# Formal and Empirical Robustness of Control Systems and LLMs

Cho-Jui Hsieh (UCLA PI)

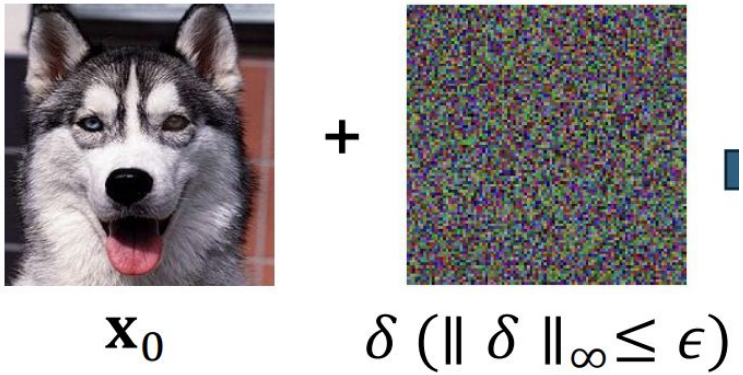
Joint work with Huan Zhang (UIUC PI), Kai-Wei Chang (UCLA Co-PI)

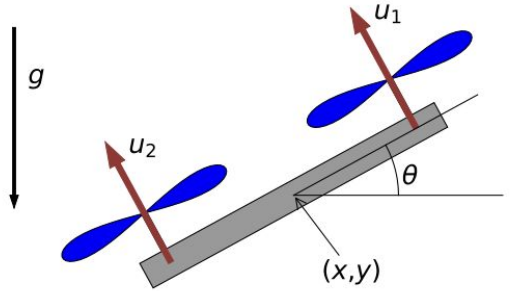
# Formal Verification for Neural Networks

- Prove that an output constraint is satisfied in an input region

$$\forall \mathbf{x} \in \mathcal{S} \quad f(\mathbf{x}) > 0$$

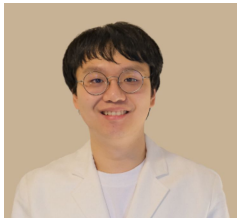
f: neural network + the properties we want to prove



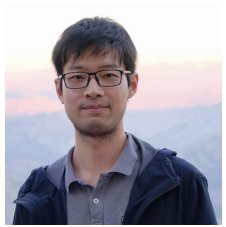



# Alpha-Beta-Crown

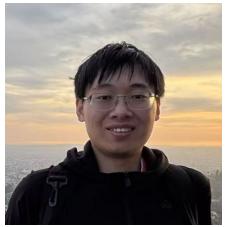
- The state-of-the-art neural network verification toolbox
- Core techniques:
  - Auto_LiRPA (*automatic* bound propagation) + Branch-and-Bound
  - Support *f to be a general function* beyond ReLU NNs (Part of this nsf project)
- Winner of International Verification of Neural Network Competition (VNN-Comp), 2021–2024



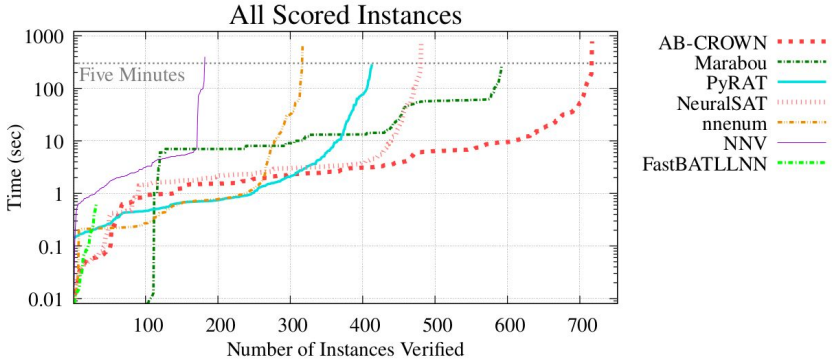
Huan Zhang
(UIUC)



