

Yash Patil

Santa Clara, CA | Open to relocation

| yashpatil582@gmail.com | linkedin.com/in/yashpatil23/ | [Github](https://github.com)

EDUCATION

Santa Clara University

Master of Science in Computer Science and Engineering (GPA: 3.63 / 4.00)

Coursework: Artificial Intelligence, Design and Analysis: Algorithms, Operating Systems, Computer Architecture

September 2023 – June 2025

Santa Clara, CA

Rajarambapu Institute of Technology

Bachelor Of Technology in Computer Science and Engineering

Coursework: Data Structures, Discrete Mathematics, Advanced Databases, Machine Learning

August 2016 – October 2020

Sangli, India

TECHNICAL SKILLS

Languages: Python, C++, Java, C, SQL, R, JavaScript, Bash

ML and AI: NLP, Transformers, RAG, embeddings, prompt engineering, retrieval, ranking, sentiment analysis, forecasting, search quality

Frameworks and Tools: PyTorch, TensorFlow, LangChain, FAISS, ChromaDB, Pinecone, Weaviate, Ollama, Apache Spark, Hadoop, Hive, Databricks, Flink

MLOps and Deployment: Docker, Kubernetes, REST APIs, CI/CD, unit testing, integration testing, Git, observability, model monitoring, Kafka, Streamlit, Tableau, Power BI

Cloud and Infrastructure: AWS, GCP, Azure, Amazon Bedrock, Azure AI Foundry

Databases and ETL: SQL Server, Redis, batch automation, shell scripting, data APIs

Certifications: Generative AI with Large Language Models (AWS, DeepLearning.AI), AWS Certified Data Engineer

EXPERIENCE

Tip Top Technologies

AI/ML Engineer

October 2024 – Present

Sunnyvale, CA

- Shipped low latency AI services by deploying Dockerized microservices on Kubernetes in AWS, with CI/CD, observability, and reliability hooks.
- Redesigned the real time recommendations ranking pipeline (retrieval + re-ranking) for News360, improving NDCG@10 by 15% and reducing p95 latency from 120 ms to under 50 ms through profiling, caching, and serving optimizations.
- Built an embedding based semantic retrieval system (FAISS + Pinecone) that increased recall@100 by 12%, directly improving search relevance while supporting 2M+ queries per day with production monitoring and evaluation.
- Predicted near term trend movement (1, 3, 6, 9 hour horizons) by forecasting over historical trend signals for ranked terms.
- Improved search quality with misspelling handling and term auto correction, reducing unresolved queries by 30% in COM360 via knowledge graphs and context aware retrieval.
- Building an agentic meeting copilot extension for COM360 using a graph database for long term memory, plus a voice interface (text to speech via Amazon Polly) to join meetings, answer questions in real time, and draft responses with explicit consent and guardrails.

Samvid

AI Intern

July 2024 – September 2024

Remote

- Built LLM pipelines for API orchestration, authentication, and real time monitoring to support production workflows.
- Benchmarked model quality and latency across managed APIs (GPT 4o, Gemini) and AWS Bedrock models (Claude, Llama 3.1, Mistral) to inform production readiness decisions.
- Automated data validation workflows for ML pipelines (schema, nulls, distributions), reducing cross-team sync meetings by 6 hours per week and accelerating model iteration cycles by 40%.

Searchspring

Data Scientist

April 2023 – August 2023

Remote

- Improved search relevance by 30% by developing spell correction NLP models, and built Databricks PySpark ETL pipelines that increased analytics throughput by 40%.

Anju Life Sciences Software

Data Scientist

March 2021 – December 2022

Pune, India

- Led HIPAA compliant analytics across clinical trials, claims, and EHR datasets, increasing ingestion throughput by 45% via ETL pipelines and optimized SQL Server stored procedures and APIs.

PROJECTS

Resume Database Chatbot Using Local LLM | Ollama, LangChain, Streamlit, ChromaDB

- Built an ATS-optimized resume analysis and chatbot system using RAG for document and job alignment.

Databricks Assistant Smart Merge Patch Prototype | Python, Streamlit, Google Gemini API, Prompt Engineering, diffliB (inline diff), GitHub

- Reproduced a real Databricks Assistant Diagnose Error issue where partial LLM fix snippets overwrite the entire notebook cell, deleting working code context.
- Built a Streamlit plus Gemini prototype that applies fixes as a safe patch (smart merge insertion before the failing write) and shows an inline diff preview before updating code.