Swastik Agrawal

+1 (240) 501-5034 | swastik3@terpmail.umd.edu | linkedin.com/in/sw-agrawal/ | github.com/swastik3

Education

University of Maryland, College Park College Park, MD

Bachelor of Science in Computer Science and Mathematics: GPA 3.969 (Dean's List) *Aug 2023 – Dec 2026*

Relevant coursework: Advanced Numerical Optimization (grad course), Proof-based Linear Algebra, Advanced Calculus, Computational Methods, Applied Linear Algebra, Introduction to AI, Algorithms, Computer Systems, Discrete Structures

Technical Skills

Languages: Python, JavaScript/TypeScript, Java, C, OCaml, Rust
Frameworks & Tools: PyTorch, NumPy, Flask, React, Pandas, LangGraph, Git, Firebase, MongoDB, PostgreSQL

Work Experience

AI Product Intern June 2026 - Present

Boston Consulting Group (BCG) X *San Francisco, CA*

- Developing a multi-agent workspace to accelerate client proposal development, spanning institutional knowledge retrieval, expert search, fanned-out research, and competitor analysis
- Deploying the integration company-wide via a Claude-native plugin and MCP server, driving a 20+ hour reduction in proposal turnaround and shifting team focus toward strategic differentiation and pitch quality

SWE Intern Feb 2026 – May 2026

Stealth Startup *Remote*

- Built competitor performance metrics with real-time alerts to drive social media marketing decisions
- Developed a fully end-to-end agentic microservice for video creation using Mastra, automating the full workflow from content planning to final delivery across 15+ video formats
- Managed cloud deployments for these features using Terraform across AWS services including Step Functions, Lambda, ECS Fargate, and S3

SWE Intern May 2025 – Dec 2025

Mindgrasp Inc. *Bethesda, MD*

- Developed an in-house graph framework for LLM pipelines to implement complex multi-threaded graphical workflows, enabling improved tool calling and response streaming using SSE (reduced user latency by $\sim 90\%$)
- Migrated the codebase from multiple distributed services to a monorepo architecture (reducing operational costs by $\sim 40\%$) and enhanced database schemas to enable feature improvements across 4 company products
- Refined signup and payment funnels through A/B testing resulting in $\sim 80\%$ increase in registration rate
- Deployed AWS lambda functions for audio transcriptions and file processing; contributed 500+ unit tests

AI Engineer Intern Jan 2025 – May 2025

Fetch.ai *Remote*

- Developed 10+ agents with sophisticated LangGraph workflows equipped with web search to compile travel options through natural language derived API calls to multiple third party services

Co-Founder and Head of Tech Jul 2024 – Aug 2025

AI/ML at UMD *University of Maryland, College Park, MD, MD*

- Milestones: 600+ members, 5 workshops, 5+ company collaborations, 2x UMD hackathon judge
- Managed the development of a recommendation engine for WISE Cities LLC for website listings using composite scoring based on popularity data, matrix factorization, and keyword search
- Led a research project aimed at building a real-time context-injection pipeline with lecture transcriptions to answer queries using a dynamically updated knowledge base (piloted in 4 UMD classes)

SWE Intern Sep 2024 – Dec 2024

Untethered Labs Inc. *Greenbelt, MD*

- Migrated the Node.js company portal codebase from node v12 to v16 incrementally by refactoring code, minimizing clutter, reducing initial load time by 12%
- Implemented an analytics dashboard to monitor tenant usage across enterprise deployments, for real-time metric analysis of key products

Research Experience

Student Researcher | **ICML 2026 Mech Interp Workshop** — Feb 2026 – Present
Dr. Sarah Wiegreffe Research Group — *University of Maryland, College Park, MD*
- Investigating the performance gap in activation steering (Contrastive Activation Addition) by controlling for steering coefficient, layer, token position and dataset formulation
- Exploring a geometric proxy using the difference-of-means line to estimate test-time steering success in mutliple-choice and open-ended generation settings