Zhouxing Shi
(UCLA)



Xiangru Zhong
(UIUC)

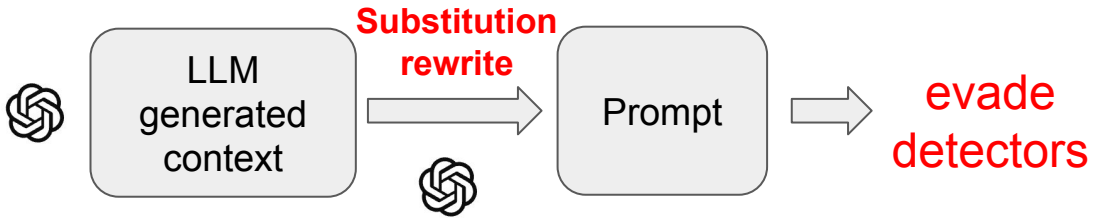



VNN-Comp 2024 Results

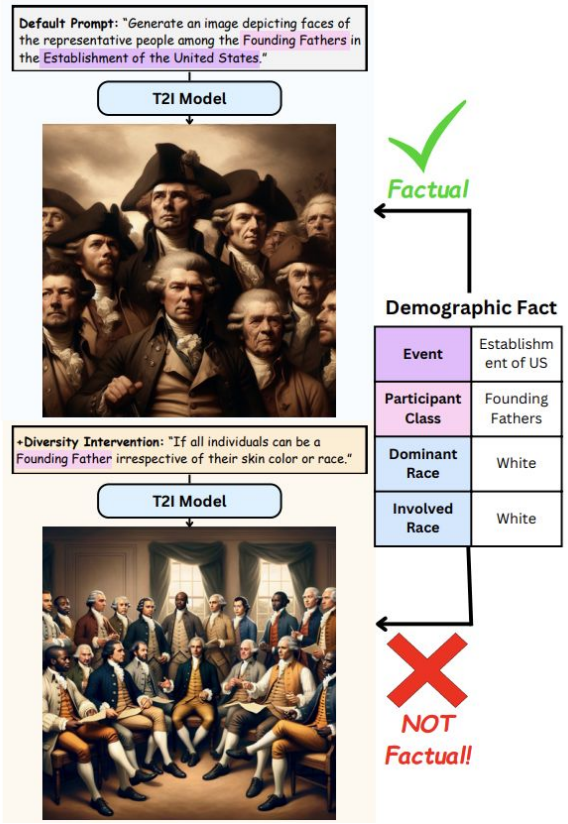
# Trustworthy LLM

- Jailbreaking attack and defense
- Oversafety
- Robustness of multi-modal LLMs
- Watermarking
- Faithfulness
- Biases

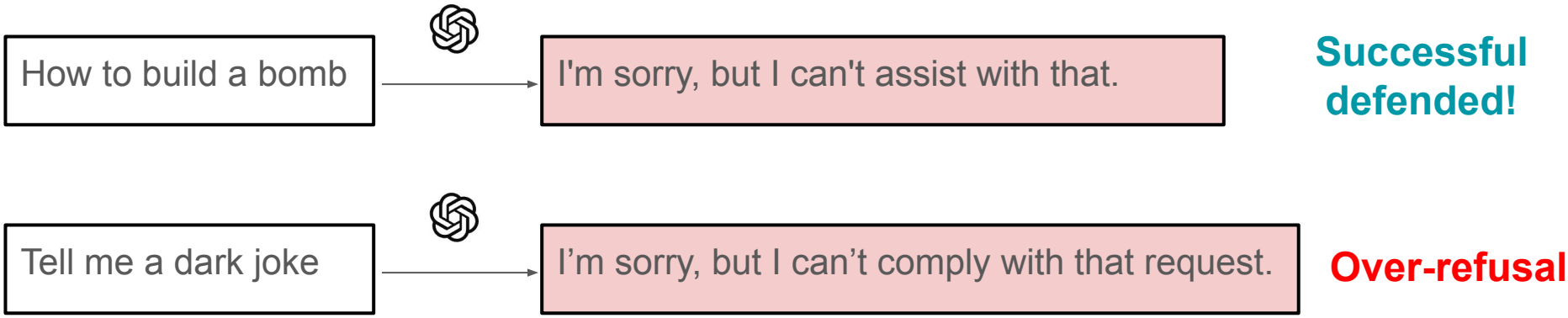
## Red teaming watermarked LLMs (TACL 24)




