Xiaoyu Yuan

Researcher in NLP & ML | LLM-human Hybrid Detection, AI Safety & Trustworthiness, Explainability

xiaoyuyuan19@gmail.com | xiaoyuyuan19.github.io/portfolio | Github | Linkedin

Education

University of Helsinki Helsinki, Finland

Master in Computer Science Track: Algorithms and Machine Learning *Aug. 2023 – Jun. 2025*

GPA: 4.38 Awarded **Full Scholarship** for the Master's Programme

Key Courses: Neural Networks & Deep Learning, Machine Learning in Molecular Biology, Advanced Course in Machine Learning, Information Retrieval, Data Analysis with Python, Big Data Platforms, Academic Writing (CEFR C1)

University of Oulu Oulu, Finland

Bachelor in Software Engineering (Double Degree) *Aug. 2019 – Jun. 2023*

GPA: 4.01 Conferred jointly by Nanjing Institute of Technology & University of Oulu

Key Courses: AI and Software Engineering, WEB Development Techniques, Software Modeling and Design

Work Experience

Human-AI RLHF Evaluation Tasker (Freelance) **Outlier.ai**

Remote (Freelance) *Jan. 2024 – Present*

- Conducted large-scale preference **evaluation of LLM outputs** across tasks (e.g., summarization, Q&A).
- Specialized in **detecting hallucination**, **evaluating reward models**, and prompting refinement.
- Contribute to **RLHF** and comparison alignment pipelines, focusing on Reward Modeling and cue-response pairing.
- Gained firsthand insight into weak supervision signals such as noisy rankings and response edits.
- Special focus on hybrid LLM-human text assessment, aligning with AI Safety and Trustworthiness goals.
- Developed prompt-based pipelines for distinguishing human vs. LLM generations.

Software Test Engineer Intern ***Nanjing Tongdahai Information Technology Co., Ltd.***

Nanjing, China *Jun. 2021 – Jul. 2021*

- Conducted **black-box testing** and analyzed 2000+ **UX feedbacks**
- Collaborated in **Agile sprints** and resolved bugs across platforms

Research Experience

High-Precision Visual SLAM for GI Gastrointestinal Navigation Helsinki, Finland

Master Thesis Research, University of Helsinki *Aug. 2024 – Present*

- Developed a localization framework in **GI environments**, analogous to anatomical tracking in medical imaging.
- Enhanced ORB-SLAM3 with deep learning for **biomedical image-based localization** in low-light, tissue-deformable **endoscopy scenes**—analogous to **anatomical tracking in pre-radiotherapy imaging**.
- Integrated Transformer depth estimation**,** leveraging **Foundation Models** for robust trajectory prediction.
- Developed a **dual-attention U-Net**, refining **SLAM-based** depth perception and localization accuracy.
- Experience with remote **GPU and HPC services**.

Ancient Character Recognition Website with Transformer-based OCR Nanjing, China

Bachelor Thesis Research, Nanjing Institute of Technology & University of Oulu *Jun. 2021 – Jun. 2023*

- Developed a **Transformer-based OCR system** with a custom dataset to improve model accuracy.
- Built a **real-time OCR pipeline** at HoumaOCR [YouTube Demo Link], with **Flask & JS-based interfaces** for live uploads and visualization.
- Experience in dataset annotation pipelines and OCR text authenticity checks, relevant to text detection workflows.
- Contributed to **two international conference papers, one software copyright, and one patent**.
- Recognized with the **1st Prize** in the Chinese Collegiate Computing Competition and the **2nd Prize** in the Jiangsu Province Bachelor Thesis Award.

Projects Experience

Explainable AI & LLM Alignment Projects

Preference Optimization in LLM: RLHF & DPO Exploration *Self-Initiated Research Project* 2025

- Explored **Direct Preference Optimization (DPO)** as an efficient alternative to reinforcement learning in **Large Language Models (LLM) alignment**.
- Conducted **mathematical derivations** of the DPO loss function, visualized **gradient flow and parameter updates**, and connected DPO with logistic regression.
- Compared the **DPO** pipeline with **RLHF + PPO** in terms of implementation complexity and stability.
- Created a **video presentation and demo slides** demonstrating model behavior and potential for deployment in human-in-the-loop settings [Video link].
- Insights gained in detecting preference consistency and hallucinations across hybrid LLM-human outputs, directly relevant to machine-generated text detection.

NLP: Sedimentological NER based on CRF, BiLSTM and BILSTM-CRF 2022

- Collaborated on the development of a **NER system** for specialized texts, extracting historical locations, geological periods, and material entities from the scientific literature.
- Benchmarking **BiLSTM-CRF & Conditional Random Fields (CRF)** for structured text recognition.
- Contributed to **dataset annotation** and **model evaluation**, improving entity recognition accuracy.
- Awarded **1st Prize** in the 1st "Prospective Cup" Meta-Intelligent Data Challenge (Track: Entity Recognition in Sedimentology Knowledge Map), held at the CENet2022 Conference, Haikou, China (Nov. 2022).

