Saksham Gupta

higgsboson1209@gmail.com | linkedin.com/in/higgsboson1209 | +91-8178769108

Education

Vellore Institute Of Technology — Vellore, TN
Bachelor of Technology in Computer Science — *July 2018 – May 2022*

- GPA: 8.96/10.00
- Relevant Coursework: Introduction to Neural Networks, Machine Learning, Image Processing, Introduction to AI, Linear Algebra, Graph Theory

Publications

Personalized action suggestions in low-code automation platforms — Dec 2022
Saksham Gupta, *Gust Verbruggen, Mukul Singh, Sumit Gulwani, Vu Le* — *ICSE 2023; Industry Track*

Experience

Research Engineer II — Apr 2024
Zomato — *Delhi, IN*

- Part of the Generative AI team at Zomato, leading the development of Natural language search
- Fine-tuned and quantized Llama 3.1 8B to generate relevant dishes in under 500 ms using skip-forward decoding
- Trained different embedding models using Torch FSDP to reduce training time by 92% to power semantic search
- Increased the searchability of more than 33% catalog items by training RoBERTa-based model using triplet loss
- Created automated user-feedback driven pipeline for retraining embedding models
- Increased retrieval recall by 23% by performing automated catalog tagging using LLM-powered pipelines
- Set up fully automated data-generation pipelines to generate synthetic training data using OpenAI and asyncio

Member of Technical Staff — Sep 2023 - Mar 2024
Boson AI — *Remote, IN*

- Worked on enhancing role-playing capabilities of a custom LLM as a research engineer under Dr. Alex Smola
- Built a pipeline for image data extraction using Nougat with SGLang to achieve a 30x speed-up on 8 x A-100s
- Scraped 200,000+ game scripts from various online sources; performed data cleaning and labeling for training
- Deployed an in-house data labeling solution hosted on AWS EC2 with PostgreSQL for data storage

Pre-Doctoral Research Fellow — Sep 2022 - Sep 2023
Microsoft Research — *Remote, IN*

- Research Fellow at the PROSE group working on AI4Code led by Dr. Sumit Gulwani
- Trained CodeBERT model using triplet loss for performing fine-grained code comparison and achieving 28% improvement over SOTA CodeBERTScore
- Created benchmark dataset for fine-grained code comparison using Tree-sitter and built pipelines for perturbing and paraphrasing code snippets
- Published work on personalizing language models using user history at ICSE 23, leading to an increase in prediction accuracy of 22%

Software Engineer — Jan 2022 - Aug 2022
Prodigal YC S18 — *Remote, IN*

- Part of the Infra and Data team at Fintech SaaS startup backed by Y Combinator, Accel, and Menlo Ventures
- Built Cupid, a service to match metadata files to audio files as REST API in Go. Reduced downstream processing delays by 7 times with the ability to match and move 1k audio files to S3 in under 1 second
- Built Hermes, a real-time service to move files from SFTP to S3, and reduced audio processing start time delays by 99.6%, and built a REST API in Go to fetch call metadata and modify filenames in AWS RDS
- Built aggregator service to update the call counter of all tenants in REDIS and by processing messages from SQS using Goroutines. Achieved processing speed of 1600 messages/sec, providing 10X speed in comparison to sequential code

Technical Skills

Languages: Python, C++, Go
Frameworks: Pytorch, Huggingface accelerate, PyTorch FSDP
Databases: S3, MySQL, MongoDB, DynamoDB
Developer Tools: Git, Vim, Shell & Bash Scripting