

Prathamesh Devadiga

CS PhD Student, JASPR Lab, Dartmouth College

prathamesh.p.devadiga.gr@dartmouth.edu | linkedin.com/in/prathamesh-devadiga | prathameshdevadiga.vercel.app

RESEARCH INTERESTS

Machine learning security and privacy, memorization and training-data extraction in LLMs, machine unlearning and model editing auditing, adversarial robustness, LLM jailbreaks and defenses, production AI systems

EDUCATION

Dartmouth College, Hanover, NH, USA

Aug 2026 – Present

Doctor of Philosophy (PhD) in Computer Science

Advisor: Prof. Shawn Shan, JASPR Lab (Joint AI Security Privacy Research)

Research Interests: Trustworthy Machine Learning, AI Security and Privacy, Large Language Models, Memorization and Generalization Theory, Adversarial Machine Learning, Responsible AI Systems

PES University, Bangalore, India

Aug 2022 – May 2026

Bachelor of Technology in Computer Science Engineering

Cumulative GPA: 8.98/10.00

Relevant Coursework: Machine Learning, Deep Learning, Generative AI, Linear Algebra and its Applications, Statistics for Data Science, Engineering Mathematics I & II, Data Structures and Algorithms, Operating Systems, Computer Networks, Distributed Systems

RESEARCH EXPERIENCE

Dartmouth College

Remote

Research Assistant (ML Security & LLM Privacy) — Advisor: Prof. Shawn Shan

July 2025 – July 2026

- Co-developed **CHASE (Correlated Hierarchical Adaptive Search for Extraction)**, a Gaussian Process bandit framework for query-efficient memorization extraction from instruction-tuned large language models
- Discovered that the prompt space for memorization extraction exhibits a **low-rank, kernel-correlated structure**, enabling effective prompt optimization with as few as **100 API queries** per target passage
- Evaluated CHASE across **9 instruction-tuned LLMs from 7 model families**, including production APIs, demonstrating up to **2.9×** higher extraction success than random search while maintaining over **10×** gains under seven defense strategies
- Showed that ordinary academic and continuation prompts consistently outperform adversarial jailbreaks for memorization extraction, highlighting limitations of existing prompt-level safety defenses

Adhāra AI Labs

Remote

Founder & Lead Researcher

Feb 2025 – Present

- Founded independent research laboratory focused on Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and deep learning systems
- Leading research initiatives from concept to production, ensuring results translate into scalable, real-world applications
- Research outputs include PyraFuseNet (ICIAI), SLMs as Compiler Experts (NeurIPS Workshop), RegimeNAS, MorphNAS, and a low-latency jailbreak prevention framework

Lossfunk

Bangalore, India

AI Research Intern

Sep 2025 – June 2026

- Built an Indic LLM pipeline for extremely low-resource languages (Tulu: ~0.001% of typical training data)
- Developed a framework using hard negative constraints to reduce catastrophic language leakage from 80% to 5%, demonstrating that explicit prohibitions outperform positive instructions
- Paper accepted at the EACL 2026 Workshop on Language Models for Low-Resource Languages (LoResLM): “Making Large Language Models Speak Tulu: Structured Prompting for an Extremely Low-Resource Language”

Nokia

Bengaluru, India

Intent-Based Network Management System (3GPP SON)

Aug 2025 – Oct 2025

- Designed an intent-based self-organizing network (SON) system enabling operators to express high-level network goals in natural language, automatically translated into 3GPP-compliant intents
- Built an LLM-powered intent classification pipeline using Retrieval-Augmented Generation (RAG) to dynamically map natural language inputs to structured intent schemas
- Architected an intent orchestration layer supporting synchronous and asynchronous execution, conflict detection, duplicate intent suppression, and intent expiry management
- Developed a REST-based workflow integrating an interactive web frontend, intent classifier, orchestrator, and Auto Pilot (AP) APIs for automated network configuration

Google Summer of Code 2025

University of California Santa Cruz

Remote
May 2025 – Aug 2025

- Selected for highly competitive global program (acceptance rate less than 10%)
- Built billion-scale vector embedding benchmarks (768, 1024, 2048 dimensions) from open-source codebases
- Developed infrastructure for evaluating Approximate Nearest Neighbor (ANN) algorithms at scale under realistic workloads

Indian Institute of Technology, Indore

Undergraduate Research Intern (Deep Learning) — Advisor: Prof. Nagendra Kumar

Remote
Dec 2023 – Aug 2024

- Designed KASPER framework achieving 99.5% accuracy in PDF malware detection with adversarial robustness
- Developed custom malware injection pipeline for realistic payload embedding and comprehensive model evaluation
- Maintained 99.5% detection accuracy against FGSM and PGD adversarial attacks
- Led benchmarking and explainability analysis using Kolmogorov-Arnold Networks (KANs)
- Contributions formed core technical implementation for paper accepted at Applied Soft Computing (Q1 Journal)

