PRAHALADH CHANDRAHASAN

+1 412-339-7156|prahald92@gmail.com|Github| LinkedIn| Website

WORK EXPERIENCE

Circle AI *Jan 2026 - Present*
Founding Forward Deployed Engineer *San Francisco, CA*

• Designed a container orchestration system for AI agent execution on **Azure Container Apps**, implementing a **Redis-coordinated warm pool** that reduced task startup latency from ~30s to ~2s, **WebSocket**-based bidirectional messaging with **stream rate limiting**, and **mid-execution message injection** enabling dynamic task steering

•Engineered an **email processing pipeline** with a thread consolidation engine, handling parallel attachment uploads, and follow-up depth tracking to unify multi-email insurance workflows into a single agent execution context

• **Created the pre-production environment** end-to-end — Azure AD app registration, Supabase instance, Vercel preview workflows, and Service Bus routing — then led migration of all container apps into a **unified VNet with static IP allocation**, **RBAC-based Key Vault access**, managed identity auth, and **Service Bus queue-depth autoscaling**

• **Optimized agent performance** by pre-baking skill definitions and system dependencies into Docker images, **parallelized Redis batch operations** for container pool management, implemented **command-level sandboxing** restricting agent bash/file access, and added per-environment rate limiting

• Built cross-brokerage task sharing with **RLS-enforced access control,** and PostgreSQL RPC-based discovery, enabling multi-user collaboration while maintaining **SOC2-compliant** brokerage data isolation

• Migrated an entire service to structured Python logging with context fields, built a **DLQ reprocessor cron job** for automatic recovery of failed agent tasks with status-aware filtering and retry tracking, and instrumented approval workflows with execution-type attribution in **Sentry**

Language Technologies Institute *Jan 2025 - Dec 2025*
Machine Learning Engineer *Pittsburgh, PA*

• 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** as a cloud-native application to host baseline and participant VideoRAG submissions, developed baseline VideoRAG systems, and a fast evaluation pipeline for submissions on ***VBench***

• Co-developed **RefusalBench Studio**, an end-to-end interactive platform for generating dynamic, linguistically controlled evaluations that measure an LLM's ability to detect and perform selective refusal in RAG settings, featuring an Inference Lab where user-defined evaluation workflows are executed by a ReAct-style orchestrator agent with tool-augmented reasoning

• Evaluated nine frontier models (**Leaderboard**) on 1,600 RefusalBench-NQ instances under vanilla and agentic self-correction workflows, orchestrating parallel multi-model verification with four verifiers against an eight-point checklist and streaming results to a real-time dashboard; containerized the full pipeline for reproducible, scalable execution

Bank of America Continuum India *Jul 2022 - Jul 2024*
Software Engineer *Chennai, India*

• Architected **Tosca-based** end-to-end **automation** for payment flows across five landscapes, eliminating 1,000+ manual regression hours annually and improving release velocity by 65%, while detecting **critical defects** and filing a **patent** for a payments fraud detection algorithm

RedHat *Jan 2022 - Jul 2022*
Software Engineer Intern *Bangalore, India*

• Engineered two features (ENTESB-18633 and ENTESB-18785) shipped in **Hawtio release 7.11** using **AngularJS** and **PatternFly** framework, enhancing real-time monitoring capabilities for RedHat Fuse environments and improving enterprise customer experience

SKILLS

Programming Languages: Python, JavaScript, TypeScript, Bash, SQL, C++, Java

Frameworks & Libraries: PyTorch, FastAPI, Node.js, Express.js, Hugging Face, OpenAI, TensorFlow, vLLM, W&B, LangChain, LlamaIndex, Anthropic Agent SDK, React, LlamaParse, Sentry, SGLang

Cloud & DevOps: AWS (EC2, S3, EKS, ECR, RDS, IAM, Bedrock, SageMaker), Azure(Container Apps, Service Bus, IAM, AI Foundry), Redis, Kubernetes, Docker, GitHub Actions

EDUCATION

CARNEGIE MELLON UNIVERSITY (SCHOOL OF COMPUTER SCIENCE) *Pittsburgh, PA*
Master's in Information Technology: Privacy Engineering (GPA : 3.92/4.0) *Aug 2024 - Dec 2025*

PROJECTS

Prompt based steering for AI Safety

•Developed **BERTSteer**, a prompt-conditioned activation steering method for machine unlearning that uses a **fine-tuned DistilBERT** intent classifier to selectively clamp top-k harmful latents from 16K **sparse autoencoder** features on Gemma-2-2B at inference time, suppressing hazardous knowledge while preserving general capabilities; generated synthetic training data using Llama 3.3-70Band ran feature extraction, SAE interventions, and ablations on g5.4xlarge instances with NVIDIA A10 GPUs