Ruijie(Collin) Zhang

https://scholar.google.com/citations?user=EH6bkiwAAAAJ
rz454@cornell.edu

EDUCATION

Cornell University
Sep 2023 - Now, Ithaca & New York
Ph.D Computer Science

New York University
Sep 2019 - May 2023, New York
B.S. Computer Science & Econometrics.

INTERNSHIPS

Qwen Team, Research Intern
LLM language-mixing intervention
Developed an efficient intervention mechanism to detect and prevent language-mixing errors, where LLMs inadvertently switch languages during generation—a prevalent issue across all state-of-the-art LLMs. Deployed in Qwen Chat and Qwen's API to mitigate language mixing in model outputs without affecting performance on any other tasks.

LLM repetition intervention
Designed and implemented a detection and intervention system that preemptively prevents infinite repetition loops in LLM generation. Deployed in Qwen's API, successfully preventing 80% of endless repetition patterns with zero false positives.

Selected Preprints

Language Confusion Gate: Language-Aware Decoding Through Model Self-Distillation
Collin Zhang, Fei Huang, Chenhan Yuan, Junyang Lin

Adversarial Decoding: Generating Readable Documents for Adversarial Objectives
Collin Zhang, Tingwei Zhang, Vitaly Shmatikov

PUBLICATIONS

Extracting Prompts by Inverting LLM Outputs
Collin Zhang, John X. Morris, Vitaly Shmatikov
EMNLP 2024 Oral

Harnessing the universal geometry of embeddings
Rishi Jha, **Collin Zhang**, Vitaly Shmatikov, John X Morris
NeurIPS 2025

Zombie: Middleboxes that Don’t Snoop
Collin Zhang, Zachary DeStefano, Arasu Arun, Joseph Bonneau, Paul Grubbs, Michael Walfish
NSDI 2024

Open-Source Projects

output2prompt 47 stars
Code for *Extracting Prompts by Inverting LLM Outputs*

adversarial_decoding 20 stars
Code for *Controlled Generation of Natural Adversarial Documents for Stealthy Retrieval Poisoning*

FastDraw 59 stars
A Fast and Complete Swift Drawing Library

Talks

EMNLP 2024, Nov 12, 2024
Extracting Prompts by Inverting LLM Outputs [**link**]