Vaibhav Sahu

2679280709 | vaibhavs@seas.upenn.edu | linkedin.com/in/vaibhav | vsa1920.github.io

Education

University of Pennsylvania — Philadelphia, PA
Master of Science in Scientific Computing - GPA: 3.83/4 — *Aug. 2022 – May 2024*

Indian Institute of Science — Bangalore, India
Bachelor of Science in Physics - GPA: 8.2/10 — *Aug. 2016 – June 2020*

Experience

AI Engineer — November 2024 – Present
IgniteIQ.ai — *Remote, United States*

- Made a dynamic and customizable auto-tagging model utilizing **Pydantic** based **OpenAI Structured Outputs** and **GPT-4o** to retrieve information fields/tags based on provided schema and descriptions.
- Designed a **clause detection agent** using GPT to match clause citations to agreement documents with 95% recall.
- Implemented an algorithm to connect letters with other related letters in different conversation chains using **Semantic Search** (using cached vector embeddings and **FAISS**), Regex, and Breadth-First Search.
- Made an **Agentic Chunking** tool using LLMs that can chunk large documents and preserve contexts across pages.
- Implemented an **OCR-based text extraction tool** using **Gemini 2.0 Flash**
- Designed workflows on file ingestion and processing for a database on a **PostgreSQL** server using the AI tools mentioned above
- Skills: PostgreSQL, GPT-4o, Semantic Search, FAISS, RAG, Agentic Workflows, Python, Pydantic, FastAPI, Gemini

Research Engineer (HPC and AI) — April 2021 – July 2022
Simyog Technology — *Bangalore, India*

- Performed a hot-spot analysis on the **C++** Finite Element Analysis computational solver identifying performance bottlenecks and optimized them with **Intel MKL**, and **Multi-threading** parallelization using **OpenMP**, achieving a **22%** speedup.
- Parallelized 1000s of Matrix-Vector operations and implemented a concurrent GMRES algorithm with optimal contiguous memory usage, boosting solver performance by **40%**.
- Trained models in **Python** for predictive modeling of different ICs using **TensorFlow** and **packaged and deployed** it into the simulation application.
- Skills: C++, OpenMP, Intel MKL, Intel Vtune, Python, TensorFlow, pandas

Projects

AgentsMD: Agentic AI for ER Triage | *Python, OpenAI, AssemblyAI, Flask, SQLite*

- Developed a **multi-agent system** for emergency room triage using **competing LLM agents** to generate differential diagnoses in real-time
- Implemented **speech-to-text** transcription and real-time diagnostic assessment pipeline using the **OpenAI** and **AssemblyAI** APIs
- Built an interactive **web interface** with **Flask** for medical staff to manage patient assessments, record conversations, and view AI-generated diagnoses

Playing Hangman with Transformers | *PyTorch, Hugging Face, NumPy*

- Formulated a **pre-training** task in the form of a multiclass classification for the transformer to learn guessing letters in hangman and generate data using **Biased Random Sampling** Simulations
- Pre-trained the **Google's CANINE character-level LLM** on a corpus of **380k** words to achieve an optimal performance of **56%** game winning accuracy which is on-par with Human-level Performance (**HLP**)
- Implemented a novel **novel self-play fine-tuning** task to achieve **63%** winning accuracy within 6 wrong guesses and **87%** accuracy within 10 wrong guesses

Technical Skills

Languages: Python, C/C++, MATLAB
ML Frameworks: PyTorch, TensorFlow, PyTorch, Scikit-learn, Hugging Face, OpenAI FAISS
Other Frameworks: pandas, NumPy, Matplotlib, Seaborn, OpenMP, AWS, Azure, FastAPI, SQL, Spark, Pydantic
Courses: Deep Learning, Computer Vision, Big Data Analytics, Applied Machine Learning, Natural Language Processing
Conceptual Skills: Machine Learning, Large Language Models, Fine-tuning, RAG, Natural Language Processing, Parallel Computing, Data Science, Speech-to-text

Certifications

Generative AI with Large Language Models — Certificate
Introduction to Machine Learning in Production — Certificate
DeepLearning.AI: Deep Learning Specialisation — Certificate
Fundamentals of Parallelism on Intel Architecture — Certificate