Min (Henry) Cai

✉ caimin2021@email.szu.edu.cn • github.com/HenryCai11/

Education

Shenzhen University — **Shenzhen, China**
M.S. in ***Computer Science*** — *2021 – now*

Beijing Language and Culture University — **Beijing, China**
B.A. in ***Linguistics (Translation)*** — *2016 – 2021*

Research Interests

I have broad interests in ML and NLP, particularly in understanding the mechanisms of neural Language Models (LM), and in turn, helping to better understand human nature.
Specifically, my current research focuses on:

- **interpreting and controlling LLM behaviors**, and fine-tuning algorithms to align with human values.
- **LLM Agents** capable of solving complex tasks, e.g. multi-agent social deductive games (Avalon).

Publications

1. **SELF-CONTROL of LLM Behaviors by Compressing Suffix Gradient into Prefix Controller**
   - Min Cai, Yuchen Zhang, Shichang Zhang, Fan Yin, Difan Zou, Yisong Yue, Ziniu Hu
   - Submit to NeurIPS 2024
2. **STRATEGIST: Learning Strategic Skills by LLMs via Bi-Level Tree Search**
   - Jonathan Light, Min Cai, Weiqin Chen, Guanzhi Wang, Xiusi Chen, Wei Cheng, Yisong Yue, Ziniu Hu
   - Submit to NeurIPS 2024
3. **AVALONBENCH: Evaluating LLMs Playing the Game of Avalon**
   - Jonathan Light*, Min Cai* (equal contribution), Sheng Shen, Ziniu Hu
   - NeurIPS 2023 workshop, Foundation Models for Decision Making
4. **Self-Convinced Prompting: Few-Shot Question Answering with Repeated Introspection**
   - Haodi Zhang, Min Cai, Xinhe Zhang, Defu Lian, Rui Mao, Kaishun Wu
   - arXiv preprint: 2310.05035

Work and Research Experience

LLM Control and Mechanistic Intepretability — **Remote**
Advisor: Ziniu Hu and Shichang Zhang — *Jan. 2024 – Present*

- We propose SELF-CONTROL, a novel method utilizing suffix gradients to control the behavior of large language models (LLMs) without explicit human annotations.
- Given a guideline expressed in suffix, SELF-CONTROL computes the gradient of this self-judgment with respect to the model's hidden states, directly influencing the auto-regressive generation towards desired behaviors.
- Further introduce SELF-CONTROL$_{\text{PREFIX}}$ that use LoRA to compress self-collected <query, representation> pairs into a prefix controller for efficient inference-time control.
- Lead the whole project, implement the codes at **Github: https://github.com/HenryCai11/LLM-Control**

LLM Agent Playing Avalon — **Remote**
Advisor: Ziniu Hu — *Jun. 2023 – May 2024*

- We propose AVALONBENCH, a game environment tailored for evaluating multi-agent LLM Agents.
- Collaboratively build the game engine for 'The Resistance: Avalon' and integrate it into AgentBench (2k stars on Github). Specifically, I implement an asynchronous multi-agent module for AgentBench.
- We further propose STRATEGIST that learns skills via bi-level tree search, and integrate it into AVALONBENCH as an advanced agent.

LLM for Complex Reasoning **Shenzhen University**
Advisor: Haodi Zhang *Jun. 2021 – Jun. 2024*

- We propose Self-Convince, a self-refine framework that leverages self-generated signals, i.e. correctness, analysis to iteratively improve LLMs' reasoning ability.

Teaching Experience

- **Teaching Assistant for "Compliers" at Shenzhen University, Spring 2023**.
  - Hold a 40-student on-campus programming tutorial (2 hours per week), help students with their homework, go through and check their codes.

Selected Hornors and Awards

2021: Outstanding Student Scholarship, the second prize, Shenzhen University

2018: Outstanding Student Scholarship, the third prize, Beijing Language and Culture University

Core Skills

- Pytorch (Huggingface Transformers)
- Parameter-efficient fine-tuning (LoRA, QLoRA, Prefix-Prompt-Tuning, etc.)
- Asynchronous programming. Using Python's `asyncio` to build IO-bound systems, e.g. AgentBench.