

Pankayaraj Pathmanathan

Curriculum Vitae

Website | p.pankayaraj@gmail.com | LinkedIn | Github | Full Resume

WORK EXPERIENCE (SELECTED)

CURRENT, FROM SEPT 2022 (FULL TIME)

Research Assistant, Teaching Assistant
University of Maryland College Park

During this time, I primarily worked on LLM poisoning attacks, including RLHF poisoning, backdoor poisoning, and copyright poisoning. These works have been **published on Neurip and ICML 2024 workshops** and are under review for AAAI, ICLR, ACL 2025.

FEB 2022 – AUG 2022 (FULL TIME)

Research Engineer
Singapore Management University

During this time, I had worked on constraint reinforcement learning methods that exploit the hierarchical reinforcement learning paradigm to better satisfy long horizon constraints in an effective manner. This work was **published at AAAI 2023**

AUGUST 2020 – AUGUST 2021 (FULL TIME)

Research Assistant + Collaborator
SLTC, QBITS Lab + Flowers Laboratory, ENSTA Paris

During this time, I had worked on mitigating catastrophic forgetting in continual reinforcement learning with the use of curiosity. This work was **published at Cognitive Computational Journal 2023**

FEB 2019 – AUGUST 2019 (FULL TIME)

Research Intern
SLTC, QBITS Lab

This was my undergraduate research internship during which I had worked on multi agent multi arm bandit algorithms. These works were **published at IEEE CDC 2020 and European Control Conference 2020**

PUBLICATIONS (SELECTED)

Pankayaraj P, Varakantham, P. (2022). Constrained reinforcement learning in hard exploration problems [Poster], In 37th **AAAI** Conference on Artificial Intelligence Washington, D.C. USA

Pankayaraj P, Rodríguez, N. D., Ser, J. D. (2023). Using curiosity for an even representation of tasks in continual offline reinforcement learning, In **Cognitive Computation** Journal 2023

Pankayaraj P, Maithripala, D. H. S. (2020) A decentralized communication policy for multi agent multi armed bandit problems [Presented], In **European Control Conference** 2020, Saint Petersburg, Russia

Pankayaraj P, Maithripala, D. H. S., Berg, J. M. (2020). A decentralized policy with logarithmic regret for a class of multi-agent multi-armed bandit problems with option unavailability constraints and stochastic communication protocols [Presented], In 59th **IEEE Conference on Decision and Control**, Jeju Island, Republic of Korea

Panaitescu-Liess, M.-A., Che, Z., An, B., Xu, Y., **Pankayaraj P**, Chakraborty, S., Zhu, S., Goldstein, T., Huang, F. (2024). Can watermarking large language models prevent copyrighted text generation and hide training data?, In 39th **AAAI** Conference on Artificial Intelligence

EDUCATION

CURRENT **PhD computer science**

ADVISOR: FURONG HUANG
University of Maryland College Park.

2015-2020 **BSc Computer Science**

UNIVERSITY OF PERADENIYA, SRI LANKA

AWARDS

2022-2024 **Dean's Fellowship**

University of Maryland

2020 **Best Paper Award**

ESCaPe 2020, Symposium, Sri Lanka

REFERENCES

NAME **Prof. Furong Huang**
EMPLOYER University of Maryland College Park

NAME **Prof. Pradeep Varakantham**
EMPLOYER Singapore Management University

NAME **Dr. Mugalan Maithripala**
EMPLOYER University of Peradeniya

WORKSHOPS (ACCEPTED)

Pankayaraj P, Chakraborty, S., Liu, X., Liang, Y., Huang, F. (2024). Is poisoning a real threat to LLM alignment? maybe more so than you think, In **ICML 2024** Workshop on Models of Human Feedback

Pankayaraj P, Sehwal, U. M., Panaitescu-Liess, M.-A., Huang, F. (2024a). Advbdgen: Adversarially fortified prompt-specific fuzzy backdoor generator against llm alignment, in the **Neurips** Safe Generative AI Workshop 2024

Panaitescu-Liess, M.-A., **Pankayaraj P**, Y. K., Che, Z., An, B., Zhu, S., Agrawal, A., Huang, F. (2024). Poisonedparrot: Subtle data poisoning attacks to elicit copyright-infringing content from large language models, in the **Neurips** Safe Generative AI Workshop 2024

Panaitescu-Liess, M.-A., Che, Z., An, B., Xu, Y., **Pankayaraj P**, Chakraborty, S., Zhu, S., Goldstein, T., Huang, F. (2024). Can watermarking large language models prevent copyrighted text generation and hide training data?, In **NeurIPS 2024** Workshop AdvML-Frontiers **Best Paper**

Pankayaraj P, Sumanasekera, Y., Samarasinghe, C., Elkaduwe, D., Jayasinghe, U., Maithripala, D. H. S. (2020). Multi-agent reinforcement learning in sparsely connected cooperative environments [Presented, awarded the **Best Research Paper**], in ESCaPe 2020, Sri Lanka.