Zihao Fu

Research Assistant Professor, Department of LML, CUHK
Affiliated Lecturer, Faculty of MMLL, University of Cambridge
Natural Language Processing / Large Language Models / Machine Learning / Explainability / Responsible AI

fuzihaofzh@gmail.com | fuzihaofzh.github.io | github.com/fuzihaofzh | +44 787 1314 588
| fuzihaofzh.github.io/blog | GoogleScholar

Education & Research Career

Research Assistant Professor, Department of LML, CUHK 2025–Now

Serve as a Research Assistant Professor in the Department of Linguistics and Modern Languages at The Chinese University of Hong Kong. Focuses on large language models, natural language processing, interpretability, and their interdisciplinary applications.

Affiliated Lecturer, Faculty of MMLL, University of Cambridge 2024–Now

Serve as an Affiliated Lecturer in the Faculty of Modern and Medieval Languages and Linguistics, collaborating with Prof. Nigel Collier. Deliver guest lectures for the Li18 course on Computational Linguistics. Supervise students conducting research in Natural Language Processing and Large Language Models.

Postdoc, Oxford Internet Institute, University of Oxford 2024–2025

Collaborate with Prof. Chris Russell, Prof. Brent Mittelstadt, and Prof. Sandra Wachter on research focuses on the policy and fairness of large language model applications, we have developed a toolkit to ensure these models adhere to ethical standards and promote fairness. We enforce fairness in arthritis diagnosis using these tools. Additionally, we develop trustworthy watermarking tools for large language models, which can be advantageous for policymaking and surveillance.

Affiliated Data Scientist, Oxford University Hospitals, NHS Trust 2024–2025

Collaborate with Prof. Raashid Luqmani. Using hospital data, analyze the biases present in the diagnosis of rheumatoid arthritis; utilize the tools we developed to eliminate biases in the diagnosis.

Affiliated Postdoc, Exeter College, University of Oxford 2024–2025

Postdoc, The Language Technology Lab, University of Cambridge 2022–2024

Collaborated with Prof. Nigel Collier; Conducted research on the theoretical analysis of language models, specifically focusing on the stability analysis of fine-tuning and the effectiveness of parameter-efficient models; Investigated AI for science applications in the biomedical field, such as biomedical named entity recognition/linking, epidemic monitoring, and etc.

Affiliated Postdoc, Trinity College, University of Cambridge 2022–Now

Ph.D., Faculty of Engineering, The Chinese University of Hong Kong 2017–2021

Supervised by Prof. Wai Lam; Conducted research on text generation, with a particular emphasis on the theoretical analysis of repetition in language models, retrieval-enhanced text generation, and exploring innovative methods in distant and semi-supervised text generation, and etc.

Visiting Scholar, NLP Lab, Department of Computer Science, Tsinghua University 2020–2021

Supervised by Assoc. Prof. Zhiyuan Liu; Conducted research on language model designs, particularly identifying the key differences between the usual encoder-decoder structure and the emerging decoder-only structure.

M.Eng., National Lab for Aeronautics and Astronautics, Beihang University 2012–2015

Supervised by Prof. Guanghong Gong; Conducted research in Computational Fluid Dynamics, specifically focusing on uncompressible potential flow solvers. My contributions include the development of a panel method PDE solver and the proposition of a novel explicit moment integration algorithm; Outstanding postgraduate student.

B.Eng., School of Automation Science and Electrical Engineering, Beihang University 2008–2012

Top 10%; Outstanding undergraduate student; Recommended for graduate school without examination.

Working Experience

Machine Learning Algorithm Engineer, IDST, Alibaba Cloud, Beijing 2015–2017

Developed several distributed parallel algorithms that run on Alibaba Cloud's PAI platform, including summarization, Word2Vec, TFIDF, Sentence Distance, keywords extraction, PMI, Group Knapsack, and etc. These algorithms are still serving millions of customers every day; Obtained eight patents.

