

Pankayaraj . P - 3rd Year PhD Student(GPA 3.8).

✉ p.pankayaraj@gmail.com

🌐 <http://pankayaraj.github.io>

🌐 <https://www.linkedin.com/in/pankayaraj-pathmanathan-259926119/>

Employment History

- 2023-2023 **Research Assistant** Advisor: Prof. Furong Huang.
Department of Computer Science, University of Maryland, USA
- Worked on diverse reinforcement learning.
- Worked on RLHF poisoning in LLMs, Copyright Poisoning and Backdoor poisoning in LLMs.
- Worked on robust RAG systems
- Worked on robust reward modelling for LLM alignment
- Works published in ICML, ICLR and Neurips 2024 workshops and AAAI conference 2025. Some works currently under submission
- 2022-2022 - 7 months **Research Engineer** Supervisor: Prof Pradeep Varakantham
Singapore Management University (SMU)
- Worked on Constraint Reinforcement Learning in Hierarchical Settings
- Work was published on AAAI 2023.
- 2020-2021 - 12 months **Research Assistant** SLTC, QBITS Lab. + **Collaborator** Flowers Laboratory, ENSTA Paris.
- Worked on the ways to improve continual offline reinforcement learning with artificial curiosity
- Work was published on Cognitive Computational Journal 2023
- 2019-2019 - 5 months **Research Assistant Intern** SLTC, QBITS Lab.
- Worked on devising communication strategies for multi agent multi arm bandit problems in both normal and delayed reward settings.
- Works were published on IEEE CDC 2020 and ECC 2020 respectively.

Research Publications

Conferences, Journals and, Workshops (Published)

- 1 Panaitescu-Liess, M.-A., Che, Z., An, B., Xu, Y., **Pankayaraj P**, Chakraborty, S., Zhu, S., Goldstein, T., & Huang, F. (2025). Can watermarking large language models prevent copyrighted text generation and hide training data?, In *In the The 39th Annual AAAI Conference on Artificial Intelligence 2025 and 3rd NeurIPS 2024 Workshop AdvMLFrontiers [best paper]*. [🌐 https://arxiv.org/abs/2407.17417](https://arxiv.org/abs/2407.17417)
- 2 Panaitescu-Liess, M.-A., **Pankayaraj P**, Y. K., Che, Z., An, B., Zhu, S., Agrawal, A., & Huang, F. (2025). Poisonedparrot: Subtle data poisoning attacks to elicit copyright-infringing content from large language models in the **63rd Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025)** [oral], In *Naacl 2025*.
- 3 **Pankayaraj P**, Chakraborty, S., Liu, X., Liang, Y., & Huang, F. (2025b). Is poisoning a real threat to LLM alignment? maybe more so than you think, In *In the The 39th Annual AAAI Conference on Artificial Intelligence 2025 andn ICML 2024 Workshop on Models of Human Feedback for AI Alignment*. [🌐 https://arxiv.org/abs/2406.12091](https://arxiv.org/abs/2406.12091)

- 4 **Pankayaraj P**, Schwag, U. M., Panaitescu-Liess, M.-A., & Huang, F. (2024). Advbden: Adversarially fortified prompt-specific fuzzy backdoor generator against llm alignment, In *Neurips Safe Generative AI Workshop 2024*. <https://openreview.net/forum?id=FdQBJu2e4d>
- 5 **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 **Impact Factor : 5.4**.
- 6 **Pankayaraj P**, & Varakantham, P. (2022). Constrained reinforcement learning in hard exploration problems [Poster], In *37th AAAI Conference on Artificial Intelligence Washington, D.C. USA* **Acceptance rate : 19.6 %**.
- 7 **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* **Acceptance rate : 58%**.
- 8 **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* **Acceptance rate : 52.7%**.
- 9 Jayatilaka, G., Weligampola, H., Sritharan, S., **Pankayaraj Pathmanathan**, Ragel, R., &], I. N. [(2019). Non-contact infant sleep apnea detection [Presented], In *ICIIS 2019 Sri Lanka*.

Under Review (Current)

- 1 **Pankayaraj P**, & Huang, F. (2025a). *RAGPart & RAGMask: Retrieval-stage defenses against corpus poisoning in retrieval-augmented generation* **Under Review**, in the *2 nd Conference on language modelling (COLM 2025)*.
- 2 **Pankayaraj P**, & Huang, F. (2025b). *Reward models can improve themselves: Reward-guided adversarial failure mode discovery for robust reward modeling* **Under Review**, in the *Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*.
- 3 **Pankayaraj P**, Schwag, U. M., Panaitescu-Liess, M.-A., & Huang, F. (2025a). *AdvBDGen: Adversarially fortified prompt-specific fuzzy backdoor generator against llm alignment* **Under Review**, in the *39 th Annual Conference on Neural Information Processing Systems(NeurIPS 2025)*.




Symposiums (Published)

- 1 **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**, *[best research paper]*], in *ESCaPe 2020, Sri Lanka*.