Enhancement Proposal for PANNZER via Protein Language Model *Self-Initiated Literature-Based Project* 2025

- Conducted an independent literature-driven dissection of the PANNZER pipeline for Protein Function Prediction.
- Analyzed the integration of **Suffix-Array-based** retrieval with **GO enrichment mechanisms**.
- Proposed a future research agenda focusing on representation learning (e.g., **ProtBERT, ESM**) and **IR techniques** (e.g., **Transformer IR, Burrows-Wheeler Transform optimizations**).
- Mapped workflow logic with visual **annotations**, raising questions for potential modular enhancements.
- Provided constructive suggestions to the original authors and discussed with Petri Törönen & Liisa Holm.
- This reflective study enhanced my **interdisciplinary understanding** of bioinformatics pipelines and provided a conceptual foundation for **designing interpretable models** in **Cancer Microenvironment** analysis.

Scalable Inference in Extreme Multi-Label Classification (XMC) 2025

- Replications and benchmarks of the **SPARTEX** and **SOTA XMC models** were performed to analyse performance **under memory and latency constraints**.
- **Sparse modes (block sparse softmax, fan-in sparse auxiliary header)** and their impact on the representation capability were investigated.
- Adaptive semantic **bottleneck** designs are proposed to improve the efficiency and generality of sparse inference.

Rule-Based Conversational Agent with RASA *Course Project – Software Engineering* 2021

- Led the design and implementation of a **RASA-based dialogue agent** for travel planning, with multi-intent support and **slot-based memory modeling**.
- Developed custom **NLU pipelines**, dialogue management stories, fallback and **edge-case handling policies**.
- Coordinated full-cycle documentation, from **stakeholder requirements** and slot mappings to evaluation reports.
- Delivered a 76-page engineering report outlining the **complete agent development lifecycle**.

Other Web Projects Practice

AI+BII: Generative AI for Architectural Image Inpainting 2024

- Designed a **Stable Diffusion-based generative inpainting system** to reconstruct missing structures in historical architecture images.
- Integrated **VAE and U-Net** backbones in a latent compression pipeline to **boost fidelity and resolution**.
- Enabled multimodal generation via **CLIPText-guided conditioning** and **parameterized generation logic**.
- **Deployed** an **interactive demo on Hugging Face Spaces**, supporting custom mask drawing, prompt input, and guided sampling control for real-time generation.
- [YouTube Demo Link] & [GitHub Link]

DSPaperUniverse Platform, Academic Literature Exploration and Visualization 2023

- Designed an **Interactive Graph platform** to visualize citation networks and model keyword co-occurrence.
- Applied network analysis techniques to **explore structural properties** (e.g., edge clustering) and topic diffusion.
- The approach is extensible to **biological networks** (e.g., **PPI, GO DAG**) in bioinformatics applications.
- **Web-based service** design for scalable data visualization, extensible to data collection and management for NLP.
- [YouTube Demo Link] & [GitHub Link]

Publications & Awards

- **X. Yuan**, Z. Zhang, Y. Sun, Z. Xue, X. Shao, & X. Huang. (2023). A new database of Houma Alliance Book ancient handwritten characters and its baseline algorithm. In Proc. of the 8th Int. Conf. on Multimedia Systems and Signal Processing (ICMSSP '23). ACM. DOI: 10.1145/3613917.3613923
- Z. Zhang, X. Huang, **X. Yuan**, & Y. Sun. (2023). HABFD: Houma Alliance Book facsimiles database. In Proc. of the IEEE Int. Conf. on Image, Vision and Computing (ICIVC '23). IEEE. DOI: 10.1109/icivc58118.2023.10269984
- 1st Prize, Chinese Collegiate Computing Competition (National-Level, China), 2023
- 2nd Prize, Jiangsu Province Bachelor Thesis Award (Provincial-Level, China), 2023
- **3rd Prize, Jiangsu Province University Mathematical Modeling Competition, 2021**
- 1st Prize in the 1st "Prospective Cup" Meta-Intelligent Data Challenge (Track: Entity Recognition in Sedimentology Knowledge Map), held at the CENet2022 Conference, Haikou, China (Nov. 2022).

Technical Skills

NLP & Transformer Models: Transformers, BERT, ProtBERT, ESM, LangChain, Dialogue State Tracking

LLM Alignment & Detection: Prompt Engineering, GPT-4 API, RLHF, DPO, Human-AI Hybrid Text Evaluation

Web Development & Deployment: Flask, JavaScript, Node.js, TypeScript, Basic React (transferable to Vue.js), Nginx Deployment, RESTful API Design, HuggingFace

Explainability & Weak Supervision: Attention Attribution, Label Denoising, Consistency Regularization

Programming & Frameworks: Python, PyTorch, TensorFlow, Scikit-learn, Git, Docker, Linux

Probabilistic & Scientific ML: Bayesian Modeling, PyMC3, Simulation-Based Inference

Software Engineering: Agile Development, Version Control, Testing & Debugging