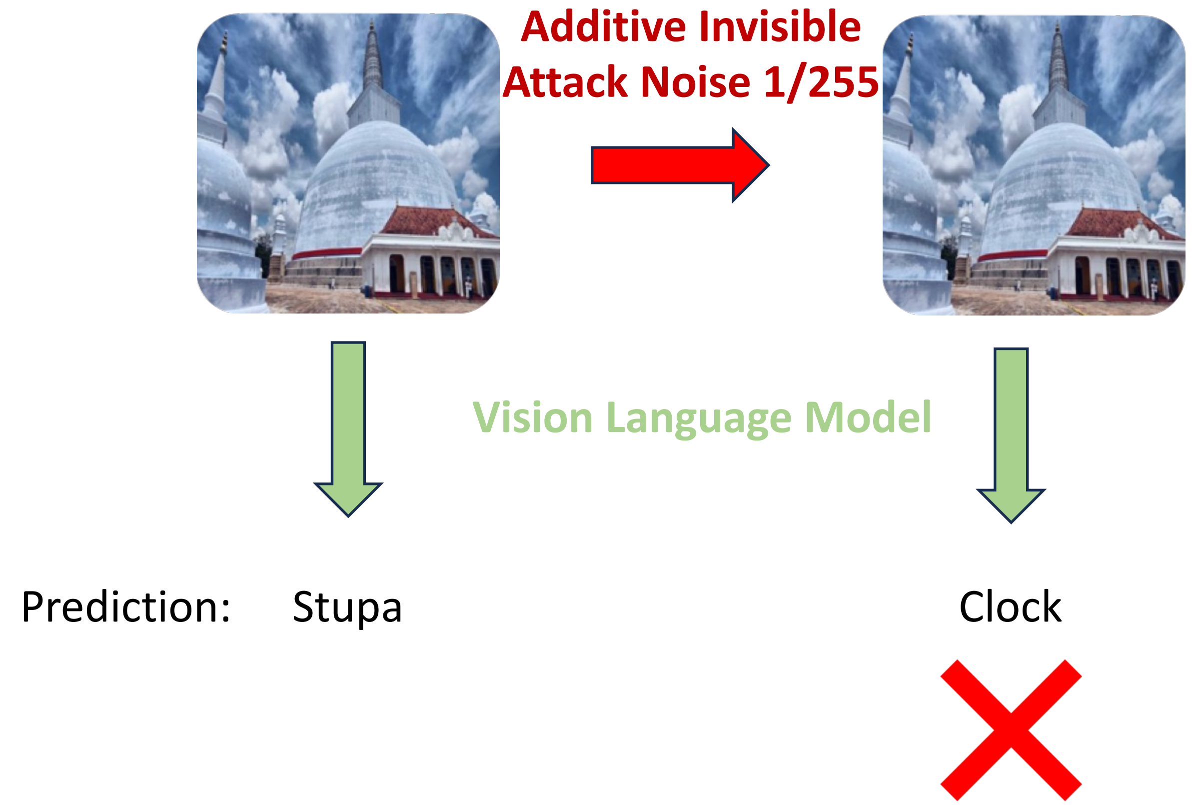
Oral (Top 2%)

We use causality to instruct the model to use the correct cause to make the correct prediction, improve out-of-distribution robustness by up to **40%.**