Preprints

- 1 **Pankayaraj P**, Sumanasekera, Y., & Samarasinghe, C. (2019). *A review on reinforcement learning based autonomous quadcopter control*.




Education






2022-2027	 PhD Computer Science , Department of Computer Science, University of Maryland, USA. GPA: 3.8 out of 4.0 ADVISOR: Prof Furong Huang
2015-2020	 B.Sc. Computer Engineering University of Peradeniya, Sri Lanka GPA: 3.5 out of 4.0
English Proficiency	 TOFEL: 112 , Reading: 29 , Listening: 30 , Writing: 26 , Speaking: 27

Academic Volunteering


- 2020  **Peer Reviewer : Journal** IEEE Transactions on Communications
- Impact Factor : 5.69 (2018)

Projects





- 2023  **Efficient Exploration in Reinforcement Learning** Improving the sample efficiency in RL where entropy of occupancy measure is used as an exploration mechanism
Report <https://drive.google.com/file/d/1ESQZunSYI8WegsgiQ63HJmHgUXHIF1LU/view?usp=sharing>
-  **Virtual Maze Navigation Using Different Locomotion Techniques** Analysing the effects of Redirected Walking and Steering in VR environments and proposing a hybrid locomotion technique.
Report <https://drive.google.com/file/d/11JiiJo0ZzJLbtfumzjThmq3Yp0BEG1pF/view?usp=sharing>
- 2019  **Reinforcement Learning Based Autonomous Quadcopter Control**
Using Reinforcement Learning algorithms to make Quadcopter control decisions on an AirSim simulated environment

REPORT: <https://drive.google.com/drive/folders/16Ej8XL4SRrthL58FsMMnYDaD5WYWruF9>
- 2018  **A user recommendation method using Bayesian Reinforcement Learning**
Github Link : <https://github.com/pankayaraj/sitnshop/tree/BackEndAlgorithm>
REPORT : <https://drive.google.com/drive/folders/19Sq1UEXeANUWQGGVf8EXBCrPRntAFvaF>
- 2017  **Creating a python based library with a Tensor flow back end for Bayesian Optimization and Multi Arm Bandit Problem**
GitHub Link : <https://github.com/pankayaraj/Multi-Arm-Bandit-Library>
PyPi link : <https://pypi.python.org/pypi/mabandit/1.3>
REPORT: <https://drive.google.com/drive/folders/1H2Pcbfj825LPbYj03rnKLY0lgRQCFwXb>
- 2018  **SitnShop– An Advertising platform for shops** An advertising platform for anykind of shop and it also helps the customers of the shops to easily find related shops.
Github Link : <https://github.com/pankayaraj/sitnshop/>
REPORT : <https://drive.google.com/drive/folders/1ZcsJkFPDCJhvh8k0yt0LjnxBkJ1cAgaB>
-  **Infant Sleep Apnea detection system : A portable device that can detect Sleep Apnea condition in infants using techniques such as Optical flow, Edge detection, Fourier analysis with python, Raspberry pi.**
Github Link : https://github.com/pankayaraj/Sleep_Apnea_Detection
REPORT: https://drive.google.com/drive/folders/17fLXhj1uxl5MuqNqEM46_tYuRsWoXC2Z
-  **Making a central server for the sleep apnea problem**
Technologies : Python, Django, Django rest framework, HTML/CSS,JS
Github Link: https://github.com/pankayaraj/Django_Server_Sleep_Apnea

Projects (continued)




- 2017  **ExpertMiner :**
Earth resource location prediction using Hyper Spectral Images from satellites Tech-
niques : Pattern Recognition, Correlation Mapping
Github Link : https://github.com/pankayaraj/HSI_Project

Skills

- Languages  English (TOFEL: **112**, Reading: **29**, Listening: **30**, Writing: **26**, Speaking: **27**),
Tamil(Native).
- Programming  Python , \LaTeX , C++,
- Libraries  Tensorflow, Pytorch, Kivy, Numpy, Scipy
- Web Dev  Django

Miscellaneous Experience

Awards and Achievements

- 2017  **ACES Hackathon 2017(Intra university hackathon) :** Project:Expert miner: Software section
winners, Best idea of the competition
- 2016  **ACES Coders v6.0** (Inter university programming competition):Country Rank : 4th
-  **IEEEExtreme 10.0 Programming competition**(24 hour Global Programming competition
Country Rank: 38th
World Rank: 330th

References

Prof Furong Huang

Assistant Professor

Department of Computer Science, University of
Maryland,
USA.

furongh@umd.edu

Prof Pradeep Varakantham

Professor of Computer Science

School of Computing and Information Systems, Sin-
gapore Management University,
Singapore.

pradeepv@smu.edu.sg

Dr D.H.S Maithripala

Senior Lecturer

University of Peradeniya,
Peradeniya, Sri Lanka.

smaithri@pdn.ac.lk

mugalan@gmail.com