PRANAV DULEPET

+1(925) 997-0461 ◇ San Ramon, CA
ps.dulepet@gmail.com ◇ linkedin.com/in/pranavdulepet ◇ pranavdulepet.github.io

EDUCATION

M.S.E. in Computer Science, Johns Hopkins University — Aug 2026
Specialization: Human Language Technology
B.S. in Computer Science, University of Maryland, College Park — May 2025
Specialization: Machine Learning
Honors: Dean's List, QUEST (Consulting) Honors Program

PUBLICATIONS

- A. Gerami, M. Hoover, **P. Dulepet**, R. Duraiswami FAST: Factorizable Attention for Speeding up Transformers. arXiv preprint arXiv:2402.07901, 2024.
- S. Schulhoff, M. Ilie, N. Balepur, **P. Dulepet** *et al.* The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv preprint arXiv:2406.06608, 2024.

WORK EXPERIENCE

Software Engineer Intern (AI/ML)
Apple (Apple Intelligence) — May 2025 - Aug 2025

- **AI/LLM infrastructure** for Apple Intelligence, all work is under **NDA**
- Developed feature related to making LLM outputs more useful in a variety of applications
- Integrated feature into **internal AI evaluation and monitoring tool** for user feedback and improving the developer process
- Built prototypes and presented to various **app/feature teams** and **senior leadership**

Software Developer Intern (AI/ML)
Amazon (Alexa) — Jun 2024 - Aug 2024

- Received a **full-time offer** based on the quality and impact of work as an intern
- Developed an **end-to-end recommendation pipeline** using customer-Alexa interaction data
- Built a data pre-processing framework with PySpark for over **60TB** of interaction data
- Prompted and built with **Claude 3 Sonnet through AWS Bedrock** to generate structured and cohesive outputs from customer interaction data with an **acceptance rate of 94%**
- Experimented and implemented custom **evaluation techniques derived from research papers**, including **RAG** and **LLM-as-a-judge** approaches
- Collaborated with Alexa Applied Scientists to use their generated user summary methods to reduce our team's costs and streamline the recommendation process
- Provided **detailed documentation and productization steps** which are currently being implemented

Student Consultant
University of Maryland — Aug 2022 - May 2024

- Developed a **graph-based ML tool** to automate medical tool and labor pricing for Capital i, reducing their pricing processing time from 10 days to 10 seconds
- Worked with the **CEO of Capital i** and the main technical lead, a PhD, to adhere to business needs and technological limitations
- Recommended a new pricing strategy using a **linear regression model** for shipping and total cost for the non-profit, Firstbook
- Explored various techniques such as **Random Forest, Decision Trees, and Support Vector Regression** before settling on a linear regression-based model
- Worked with UMD's Office of Student Conduct to analyze and recommend solutions and metrics for understanding and improving mental health on campus
- Conducted surveys with students, faculty, and administrators to develop an assessment and combination of resources to quantify overall 'student happiness' on campus

Software Engineer Intern
Fidelity Investments — Jun 2023 - Aug 2023

- Built a LinkedIn-like MyNetwork recommendation engine for an internal Fidelity app for **80k employees**
- Met with engineers, product managers, and the Vice President of Software Engineering to include necessary downstream tasks

- Explored various machine learning and deep learning techniques before identifying a **graph neural network for link prediction** as the ideal framework
- Implemented a prototype of **Reinforcement Learning with Human Feedback** to constantly improve the recommendation engine as more employees use the feature
- Achieved recommendations with a **98% satisfaction rate** during initial user testing
- Used Python, PyTorch, DGL, Swift to build the custom Graph Neural Network to train and inference
- Identified bugs/improvements in the internal app and increased code coverage by 50%

Machine Learning Intern

Capital One Jan 2023 - May 2023

- Implemented NMSLIB similarity search frameworks on financial graph embeddings as part of the Enterprise Graph Services Team to **detect transaction fraud**
- Implemented **approximate nearest neighbor (ANN) search algorithms using NMSLIB**, optimizing large-scale search efficiency with graph-based data structures
- Identified **HNSW (Hierarchical Navigable Small World) algorithms** to improve the speed and scalability of similarity search for fraud detection
- Built and fine-tuned **graph neural network (GNN) models** to create robust embeddings for financial transaction data, enhancing machine learning predictions in risk management
- Applied to samples of up to 5 million in size with high-dimensional outputting **over 90 recall (success rate)**
- Tested framework with Merchant-Account data resulting in similar recall