Independent Researcher | **ICML 2026 Mech Interp Workshop** — Feb 2026 – Present
Supervised Program for Alignment Research (SPAR) — *Remote*
- Developed Circuit Oracle: an agentic pipeline for automated end-to-end circuit discovery, interpretation, and causal hypothesis generation in LLMs
- Conducted jailbreak circuit discovery on Qwen-3-4B and Gemma-2-4b-it against GCG and AutoDAN attacks

Student Researcher | **COLM 2026** — Jan 2026 – June 2026
Dr. Fardina Alam Research Group — *University of Maryland, College Park, MD*
- Co-led research on training-free AI-generated text detection using signals from letter distributions
- Built the MDTA benchmark (642,274 prompt-aligned samples across 4 models, 5 domains, 3 temperatures, 3 adversarial strategies including lipogrammatic attacks), expanding HC3
- Introduced LD-Score, a orthogonal detection signal when integrated with perplexity based detectors like Binoculars and DNA-DetectLLM improves AUROC and F-1 Score

Research Assistant — Sep 2025 – Dec 2025
Tubaldi Lab — *University of Maryland, College Park, MD*
- Developed a communication framework for the UR3e robotic arm, implementing automated trial randomization, sorting operations, and streamlined data collection scripts
- Designed a bimodal object classification model integrating temporal pressure sensing and visual input

Student Researcher | **Paper: Link** — Feb 2024 – Aug 2025
Risk-Informed Solutions in Engineering Laboratory — *University of Maryland, College Park, MD*
- Engineered a temporal deep-learning model to impute unavailable data on tropical cyclones, improving the performance of existing shallow approaches by 15%
- Tested temporal model architectures, feature sets, physics-driven approaches and fine-tuning pipelines
- Presented at the ICOSSAR 2025, AGU Fall meeting 2024, and at UMD CEE symposium 2024

Projects

Winnow | *PyTorch, Transformers, Modal, FastAPI* — June 2026
- Developed a long-context handling pipeline combining soft activation-space compression, hard token pruning, and KV-cache quantization, relatively nascent research directions, into a unified system for faster long-context inference
- Implemented token pruning by fusing two importance signals: LLMLingua-2 (token-level classification) and AttentionRAG (sentence-level attention scoring)
- Applied Latent Context Language Models (LCLM) with TurboQuant KV-cache quantization for A100 GPUs

L-LMS (Winner HackMIT) | *React.js, Fetch.ai, Iris DBMS, LangGraph, Flask* — Sep 2024
- Developed a personalized agentic AI study assistant that leverages official course information from Canvas LMS, and user uploaded documents
- Implemented a multi-agent RAG pipeline by curating a LangGraph-based AI engine, minimizing hallucinations
- Designed 2 fetch.ai agents: one allowing users to OCR handwritten notes + miscellaneous study material and the other to send emails to professors using the Gmail API

DenseTeX | *PyTorch, Transformers, Wandb, Git* — May 2024 – Jul 2024
- Designed a novel encoder-decoder transformer architecture to convert images of scientific math equations to LaTeX
- Developed a DenseNet encoder with 2D Positional Encoding, paired with a custom GPT-2 based transformer and an extended RoBERTa-based tokenizer
- Integrated wandb to track test metrics (pplx, BLEU, F1, etc) for the training process

PocketPeds (App Dev Club, UMD) | *React-native, Tailwind, Flask, PostgreSQL, Git* — Feb 2024 – May 2024
- Led the development of PocketPeds, an application to improve pediatric healthcare by helping caretakers calculate safe dosages for pediatric OTC medications using a mobile device
- Developed a robust backend integrating MongoDB, PostgreSQL, JWT authentication and the FDA NDC API

Misc

7x Hackathon Winner: MIT, Harvard, GaTech, UMD, Georgetown University, Stony Brook, Nucleate@Pitt