Vishnu Vinod

vishnuvinod2001 | vishnuvind | vishnuvind.github.io | vishnuvind | Vishnu Vinod

Education

Indian Institute of Technology Madras — *2019 - 2024*

Dual Degree (B.Tech + M.Tech) in Computer Science & Engineering — **8.8/10.0**

Key Coursework: Multi-Arm Bandits, Reinforcement Learning, Stochastic Optimization, Modern Computer Vision, Mathematical Foundations of Data Science, Basic Graph Theory, Probability Statistics & Stochastic Processes, Data Analytics, Deep Learning, Pattern Recognition and Machine Learning, Advanced Graph Algorithms, GPU Programming, Design and Analysis of Algorithms, Structural Graph Theory, Operating Systems, Computer System Design, Computer Organization and Architecture, Quantum Mechanics, Quantum Computation & Quantum Information

Publications

InvisibleInk: High-Utility and Low-Cost Text Generation with Differential Privacy — *NeurIPS 2025*

***Vinod, V.**, Pillutla, K., Thakurta, A.*

Preserving Expert-Level Privacy in Offline Reinforcement Learning — *TMLR 2025*

Sharma, N., **Vinod, V.***, Thakurta, A., Agarwal, A., Balle, B., Dann, C. & Raghuveer, A.*

Generating Universal Adversarial Perturbations for Quantum Classifiers — *AAAI 2024*

Anil, G., **Vinod, V.*** & Narayan, A.*

Research Experience

Post-Baccalaureate Fellow, CeRAI, IIT Madras† — *Jul '24 - Present*

Mentored by **Prof. Krishna Pillutla**

InvisibleInk: Low-Cost Private Text Generation using LLMs

- Output text of an LLM can leak information present *in-the-context* of the generation; DP can mitigate privacy leakage at inference-time.
- Existing methods require high computational overhead ($\geq 100\times$) to privatize output text and have low data yield rates ($\leq 1\%$).
- Introduced *InvisibleInk* for high-utility and low-cost text-generation from LLMs under differential privacy; *Paper accepted at NeurIPS 2025*.
- *DClip*, isolates and clips *only* sensitive information in model logits, and *Top-k+ sampling* from a tight superset of the top-k private tokens.
- Empirical evaluation on medical, commercial, and legal datasets (MIMIC Notes/Yelp Reviews/TAB-ECHR); Additional ablation analyses.
- Reduced computational overhead by a **factor of 8x** and boosted data yield rates **to over 10%** for similar privacy and utility levels (vs. SOTA).
- Proposed practical recommendations for **optimal hyperparameter selection** in compute-constrained settings.

Auditing Differentially Private Text generated by LLMs *(ongoing)*

- Privacy audits empirically estimate the privacy-level of "private" algorithms; essential to test the correctness of privacy guarantees.
- Text generation using the exponential mechanism yields tight Gaussian DP guarantees; Existing work focuses on strong membership inference.
- Aim: Develop tight f-DP auditing schemes tailored for auditing LLM-generated text in both black-box and white-box settings.

Differentially Private Long-form Retrieval Augmented Generation *(ongoing)*

- LLMs can be used for query-answering based on a sensitive reference dataset using the RAG framework.
- Existing Private RAG methods are split into two modules: private retrieval and private generation; Small generation lengths ($<$50 tokens).
- Aim: Adapting InvisibleInk with better private retrieval strategies to allow long-form RAG for use in correctness-sensitive settings.

Student Researcher, Google Research India — *Nov '23 - Apr '24*

Mentored by **Dr. Aravindan Raghuveer** & **Prof. Balaraman Ravindran**

Expert-Level Differentially Private Offline Reinforcement Learning

- Offline RL algorithms train on data contributed by behavioural policies (experts); Learnt policy can leak the privacy of experts.
- Existing methods operate under strict assumptions: linear function approximators, tabular settings, or trajectory-level privacy guarantees.
- Proposed offline RL training paradigm with expert-level privacy guarantees; compatible with non-tabular, deep offline RL settings.
- Used *expert-consensus* to find *stable trajectory* prefixes for noise-free training, in conjunction with adapted *expert-level DP-SGD* for the tails.
- Evaluated across environments against an *expert-level DP-SGD* baseline; Consistent **performance gains across all settings**!
- Training framework compatible with *all SOTA off-the-shelf* gradient-based offline RL and user-level privatization algorithms.

Research Intern, University of British Columbia — *May '22 - Aug '23*

Mentored by **Prof. Apurva Narayan**

Generating Universal Adversarial Perturbations for Quantum Classifiers

- Existence of UAPs demonstrated for classical models; Notion of UAPs ill-defined in the realm of Quantum Machine Learning.
- Conceptualized additive & unitary Universal Adversarial Perturbations (UAPs) for parametrized quantum circuit-based quantum classifiers.
- Showed theoretical guarantees for additive UAPs on amplitude-encoded data; Proposed *QuGAP-A* to generate additive UAPs.
- Proposed *QuGAP-U* to construct unitary UAPs for perturbing quantum data; Trained using a novel *fidelity-based loss*.
- Experimentally validated the proposed methods on multiple datasets (MNIST, FMNIST, TIM) for binary and 4-class classification.
- QuGAP-U achieved **full misclassification** at over **20% higher quantum state fidelity** compared to previous SOTA methods.