Software Engineer (Tech Lead)

Hack4Impact Sep 2021 - May 2023

- Revamped the non-profit, Edu-Futuro's, **internal website** to include Case & Service Management and Beneficiary Creation workflows
- Developed a **Dashboard and messaging portal** for non-profit, Step Up Tutoring to help tutors better connect with students and parents
- Utilized **React, Node.js, and Firebase**, managing a team of 5 software engineers and 2 designers with the help of a co-tech lead and 2 product managers

Software Engineer Intern

Evozyne Jun 2022 - Aug 2022

- Developed SMT solvers (Z3) in Python to decrease the runtime of modeling the Gene Synthesis process by 5x while maintaining precision
- Researched and implemented **Z3 SMT solver by Microsoft Research** to solve NP-complete problems in gene synthesis
- Gained deep understanding of **first-order logic and its application in computational biology** to improve the gene cloning process
- Conducted **exploratory analysis on SMT solvers**, comparing performance and feasibility for biological datasets
- Explored SMT's potential use cases in Gene Regulation Networks, Reversing Genomes, and Protein Folding

RESEARCH EXPERIENCE

Research Assistant

Johns Hopkins University - CLSP (PI: Professor Benjamin Van Durme) Aug 2025 - Present

- LLM cost and confidence calibration for high-stakes domains

Technology Policy Fellow

Paragon Policy Fellowship Sep 2024 - May 2025

- Designing a **streamlined AI model approval** process for the Santa Clara County Government, reducing redundant labor by consolidating three separate interviews into a centralized questionnaire
- Developed and implemented an **AI Usage Guidelines document** to assist clients in accurately and comprehensively submitting GenAI applications, based on research into optimal submission strategies and existing model approvals
- Conducted in-depth research on **GenAI usage trends**, **regulatory challenges**, and **ethical considerations** in public sector applications to guide policy and risk management strategies for Santa Clara County
- Analyzed GenAI implementation practices across government entities, focusing on regulatory frameworks, ethical challenges, and public perceptions, to provide **data-driven recommendations** for Santa Clara County
- Providing insights into AI best practices in the public sector, recommending tailored strategies for ethical, secure, and effective GenAI deployment in government settings

Undergraduate Researcher
University of Maryland - PIRL (PI: Professor Ramani Duraiswami) Jan 2023 - Jan 2025

- Helped **develop a factorable attention mechanism** reducing transformers' complexity to $O(N)$, inspired by fast multipole and Gauss transform methods (view paper)
- Ensured this streamlined process still captures complete data relationships, avoiding data loss often seen with similar methods
- **Created experiments** for Tiny Shakespeare, MNIST, and the Long Range Arena (LRA) datasets and benchmark
- Structured experiments as **Slurm jobs and integrated Weights & Biases** to monitor and evaluate results
- Assisted in creating figures, particularly the **attention matrices**
- Also worked with Swift, LiDAR, and Autonomous Reinforcement Learning simulations

Undergraduate Researcher
University of Maryland - CLIP (NLP Lab) Jan 2024 - Jun 2024

- Contributed to The Prompt Report: A Systematic Survey of Prompting Techniques (view paper)
- Collaborated with researchers from the **University of Maryland, Stanford, OpenAI, Princeton, Microsoft**, and more
- Led and authored the **meta-analysis** of the Multi-modal, Evaluation, and Chain-of-Thought prompting sections
- Conducted a **literature review** and explored various prompting techniques on open and closed-source language models to write **custom definitions and analyses**

Undergraduate Researcher
University of Maryland - GAMMA (PI: Dinesh Manocha) Jan 2024 - May 2024

- Joined the GAMMA lab as part of CMSC 499A - Research with Professorial Faculty, a class part of the Computer Science Honors Program
- Developed a **pipeline for camera-controlled view synthesis using Stable Diffusion and Zero123++**, extending the Hawkl framework for text-controlled aerial view synthesis
- Integrated **mutual information guidance** from input and Zero123++ models, experimenting with homography through summation, averaging, and weighted combinations
- Achieved background manipulation while maintaining foreground consistency in aerial images, exploring various strategies for camera angle stability
- Worked on ensuring **temporal consistency** in video generation, applying the developed techniques across multiple frames
- Tested variations of **adding noise to latent spaces**, experimenting with homography-based transformations in Zero123++

