Zekun Wu

AI Research Scientist · Safety, Auditing, and Governance of Agentic AI Systems

Canary Wharf – E22 2AD – London, UK

+44 07541369300 wu981526092@163.com

981526092.github.io/zekunwu.github.io zekun-wu-1190091a3 xurQ_DgAAAAJ

Work Experience

Holistic AI **London, UK**

AI Research Scientist *(Full Time)* *2023.05–Present*

- Lead architect of the **Agent Graph project** for auditing multi-agent AI systems, implementing a monitoring platform that enables prompt simulation, adversarial testing, and causal analysis through knowledge graph generation, with interactive visualization of system components and automated log analysis capabilities.
- Developed **Agentic AI Safety tools** for auditing, monitoring, redteaming and benchmarking; integrated Agentic AI driven automated workflows into the AI Governance Platform.
- Supervised **20+ projects** and a **10+ member team** across Holistic AI, Stanford, Oxford, UCL, and UC Berkeley—resulting in **15+ publications** (NeurIPS, EMNLP, NAACL, etc.).
- Delivered AI audits for **10+ clients** (e.g., Unilever, Michelin, Writer AI), covering compliance with the EU AI Act, NYC Bias Audit Law, and more.
- Main Organizer of the **Holistic AI x UCL Hackathon**, a two-day in-person hackathon for 100+ participants. hackathon.holisticai.com/
- Hosted the **"LLMs Bias Detection" webinar** with 300+ attendees. holisticai.com/events/bias-detection-in-large-language-models | youtube.com/watch?v=F9UecDrDo_8

OECD.AI **Remote**

AI Policy Researcher *(Part Time)* *2024.08–Present*

- Conducting R&D on a **Policy Research RAG System, AIR**, focusing on architecture enhancement, evaluation framework, and drafting technical report. oecd.ai/en/about-air
- Leading automated validation and enhancement of the **OECD.AI Catalogue of Tools & Metrics** for Trustworthy AI in collaboration with OECD experts. oecd.ai/en/catalogue/overview

University College London **London, UK**

Postgraduate Teaching Assistant *2024.09–2025.01*

- Teaching Assistant for **COMP0173 "AI for Sustainable Development"** (Master's Level 7), delivering lectures and leading lab sessions.
- Guest lecturer for **COMP0195 "Accountable, Transparent, and Responsible AI"** (Master's Level 7), presenting on Generative AI Audit.

EU AI Office GP AI Code of Practice **Remote**

Drafting Committee Member *2024.09–Present*

- Drafting the inaugural **EU General-Purpose AI Code of Practice**, contributing to Working Group 2 on systemic risk and participating in plenaries and multi-stakeholder consultations focused on EU AI Act compliance.

SeeTalent **London, UK**

Founder Engineer *(Freelance)* *2023.12–Present*

- Partnered with UCL and Goldsmiths psychologists to develop **Agent-based psychometric analysis tools** and a system for dynamic group interviews simulation. seetalent.ai/#about

Seele Art Labels (Online Art Collection Studio) **Online**

Founder *2021.01–2023.12*

- Founded and operated an online art studio across Xianyu, Taobao, and WeChat, generating **2.5M+ RMB in revenue** with a **12–15% profit margin** through curated portfolios and cost-reducing studio collaborations (Xianyu: SEELE_ART).

DeepBlue Technology **Shanghai, China**

Machine Learning Engineer Intern *2020.08–2020.09*

- Developed ML pipelines for vending machines using MQTT and the MVP framework.

Jin Xin Rong Tian Technology **Shenzhen, China**

Software Engineer Intern *2020.04–2020.05*

- Developed automated data pipelines using RPA, built a Flask-based platform, and customized the Landray EKP system to streamline enterprise operations for FangDD.

Education Background

University College London **London, UK**

PhD in Computer Science *(Part-Time)* *2024.01–Present*

- Researching **"Sustainable and Responsible Agentic AI"** to address emerging sustainability and ethical challenges in Agentic AI. Co-supervised by Dr. María Pérez-Ortiz, Dr. Adriano Koshiyama, and Dr. Sahan Bulathwela.
- Served as **reviewers** for ICLR 2025, ACL ARR MAY/FEB 2025, COLM 2025, and **Program Committee** for NeurIPS 2023/2024 (SoLaR Workshop), ICML 2024 (TiFA Workshop), EMNLP 2024 (CustomNLP4U Workshop).
- Presented research on LLM stereotypes at **UNESCO IRCAI 2024**. youtube.com/watch?v=fBxdDVTEoOo
- **Oxford Workshop** "AI Accountability and Trade Policy" (Sep 2024) at Oxford's Blavatnik School.
- **Ofcom MCC Initiative** (March 2024): 30-minute presentation on auditing LLMs, Manchester, UK.

