

# Pankayaraj Pathmanathan

## Curriculum Vitae

[Website](#) | [p.pankayaraj@gmail.com](mailto: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 **AAAI main conference** and **Neurip, ICML 2024 workshops** and are under review for ICLR, ACL 2025.

MAY 2025 – AUG 2025 AND NOV 2025 - CURRENT (FULL TIME & PART TIME)

### Machine Learning Research Intern *Netflix, Los Gatos*

Worked on pretraining large language model based embeddings models

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**

## PUBLICATIONS (SELECTED)

**Pankayaraj P**, Sehwag, U. M., Panaitescu-Liess, M.-A., Cho-Yu Jason Chiang, Huang, F. (2025a). Advbdgen: Adversarially fortified prompt-specific fuzzy backdoor generator against llm alignment [Oral] 4.6%, in the In **40th AAAI - AIA Conference** on Artificial Intelligence, Singapore

**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 [Poster], In **39th AAAI - AIA Conference** on Artificial Intelligence Philadelphia, Pennsylvania, USA

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

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 [Oral] , in the **NAACL 2025**

**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

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 Philadelphia, Pennsylvania, USA

**Pankayaraj P**, Maithripala, D. H. S. (2020) A decentralized communication policy for multi agent multi armed bandit problems [Oral], 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 [Oral], In **59th IEEE Conference on Decision and Control**, Jeju Island, Republic of Korea

## EDUCATION

CURRENT	<b>PhD computer science</b> ADVISOR: FURONG HUANG University of Maryland College Park.
2015-2020	<b>BSc Computer Science</b> UNIVERSITY OF PERADENIYA, SRI LANKA

## AWARDS

2022-2024	<b>Dean's Fellowship</b> <i>University of Maryland</i>
2024	<b>Best Paper Award</b> <i>Neurips AdvML-Frontiers</i>
2020	<b>Best Paper Award</b> <i>ESCaPe 2020, Symposium, Sri Lanka</i>

## REFERENCES

NAME	<b>Prof. Furong Huang</b>
EMPLOYER	University of Maryland College Park
NAME	<b>Ashish Rastogi</b>
EMPLOYER	Netflix
NAME	<b>Hafez Asgharzadeh</b>
EMPLOYER	Netflix
NAME	<b>Prof. Pradeep Varakantham</b>
EMPLOYER	Singapore Management University

## WORKSHOPS (SELECTED)

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 [**Best Paper**] **NeurIPS Workshop AdvML-Frontiers** 2024

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

**Pankayaraj P**, Huang, F (2025). Reward Models Can Improve Themselves: Reward-Guided Adversarial Failure Mode Discovery for Robust Reward Modeling **AAAI Workshop on Trust and Control in Agentic AI** 2026