PRAHALADH CHANDRAHASAN

(412) 339 - 7156|prahald92@gmail.com|Github| LinkedIn| Website

EDUCATION

CARNEGIE MELLON UNIVERSITY (SCHOOL OF COMPUTER SCIENCE) *Pittsburgh, PA*

Master's in Information Technology (GPA : 3.92/4) *Aug 2024-Dec 2025*

WORK EXPERIENCE

Research Assistant - LTI Carnegie Mellon University *Pittsburgh, PA*

Machine Learning Engineer *Jan 2025 - Present*

- Engineered and deployed **DeepResearch Comparator**, a large-scale agentic evaluation platform on an **EKS Cluste**r, enabling **fine-grained human feedback collection** and benchmarking of closed- and open-source agents.
- Productionized **deep research agents** and **multimodal RAG systems** by exposing them as **scalable APIs**, integrating visualization and monitoring of outputs and metrics using **ZenoML**.
- Benchmarked DeepResearch agents (e.g., WebThinker, custom agents) on ***BrowseComp*** and ***HealthBench*** using **HPC clusters**, optimizing throughput with the **vLLM** library by tuning model serving parameters for various GPU configurations.
- Built a live arena-style platform for **NeurIPS MMU-RAG (Text-to-Video track)** as a cloud-native application to host baseline and participant VideoRAG submissions, developed baseline VideoRAG systems, and a fast evaluation pipeline for submissions on ***VideoBench.***

Bank of America Continuum India *Chennai, India*

Software Engineer *Jul 2022-Jul 2024*

- Accelerated testing efficiency by implementing Tosca-based automation for comprehensive end-to-end payment flows from initiation to clearing, eliminating **1,000+** manual regression testing hours annually and improving release velocity by **65%**
- Architected reusable Tosca UI and API modules deployed across 5 regional payment landscapes, reducing development cycles by **40%**
- Detected and resolved 5+ critical production defects through rigorous testing protocols, preventing potential revenue loss of **$5 million**
- Served as on-call engineer for daily sanity runs, collaborating with release and environment management teams to ensure seamless deployments and rapid incident resolution
- Spearheaded innovation by developing and filing a **patent** for a payments fraud detection algorithm and presented work across various business verticals

RedHat *Bangalore, India*

Software Engineer Intern *Jan 2022-Jul 2022*

- Engineered and delivered two critical features (ENTESB-18633 and ENTESB-18785) successfully integrated into **Hawtio release 7.11** using **AngularJS** and **PatternFly** framework, enhancing real-time monitoring capabilities for RedHat Fuse environments and improving enterprise customer experience
- Managed and maintained three core Hawtio repositories (hawtio,hawtio-core, hawtio-integration) as a primary contributor, improving code quality through unit tests, enhanced documentation, and package upgrades.
- Implemented automated CI/CD workflows using **GitHub Actions** across the entire Hawtio project ecosystem, streamlining issue management processes and reducing maintenance overhead by 15% for the development team

PROJECTS

Language Model Implementation from Scratch

- **Implemented sparse attention mechanisms and LoRA parameter-efficient fine-tuning** to enhance Llama2 model efficiency and performance on sentiment classification tasks, achieving improved computational scalability while maintaining model expressiveness across SST-5 and CFIMDB datasets
- **Extended the baseline architecture with differential attention patterns** as an advanced attention variant, systematically comparing sparse attention, standard attention, and differential attention mechanisms to demonstrate measurable performance improvements in both zero-shot prompting and fine-tuning scenarios
- **Developed a comprehensive analysis of attention pattern efficiency** by implementing multiple decoding strategies (beam search, nucleus sampling, top-k sampling) in conjunction with sparse attention, demonstrating how attention sparsity affects text generation quality and computational overhead in sentiment-aware language modeling tasks

Evaluation methods for identifying gender bias in LLMs

- **Investigated gender bias in Large Language Models (LLMs)** by analyzing personality trait expression across GPT-4o and Llama-3.1-8B-Instruct, introducing metrics such as Stereotype Alignment Score, Empirical Alignment Score, and Bias Amplification Factor
- Extended baseline research on AI behavioral similarity by implementing **gender-specific prompting strategies**, conducting comparative analysis with human personality trait distributions, and uncovering how LLMs amplify or diverge from societal stereotypes
- Identified that GPT-4o demonstrates significantly higher stereotype alignment in Agreeableness and Conscientiousness dimensions, while Llama-3.1-8B-Instruct exhibits minimal gender differentiation but unrealistic personality profiles overall

SKILLS

Programming Languages : C, C++, Python, Java, SQL, JavaScript, HCL, Shell Scripting (Git & Bash)

Frameworks and Libraries : Pytorch, FastAPI, Flask, HuggingFace, OpenAI, Tensorflow, vLLM, wandb,MLFlow, LangChain, LlamaIndex, XGBoost, Scikit-learn, Pandas, Numpy, Scipy, Pysyft

Cloud & Devops : AWS(EC2, S3, EKS, ECR, RDS, IAM, Bedrock, SageMaker), Kubernetes, Docker, Terraform, Github Actions