Zekun Wu

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

in *zekun-wu-1190091a3*
xurQ_DgAAAAJ

Work Experience

2023.05– Present **AI Research Scientist (Full Time)**, *Holistic AI*, London, UK
- 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 for 100+ participants. Event Page
- Hosted the "LLMs Bias Detection" webinar with 300+ attendees. Event Page

2024.08– Present **AI Policy Researcher (Part Time)**, *OECD.AI*, Remote
- Conducting R&D on a Policy Research RAG System, AIR, focusing on architecture enhancement, evaluation framework, and drafting technical report. Report Link
- Leading automated validation and enhancement of the OECD.AI Catalogue of Tools & Metrics for Trustworthy AI in collaboration with OECD experts.

2024.09– 2025.01 **Postgraduate Teaching Assistant**, *University College London*, London, UK
- 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.

2024.09– Present **Drafting Committee Member**, *EU AI Office GP AI Code of Practice*, Remote
- 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.

2023.12– Present **Founder Engineer (Freelance)**, *SeeTalent*, London, UK
- Partnered with UCL and Goldsmiths psychologists to develop Agent-based psychometric analysis tools and a system for dynamic group interviews simulation.

2021.01– 2023.12 **Founder**, *Seele Art Labels (Online Art Collection Studio)*, Online
- 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).

2020.08– 2020.09 **Machine Learning Engineer Intern**, *DeepBlue Technology*, Shanghai, China
- Developed ML pipelines for vending machines using MQTT and the MVP framework.

2020.04–2020.05 **Software Engineer Intern**, *Jin Xin Rong Tian Technology*, Shenzhen, China
- 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

2024.01–Present **PhD in Computer Science (Part-Time)**, *University College London*, London, UK
- Researching "Sustainable and Responsible Agentic AI" to address emerging sustainability and ethical challenges in Agentic AI. This research is co-supervised by Dr. María Pérez-Ortiz, Dr. Adriano Koshiyama, and Dr. Sahan Bulathwela.
- Served as reviewers for ICLR 2025 Main Conference, ACL ARR FEB 2025, COLM 2025, and Program Committee for NeurIPS 2023 and 2024 (SoLaR Workshop), ICML 2024 (TiFA Workshop), and EMNLP 2024 (CustomNLP4U Workshop).
- Presented research on LLM stereotypes at UNESCO IRCAI 2024. Video Link.
- Oxford Workshop "AI Accountability and Trade Policy" (Sep 2024) Presented at Oxford's Blavatnik School, focusing on tiered access of AI System for accountability and compliance.
- Ofcom MCC Initiative (March 2024): Delivered a 30-minute presentation on auditing LLMs as part of Ofcom's MCC Educational Initiative, Manchester, UK.

2022.09–2023.09 **MSc in Artificial Intelligence for Sustainable Development**, *University College London*, London, UK
- Achieved the highest overall grade in the program, was placed on the Dean's List (5%), and graduating with a Distinction and scores of 81.8% overall and 87% for the thesis.
- Received the £3000 Carbon Re Prize for ranking first in academic Excellence within the program, as published on the UCL website: (News Link).
- Featured as an Outstanding Graduate in a column interview on the UCL website. (Link).
- Thesis paper was accepted for a poster presentation at the NeurIPS 2023 SoLaR Workshop under the guidance of Dr. Adriano Koshiyama and Dr. Sahan Bulathwela. (Paper Link)

2019.09–2022.09 **BSc in Computer Science**, *University College London*, London, UK
- Ranked 93rd globally in the 2020 ULTRA Coding Competition as 'Active Red Giraffe'.
- Completed UCL Summer School in Data Science and Big Data Analytics (2022.07), awarded Highest Overall Grade (85.65%, Grade A).
- BSc thesis, supervised by Prof. John Shawe-Taylor, on 'Sentiment Mapping and Matching between Audio and Text Representation'—developed neural pipelines for cross-modal sentiment alignment, forming the foundation of the UCL Downstream AI Art Project. The project, in collaboration with Freya Gabie, UCL Public Art, the British Library, and the Dementia Research Institute, delivers therapeutic audio-visual interaction in MRI waiting rooms. Demo: https://downstream-022f0a0fd7e3.herokuapp.com/

2017.09–2019.06 **A Level CIE**, *Wuhan Britain-China School*, Wuhan, China
- 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

- Additional works listed at **Google Scholar Link.**
- **JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models** (First Co-author). EMNLP 2024 (Findings). Link.
- **From Text to Emoji: How PEFT-Driven Personality Manipulation Unleashes the Emoji Potential in LLMs** (Corresponding Author). NAACL 2025 (Findings) and NeurIPS 2024 Behavioral ML Workshop. Link.

- **SAGED: A Holistic Bias-Benchmarking Pipeline for Language Models with Customisable Fairness Calibration** (Corresponding Author). COLING 2025 Oral Main. Link.
- **Towards Auditing Large Language Models: Improving Text-based Stereotype Detection** (First Co-author). NeurIPS 2023 SoLaR Workshop. Link.
- **Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks** (First Co-author). NeurIPS 2024 EvalEval Workshop. Link.
- **Bias Amplification: Language Models as Increasingly Biased Media** (Corresponding and Second Author). Submitted to ACL ARR 2025 May. Link.
- **HEARTS: A Holistic Framework for Explainable, Sustainable and Robust Text Stereotype Detection** (Corresponding and Second Author). Dual Acceptance at the NeurIPS 2024 SoLaR and SafeGenAI Workshops. Link.
- **THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models** (Corresponding and Second Author). NeurIPS 2024 SoLaR Workshop. Link.
- **HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications** (Corresponding and Second Author). NAACL 2025 Industry Track and EMNLP 2024 Workshop CustomNLP4U. Link.