TEACHING EXPERIENCE

Undergraduate Teaching Assistant
University of Maryland Aug 2023 - May 2024

- Taught Python and data science topics to **90 undergraduate** students from the Computer Science, Engineering, and Business schools
- **Developed and graded problem sets and exams**, as well as tutored students during office hours
- Helped student teams communicate with their industry clients to conduct data analysis

PROJECTS

view more details

agora. *Large Language Models, LangChain, Python, SwiftUI, Swift, AWS, MongoDB, Rest APIs*
Developed an iOS app and an **agentic LLM pipeline** to provide personalized and affordable meals and recipes for students. Adapted **Stable Diffusion to generate visuals**. Integrated Amazon Fresh and Kroger API to buy provide automatically filled shopping carts for users to order. Recieved a shout-out from two University newspapers: UMD Computer Science Dept. and UMD's premier student newspaper, the Diamondback. (website link)

College RO *Swift, SwiftUI, Python, Node.js, Rest APIs, MongoDB, AWS, Google/Firebase Analytics*
Launched CollegeRO on the App Store, helping college students find **research opportunities**, reaching a peak of **2k app units**. Provides easy access to a continuously updated list of research opportunities that users can search through with structured and highly-personalized queries regarding their skills, interests, etc. (app link)

LegalAI *Python, scikit-learn, spaCy, Elasticsearch, Textacy, Blackstone, pytextrank*
Trained and tested documents from the Supreme Court and other legal groups to apply **NLP techniques** such as Classification, Similarity, and Summarization. Implemented **TF-IDF, LDA, BM25, textrank, etc.** (GitHub link)

Things Near Me *Full-Stack iOS Development, Swift, UIKit, Firebase*
Developed Things Near Me, for people to share the **availability of supplies** in the neighborhood, reaching a peak of **1.6K app units**. Users can post that they have or need certain supplies. Users can also search on a map interface to help or pick up supplies they need. (app link)

Aerial Object Detector *Python, YOLOv5, PyTorch, Google Colab, GitHub*
Developed a prototype of a model that **classifies harmful and non-harmful objects** in the air. Applied transfer learning to YOLOv5 to detect harmful balloon-shaped objects in images and videos. Won **1st place** at the Northrop Grumman Innovation challenge at the University of Maryland. (GitHub link)

COURSEWORK

Graduate Level: Natural Language Processing (JHU), Intro to Human Language Technology (JHU), AI Safety Alignment & Governance (JHU), LLM Security & Privacy, Long-Context LLMs, Systems for Machine Learning

Artificial Intelligence: Intro to Artificial Intelligence, Intro to Deep Learning, Intro to Machine Learning, Machine Learning Research

Computer Science Core: Algorithms, Data Science, Advanced Data Structures, Discrete Structures, Introduction to Computer Systems, Object Oriented Programming I & II

Specialized Topics: Computer Vision, Human-Computer Interaction, Networks & Security, Organization of Programming Languages, Undergraduate Research, Undergraduate Honors Seminar, Linear Algebra, Applied Statistics & Probability, Calculus I & II

LINKS

- **Email:** ps.dulepet@gmail.com
- **LinkedIn:** https://www.linkedin.com/in/pranavdulepet/
- **Portfolio Website:** https://pranavdulepet.github.io/
- **FAST Paper:** https://arxiv.org/abs/2402.07901
- **Prompt Report Paper:** https://arxiv.org/abs/2406.06608
- **Mental Health Assessment Resource:** https://mentalhealth.umd.edu
- **Enterprise MyNetwork Platform Medium Article:** https://medium.com/@pdulepet/enterprise-mynetwork-platform-c138f7e98537
- **Capital One NMSLIB Similarity Search Frameworks:** https://www.capitalone.com/tech/machine-learning/sim search-graph-embeddings/
- **SMT in Computational Biology Medium Article:** https://medium.com/@pdulepet/smt-in-computational-biology-dccf006eb397
- **agora. Website:** https://master.d1frbpmrrocpzu.amplifyapp.com/
- **College RO App Store Link:** https://apps.apple.com/us/app/college-ro/id1577113429
- **LegalAI GitHub Repository:** https://github.com/pranavdulepet/legalai
- **Things Near Me App Store Link:** https://apps.apple.com/us/app/things-near-me/id1506053357
- **Aerial Object Detector GitHub Repository:** https://github.com/pranavdulepet/aerial-object-detection