HAOCHEN LI

haochen003@e.ntu.edu.sg | https://alex-haochenli.github.io/ | +65 98131304

EDUCATION

Nanyang Technological University, Singapore Jan. 2022 - Dec. 2025 (Expected)
Ph.D Student in School of Computer Science and Engineering
Supervisor: Prof. Chunyan Miao

Beihang University, Beijing, China Sept. 2017 - Jun. 2021
B.Eng in Electrical Engineering

RESEARCH INTERESTS

My PhD research aims to bridge programming language and natural language through language models. Specifically, I focus on code search (code retrieval), code generation, and the synergy of the two through the Generation-Augmented Retrieval and Retrieval-Augmented Generation framework.

SELECTED PUBLICATIONS

[**ACL'25**] **GiFT: Gibbs Fine-Tuning for Code Generation**
Haochen Li, Wanjin Feng, Xin Zhou, Zhiqi Shen
- Inspired by Gibbs Sampling, we propose a new self-training method that allows self-generated data to be drawn from the marginal distribution of the joint NL-code space.
- We theoretically demonstrate that generating synthetic data by marginal sampling reduces bias and increases diversity compared to conditional sampling, as done by current self-training methods.
- We propose a perplexity-based data selection method to mitigate imbalance in synthetic data.
- Our method achieves superior performance, particularly on more challenging benchmarks.

[**ACL'24 Oral**] **Rewriting the Code: A Simple Method for Large Language Model Augmented Code Search**
Haochen Li, Xin Zhou, Zhiqi Shen
- We propose a simple yet effective extension on Generation-Augmented Retrieval framework, which rewrites codes in codebase for style normalization.
- Our method significantly boost the performance of sparse retrieval systems (up to 35.7%) and dense retrieval systems in both zero-shot (up to 27.6%) and fine-tuning (up to 23.6%) settings.
- With our method, non-neural model is comparable to neural models under the zero-shot setting; neural models under the zero-shot setting is comparable to fine-tuned ones.
- We are the first to propose a novel evaluation metric, dubbed Code Style Similarity, to quantitatively measure the disparity in code style.

[**Preprint'24**] **Towards Goal-oriented Prompt Engineering for Large Language Models: A Survey**
Haochen Li, Jonathan Leung, Zhiqi Shen
- We survey 35 studies on goal-oriented prompt engineering for LLMs.
- We taxonomize existing works into a 5-stage framework, including goal decomposition, action selection, action execution, sub-goal result evaluation, and sub-goal selection.

[**EMNLP'23**] **Rethinking Negative Pairs in Code Search**
Haochen Li, Xin Zhou, Luu Anh Tuan, Chunyan Miao
- We propose Soft-InfoNCE, a novel contrastive loss that explicitly models relations among negative

samples by adding a weight to the InfoNCE loss.
- We theoretically prove that Soft-InfoNCE can control the representation distribution of negative samples and reduce bias in mutual information estimation.

[**EMNLP'22**] **Exploring Representation-level Augmentation for Code Search**
Haochen Li, Chunyan Miao, Cyril Leung, Yanxian Huang, Yuan Huang, Hongyu Zhang, Yanlin Wang
- We unify existing representation-level augmentation methods into a general format and propose three novel augmentation methods based on the general format.
- We theoretically prove that augmentations following the general format yield tighter mutual-information bounds during contrastive learning.

WORKING EXPERIENCE

Microsoft Research Asia Nov. 2021 - Dec. 2021
Research Intern | Advisor: Prof. Yanlin Wang
Topic: Code Search

Mech-mind Robotics Technologies Co., Ltd. Dec. 2020 - Apr. 2021
Computer Vision Algorithm Engineer Intern
Topic: Object Detection, Image Segmentation and Anomaly Detection

Institute of Software, Chinese Academy of Sciences Jan. 2020 - Aug. 2020
Research Intern | Advisor: Prof. Jingzheng Wu
Topic: Code Representation Learning

ACADEMIC ACTIVITIES

- Program Committee / Reviewer:
  - 2025: ICML, NeurIPS, ACL ARR
  - 2024: NeurIPS, ICLR, ACL ARR (EMNLP 2024 Outstanding Reviewer)
  - 2023: ACL, EMNLP
  - 2022: EMNLP
- Teaching Assistant:
  - NTU CE1103 Introduction to Computational Thinking & Programming
  - NTU CE1115 Introduction to Data Science & Artificial Intelligence