Abhinand Jha

412-519-8559 | reachabhinandjha@gmail.com | linkedin.com/in/abhinandj/ | abhinand20.github.io

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Computer Engineering (Machine Learning minor) – GPA 4.0/4.0 — *Jan 2022 – Dec 2023*

Manipal Institute of Technology — Karnataka, India
Bachelor of Technology in Electrical Engineering – GPA 4.0/4.0 — *May 2016 – Aug 2020*

Experience

Google, Search — Mountain View, CA
Software Engineer L4, Multimodal AI — *Sept 2025 – Present*

- [Blog] Contributing to Google Search's Multimodal AI team to make Search products like **AI Mode, Lens, and Search Live** natively multimodal – spanning infrastructure, frameworks, and user-facing experiences
- Partnering with Google DeepMind researchers, Search engineers and PMs to accelerate development and launch cycles for new multimodal Search capabilities in Google products powered by Gemini model advancements

Google, Cloud Compute — Seattle, WA
Software Engineer L3–L4, GCE Fleet Deployment Platform — *Feb 2024 – Aug 2024*

- [Press release] **Led and delivered** a high-priority Sovereign Cloud initiative by designing and implementing distributed systems enabling large-scale rollouts across Google Compute Engine (GCE) fleets of VMs, GPUs, and TPUs
- [Blog] Architected and developed the next-generation GCE Fleet Release Management Infrastructure, improving deployment scalability, release reliability, and support for frequent NPI launches and maintenance-sensitive hardware (GPUs, TPUs) for GCP's **AI Hypercomputer HPC** offering
- Improved team's oncall operational burden by **developing an LLM-powered CLI tool** substantially reducing operational toil and accelerating issue debugging and mitigation times

Software Engineer Intern, GCP Vertex AI — *May 2023 – August 2023*

- Designed and developed a full-stack web application that uses Large Language Models and Recommender systems to empower Google Cloud Support Engineers, resulting in a **30% reduction in response time** for customer cases
- Built pipelines for scheduled data ingestion of over 1 million records weekly, into vector databases and model fine-tuning using GCP, resulting in improved efficiency of the data pipeline and reduced data ingestion time

Deloitte — India
Data Scientist — *Sept 2020 - Dec 2021*

- Secured 1st position out of 30 teams in a firm-wide Hackathon by developing a tool that generates ontologies from text documents, deployed in production – eliminating the operational burden of **100 swe-hours/year**
- Designed and implemented full-stack AutoML pipelines integrating enterprise data sources, and established foundational CI/CD automation, improving model deployment reliability and development velocity

Publications

KOALA: Knowledge Conflict Augmentations for Robustness in Vision Language Models
International AAAI Conference on Web and Social Media 2025 [paper] [code]

Quantifying Memorization and Retriever Performance in Retrieval-Augmented Vision-Language Models
ACL Workshop on Large Language Model Memorization (L2M2) 2025 [paper]

Chaotic clock driven cryptographic chip: Towards a DPA resistant AES processor
IEEE Transactions on Emerging Topics in Computing. 10(2):792–805, 2022 [article]

Technical Skills

Languages: Python, Golang, C++, Java, Scala, Kotlin, TypeScript, SQL
Tools & Frameworks: Tensorflow, Pytorch, Kubernetes, Spanner, GCP, Azure, AWS, Helm, Docker, CI/CD, Obervability and Alerting