Moy Zhangdie Yuan

www.moyyuan.com
LinkedIn | Google Scholar

moy@moyyuan.com
+44 (0) 750-881-8988

Education

- **University of Cambridge** — Cambridge, UK
  PhD, MPhil in Computer Science (Distinction ≈ 4.0 GPA) — *Oct 2020 – Oct 2025*
  - Supervisor: Prof. Andreas Vlachos
  - PhD Thesis: *Language Models as Forecasters*
- **University of Edinburgh** — Edinburgh, UK
  BSc in Computer Science (1st Class Honors ≈ 4.0 GPA) — *Sep 2016 – Jun 2020*

Technical Skills

- **Languages/Frameworks**: Python, Bash, PyTorch, Hugging Face (Transformers), LangChain, Pandas, NumPy, SQL
- **Infrastructure**: AWS (SageMaker, Bedrock, EC2, S3), Docker, Kubernetes, Slurm, Git, TensorBoard, Flask, Streamlit
- **Research Areas**: Large Language Models (LLM), Multimodal LLMs, Reasoning, Alignment (RLHF/DPO), Event Forecasting, AI Agents, Conversational AI, RAG, LLM Benchmarking, Hallucination Mitigation, Knowledge Graphs

Professional Experience

- **Amazon AGI** — Cambridge, UK
  Applied Scientist — *Oct 2025 – Present*
  - Spearheading the post-training speech modality architecture for Amazon's unified foundation models (powering products including Nova and Alexa+), enabling direct End-to-End Speech-to-Speech (S2S) and TTS capabilities.
  - Developing novel alignment pipelines, including SFT and DPO, and inference-time guidance strategies to mitigate cross-modal hallucinations and ensure fine-grained control over speech prosody and style.
  - Defining the research roadmap for large-scale audio generation like zero-shot voice cloning by conducting rigorous feasibility studies to select optimal methodologies, while collaborating with engineering teams to optimize the trade-off between performance, streaming latency, and engineering simplicity for scalable production deployment.
- **AWS AI** — California, USA
  Applied Scientist Intern — *Oct 2024 – Jan 2025*
  - Engineered a robust RAG and Chain-of-Thought (CoT) pipeline for the Health AI team to automate ICD-10 medical coding, integrating verification steps on top of SFT to mitigate critical hallucinations.
  - Achieved a 15% accuracy improvement while reducing inference latency by 70% by optimizing prompts and successfully distilling capabilities into smaller models compared to larger state-of-the-art baselines.
  - Validated system reliability and clinical robustness using MIMIC benchmarks and proprietary expert datasets.
- **A*STAR** — Singapore
  Research Assistant Intern — *Jun 2019 – Oct 2019*
  - Developed a controllable Seq2Seq paraphrasing system from scratch, creating a custom dataset and using novel codebook-based conditioning to rigorously evaluate tone and politeness modulation.
- **Chinese Academy of Sciences** — Chengdu, China
  Software Engineer Intern — *Jun 2017 – Oct 2017*
  - Engineered the backend of an internal management system using ASP.NET and optimized SQL retrieval pipelines, reducing query latency by over 10% for resource allocation tasks.

Honors, Awards & Service

- **Research & Academic**: AVeriTeC ERC Fully-funded PhD Grant (2021), A*STAR SIPGA Award (2019), Girton College Prize & Scholarship (2021), Trinity Hall Research Grant (2021).
- **Entrepreneurship**: Trinity Hall Lee-Yung Family Fund (2023), Trinity Hall Experiencing Entrepreneurship Award (2023), Cambridge Judge Business School EnterpriseTECH Alumni (2023).
- **Professional Service**: Reviewer/PC Member: NeurIPS, ICLR, ACL, EMNLP, NAACL, EACL, AKBC, and ARR (2022–Present); Invited speaker at academic and industry venues covering LLM Reasoning and AI Commercialization.

Academic Research Experience

- **CambridgeNLP, University of Cambridge** — Cambridge, UK
  Research Assistant — *Oct 2021 – Sep 2025*
  - Spearheaded research on LLM forecasting, investigating how models update probabilities via RAG and Bayesian reasoning; revealed critical failure modes including belief inertia and recency bias despite access to new information. Created comprehensive evaluation suites for model confidence and plausibility, revealing that calibration and reasoning reliability do not automatically improve with model scaling.
  - Developed automated fact-checking systems using Knowledge Graphs for zero-shot verification and designed *varifocal* question generation methods to decompose complex claims into verifiable sub-questions.
  - Released a large-scale dataset revealing that LLMs fail to handle surface perturbations in deductive reasoning, and proposed novel symmetry-aware objectives that significantly improve logical consistency.
  - Supervised and demonstrated for *Machine Learning and Real-world Data* and *Data Science* courses, providing in-depth technical mentorship on data pipelines and model implementation to undergraduates.
- **EdinburghNLP, University of Edinburgh** — Edinburgh, UK
  Research Assistant — *Jun 2018 – Oct 2018*
  - Conducted research applying language models to structure parsing including Abstract Meaning Representation in cross-lingual settings and launched the first public demo for the group's Discourse Representation Structure parser.
  - Taught *Processing Natural Languages* as a tutor, explaining theoretical concepts in logic and automata theory.

Selected Publications

** denotes equal contribution.*

LLM for Forecasting

- **[2025] Assessing Large Language Models in Updating Their Forecasts with New Information.**
  Zhangdie Yuan, Zifeng Ding, Andreas Vlachos.
  Submitted to The Forty-Third International Conference on Machine Learning (ICML).
- **[2025] FOReCAst: The Future Outcome Reasoning and Confidence Assessment Benchmark.**
  Zhangdie Yuan, Zifeng Ding, Andreas Vlachos.
  Proceedings of The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS).
- **[2024] PRobELM: Plausibility Ranking Evaluation for Language Models.**
  Zhangdie Yuan, Eric Chamoun, Rami Aly, Chenxi Whitehouse, Andreas Vlachos.
  Proceedings of The First Conference on Language Modeling (COLM).

LLM for Reasoning

- **[2025] Toward Reliable Clinical Coding with Language Models: Verification and Lightweight Adaptation**
  Zhangdie Yuan*, Han-Chin Shing*, Mitch Strong, Chaitanya Shivade.
  Proceedings of The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- **[2025] TCP: A Benchmark for Temporal Constraint-Based Planning.**
  Zifeng Ding*, Sikuan Yan*, **Zhangdie Yuan***, Xianglong Hu, Fangru Lin, Andreas Vlachos.
  Proceedings of The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- **[2024] Zero-Shot Fact-Checking with Semantic Triples and Knowledge Graphs.**
  Zhangdie Yuan, Andreas Vlachos.
  Proceedings of Knowledge Graph and Large Language Models at the 62nd Annual Meeting of the Association for Computational Linguistics (ACL).
- **[2023] Capturing (Anti-)Symmetry in Language Models with Symmetry-Aware Objectives.**
  Zhangdie Yuan, Andreas Vlachos.
  Computing Research Repository (CoRR).
- **[2023] Can Pretrained Language Models (Yet) Reason Deductively?**
  Zhangdie Yuan*, Songbo Hu*, Ivan Vulić, Anna Korhonen, Zaiqiao Meng.
  Proceedings of The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
- **[2022] Varifocal Question Generation for Fact-Checking.**
  Nedjma Djouhra Ousidhoum*, **Zhangdie Yuan***, Andreas Vlachos.
  Proceedings of The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).