Rao Ziyang

PhD Candidate in Natural Language Processing
International Max Planck Research School / University of Tübingen
ziyang_rao@outlook.com
https://www.ziyangrao.com/

ABOUT ME

I am an incoming PhD candidate in the Health NLP Lab of the University of Tübingen, working under the supervision of Prof. Carsten Eickhoff and Dr. Harry Scells. Previously, I got my MPhil. degree in Artificial Intelligence from HKUST (GZ) supervised by Prof. Hui Xiong and Dr. Xuming Hu.
My research interests lie in **information retrieval, mechanistic interpretability, and multimodal LLMs**. I am exploring the internal knowledge representation and training dynamics of LLMs. I also explore the applications of modern IR methods in high-stakes domains such as healthcare and medicine. Beyond academic pursuits, I actively engage in AI startups, industrialization, and knowledge transfer.
I am seeking a short-term internship opportunity since my PhD enrollment has been deferred due to visa check.

EDUCATION

International Max Planck Research School for Intelligent Systems Tübingen, Germany
PhD candidate in Natural Language Processing (deferred)

- Jointly hosted by the Max Planck Institute and the University of Tübingen
- **Supervisors:** Prof. Carsten Eickhoff and Dr. Harry Scells

Hong Kong University of Science and Technology (Guangzhou) Guangzhou, China
MPhil. in Artificial Intelligence Sep. 2023 — Oct. 2025

- **CGPA:** 4.0 / 4.3
- **Supervisor:** Prof. Hui Xiong, Associate Vice President, AAAI, AAAS, and IEEE Fellow
- **Co-supervisor:** Dr. Xuming Hu, Assistant Professor
- Postgraduate scholarship award

Renmin University of China Beijing, China
BSc. in Economic Statistics Sep. 2019 — Jul. 2023

- **CGPA:** 3.4 / 4.0
- Outstanding Graduation Thesis of Renmin University of China **(top 5 of the cohort)**

National University of Singapore Singapore, Singapore
Summer School May. 2022 — Jul. 2022

- Specialised in data mining and visual computing

RESEARCHES

A Mechanistic View on Knowledge Permeation in LLMs Sep. 2025 — present
Co-Lead Researcher

- Revealed a consistent pattern in finetuning LLMs on new knowledge: generalization requires significantly more training than memorization, a phenomenon similar to **grokking**.
- Hypothesis: fine-tuning does not create novel reasoning circuits. Instead generalization emerges when new factual knowledge successfully ”permeates” into and activates pre-existing computational circuits.
- Supported by **mechanistic interpretability** evidence, e.g., activation patching, and the behaviors of attention heads.

Modeling and Interpreting Information Flow in Visual Language Models Apr. 2025 — Sep. 2025
Lead Researcher

- MPhil. thesis, preparing for conference submission.
- Employed **sparse autoencoders (SAEs)** in the residual stream of multiple VLMs to model the internal information flow, extracted **human-understandable** visual and text concepts and revealed the concept hierarchy inside VLMs.
- Identified **modality-specific features** with entropy metrics and interpreted the dynamics of **modality gap** between visual and text with feature behavior.
- Located the fine-grained origin of **visual hallucinations** with activation patching on extracted features.

A Dataset and Benchmark for Event Camera ISP May. 2024 — Sep. 2024
Contributing Researcher

- Accepted by **ICLR 2025**.
- Proposed a new task of event camera guided image signal processor (event-ISP), collected a large scale pixel-level aligned RGB-event signal paired dataset and benchmarked SOTA baselines on our dataset.
- **Contributions:** collection and organization of the dataset, development of benchmarking pipeline and reproduction of baseline models, writing of the experiment and analysis section.

WORK EXPERIENCES

Guangzhou QiWu Co.,Ltd. an AI startup Guangzhou, China
Founding member Aug. 2024 — present

- Responsible for **business plan pitching** and product-market fit.
- Provide AI solutions including 3D design and smart home system or our business partners.

HKUST Fok Ying Tung Research Institute Guangzhou, China
Research Assistant Oct. 2022 — Aug. 2023

- **Advisor:** Prof. Hui Xiong, Associate Vice President, AAAI, AAAS, and IEEE Fellow.
- Participated in a **big data mining** project to develop a large-scale academic talent network sponsored by Tencent.
- Developed an information parser with **NLP techniques** (e. g. word2vec, BERT-based models) to retrieve structured data from the raw crawl results. **before ChatGPT api launch*

PROJECTS

Knowledge Graph Analysis on Bloomberg Tech News Oct. 2023 — Nov. 2024
Data Mining Project

- Crawled Bloomberg news articles and built a **knowledge graph** from raw crawl.
- Performed data mining on the graph (e. g. communities detection, shortest path search) and analyzed with finance domain expert for further interpretation.
- Explored the potential capabilities of LLMs to reason and perform analysis on knowledge graph.

Intelligent Production Line Control System Aug. 2023 — Aug. 2024
Industrial Project

- Cooperated with GJSS Co, Ltd. (a Japanese steel sheet manufacturer) to improve the efficiency of the steel sheet production line.
- Developed multiple models for different stages of production and deployed the solution on real production environment.

Large Language Model for Spatio-Temporal Graph Oct. 2023 — Mar. 2024
Unsuccessful Research Project

- Explored the potential of LLMs to handle **spatio-temporal graph (STG)** with no pretrained encoder available.
- Trained a new **STG tokenizer** to encode STGs into the token space of GPT-2, then finetuned GPT-2 with **soft prompt** and **adapter** modules.
- Performance not satisfying, but gained key insights about multi-modality potential of LLMs for further research.

A Simple Framework for Contrastive Learning on Spatio-Temporal Graph Prediction Sep. 2022 — Mar. 2023
BSc. Graduation Thesis

- Identified the difficulties of designing positive pairs for **contrastive learning** on spatio-temporal graph (STG) models.
- Designed a novel contrastive framework for STG models which utilize a simple Dropout layer instead of carefully designed augmentation methods to automatically generate positive pairs for contrastive task.
- The framework could be integrated in a plug-and-play manner and significantly boost the performance of most baselines.

ACADEMIC SERVICE

- **Conference Reviewer:** ICLR 2025