Shuang Ao

shuang.ao@open.ac.uk @ShuangAo92 LinkedIn
https://github.com/AoShuang92 https://aoshuang92.github.io/

Areas of Expertise

VLLMs | Responsible AI | Trustworthy LLMs| Uncertainty Quantification | Automatic Failure Detection

Employment History

Sep 2024 – Present **Post-doctoral Researcher in AI for Good,** University of Southampton, Southampton, UK.

Jul 2020 – Jul 2021 **NLP Research Intern,** Doti Health Limited, London, UK.

Sep 2020 – Dec 2020 **Data Science Intern,** Fellowship.ai, London, UK.

Mar 2017 – Jan 2020 **Linguistic Specialist for Curriculum Design,** EduGrove Mandarin Enrichment Centre, Singapore.

Research Experience

Oct 2024 – Present **Research Fellow** Responsible AI, University of Southampton, Southampton, UK.
Research: Responsible LLM-based Multi-Agent System.
(i) Proposing trustworthy LVLM for better failure detection and self-reasoning for LLMs in medical visual question-answering task.
(ii) Designing robust distance-guided method for safety alignment in adaptation for LLMs.
(iii) Proposing novel methods for uncertainty quantification for LLM-based Multi-Agent System in topic of climate change and finance.

Oct 2021 – Sep 2024 **PhD Student,** Knowledge Media Institute (KMi), The Open University, Milton Keynes, UK.
Research: Responsible AI with large vision-language models for safety-critical tasks.
(i) Designing contrastive semantic similarity between sampled generations by LLMs via utilizing CLIP, to estimate the uncertainty of LLMs (e.g., ChatGPT, LLaMA, OPT).
(ii) Utilizing miscalibration information to integrate into state-of-the-art calibration techniques with training and post-hoc methods, for better trustworthiness and automatic failure detection for deep learning models.
(iii) Designing novel evaluating metrics as the compensation of accuracy to better quantify model uncertainty.
Supervision: 1 PhD student and 1 BSc student in the project of GenAI in Education, and 1 BSc student at Summer Scholarships for Black Students.
Collaboration: DynAIkon Ltd, Royal Society for Blind Children; The University of Edinburgh.

Research Experience (continued)

Jul 2020 – Jul 2021 **NLP Healthcare Research Intern,** Doti/Datum Health Limited, London.
Research: Developing NLP QA datasets and model with transformer.
(i) Designing well-calibrated inductive transfer learning language modeling to improve the feature representation and extract more distinctive information.
(ii) Building professional clinical dialogue dataset from scratch for prompting responses on medical pain treatment for back pain.
Collaboration: Clinicians from Babylon Health Ltd.

Sep 2020 – Dec 2021 **Data Science Intern,** Fellowship.ai, London.
(i) Proposing a novel loss function by smoothing the label to tackle the class imbalanced issue.
(ii) Creating a joint representation with multi-label classification networks on fashion images.

Education

Oct 2021 – Present **Ph.D., The Open University** Knowledge Media Institute (KMi).
Thesis title: *Developing Trustworthy AI: Uncertainty Quantification and Failure Detection in Large Vision-Language Models.*

2020 – 2023 **Summer School**
Oxford Machine Learning Summer School (OxML), 2023, 2020.
Eastern Europe Machine Learning Summer School (EEML), 2021-2022

2019 – 2022 **MOOC**
Research Software Engineering with Python, **Alan Turing Institute**.
Build a Data Science Web App with Streamlit and Python, Databases and SQL for Data Science, Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization, Deep Neural Networks with PyTorch, AI Capstone Project with Deep Learning, Machine Learning with Python, Python Data Structures, Programming for Everybody, Introduction to Programming with MATLAB, **Coursera**.
Learn PyTorch for Natural Language Processing, Learn HTML, **Udemy**.

Aug 2015 – Jun 2016 **Master in Language Studies**, National University of Singapore, Singapore.
Thesis title: *A Corpus Linguistic Study on the Influence of Language Policies in Singapore*
Major in: (i) Theoretical Linguistics; (ii) Corpus Linguistics; (iii) Acoustic Analysis.