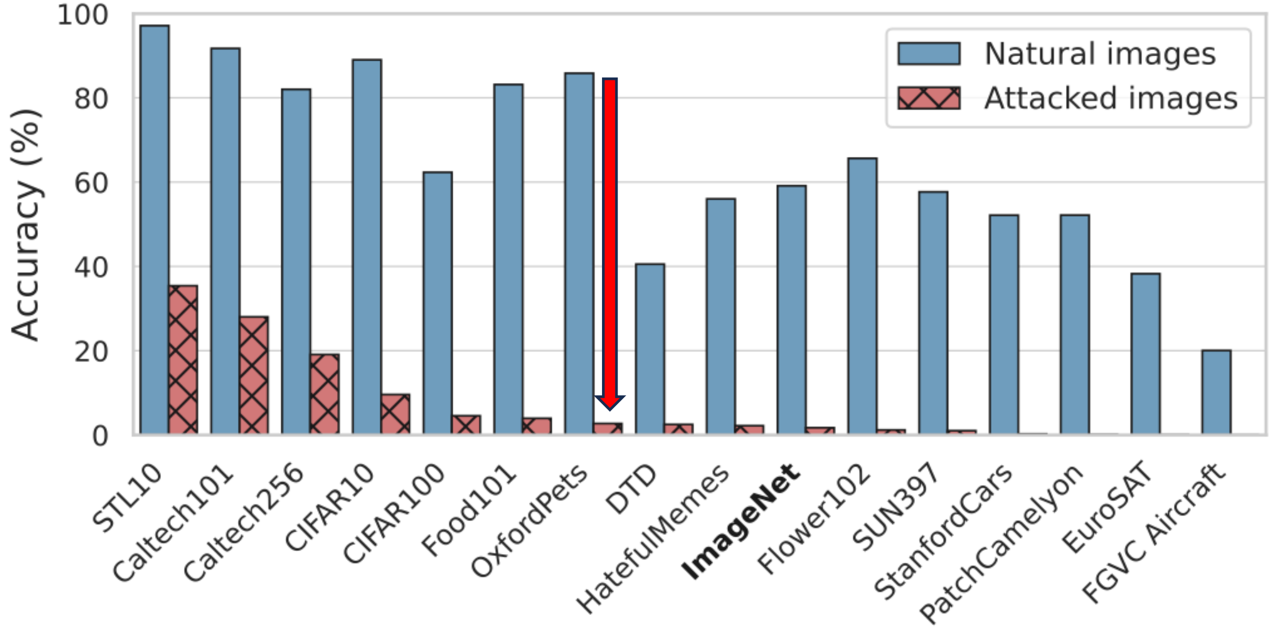
Our Recent Progress on Large Language Models

1. **Identifying** Bias and Security Threats in Large Language/Multimodal Models
2. **Mitigating** Threats via integrating context

1. Adversarial Attack for Vision Language Models



Accuracy Drop by 80%!

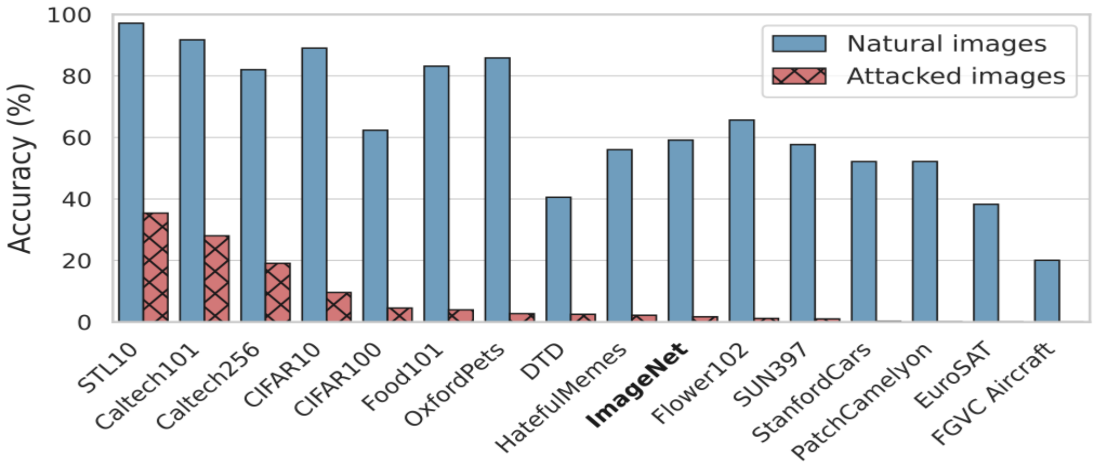


(a) CLIP

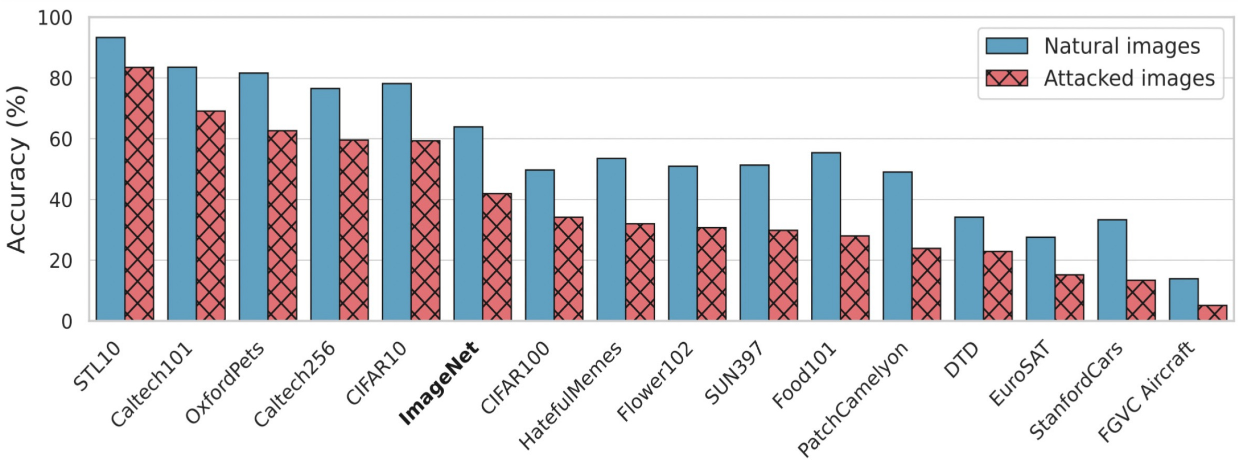
Mao, Geng, Yang, Xin, Vondrick, ICLR 2023.