Research Topics

Responsible AI: Fairness and Ethical AI, Large Language Model Governance, Watermarking

Natural Language Processing: Large Language Models, Text Generation, Biomedical NLP, Named Entity Recognition, Knowledge Integration

Machine Learning: Stability Analysis, PAC Theory, Explainable Machine Learning, Language Model Analysis, Regularization, Optimization

Biomedical Applications: Digital Health, Disease Surveillance, Epidemiology

Publications

Under Review

Towards Trustworthy Watermarking for Large Language Models
Ziaho Fu, Chris Russell
Under Review. 2024

A Stability Analysis of Fine-Tuning a Pre-Trained Model
Zihao Fu, Anthony Man-Cho So, Nigel Collier
arXiv preprint arXiv:2301.09820. 2023

Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder
Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, Nigel Collier
arXiv preprint arXiv:2304.04052. 2023

Peer Reviewed

Evaluating Model Explanations without Ground Truth
Kaivalya Rawal, **Zihao Fu**, Eoin Delaney, Chris Russell
Proceedings of FAccT. 2025

OxonFair: A Flexible Toolkit for Algorithmic Fairness
Eoin Delaney, **Zihao Fu**, Sandra Wachter, Brent Mittelstadt, Chris Russell
Proceedings of NeurIPS. 2024

BAND: Biomedical Alert News Dataset
Zihao Fu, Meiru Zhang, Zaiqiao Meng, Yannan Shen, David Buckeridge, Nigel Collier
Proceedings of AAAI. 2024

Biomedical Named Entity Recognition via Dictionary-based Synonym Generalization
Zihao Fu, Yixuan Su, Zaiqiao Meng, Nigel Collier
Proceedings of EMNLP. 2023

On the Effectiveness of Parameter-Efficient Fine-Tuning
Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, Nigel Collier
Proceedings of AAAI. 2023

Repetition In Repetition Out: Towards Understanding Neural Text Degeneration from the Data Perspective
Huayang Li, Tian Lan, **Zihao Fu**, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, Yixuan Su
Proceedings of NeurIPS. 2023

COFFEE: A Contrastive Oracle-Free Framework for Event Extraction
Meiru Zhang, Yixuan Su, Zaiqiao Meng, **Zihao Fu**, Nigel Collier
Proceedings of ACL MATCHING. 2023

BioCaster in 2021: Automatic Disease Outbreaks Detection from Global News Media
Zaiqiao Meng, Anya Okhmatovskaia, Maxime Polleri, Yannan Shen, Guido Powell, **Zihao Fu**, Iris Ganser, Meiru Zhang, Nicholas B King, David Buckeridge
Bioinformatics. 2022

Open Domain Text Generation
Zihao Fu
PhD Thesis. PQDT - Global. 2021

A Theoretical Analysis of the Repetition Problem in Text Generation
Zihao Fu, Wai Lam, Anthony Man-Cho So, Bei Shi
Proceedings of AAAI. 2021

Open Domain Event Text Generation
Zihao Fu, Lidong Bing, Wai Lam
Proceedings of AAAI. 2020

Partially-Aligned Data-to-Text Generation with Distant Supervision
Zihao Fu, Bei Shi, Wai Lam, Lidong Bing, Zhiyuan Liu
Proceedings of EMNLP. 2020

Dynamic Topic Tracker for KB-to-Text Generation
Zihao Fu, Lidong Bing, Wai Lam, Shoaib Jameel
Proceedings of COLING. 2020

Unsupervised KB-to-Text Generation with Auxiliary Triple Extraction using Dual Learning
Zihao Fu, Bei Shi, Lidong Bing, Wai Lam
Proceedings of AACL. 2020

Fact Discovery from Knowledge Base via Facet Decomposition
Zihao Fu, Yankai Lin, Zhiyuan Liu, Wai Lam
Proceedings of NAACL. 2019

Word Embedding as Maximum a Posteriori Estimation
Shoaib Jameel, **Zihao Fu**, Bei Shi, Wai Lam, Steven Schockaert
Proceedings of AAAI. 2019