University College London **London, UK**

MSc in AI for Sustainable Development *2022.09–2023.09*

- Graduated **ranked #1** in the program with **Distinction** (81.8% overall, 87% thesis), awarded **Dean's List** recognition for ranking within **top 5%** of Engineering Faculty.
- Received **£3000 Carbon Re Prize** for ranking first in academic excellence. ucl.ac.uk/computer-science/news/2024/may/top-students-ai-sustainable-development-msc-win-award-academic-excellence
- Featured as **Outstanding Graduate** in UCL website interview. ucl.ac.uk/computer-science/zekun-wu
- Thesis accepted for **NeurIPS 2023 SoLaR Workshop** poster presentation. nips.cc/virtual/2023/78925

University College London **London, UK**

BSc in Computer Science *2019.09–2022.09*

- Ranked **93rd globally** in 2020 ULTRA Coding Competition as 'Active Red Giraffe'.
- UCL Summer School in Data Science, awarded **Highest Overall Grade (85.65%, Grade A)**.
- BSc thesis on **'Sentiment Mapping and Matching between Audio and Text Representation'** supervised by Prof. John Shawe-Taylor—foundation of UCL Downstream AI Art Project delivering therapeutic audio-visual interaction in MRI waiting rooms. downstream-022f0a0fd7e3.herokuapp.com/

Wuhan Britain-China School **Wuhan, China**

A Level CIE *2017.09–2019.06*

- **A*, A*, A** in Economics, Mathematics, and Computer Science.
- **Super Econ Prize**, 2019 National Economics Challenge Finals; invited speaker, 2020.
- IB Economics TA; Co-led Badminton Club; Fundraising Director of Fungi Charity Club.

Academic Publications

- **JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models** | ***First Co-author*** | *EMNLP 2024 (Findings)* | DOI: 10.18653/v1/2024.findings-emnlp.184
- **From Text to Emoji: How PEFT-Driven Personality Manipulation Unleashes the Emoji Potential in LLMs** | ***Corresponding & Second Author*** | *NAACL 2025 (Findings) & NeurIPS 2024 Workshop* | aclanthology.org/2025.findings-naacl.265/
- **SAGED: A Holistic Bias-Benchmarking Pipeline for Language Models with Customisable Fairness Calibration** | ***Corresponding Author*** | *COLING 2025 (Oral Main)* | aclanthology.org/2025.coling-main.202/
- **HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications** | ***Corresponding & Second Author*** | *NAACL 2025 Industry Track & EMNLP 2024 Workshop* | aclanthology.org/2025.naacl-industry.79/
- **LibVulnWatch: A Deep Assessment Agent System and Leaderboard for Uncovering Hidden Vulnerabilities in Open-Source AI Libraries** | ***First Co-author*** | *ICML 2025 TAIG Workshop* | arxiv.org/abs/2505.08842
- **Towards Auditing Large Language Models: Improving Text-based Stereotype Detection** | ***First Co-author*** | *NeurIPS 2023 SoLaR Workshop* | arxiv.org/abs/2404.01768
- **Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks** | ***First Co-author*** | *NeurIPS 2024 EvalEval Workshop* | arxiv.org/abs/2410.11059
- **Bias Amplification: Language Models as Increasingly Biased Media** | *Corresponding & Second Author* | *Submitted to ACL ARR 2025* | arxiv.org/abs/2410.15234
- **HEARTS: A Holistic Framework for Explainable, Sustainable and Robust Text Stereotype Detection** | *Corresponding & Second Author* | *NeurIPS 2024 SoLaR & SafeGenAI Workshops* | arxiv.org/abs/2409.11579
- **THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models** | *Corresponding & Second Author* | *NeurIPS 2024 SoLaR Workshop* | arxiv.org/abs/2409.11353
- **Eliciting Personality Traits in Large Language Models** | *Third Author* | *arXiv preprint* | arxiv.org/abs/2402.08341
- **Advancing Multimodal Data Fusion in Pain Recognition: A Strategy Leveraging Statistical Correlation and Human-Centered Perspectives** | *Fourth Author* | *ACII 2024 AHRI Workshop* | arxiv.org/abs/2404.00320
- **CauSkelNet: Causal Representation Learning for Human Behaviour Analysis** | *Fourth Author* | *IEEE FG 2025 (Oral)* | arxiv.org/abs/2409.15564

🎓 **Complete publication list:** scholar.google.com/citations?user=xurQ_DgAAAAJ

Research Skills & Technical Expertise

- **Programming:** Python, R, TypeScript, SQL, CUDA
- **AI/ML Frameworks:** PyTorch, Transformers, HuggingFace, LangChain, LangGraph, CrewAI
- **Web & Database:** React, Shadcn, PostgreSQL, SQLite
- **Research Infrastructure:** GCP, Azure, Docker, MLOps, Weights & Biases, SLURM
- **Statistical Analysis:** SciPy, SPSS, A/B testing, Causal inference, Bias detection
- **Languages:** Native Chinese, Fluent English, Intermediate Japanese