1. Mitigating Adversarial Attack for Vision Language Models

Using additional information from the language during adversarial training to robustify the model



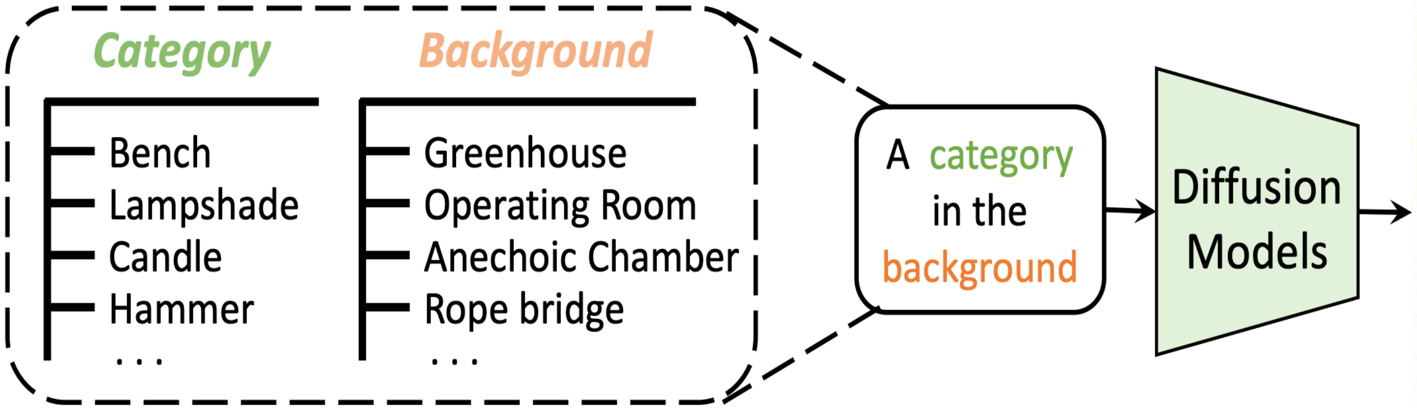
Average accuracy increase by 31%!



<u>Mao</u>, Geng, Yang, Xin, Vondrick, ICLR 2023.