Learning Domain-Sensitive and Sentiment-Aware Word Embeddings
Bei Shi, **Zihao Fu**, Lidong Bing, Wai Lam
Proceedings of ACL. 2018

Learning Sentimental Weights of Mixed-gram Terms for Classification and Visualization
Tszhang Guo, Bowen Li, **Zihao Fu**, Tao Wan, Zengchang Qin
Proceedings of PRICAI. 2016

Pilot Behavior Modeling Using LSTM Network: A Case Study
Yanan Zhou, **Zihao Fu**, Guanghong Gong

Proceedings of ASC. 2016

Explicit Moment Integration Algorithm and Its Application
Zihao Fu, Gong Guanghong
Proceedings of JBUAA. 2015

Research on the Optimization Methods of the Blended-Wing-Body Aircraft
Zihao Fu
Master Thesis. Beihang University. 2015

Patents

Text Information Clustering Method and Text Information Clustering System
Zihao Fu, Kai Zhang, Ning Cai, Xu Yang, Wei Chu
Alibaba Group Holding Limited. WO2017148267A1. Patent, 2018

Method and Apparatus for Abnormal Access Detection
Zihao Fu, Kai Zhang, Ning Cai, Xu Yang, Wei Chu
Alibaba Group Holding Limited. WO2017124942A1. Patent, 2017

Abnormal Access Detection Method and Equipment
Zihao Fu, Kai Zhang, Ning Cai, Xu Yang, Wei Chu
Alibaba Group Holding Limited. CN106982196B. Patent, 2017

A Kind of Distribution Method and System of Virtual Resource
Xu Yang, **Zihao Fu**, Kai Zhang
Alibaba Group Holding Limited. CN109285015A. Patent, 2019

A Kind of Customer Service Dialogue Clustering Method and Device
Kai Zhang, Ning Cai, Xu Yang, **Zihao Fu**
Alibaba Group Holding Limited. CN107341157A. Patent, 2017

Method for Training Model Using Training Data and Training System
Bin Dai, Shen Li, Xiaoyan Jiang, Xu Yang, Yuan Qi, Wei Chu, Shaomeng Wang, **Zihao Fu**
Alibaba Group Holding Limited. WO2017143914A1. Patent, 2017

A Kind of Feature Selection Method and Device
Yan Xi, Ke Zhang, Shukun Xie, Jun Huang, **Zihao Fu**, Qiangpeng Yang, Wenpeng Li, Xiaoguang Wang, Zhouhua Yu
Alibaba Group Holding Limited. CN107169571A. Patent, 2017

Feature data processing method and device
Bin Dai, Shen Li, Xiaoyan Jiang, Xu Yang, Yuan Qi, Wei Chu, Shaomeng Wang, **Zihao Fu**
Alibaba Group Holding Limited. US11188731B2. Patent, 2018

Invited Talks

OxonFair: A Flexible Toolkit for Algorithmic Fairness. Future blood testing network+, Henley Business School. 2024
On the Effectiveness of Parameter-Efficient Fine-Tuning. OII, University of Oxford. 2024
BAND: Biomedical Alert News Dataset. World Health Organization (WHO). 2023
Introduction to Large Language Model: Technology, Challenges, and Prospects. CSSA Cambrige. 2023
Towards Trustworthy Language Models. University of Sheffield. 2023
Retrieval-Augmented Generation. University of Glasgow. 2023
Language Model for Science. Cambridge Centre for Data-Driven Discovery. 2023

Peer Reviewer

Reviewer: AAAI, ACL, CVPR, EACL, EMNLP, ICASSP, ICDM, ICML, NeurIPS, TKDE

External Reviewer: AAAI, CIKM, COLING, EMNLP, ICDM, IJCAI, SIGIR, TACL

Teaching Experience

University of Cambridge, Li18, Computational Linguistics, Guest Lecturer 2022–2023

- Engaged with 40 undergraduate students
- Delivered lecture on "Finite State Techniques" (2022), "N-Gram Model" (2023) and facilitated in-class Q&A

