Jinqiang Yu

Chatswood NSW 2067, Australia

✉ jinqiang.yuuu@gmail.com | linkedin.com/in/jinqiang-yu-404bb0187 | 🎓 Jinqiang Yu

Summary

I am passionate about machine learning and AI areas, with experience in the fields of machine/deep learning , NLP, CV and explainable AI for ML, in particular LLMs. With publications as the lead author in renowned AI conferences and journals like AAAI and JAIR, I have also served as a reviewer for top AI conferences such as AAAI and IJCAI. Stay updated on cutting-edge AI techniques like generative AI and large language models, I am expecting to leverage my technical skills to contribute to AI/ML projects.

Education

Monash University *Melbourne, Australia*

PhD in Artificial Intelligence *Feb 2021 - Aug 2024*

- **Thesis Topic**: Explainable AI for Machine Learning
- **Thesis Description**: Th PhD project aims to deal with explainable AI and machine learning problems, including training interpretable machine learning (ML) models, generating accurate and concise explanations to explain ML models in various domains like NLP, CV, healthcare and general classification tasks, and applying explainable AI for ML, in particular LLMs, such improving LLM performance and enabling LLMs to generate explainable outputs.

Monash University *Melbourne, Australia*

Master of Information Technology *Mar 2019 - Dec 2020*

- Graduated with H1
- **Minor Thesis Topic**: Computing optimal interpretable machine learning models.
- **Thesis Description**: The thesis focuses on interpretable models, aiming at developing advanced approaches to computing machine learning models that are both accurate and interpretable.

Experience

Tiktok *Sydney, Australia*

Machine Learning Engineer *Jul 2024 - Present*

- My work focuses on comment moderation at Tiktok with the use of LLMs. This includes developing and refining prompt engineering techniques to ensure accurate AI labeling, deploying models for moderation, and conducting research to optimize model performance. My role involves balancing the technical demands of building effective moderation systems, cost, and user experience.

Monash University *Melbourne, Australia*

Teaching Associate *Jun 2021 - Nov 2021*

- Tutoring and grading students in tutorials, assignments, and final exams.
- **Unit**: FIT5220 - Solving discrete optimisation problems.
- **Topics**: Constraint Programming, Mixed Integer Programming, Boolean Satisfiability (SAT) Solving, Large Neighbourhood Search.

Professional Projects

Improving LLM model performance on comment moderation tasks *Sydney, Australia*

Tiktok *Jul 2024 - present*

- Comment moderation is essential today for maintaining safe online platforms and fostering user trust. In this project, I leverage the capabilities of LLMs to effectively address comment moderation tasks. My focus is on improving the model performance to enhance accuracy and efficiency. To achieve this, I applied various techniques, including data augmentation, masking loss of trivial tokens, and experimenting with different fine-tuning strategies. These methods help optimize the model's ability to detect and manage harmful or inappropriate content while maintaining a positive user experience.
- **Expertise: Data Augmentation, Mask Loss, Instruction fine-tuning, Chain-of-thought.**

Improving AI labeling with prompt engineering *Sydney, Australia*

Tiktok *Jul 2024 - present*

- Data is crucial as it forms the foundation of ML and AI systems. However, data collection is often challenging, and a significant portion of large datasets remains unlabeled. AI-driven labeling plays a critical role in addressing this gap by providing accurate labels for unlabeled data. In this project, my focus is on enhancing AI labeling processes by leveraging advanced prompt engineering techniques to ensure the collection of accurate and comprehensive data for more effective model training and performance.
- **Expertise: Transformers, Generative AI, Prompting Engineering, Chain-of-thought.**

Applying Explainable AI in Machine Learning and LLMs

Melbourne, Australia

Monash University

Oct 2023 - Aug 2024

- Conducting research on LLMs with the use of explainable AI. Aiming to improve LLMs' performance, help others understand the models' behavior, and build trust of the models.
- Developing approaches to improving the performance of LLMs with the use of Parameter-Efficient Fine-Tuning (PEFT) and Prompt Engineering given explanations.
- Developing methods to generate more explainable outputs of LLMs for different users with the use of Instruction Fine-Tuning and Reinforcement Learning with Human Feedback (RLHF).
- **Expertise: Machine Learning, LLMs, Transformers, Generative AI, Prompting Engineering, PEFT, Instruction Fine-Tuning, RLHF, Explainable AI.**

Explainability and Its Application in NLP and CV

Melbourne, Australia

Monash University

Dec 2022 - present

- Conducting research on explainability in NLP and CV domains.
- Developing approaches to generating feature-attribution and feature selection based explanations in NLP/CV domains.
- Applying explainable AI methods for healthcare images.
- **Expertise: LLMs, Transformers, CNNs, Natural Language Processing, Computer Vision, Explainable AI.**

Computing Succinct and Accurate Explanations for ML

Melbourne, Australia

Monash University

Feb 2021 - Apr 2023

- Conducted research on generating more accurate and concise post-hoc (*abductive*) explanations answering why the prediction is made and (*contrastive*) explanations targeting why not the prediction is made.
- Developed approaches to mining background knowledge and applying them to generating accurate and concise explanations in diverse domains.
- Developed approaches to generating feature attribution indicating the contribution of a feature.
- **Expertise: Machine Learning, Data Mining, Explainable AI.**

Selected Publications

Anytime Approximate Formal Feature Attribution
Jinqiang Yu Graham Farr, Alexey Ignatiev, Peter J. Stuckey
arXiv preprint arXiv:2312.06973, 2023

On Formal Feature Attribution and Its Approximation
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey
arXiv preprint arXiv:2307.03380, 2023

From Formal Boosted Tree Explanations to Interpretable Rule Sets
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey
29th International Conference on Principles and Practice of Constraint Programming (CP), 2023

Eliminating the Impossible, Whatever Remains Must Be True: On Extracting and Applying Background Knowledge in the Context of Formal Explanations
Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, Joao Marques-Silva
37th AAAI Conference on Artificial Intelligence (AAAI), 2023

Skills

Programming Languages	Python, Java, R, C++, MATLAB
DS/ML/AI Technologies	Domains - NLP, CV, LLM, Generative AI, Explainable AI ML/DL/GenAI Models - LLM, Transformer, CNN, Diffusion model, GAN, VAE, etc LLM Technologies - LLM framework (LangChain), RAG, Parameter-Efficient Fine-Tuning (PEFT), Instruction Fine-Tuning, RLHF, and Prompt Engineering Cleansing and Wrangling - Pandas, Numpy, NLTK Visualization - Matplotlib, Seaborn Big Data Management - MongoDB, SQL, Spark
Miscellaneous Technologies	Github, Docker, AWS, Slurm, Linux

Awards and Scholarships

2021-2024	**Monash Graduate Scholarship**	Scholarship covers living expenses and tuition
2020	**Best Paper Award**	Our paper "Computing Optimal Decision Sets with SAT" has been selected for the Best Paper Award for the CP/ML Track of CP 2020.

Scientific Activities

2024	PC member of the AAAI Conference on Artificial Intelligence(AAAI-2024)
2024	PC member of the International Joint Conferences on Artificial Intelligence (IJCAI-2024)