Other Major Projects

Visual Explanations for Drug-Target Affinity Prediction — *Spring '24*

CS6024 - Algorithmic Approaches to Computational Biology under Prof. Manikandan Narayanan

- Trained multi-scale GNN models for *Drug-Target Affinity Prediction* with input drug molecular structure and target protein sequence.
- Modified existing GradCAM-based approaches to generate saliency maps and identify active regions in the drug molecules and target proteins.
- Conducted quantitative analysis of drug and target saliency maps by occluding salient and non-salient regions in either map.
- Proposed new metric to assess the conformity of drug and target explanations with each other.

Empirical Study of Variance Reduced Methods in Machine Learning — *Spring '23*

CS6515 - Stochastic Optimization under Prof. Prasanth LA

- Empirically studied variance-reduced optimization methods (SAG/SAGA/SDCA/SVRG) in convex and non-convex optimization settings.
- *Convex*: Used *sklearn-lightning* to test logistic regression convergence (*loss vs. CPU time*) against SGD and Adagrad on multiple datasets.
- *Non-convex*: Used MNIST and CIFAR-10 datasets to compare convergence of SVRG with SGD and Adam optimizers.
- Studied effect of regularization and model complexity on the performance (*loss vs. epoch*) of each algorithm.

Reinforcement Learning Methods — *Spring '23*

CS6700 - Reinforcement Learning under Prof. Balaraman Ravindran

- Experimentally studied multiple reinforcement learning algorithms in different use cases.
- Studied *SARSA* and *Q-Learning* on a modified grid-world across a large number of experimental configurations.
- Implemented *DQN* and *Actor-Critic* methods (n-step/full return) on OpenAI gym environments (CartPole/AcroBot/MountainCar).
- Carried out comparative study of *SMDP* and *Intra-option Q-Learning* on the Taxi environment in OpenAI gymnasium.

Geospatial Applications of Machine Learning — *Summer '21*

Data Science Intern at GalaxEye Space Solutions Pvt. Ltd., mentored by Kishan Thakkar

- Used multiple *semantic segmentation* methods for Building Footprint Extraction from multi-spectral satellite images (SpaceNet dataset).
- Evaluated models on augmented multispectral satellite images of Rotterdam; Improved model predictions by post-processing.
- Carried out Land-Use-Land-Cover (LULC) classification using an ensemble of gradient boosting methods.
- Preprocessed raw optical data from the Sentinel-2 dataset using EO-Learn library.

Technical Skills

Languages	***Proficient:*** Python | C/C++ | Bash ***Familiar:*** MATLAB | OCaml | x86 Assembly | HDL (*nand2tetris*)
Libraries	PyTorch | TensorFlow | Keras | JAX | PennyLane | Qiskit | Pandas | Acme | Gymnasium | OpenCV | RasterIO
Frameworks	Jupyter | Git | LaTeX | Google Colab | Amazon Sagemaker

Achievements

- **2024** Selected for the Post-Baccalaureate Fellowship offered by CeRAI and the Wadhwani School of Data Science & AI, IIT Madras.
- **2023** Selected for a 6-month student researcher internship with the Advertising Sciences team at Google Research India.
- **2022** Selected for 12-week MITACS Globalink Research Internship at the University of British Columbia.
- **2019** Secured **All India Rank 35** among 1.1 million candidates in the JEE (Main) examination, placing in the top 0.005%tile.
- **2019** Secured **All India Rank 90** among 200,000 candidates in the JEE (Adv.) examination, placing in the top 0.05%tile.
- **2017** Secured **All India Rank 21** (SA stream); received the KVPY 2017 Fellowship from the Dept. of Science and Technology, GoI.

Talks and Presentations

- **2025** **Poster at NeurIPS 2025** (scheduled) @ San Diego Convention Centre, San Diego, CA, USA.
- **2025** **Poster at Conclave on AI Governance** @ IIT Madras., Pre-Summit event of the AI Impact Summit 2026
- **2025** **Talk at Academic Summit 2025** @ Microsoft Research India, Bangalore, India.
- **2025** **Poster at WSAI Annual Research Showcase** @ Indian Institute of Technology Madras.
- **2024** **Poster at AAAI 2024** @ Vancouver Convention Centre, Vancouver, BC, Canada.

Co-Curriculars & Volunteering

- **2025** Teaching Assistant @ IIT Madras, Reinforcement Learning, lectured by Prof. Balaraman Ravindran.
- **2025** Teaching Assistant @ NPTEL, Introduction to Machine Learning (link), lectured by Prof. Balaraman Ravindran.
- **2024** Volunteer @ 38th Annual AAAI Conference on Artificial Intelligence, Vancouver, Canada.
- **2024** Teaching Assistant @ IIT Madras, Reinforcement Learning, lectured by Prof. Balaraman Ravindran.
- **2023** Teaching Assistant @ IIT Madras, Foundations of Machine Learning, lectured by Prof. Balaraman Ravindran.
- **2021** Academic Mentor @ Student Mentorship Cell, IIT Madras.
- **2021** Coordinator @ Shows & Exhibitions, Shaastra 2021, IIT Madras.
- **2020** Deputy Coordinator @ Placement & Internship Cell, IIT Madras.

*: equal contribution; †: Also affiliated with RBCDSAI and WSAI @ IIT Madras; SOTA: State–of–the–art