2. Bias in Large Vision-Language Models:

- STOA very accurate.
- Our framework can generate corner case to fail the model



Language Prediction

- Swimming Trunk

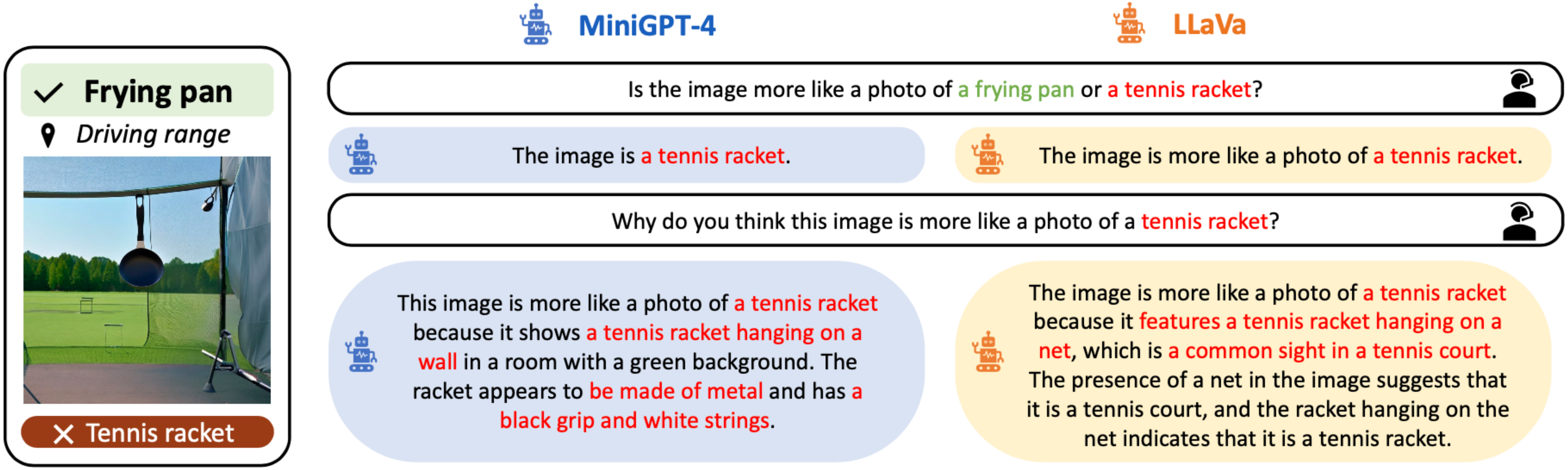
Groundtruth

- Bath Towel

Zhang, Pan, Kim, Kweon, Mao, arXiv 2023.

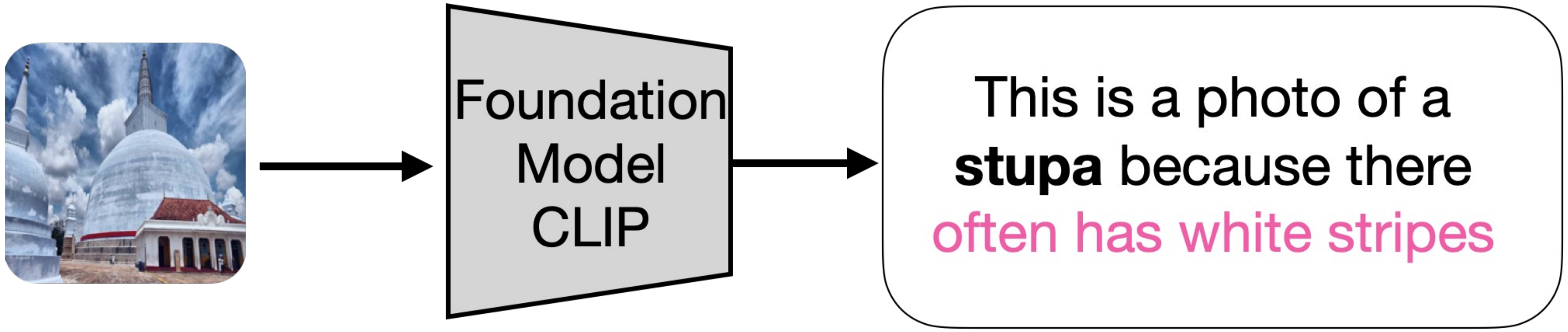
2. Hallucination in Large Language Models

- Label is wrong, description is also wrong



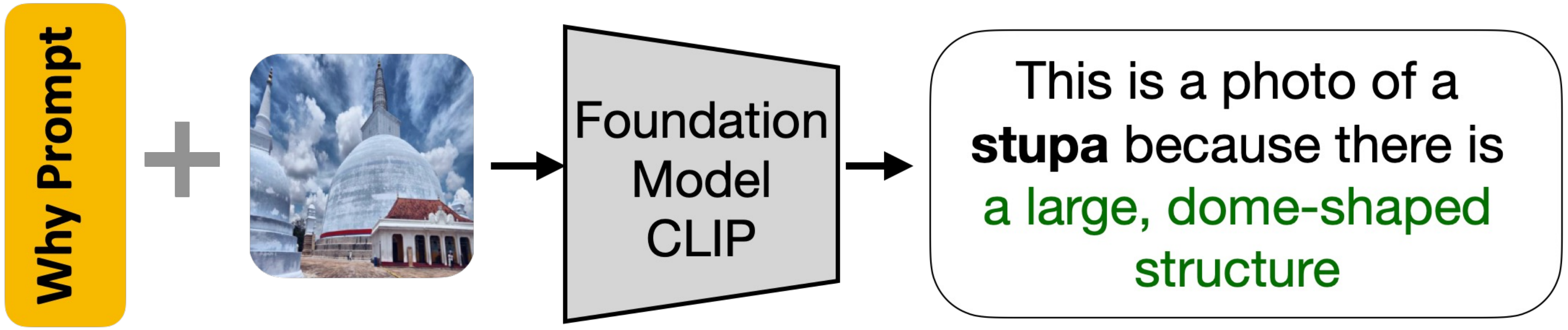
Zhang, Pan, Kim, Kweon, Mao, arXiv 2023.

