Yupeng Chen

✉ yupeng.chen@eng.ox.ac.uk ☎ +44 07344370324

Education

University of Oxford *Oct 2025 – Ongoing*
DPhil in Engineering Science

- **Research Area:** Trustworthy AI, AI Safety, Machine Learning
- **Supervisors:** Dr. Adel Bibi, Prof. Philip Torr

The Chinese University of Hong Kong, Shenzhen *Sept 2021 – May 2025*
BEng in Computer Science and Engineering

- **GPA:** 3.85/4.0 (top 2%)
- **Advisors:** Prof. Baoyuan Wu, Prof. Hongyuan Zha
- **Coursework:** Programming (A), Data Structure (A), Operating System (A), Parallel Computing (A), Calculus (A), Optimization (A), Linear Algebra (A), Probability and Statistics (A), etc.

University of Oxford *Oct 2023 – Jun 2024*
Visiting Student

- **Average Score:** 85/100
- **Coursework:** Machine Learning, Computer Vision, Artificial Intelligence, Deep Learning in Healthcare, Design and Analysis of Algorithms, Geometric Deep Learning, etc.

Research Statement

My DPhil focuses on ensuring the robustness, reliability, and ethical behavior of AI agents operating individually or in collaboration. It investigates challenges such as adversarial robustness, unintended emergent behaviors, and coordination failures. I aim to develop safety mechanisms that ensure AI agents align with human values and contribute positively to society.

Publications

(* indicates equal contribution)

1. Kangran Zhao*, **Yupeng Chen***, Xiaoyu Zhang*, Yize Chen, Weinan Guan, Baicheng Chen, Chengzhe Sun, Soumyya Kanti Datta, Qingshan Liu, Siwei Lyu, Baoyuan Wu. "Mega-MMDF and DeepfakeBench-MM: Scaling and Benchmarking Multimodal Deepfake Detection." *Under Review*, 2025.
2. **Yupeng Chen**, Penglin Chen, Xiaoyu Zhang, Yixian Huang, Qian Xie. "EditBoard: Towards A Comprehensive Evaluation Benchmark for Text-Based Video Editing Models." *Association for the Advancement of Artificial Intelligence (AAAI)*, 2025.
3. **Yupeng Chen***, Xiaoyu Zhang*, Yixian Huang, Qian Xie. "Beyond English: Unveiling Multilingual Bias in LLM Copyright Compliance" *Preprint*, 2024.

Research Experience

Research Assistant at The Chinese University of Hong Kong, Shenzhen *Shenzhen, China*
Supervisor: Prof. Baoyuan Wu *Mar 2024 – Jul 2025*

- Topic: Deepfake Detection
- Led a project to build a large-scale, diverse multimodal deepfake dataset with over 1 million synthetic videos and audio clips, aimed at supporting robust multimodal detection research.
- Developed DeepfakeBench-MM, a comprehensive benchmark for multimodal deepfake detection with in-depth experiments on 10 detectors.
- Built and launched a real-time deepfake detection website from the ground up, enabling users to conduct deepfake detection with various models.

Independent Research Project *Shenzhen, China*
Supervisor: Dr. Qian Xie *Dec. 2024 – Mar 2025*

- Topic: Copyright Compliance in LLMs

- Identified language-specific biases in copyright compliance of large language models, demonstrating that copyrighted materials in certain languages are more vulnerable to probing.
- Conducted comprehensive evaluation on major API-based and open-source large language models, including GPT-4o, Claude-3.5-Haiku, and Llama-3-70B, for multilingual copyright adherence.

Research Project at University of Oxford *Oxford, UK*
Tutor: Dr. Qian Xie *Apr 2024 – Aug 2024*

- Topic: Video Editing
- Proposed the first comprehensive evaluation benchmark for text-based video editing, addressing 4 key dimensions—fidelity, consistency, execution, and style—with a total of 9 performance metrics.
- Developed 3 novel metrics to evaluate fidelity between edited videos and original videos/prompts, designed to closely reflect human perceptual judgments.
- Systematically evaluated 5 state-of-the-art (SOTA) video editing models, applying the benchmark to derive insights into model strengths and limitations.

Research Internship at ZTE Corporation *Nanjing, China*
Supervisor: Dr. Yaofeng Tu *Aug 2023 – Oct 2023*

- Topic: Data Preprocessing for LLMs
- Preprocessed training data for large language models by retrieving, cleaning, and formatting information from 100+ PDF documents.
- Developed a custom algorithm to efficiently extract content from PDFs and standardize it into JSONL format, which became a widely adopted tool within the team for streamlined data processing.

Awards and Scholarships

- SUISF Scholarship (50,000 RMB, 2025)
- AAAI-25 Student Travel Scholarship (2,000 USD, 2025)
- SDS Student Travel Support (5,000 RMB, 2025)
- Dean's List (2021-2022, 2022-2023)
- Academic Performance Scholarship (20,000 RMB, 2021-2022)
- Second Prize in China Undergraduate Mathematical Contest in Modeling (2022)

Skills

Programming: Proficient in Python (Pytorch, OpenCV). Familiar with C/C++, Java, HTML and JavaScript
Tools: GitHub, LaTeX, VS Code, Colab, Markdown

Extracurriculars

I enjoy writing and directing drama during my free time. I was president of the phantom drama club during my undergraduate studies, where I led a diverse team in staging three plays annually that were well received by the campus community.