

# Chengzu Li

M.Phil. in Advanced Computer Science  $\diamond$  University of Cambridge

+44-07783 497657(mobile)  $\diamond$  [cl917@cam.ac.uk](mailto:cl917@cam.ac.uk)

## EDUCATION

---

**University of Cambridge**, Cambridge, UK

2022 - 2023(*expected*)

M.Phil. student in Advanced Computer Science

**Core Courses:** Machine Learning and Physical World (Distinction), Introduction to Natural Language Processing (Distinction)

**Xi'an Jiaotong University**, Xi'an, China

2018 - 2022

B.Eng. in School of Automation Science and Technology. Minor in Fintech.

GPA: 91.96/100, ranking top 5% in School of Automation Science and Technology

**Core Courses:** Complex Analysis and Integral Transformation (100), Advanced Mathematics (92), Linear Algebra and Analytic Geometry (92), Probabilistic (90), Data Structure and Algorithms (100)

**Graduated with Distinction:** Awarded as Outstanding Graduate of 2022.

**University of Hong Kong**, Hong Kong

2021.1 - 2021.6

**Exchange student** in Computer Engineering, Faculty of Engineering

GPA: 4.08/4.3

**Student Research Assistant** in Dept. of Computer Science

2021.7 - 2021.9

**Core Courses:** Big Data Analytics (A+), Pattern Recognition and Machine Intelligence (A, 1st in class), Artificial Intelligence and Deep Learning (A), Machine Learning (A)

## RESEARCH EXPERIENCES

---

### Natural Language Grounding

**Binding Language Models in Symbolic Languages**

*Paper link (ICLR2023)*

**Advisor:** Prof. Tao Yu, *The University of Hong Kong, NLP Group*

2022.5 - 2022.9

- Designing the grammar of BINDER, which combines the neural modules including language models and vision models with SQL, Python.
- Achieved SOTA with CodeX on WikiTQ as well as TabFact and comparable result on MultiModalQA by BINDER in few-shot settings with better explainability by execution.

**Structural Knowledge Grounding (SKG) with Language Model** *Paper link (EMNLP2022)*

**Advisor:** Prof. Tao Yu and Prof. Lingpeng Kong, *Shanghai AI Lab*

2021.10 - 2022.1

- Implementing different table linearization strategies (simple linearization, linearization as natural language description) on various SKG tasks with a more stable and better performance obtained.
- Integrating pretrained language models of structural input (TAPAS, TaBERT) with T5 on SKG tasks (TabFact, WikiTQ) to quantify the trade-off between structural representation and semantic shift.

**Multi-turn Context Dependent Text2SQL with GCN**

**Advisor:** Prof. Lingpeng Kong, *The University of Hong Kong, NLP Group*

2021.6 - 2021.9

- Utilizing GCN as encoder to achieve joint encoding of database schema and natural language questions based on previous work (RATSQL) and pretrained model (T5).
- Designing a novel stack-based method for dialogue tree structure generation based on which technique put forward an innovative mechanism to update encoder's parameters.
- Delivering a presentation at HKU and successfully completing the 2022 HKU CS internship program.

### Data Sparsity and Few-shot Learning

**Generating Data for Symbolic Language with Language Models** (*submitted to ACL2023*)

**Advisor:** Prof. Tao Yu, *The University of Hong Kong, NLP Group*

2022.9 - 2023.1

- Propose a postprocessing method for overgeneration and improve the zero-shot performance on Spider and MBPP by 5%
- Generate SKG data with limited annotated data and train a smaller model based on synthesized data, from which the performance has been improved by over 5% compared to original training data.

## Domain Adaptation and Distribution Shift

### Domain Description Augmented Language Models for Domain Shift

**Advisor: Prof. Tao Yu**, *The University of Hong Kong, NLP Group*

*Ongoing*

- Designing the benchmark of long instructions over different domains on semantic parsing tasks.
- Collecting natural and high-quality databases and question-SQL pairs.
- Creating the interfaces by which annotating the Text2SQL datasets and Code generation datasets.
- Using latest models such as Codex on the benchmark to test the performances.

## Other Topics

### Entity and Relation Extraction in Social Text

**Advisor: Prof. Chao Shen, Prof. Xiaoming Liu**, *Xi'an Jiaotong University* 2020.9 - 2021.4

- Innovatively introducing a critical counterfactual data generation module to weaken the impact of the dataset's particular distribution, selection bias, etc. on the final result.
- Writing of the patent: The project is expected to publish a patent with the title "**an optimization method and system of entity and relation extraction in social noise text**" (audit status).

## PROFESSIONAL EXPERIENCES

---

### Xiamen Dianchu Technology Co., Ltd.

Algorithm Engineer (Intern), Research & Development Center

2020.7 - 2020.8

Assisting in developing the abnormal indices detection system with machine learning methods.

### Shanghai Artificial Intelligence Laboratory

Trainee Researcher, NLP Group

2021.10 - 2022.1

Participation in *UnifiedSKG* project, focusing on Language Grounding on structural knowledge.

## SELECTED AWARDS

---

- 2022** CSC Masters Scholarship from Cambridge Trust (Fully-Funded)
- 2022** Outstanding Graduate of Xi'an Jiaotong University
- 2020** Excellent Youth League Member (Top 2% among 5000+ students)
- 2020** Provincial First Prize of Mathematical Contest in Modeling(CUMCM) (Top 5%)
- 2019** National Scholarship (Top 1%, highest honor for undergraduates in China)
- 2019** Excellent Student (Top 10% among 5000+ students)
- 2018** Provincial First Prize of the 34th High School Chinese Physics Olympiad (Top 2%)

## EXTRACURRICULAR EXPERIENCE

---

**Nanyang Student Union, Academic Section**, Minister

2019.8 - 2020.6

**Center of Academic Counseling and Student Development**, Vice Director

2020.9 - 2021.6

Primarily responsible for the organization and holding of debate competition, lectures, sorting out test materials, etc., with more than 150 participants attending the lectures.

## SKILLS

---

### Programming Languages and Skills

C, Python, Pytorch, Hadoop, Zookeeper, Flume, Kafka, Programmable Logical Controller, LaTeX, Raspberry Pi

### Languages

Mandarin (Native), English (Fluent)