ZIHAN ZHANG

✉ zhangzihan_96@outlook.com · ☎ (+61) 0401-115-299 · Homepage · Scholar

SUMMARY

- **Motivated PhD candidate** with expertise in Natural Language Processing (NLP) and Large Language Models (LLMs), supported by both academic research and industry experience.
- **Skilled in designing, implementing, and optimizing AI models**, with hands-on proficiency in deep learning frameworks like PyTorch, HuggingFace, Langchain, LlamaIndex, and machine learning libraries such as scikit-learn, Pandas.
- **Experienced in cloud-based AI workflows** on AWS (e.g., S3, SageMaker, Lambda) with infrastructure tools like Terraform and Docker for model training, evaluation, and deployment.
- **Knowledgeable in MLOps and CI/CD pipelines**, ensuring streamlined processes for data preparation, model training, and deployment.
- **Collaborative and independent worker** with excellent communication skills, adept at translating research into practical, production-ready AI solutions.

EDUCATION

University of Technology Sydney, Sydney, Australia — Mar 2021 – present

Industry-based (TPG Telecom) *Ph.D. student* in Computer Science, UTS NLP Group

- Research: Knowledge Updating in Large Language Models; Retrieval-Augmented Generation (RAG)
- Supervisor: Prof. Ling Chen, and collaborated with Dr. Meng Fang from University of Liverpool

University of Melbourne, Melbourne, Australia — Jul 2018 – Dec 2020

Master in Software Engineering

- GPA: 83/100 (Top-5%, First Class Honours / High Distinction)
- Awards: Dean's Honours List (2019 & 2020), Liz Haywood Award (2020)

University of British Columbia, Vancouver, Canada — Jul 2017 – Aug 2017

Summer Exchange Program in Electrical and Computer Engineering (ECE)

China Pharmaceutical University, Nanjing, China — Sep 2014 – Jun 2018

Bachelor in Information System and Information Management

WORKING EXPERIENCE

TPG Telecom Sydney, Australia — Mar 2021 – Dec 2023

Student Researcher Fulltime

Working on text-based telco data and transforming them into actionable insights to drive business.

- **NPS survey topic modelling** - propose unsupervised clustering-based topic modelling to find latent topics among customer NPS feedback using textual embeddings from **BERT**, **SBERT** and **SimCSE**
  Outcomes: research paper [5]; improved topic coherence by **12%** on public datasets and **3.7%** on proprietary data; reduced manual work in extracting keywords insights by about **80%**
- **Webchat & Call Centre dialogue analysis** - study on the transformation of millions of raw dialogue data between customers and call centre agents into AI-driven telco service chatbots
  Outcomes: research paper [4]; finetuned **GPT-2** and **T5** models as Proof of Concept (POC) chatbots; reduced human annotation of new dialogue state data by **86%** on public datasets and **61%** on proprietary data while achieving comparable dialogue state tracking performance
- **Market offer engine** - automatically scrape, collect, transform, and analyse market offer data on the Internet for promptly making in-house product pricing strategies
  Outcomes: built web scrapers with Beautiful Soup and Selenium; deployed end-to-end pipelines using AWS

Step Functions; unify **unstructured** data (HTML text, posters, PDFs) into **structured** tables using Camelot and RegEx; stored in S3 for serving downstream analysis

RESORTer Melbourne, Australia Nov 2019 – Mar 2020

Front-end Software Developer Intern

- Refactor and develop the Lesson module in the resort web application using React.js and Material-UI
- Key outcomes: simplified web UI and work-flow logic; improved rendering performance

Publications

[1] **RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering**. **Zihan Zhang**, Meng Fang, and Ling Chen. *Findings of the Association for Computational Linguistics* (**ACL, Findings**), 2024
TL;DR: An open-domain question-answering dataset is proposed for adaptive retrieval-augmented generation (RAG). We evaluate and analyse state-of-the-art models and methods and provide an improved prompting-based method without calibration or additional training.

[2] **How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances**. **Zihan Zhang***, Meng Fang*, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2023
TL;DR: A comprehensive review of recent methods in aligning large language models (LLMs) with the ever-changing world knowledge without re-training from scratch, including knowledge editing, continual learning, and retrieval-augmented generation.

[3] **CITB: A Benchmark for Continual Instruction Tuning**. **Zihan Zhang**, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. *Findings of the Association for Computational Linguistics* (**EMNLP, Findings**), 2023
TL;DR: We propose continual instruction tuning (CIT) to continuously adapt language models to new NLP tasks and facilitate knowledge transfer without catastrophic forgetting.

[4] **Turn-Level Active Learning for Dialogue State Tracking**. **Zihan Zhang**, Meng Fang, Fanghua Ye, Ling Chen, and Mohammad-Reza Namazi-Rad. *Conference on Empirical Methods in Natural Language Processing* (**EMNLP**), 2023
TL;DR: A turn-level active learning framework is proposed for efficient data annotation for task-oriented dialogue state tracking (DST). We achieve comparable DST performance using significantly less annotated data via weakly-supervised training.

[5] **Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics.** **Zihan Zhang**, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. *Conference of the North American Chapter of the Association for Computational Linguistics* (**NAACL**), 2022
TL;DR: We directly cluster high-quality sentence embeddings with the proposed word selection method for more coherent and diverse topic modelling as an alternative to traditional neural topic modelling.

Academic Services

Conference peer reviewer:

- ACL 2023, EMNLP 2022-2023, EACL 2023, ACL Rolling Review 2023-2024
- NeurIPS 2024, ICLR 2025

Skills & Certificates

Languages: Chinese (native), English (fluent)
Programming: Python, SQL, Spark, JavaScript, Java
Libraries & Services: PyTorch, HuggingFace, AWS (S3, SageMaker, Redshift), Databricks, Scikit-learn
Software & Tools & Management: Git, Linux, Docker, Agile, Scrum, Confluence, Jira, LaTeX
Certificates:

- AWS Cloud Practitioner
- (Databricks) Large Language Models: Application through Production
- Deep Learning Specialization (Coursera by Andrew Ng)

ℹ Reference

Reference available on request.