Somesh Bagadiya

someshbgd3@gmail.com (925)819-3504 someshbagadiya.dev Linkedin Github

Professional Summary

Machine Learning Researcher with expertise in **deep learning**, **scalable AI**, and **MLOps**. Skilled in **problem-solving**, **cross-functional collaboration**, and **innovation**. Passionate about delivering **high-impact solutions** through **teamwork**, **adaptability**, and building **intelligent**, **low-latency systems** across research and production environments.

Experience

SJSU Research Foundation **San Jose, CA**
Machine Learning Researcher *June 2024 – Present*

- Led development of a deep learning pipeline for **genome sequence classification**, integrating **PyTorch** and **TensorFlow** to improve inference accuracy by **15%**, supporting high-throughput screening of 1M+ DNA records.
- Designed a modular **test-time preprocessing engine** by parsing **FASTA files**, applying **k-mer encoding**, and performing **dimensionality reduction**, reducing preprocessing latency by **48%** and enabling real-time data readiness in HPC workflows.
- Deployed a **multi-GPU training architecture** using **PyTorch Distributed** on **HPC clusters**, scaling model training across 2 GPUs and decreasing epoch time from 12 to 8.4 hours (**30% speedup**), optimizing compute efficiency in large model training loops.

Artonifs **San Jose, CA**
Software Engineer Intern *May 2024 – Aug 2024*

- Fine-tuned domain-specific **LLMs** using **LoRA and QLoRA**, accelerating adaptation on retrieval-augmented datasets, leading to **28% higher inference accuracy** on classification and QA benchmarks under latency constraints.
- Engineered fault-tolerant, **asynchronous backend microservices** in **Python + FastAPI**, designed for **low-latency inference** under 100ms, improving system throughput by **35%** under simulation of 100K+ user requests.
- Built a dynamic **A/B testing pipeline** to measure inference quality vs. runtime cost using **SQL, Pandas**, and custom logging metrics, resulting in **25% higher feature adoption** in early-stage deployments.

Cognizant - COX **Pune, India**
Software Engineer *Mar 2021 – Jul 2023*

- Designed and deployed a real-time **AI inference system** for **insurance claims processing**, using **FastAPI + Python**, reducing average response time by **50%** across 5M+ predictions while maintaining strict latency SLAs.
- Developed high-throughput **ETL and training pipelines** using **Kafka, SQL**, and **distributed file systems**, reducing feature extraction latency by **40%**, enabling continuous retraining on 10M+ record datasets.
- Streamlined MLOps by automating **CI/CD for model deployment** with **Docker, Jenkins, Kubernetes**, cutting model release cycles from 5 days to 2 days (**60% faster**), enabling rapid iteration of optimized inference models.
- Led design of a **pattern-recognition-based anomaly detection module** using unsupervised learning, improving early detection of production drifts by **35%**, enhancing reliability of predictive services across live environments.

Biencaps Systems **Pune, India**
Data Engineer Intern *May 2020 – Feb 2021*

- Built a modular ETL framework for both structured and unstructured inputs using **Python, SQL**, reducing data ingestion time by **45%** and ensuring real-time availability for downstream analytics.
- Developed a **data validation system** leveraging rule-based checks and statistical anomaly detection, achieving **99.5% accuracy** in financial dashboarding across 100+ business reports monthly.

Education

San Jose State University (SJSU) **San Jose, CA**
Master of Science, Artificial Intelligence (GPA: 3.56) *Aug 2023 – May 2025*

Savitribai Phule Pune University (SPPU) **Pune, India**
Bachelor of Engineering, Information Technology (GPA: 3.59) *Aug 2017 – May 2021*

Projects

- ***CarbonSense powered by IBM WatsonX:*** Built an **AI-driven carbon footprint platform** leveraging **WatsonX**, **fine-tuned LLMs**, and **Retrieval-Augmented Generation (RAG)** for accurate impact assessments. Designed a **CrewAI multi-agent system** for data parsing, research, and analysis. Deployed **Milvus-based vector search** for scalable retrieval and developed a **FastAPI dashboard** integrating **Watson Discovery** and **IBM Cloud Object Storage**.
- ***Introspect AI - Mental Health Monitoring Platform:*** Created a **passive monitoring system** analyzing **multi-modal personal data** (health metrics, media, activity) to detect well-being shifts. Developed a **knowledge graph + LLM RAG pipeline** for context-aware insights. Automated **10K+ daily signal collection** via **Health Connect API**, **Spotify**, and **YouTube**, with visualization dashboards for **long-term behavioral analysis**.

Technical Skills

Deep Learning Frameworks: PyTorch (Distributed, TorchScript), TensorFlow, Keras, JAX, NVIDIA NeMo, Hugging Face Transformers
Inference & Optimization: LoRA, QLoRA, Torch-TensorRT, ONNX Runtime, model quantization, speculative decoding, A/B testing
Programming Languages: Python, Perl, C++, Shell Scripting, R, JavaScript, TypeScript, Java
ML Tooling & Libraries: Ray Tune, Optuna, Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, spaCy, NLTK, XGBoost
MLOps: Docker, Kubernetes, Jenkins, GitHub Actions, GitLab CI/CD, FastAPI, Flask, REST APIs, model evaluation pipelines
Distributed Systems & Data Engineering: Apache Kafka, SQL, SparkML, PostgreSQL, MongoDB, Redis, Firebase
Cloud Platforms: NVIDIA GPU environments, Azure (Functions, Blob Storage), AWS, GCP, HPC clusters