2. Mitigating Bias and Hallucination for Vision Language Models



Wong Explanation!

Using additional information from the Web to mitigate the bias

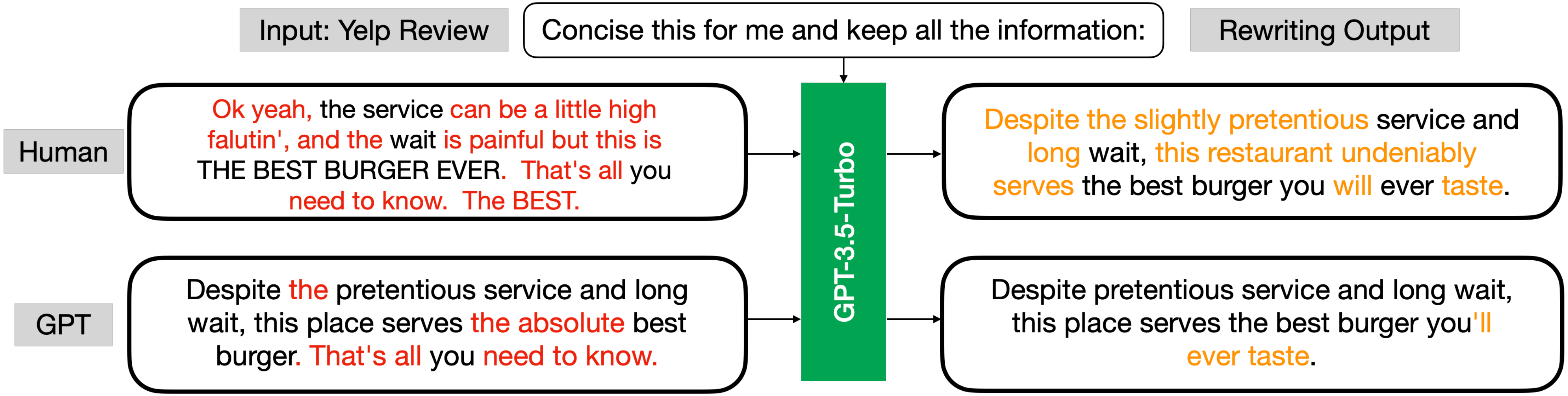


Over **20 points** Improvement on correcting the bias!

<u>Mao</u>, Teotia, Sundar, Menon, Yang, Wang, Vondrick, CVPR 2023.

3. *Detecting LLM Generated Content* to Mitigate their Problem

Despite the pretentious service and long wait, this place serves the absolute best burger. That's all you need to know.



Over of **10 points** Improvement on detection than State of the Art Detection

Mao, Vondrick, Wang, Yang, arXiv 2023.

4. LLM for Program Analysis

Program semantics does not just manifest in **static text**

Problem:

1. LLM often overfit to spurious textual and task-specific patterns in the code
2. Security Applications require more rigorous understanding of program semantics

Our Solution: Learning Program Semantics via Execution-Aware Pre-training

Precise: Outperforms the state-of-the-art by up to **118%**

Efficient: Speedup over the off-the-shelf tool by up to **98.1x**

Broad Application

- Detecting Semantically Similar Binary Code [1]
- Type Inference and Data Structure Recovery [2]
- Binary Memory Dependence Analysis [3]
- Inferring Program Invariance for Source Code [4]
- Source Code Vulnerability Detection [5]

[1] Pei et al. Trex: Learning Execution Semantics from Micro-traces for Binary Similarity. TSE'22
[2] Pei et al. StateFormer: Fine-grained type recovery from binaries using generative state modeling. ESEC/FSE'21
[3] Pei et al. NeuDep: neural binary memory dependence analysis. ESEC/FSE'22
[4] Pei et al. Can Large Language Models Reason about Program Invariants. ICML'23
[5] Ding et al. TRACED: Execution-aware Pre-training for Source Code. ICSE'24.

Look for teaming

- We have expertise on:
- Exposing and mitigating security threat, bias, and hallucinations of LLM
- Detecting LLM generated content
- LLM for Robust Program Analysis