Sep 2011 – Jun 2015 **Bachelor in Teaching English as a Second Language**, Hong Kong Baptist University, China.
Thesis title: *Most Frequent Words: the Comparison of High School English Textbook and the Natural English Corpus*
Major in: (i) Psycholinguistics; (ii) Statistics; (iii) Fairness and Ethics.

Grant Application

2024 **EPSRC Open Fellowship (in submission)**, UKRI, UK.
Interactive disease trend prediction via utilizing multimodal LLMs.

Grant Application (continued)

2023 **TIDAL Network+ Call5 (shortlisted)**, EPSRC-funded, UK.
Multi-sensory Storybook integrating AI and haptics for Children with Sight Impairment.

Skills

Coding PyTorch, Python, TensorFlow, MatLab, Java, C/C++
Databases SQL.
Web Dev HTML, JavaScript.

Awards

2023 **Student scholarship**, the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).
Best Paper Award (nominated), AISafety-IJCAI 2023.
2022-2023 **Melete Award**, the first winner of the Melete Award for PhD students at the KMi, OU.
2022 **First-award Winner of Judge's Choice of 17th Poster Competition**. Awarded by The Open University.
2021 **PhD scholarship**. Awarded by The Open University.

Invited Talk / Public Engagement

Mar 2025 **Artifitial Intelligence Trustworthiness And Risk Assessment Scientific Seminars (ATRASS)**, Paris, France (online).

Aug 2023 **AISafety–IJCAI**, Macau, S.A.R, China.
Empirical Optimal Risk to Quantify Failure Detection for Model Trustworthiness.
Jul 2023 **The 39th Conference on Uncertainty in Artificial Intelligence (UAI)**, Pittsburgh, USA.
Two Sides of Miscalibration: Identifying Over and Under-Confidence Prediction for Network Calibration.
Apr 2023 **The 45th European Conference on Information Retrieval**, Dublin, Ireland.
Building Safe and Reliable AI systems for Safety Critical Tasks with Vision-Language Processing.
Nov 2022 **The 15th International Conference on Machine Vision**, Rome, Italy.
Confidence-Aware Calibration and Scoring Functions for Curriculum Learning.
Jun 2022 **CRC PhD Student Conference by The Open University**, Milton Keynes, UK.
Confidence-Aware Model Calibration for Classification.
Oct 2021 **4th International Conference on Natural Language and Speech Processing**, Trento, Italy.
Learning ULMFiT and Self-Distillation with Calibration for Medical Dialogue System.

Research Publications

1. S. Ao, J. Hu, Y. Dong, and G. Ramchurn, "Safe pruning lora: Robust distance-guided pruning for safety alignment in adaptation of llms," in *TTransactions of the Association for Computational Linguistics (TACL)*, 2025 [In submission].
2. S. Ao, S. Rueger, and A. Siddharthan, "Css: Contrastive semantic similarity for uncertainty quantification of llms," in *The 40th Uncertainty in Artificial Intelligence (UAI)*, PMLR, 2024.

3 S. Ao, "Building safe and reliable ai systems for safety critical tasks with vision-language processing," in *The 45th European Conference on Information Retrieval*, Springer Nature Switzerland, 2023, pp. 423–428.

4 S. Ao, S. Rueger, and A. Siddharthan, "Confidence-aware calibration and scoring functions for curriculum learning," in *Fifteenth International Conference on Machine Vision (ICMV 2022)*, Proc. SPIE, vol. 12701, 2023.

5 S. Ao, S. Rueger, and A. Siddharthan, "Empirical optimal risk to quantify model trustworthiness for failure detection," in *CEUR Workshop Proceedings.*, CEUR-WS, vol. 3505, 2023.

6 S. Ao, S. Rueger, and A. Siddharthan, "Two sides of miscalibration: Identifying over and under-confidence prediction for network calibration," in *The 39th Uncertainty in Artificial Intelligence (UAI)*, PMLR, 2023, pp. 77–87.

7 S. Ao and X. Acharya, "Learning ulmfit and self-distillation with calibration for medical dialogue system," in *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021).*, 2021.