CUHK, FTEC 5510, Advanced Financial Infrastructure, Teaching Assistant 2021–2022

- Supported learning for 80 master students
- Organized final report session and conducted tutorials on financial infrastructure systems

CUHK, FTEC 5530, Quantitative and Algorithmic Trading, Teaching Assistant 2020–2021

- Guided 50 master students through 10 tutorial sessions
- Taught practical use of KDB, trading strategies, trading signals, factor model
- Created and graded homework, contributed to final exam grading

CUHK, ENGG 2780B, Statistics for Engineers, Teaching Assistant 2020–2021

- Assisted 130 undergraduate students
- Conducted statistical exercise tutorials, graded homework and final exams

CUHK, SEEM 4610, Supply Chain Management, Teaching Assistant 2020

- Facilitated learning for 30 undergraduate students
- Conducted tutorials on uncertainty management, newsvendor model, revenue management and graded final exam

CUHK, SEEM 3460, Computer Processing System Concepts, Teaching Assistant 2017–2020

- Guided 90 undergraduate students through Linux/GCC tutorials
- Designed GUI-based card playing project requiring student implementation of core code and developed corresponding grading system
- Supervised programming experiments, graded homework, and final exams

CUHK, SEEM 4540, Open Systems for E-Commerce, Teaching Assistant 2017–2019

- Assisted 15 undergraduate students through tutorials on PHP, SSL, phpMyAdmin, SQL
- Graded homework and final exams

Student Mentoring

University of Cambridge, Ph.D. student, Biomedical Event Extraction (EPI-AI Project) 2022–Now

- Conduct weekly meetings; Develop research ideas; Answer development questions; Polish papers.
- Have a paper accepted at the MATCHING workshop.

Nara Institute of Science and Technology, Ph.D. student, Generation Repetition Problem 2022–Now

- Conduct weekly meetings; Answer development-related queries; Direction suggestion; Polish papers.
- Published a research paper in NeurIPS 23.

University of Cambridge, Ph.D. student, Large Language Model for Healthcare 2023–Now

- Conduct weekly meetings; Develop research ideas; Literature review.
- Have an open source project **Visual Med-Alpaca** online; Preparing a paper for conference.

Selected Honors and Awards

AAAI Travel Grants	2021
Postgraduate Studentship	2017–2021
Outstanding Postgraduate Students	2015
Outstanding Graduate Students	2012
Second Prize of National Undergraduate Electronic Design Contest	2012
Second Prize of National Mathematical Contest in Modeling	2011
Beihang Scholarship (4 time)	2008–2012

Skills

Programming Languages: C/C++, Python, Javascript, Matlab, Fortran
Tools/Framework: Pytorch, Linux, Distributed Computing, LaTeX, Fairseq, OpenNMT, Elasticsearch, transformers, Theano
Natural Languages: English (IELTS 7.5), Chinese (Native)

Projects

Inflammatory Arthritis Prediction: Advancing machine learning to achieve real-world early detection and personalised disease outcome prediction of inflammatory arthritis. (EPSRC, EP/Y019393/1, £619,660)

Trustworthiness Auditing for AI: This project examines the social and institutional norms, legal mandates, ethical values, and technical constraints guiding the development and governance of trustworthy AI systems. Develop the necessary evidence base and tools to assess the efficacy of AI accountability toolkits across different systems, domains, and use cases.

EPI-AI: Automated Understanding and Alerting of Disease Outbreaks from Global News Media. (ESRC, ES/T012277/1, £491,373)

BioCaster: A system that tracks disease outbreaks using data from various sources and visualizes geographical disease spread.

PaperArxiv: A tool for managing academic papers, capable of extracting key information from PDFs.

StreamTask: A Python tool for managing big data multi-processing pipelines.

CAM-Tool: A tool for managing and assigning tasks across multiple machines.

BitTrader: A platform for real-time crypto coin trading, supporting backtesting, strategy development, and API operations.

CSTL: A Python wrapper for C++ STL containers, designed to mitigate memory leakage problems in Pytorch DataLoader.