Sushil Dalavi

github.com/sushildalavi — sdalavi@usc.edu — +1 (213) 691-3794 — linkedin.com/in/sushildalavi

Education

University of Southern California Aug 2024 – May 2026
Master of Science in Computer Science *Los Angeles, CA*
Relevant Coursework: Machine Learning, Deep Learning, Distributed Systems, Information Retrieval, Natural Language Processing

University of Mumbai Jun 2019 – May 2023
Bachelor of Engineering in Computer Engineering *Mumbai, India*
Relevant Coursework: Operating Systems, Distributed Systems, Computer Networks, Object-Oriented Programming, Data Structures & Algorithms

Experience

AI Engineer, USC Annenberg Norman Lear Center — Los Angeles, CA Jun 2025 – Present

- Architected an AWS data platform (S3, Glue, SageMaker, Bedrock) that ingests, deduplicates, and normalizes **1M+** multi-region records for downstream ML training and retrieval workloads.
- Shipped a multi-modal alignment system fusing audio, speaker diarization, and caption streams, reaching **99.3%** F1 and **99.9%** coverage on ground-truth evaluation.
- Developed large-scale batch pipelines processing long-form video and audio through Whisper ASR, pyannote diarization, and model-based refinement stages.
- Automated dataset QA, Unicode normalization, and deduplication in Python, lifting analysis-ready yield from **10,819** raw inputs to **9,735** records with full reproducibility.

Software Engineer, Reliance Jio Platforms Limited — Navi Mumbai, India Dec 2023 – Jul 2024

- Trained and deployed ResNet-50 and DenseNet-121 deep vision networks for medical image anomaly detection, improving recall by **35%** through transfer learning, augmentation, and loss tuning.
- Optimized quantized transformer inference (BERT, GPT-2) on GPU with batched serving, cutting p95 latency by **30%** and lifting throughput while preserving **20%** accuracy gains.
- Engineered demand-forecasting microservices (TFT, CatBoost, LSTM) over Hive SQL batch pipelines, reducing forecast MAPE by **25%** for business-critical workloads.
- Rolled out shadow-testing and canary-release workflows for **3** production ML upgrades, catching **2** latency regressions before fleet-wide deployment.

Technical Projects

JobSense — Distributed Workflow Platform | Python, FastAPI, Temporal, PostgreSQL, Redis

- Built a durable, fault-tolerant orchestration system on Temporal with **12** tool integrations, automated retries, human-in-the-loop checkpoints, and end-to-end observability.
- Designed a provider-agnostic inference gateway with multi-backend failover, Redis semantic caching, structured-output validation, and CI regression gates blocking merges on quality or cost drift.
- Implemented hybrid retrieval (BM25, dense vector, cross-encoder rerank) with Reciprocal Rank Fusion, benchmarked end-to-end against a held-out evaluation set.

ScribeAI — Inference Service with Evaluation Pipeline | Python, FastAPI, Qdrant, MLflow

- Created an async FastAPI inference service with SSE streaming, multi-backend routing (GPT-4o, Claude, fallback engine), and graceful degradation under upstream failure.
- Instrumented an MLflow-tracked evaluation harness running ROUGE, BLEU, BERTScore, faithfulness, and leakage checks on every model change, with automated regression alerts on metric drift.
- Hardened a compliance-aware data pipeline with entity-level redaction across **10+** PII types, encrypted storage via pgcrypto, and append-only audit logging for traceability.

ScholarRAG — Retrieval and Data Engineering System | Python, FastAPI, pgvector, Postgres

- Prototyped a hybrid retrieval pipeline (dense, BM25, RRF, MiniLM rerank) lifting MRR by **21.8%** and nDCG@10 by **18.0%** across a **120+** query evaluation harness.
- Reduced duplicate indexing by **50%** and re-ingestion time by **60%** via DOI/ID/title normalization and SHA-256 content hashing across heterogeneous sources.
- Improved answer grounding from **0.505** to **0.616** faithfulness and claim support from **45.4%** to **85.6%** through evidence-constrained generation and citation-aware prompting.

Technical Skills

Languages: Python, C++, Go, SQL, Bash, TypeScript, JavaScript
ML & Deep Learning: PyTorch, Hugging Face, ONNX Runtime, Quantization, LoRA/PEFT, Fine-Tuning, Distributed Training, MLflow
LLMs & Retrieval: Prompt Engineering, Function Calling, Structured Outputs, RAG, Hybrid Retrieval, Reranking, pgvector, Qdrant, RAGAS
Backend & Data Systems: FastAPI, Temporal, gRPC, REST, SSE, Kafka, Spark, PostgreSQL, Redis, MongoDB
Cloud & DevOps: AWS (S3, Glue, SageMaker, Bedrock), GCP, Docker, Kubernetes, Linux, GitHub Actions, Prometheus, Grafana, Airflow
AI-Assisted Development: Claude Code, Cursor, GitHub Copilot, Codex