BHAVESH SOOD

Pittsburgh, PA | (310) 824-3179 | bsood@andrew.cmu.edu | www.linkedin.com/in/bhavesh-sood

EDUCATION

Carnegie Mellon University — Pittsburgh, PA
MS in Electrical and Computer Engineering | GPA – 3.67 / 4.00 — May 2026
- Relevant Courses – Generative AI (10-623) , Advanced Cloud Computing (15-709), Speech Technology (11-492)
- Current Courses – Algorithms for Distributed ML & Optimization (18-667), LLM – Methods and Applications (11-667)

IIIT Delhi — New Delhi, Delhi
B.Tech Honors in Computer Science and Artificial Intelligence (CSAI) | GPA - 9.24 / 10.0 — June 2023
- Valedictorian in CSAI, Dean's List in all 4 years (2019-20, 2020-21, 2021-22, 2022-23)
- Relevant Coursework – Deep Learning , Advanced Machine Learning, Computer Vision, NLP, ML, DSA

SKILLS

Expertise Areas : Algorithms, Deep Learning, Statistics, LLMs, Transformer Models, Cloud Computing, Distributed Systems
Programming Languages : Python, C++, Java, HTML, CSS
Tools and Technologies : PyTorch, WandB, NumPy, Git, Spark, Hadoop, Terraform, Kubernetes, huggingface transformers

RESEARCH EXPERIENCE

CMU ML Department — Pittsburgh, PA
Research Project (Advised by Prof. Matt Gormley) — Aug 2025 – Present
Research Assistant (Advised by Prof. Matt Gormley) — May 2025 – Aug 2025
- Pioneered methods to improve LLM reasoning by introducing novel token-level interventions, enhancing controllability of model "thinking" behaviors leveraging the transformer architecture.
- Curated a 1B-token custom dataset from diverse code and math texts, and conducted pretraining experiments on LLaMA-3 (1B & 8B) and OLMo models through distributed GPU training with FlashAttention and FSDP, with ongoing evaluation on reasoning benchmarks (e.g., GSM8K and CommonsenseQA), all in Python.

WORK EXPERIENCE

Palo Alto Networks — Gurugram, India
Software Engineer (in ML) — Aug 2023 – Jan 2025
- Led the ML-Ops integration and transition of the Personally Identifiable Information (PII) detection pipeline from test stages to General Availability (GA) by upgrading entity support to 79 entities across multiple countries, advancing system observability with a Garuda dashboard, and improving multi-lingual keyword detection, boosting accuracy and scalability.
- Built ML models for ID-card image detection and enhanced the ML Feature Extractor Consumer service, increasing code coverage from 0 to 93%. Resolved a critical business logic issue in PII-Structured, enabling a major client to onboard successfully.

Expedia Group — Gurugram, India
ML Engineer – Intern — May 2022 – Jul 2022
- Collected / Analyzed data from various Tables and performed feature engineering to form the final dataset.
- Developed a Self-Serve Failure Prediction Model from scratch using PyTorch wrapped in PyTorch Lightning for production level code. Created a service using FastAPI to wrap and deploy the model for use in Production.

PROJECTS

Prompt-to-Prompt Image Editing with Latent Diffusion Models [CMU Gen AI Course (10-623)] — Mar 2025 – Apr 2025
- Implemented Prompt-to-Prompt image editing by leveraging cross-attention between the Bert based text encoder and the U-Net in Latent Diffusion Model (LDM), providing fine-grained control over text-driven modifications.
- Utilized VQ-VAE (Vector Quantized Variational Autoencoder) to encode images into discrete latent codes and reconstruct them post-editing, preserving structure and fine details.
- Integrated word-swapping from the Prompt-to-Prompt method by Hertz et al. to selectively modify parts of an image from one prompt to another prompt, by adjusting cross-attention maps corresponding to modified tokens.

Request-based Auto-Scaling using AWS Lambda [CMU – Adv. Cloud Computing Course (15-709)] — Feb 2025 – Mar 2025
- Automated infrastructure setup using Terraform, including deploying AWS Lambda functions and configuring security groups and load balancers.
- Configured the Dockerfile to set up the Lambda runtime, install dependencies, and auto-download the ML model from S3 for inference.
- Implemented AWS Lambda for auto-scaling inference requests in a Dogs/Cats classification service, improving scalability and performance using fine-grained capacity provisioning for individual requests.

Building My Own LLAMA-2 Model [CMU – Gen AI Course (10-623)] — Jan 2025 – Feb 2025
- Implemented two efficient techniques in modern LLMs from scratch: Rotary Position Embeddings (RoPE) to upgrade positional encoding and Grouped-Query Attention (GQA) to optimize memory usage and speed.
- Applied Instruction Fine-Tuning (IFT/SFT) employing LoRA (Low-Rank Adaptation) on the attention layers of the LM.