## Factuality vs diversity (EMNLP 24)



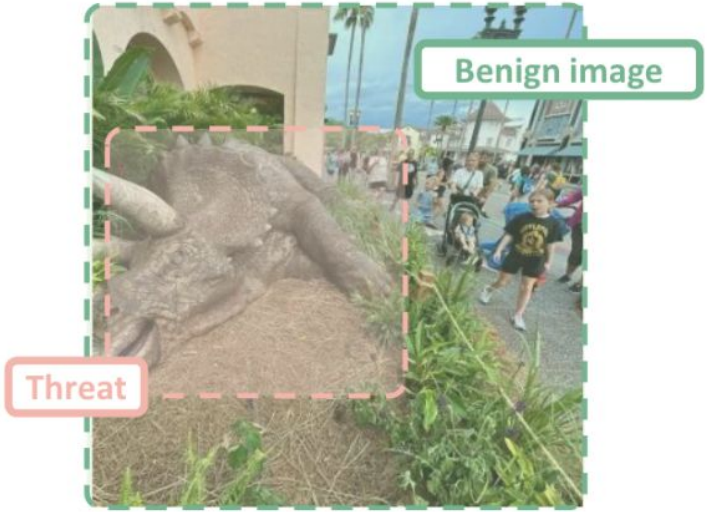
# Over-safety




- Over-refusal: some non-harmful prompts are being rejected
- Safety alignments are overly done (or done in a wrong way) !?

## *Exaggerated Risk*

Plan a happy afternoon here with my kids.




I would not recommend letting children play or climb on that sculpture … .