PUBLICATIONS

Conference and Workshop Papers

- **Devadiga, P.** (2026). *Deletion Scars: Inferring Deleted Artist Styles from Paired Diffusion Models*. *ECCV 2026 Workshop on Unlearning and Model Editing (U&ME)*.
- **Devadiga, P.,** Lakshman, A. (2026). *Resistant to Lawyers, Defeated by Disagreement: Evaluation Blindspots in Legal Language Models*. *ICML 2026 Workshop on AI for Law (AI4Law)*.
- **Devadiga, P.,** et al. (2026). *Making Large Language Models Speak Tulu: Structured Prompting for an Extremely Low-Resource Language*. *EACL 2026 Workshop on Language Models for Low-Resource Languages (LoResLM)*.
- **Devadiga, P.,** et al. (2026). *GUARDIAN: Multi-Agent Defense System for Large Language Model Security*. *International Conference on Artificial Intelligence and Information (ICIAI)*, Waseda University, Japan.
- **Devadiga, P.,** et al. (2025). *SLMs as Compiler Experts: Auto-Parallelization for Heterogeneous Systems*. *NeurIPS 2025 Workshop on Machine Learning for Systems*.
- **Devadiga, P.,** et al. (2025). *PyraFuseNet: Dual-Path Network for Resource-Constrained Vision*. *International Conference on Artificial Intelligence and Information (ICIAI)*, Nanyang Technological University, Singapore.

Preprints

- **Devadiga, P.,** et al. (2025). *SAMVAD: Multi-Agent Framework for Modeling Judicial Reasoning in Indian Jurisprudence*. arXiv:2509.03793.
- **Devadiga, P.,** et al. (2025). *RegimeNAS: Regime-Aware Neural Architecture Search for Financial Trading*. arXiv:2508.11338.
- **Devadiga, P.,** et al. (2025). *MorphNAS: Differentiable Neural Architecture Search for Multilingual Named Entity Recognition*. arXiv:2508.15836.

Media Coverage

- Featured in *The Times of India* for research on enabling large language models to generate Tulu through structured prompting (2026).
- Featured on the *AIM Network* podcast to discuss AI for low-resource languages and structured prompting techniques for Tulu (2026).

KEY PROJECTS

Cerebrum: Production LLM Training and Serving System

2025

Independent Research Project

- Built end-to-end LLM system featuring distributed FSDP/DDP training, Flash Attention v2, and Mixture-of-Experts support
- Developed **Mixture-of-Refusals (MoR)**: defense mechanism achieving 2-3x speedup on safe queries while maintaining identical safety guarantees through conditional safety routing
- Implemented vLLM-based inference engine with quantization, speculative decoding, and prefix caching; deployed with Docker, Kubernetes, and Prometheus

Arcane ML: Distributed Training Framework

2024

Open Source Project

- Built framework for distributed ML training across SSH clusters, Modal Cloud GPUs, and local setups, with PyTorch DDP support and a unified CLI

TEACHING EXPERIENCE

PES University

Teaching Assistant — Machine Learning & Deep Learning

Bangalore, India
Sep 2025 – May 2026

- Taught and mentored a cohort of 75+ students across Machine Learning and Deep Learning courses
- Prepared lecture slides, delivered tutorial sessions, and conducted weekly hands-on laboratory sessions
- Designed and graded assignments, lab exercises, and hackathons to reinforce practical understanding

PESU I/O

Subject Matter Expert — Deep Learning

Bangalore, India
Oct 2024 – Nov 2024

- Co-instructed a 30-hour course on Deep Learning from Scratch for a cohort of 40 students
- Delivered 20 hours of in-person instruction covering neural networks, CNNs, RNNs, LSTMs, GANs, and Explainable AI

AWARDS & HONORS

Amazon AI-ML Scholar

Selected from top 1,000 AI/ML students in India

2024

Oxford Machine Learning Summer School

Selected participant, University of Oxford

2024

Cohere AI Summer School

Selected participant in program on large language models and NLP

2024

Winner, Cisco ThingQbator Hackathon 6.0

National hackathon win among 1,000+ competing teams in IoT and emerging technologies

Nov 2023

Merit Scholarships (6x)

MRD Scholarship (2x), CN Rao Scholarship (2x), Distinction Scholarship (2x) for top 5–20% academic performance, PES University

2022–2025

LEADERSHIP & SERVICE

Community Mentor & Technical Speaker, Independent

2023 – Present

- Mentored 50+ teams in hackathons, research projects, and WiDS Datathon competitions
- Delivered technical talks on the DSPy framework and LLM fine-tuning with LoRA at FOSS United and GDSC events

Head of Technology, Entrepreneurship Club, PES University

2023 – 2024

- Led 10+ member technical team; developed club website and managed technical operations for 100+ members

Webmaster, IEEE Robotics and Automation Society, PES University

2023 – 2024

Core Member, Google Developer Student Clubs & Hacker Space, PES University

2022 – 2024

TECHNICAL SKILLS

- **Programming Languages:** Python, Go, Julia, SQL, Bash
- **ML/AI Frameworks:** PyTorch, TensorFlow, Hugging Face, vLLM, LangChain, DSPy, MLflow, FastAI
- **Systems & Infrastructure:** Docker, Kubernetes, Apache Spark, Apache Kafka, Hadoop, Modal, distributed training (FSDP/DDP)