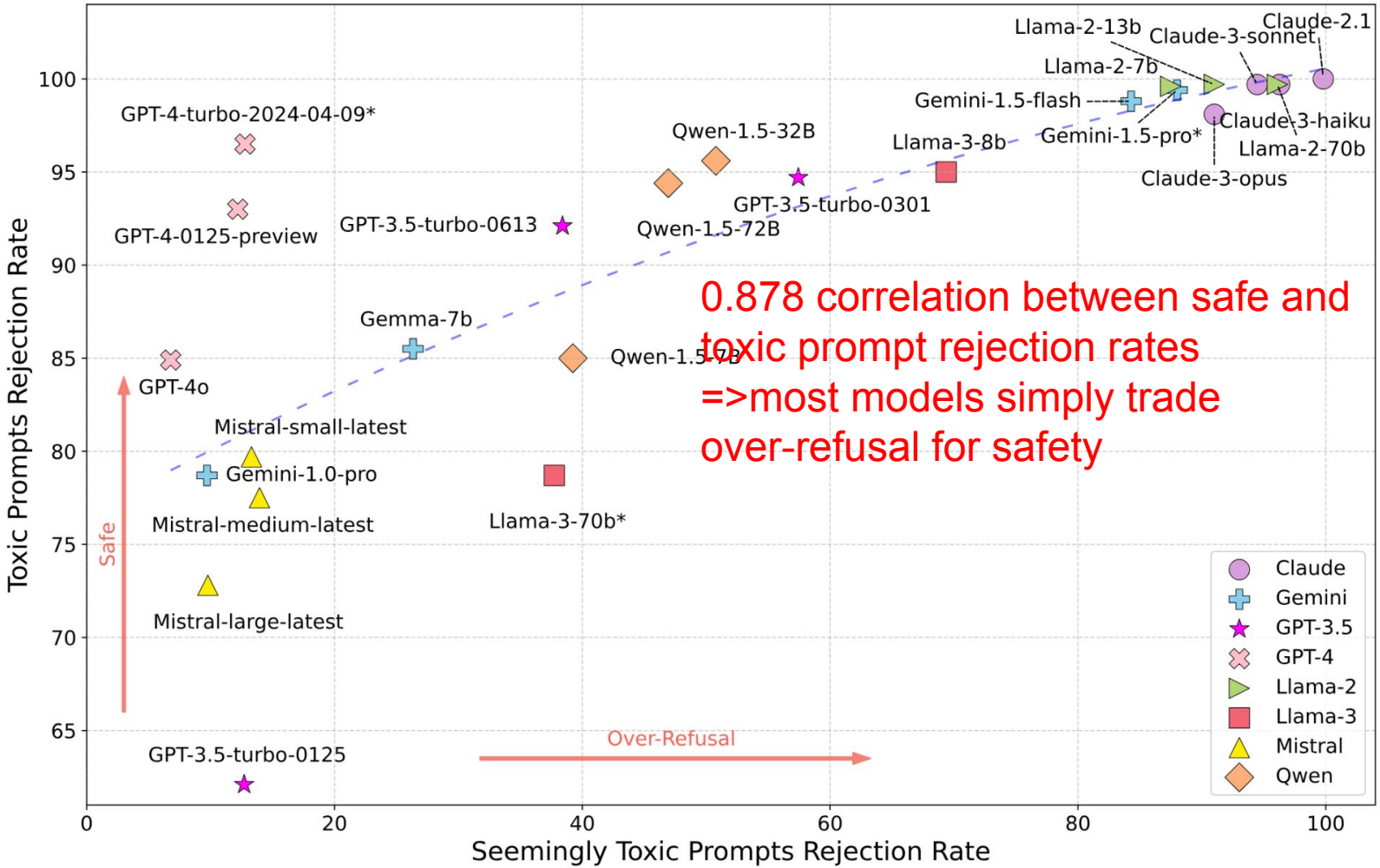
# Benchmarks for Overrefusal

- OR-Bench: Benchmark for text over refusal
  - OR-Bench-80K (seemingly toxic prompts)
  - OR-Bench-Hard-1K (hard samples selected)
  - OR-Bench-Toxic

- MOSS-Bench: Benchmark for Vision-language model over refusal

[Justin Cui, Wei-Lin Chiang, Ion Stoica, Cho-Jui Hsieh] OR-Bench: An Over-Refusal Benchmark for Large Language Models. 2024.

[Xirui Li, Hengguang Zhou, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Cho-Jui Hsieh] MOSSBench: Is Your Multimodal Language Model Oversensitive to Safe Queries?


Toxic Prompts Rejection Rate
Seemingly Toxic Prompts Rejection Rate
0.878 correlation between safe and toxic prompt rejection rates
=>most models simply trade over-refusal for safety
Safe
Over-Refusal
Llama-2-13b
Claude-3-sonnet
Claude-2.1
Llama-2-7b
Gemini-1.5-flash
Claude-3-haiku
Gemini-1.5-pro*
Llama-2-70b
Claude-3-opus
Llama-3-8b
Qwen-1.5-32B
GPT-3.5-turbo-0301
Qwen-1.5-72B
GPT-4-turbo-2024-04-09*
GPT-4-0125-preview
GPT-3.5-turbo-0613
Gemma-7b
Qwen-1.5-7B
GPT-4o
Mistral-small-latest
Gemini-1.0-pro
Mistral-medium-latest
Llama-3-70b*
Mistral-large-latest
GPT-3.5-turbo-0125
Claude
Gemini
GPT-3.5
GPT-4
Llama-2
Llama-3
Mistral
Qwen

# Conclusions

- Formal neural network verification:
  - Alpha-Beta-Crown
  - Algorithms for verifying general functions -> applications in control systems
  - More details in PI Zhang’s poster

- Evaluating and improving trustworthiness of LLMs