

Trisham Bharat Patil

Software Engineer

Pune, MH, India. ♦ +91-9175077283 ♦ trishampatil@gmail.com ♦ [LinkedIn](#) ♦ GitHub: [TrishamBP](#)

SUMMARY

Software Engineer with 3 years of experience in building end-to-end AI applications, currently pursuing an MTech in AI and ML from BITS Pilani. Skilled in developing LLM-powered systems, NLP/CV pipelines, and multilingual AI platforms that have secured multi-million dollar enterprise deals. Proficient in backend engineering, database design, and system architecture, ensuring AI applications are scalable, reliable, and optimized for real-time usage. Experienced in serverless pipelines, microservices, and real-time data processing, delivering cost-efficient, production-grade solutions across AWS, GCP, and Azure. Adept in frontend development with React.js and Next.js, creating user-friendly chat interfaces and dashboards that enhance adoption. Known for client interaction and communication skills, effectively translating requirements into technical execution. Passionate about LLM engineering, AI agents, and MLOps, with a track record of bridging cutting-edge AI research with product engineering.

EDUCATION

BITS Pilani (WILP), Pilani, RJ, India. (Currently Pursuing) Master of Technology (MTech) in Artificial Intelligence and Machine Learning Relevant Courses: Statistical Methods, Machine Learning, Artificial and Computational Intelligence, Mathematical Foundation for ML, NLP, Deep Neural Networks, Deep Reinforcement Learning, ML System Optimization, Speech Processing, Conversational AI, NLP Applications, MLOps.	October 2024 – June 2026
Worcester Polytechnic Institute (WPI), Worcester, MA, USA. Master of Science (MS) in Mechanical Engineering	August 2019 – May 2021
University of Pune (DY Patil Institute of Engg and Tech), Pune, MH, India. Bachelor of Engineering (BE) in Mechanical Engineering	August 2015 – May 2019

TECHNICAL SKILLS

Programming Languages: Python(OOP), JavaScript, TypeScript.
Frontend Development: React.js, Next.js, GradioUI.
Backend Development: Node.js, Express.js, NestJS, FastAPI(async), Celery, Alembic, Pydantic, Pytest, MongoORM, SQLAlchemy, Nginx, Cloudflare, Cloudfront, WebSockets, GraphQL, Graphene, REST, gRPC, Webhooks, SMTP, Redis, Firebase, RabbitMQ.
Database: MongoDB, PostgreSQL, MySQL.
ML Development: Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Tensorflow, Keras, Pytorch, NLP (Spacy, NLTK), OpenCV.
AI Development: Transformers, LLM's (GPT, Claude, Gemini, DeepSeek, Llama, Grok), HuggingFace, RAG, Langchain, LlamaIndex, FAISS, Weaviate, ChromaDB, FineTuning(LoRA, QLoRA).
Data Engineering & Orchestration: Apache Spark, Apache Kafka, Apache Airflow, Hadoop.
DevOps: AWS (S3, EC2, Lambda, RDS, ECR, EKS, ELB), GCP (GCS, Cloud Run), GitHub Actions, Jenkins, GitLab CI/CD, Docker.

EXPERIENCE

Gaius Networks (Remote) Senior AI Software Engineer. ParseTalent and Flipped.ai – Bootstrapped Product. - Architected and deployed a scalable RAG chatbot using OpenAI embeddings, multi-indexing, and semantic retrieval with Weaviate, enabling recruiters to parse CVs and retrieve structured hiring insights, reducing shortlisting time by 30% and doubling accuracy. - Built a hybrid SQL–LLM pipeline integrating Excel ingestion, automated PostgreSQL schema generation, and RAG-powered SQL agents, enabling conversational analytics and automated candidate outreach via email and WhatsApp. - Fine-tuned OpenAI 4.1-mini model using LoRA on 50,000+ CVs, enabling domain-specific skill extraction and role matching; boosted chatbot precision, reduced dependency on generic prompts, and improved response relevance and contextual accuracy. - Spearheaded ATS integrations (Zoho Recruit & Greenhouse) through REST APIs, expanding client compatibility and generating a new B2B revenue stream. - Built a CPU-optimized, NLP-driven CV parsing engine using spaCy, BERT, and classical models such as Markov chains and TF-IDF for vertex and domain matching, enabling efficient section detection and structured data extraction with 90%+ accuracy. - Architected and deployed AI microservices for CV parsing, JD extraction, and candidate matching on a separate EC2 cluster, using Celery workers with RabbitMQ to ensure reliable, asynchronous inter-service communication. - Migrated the CV parsing module to a fully serverless architecture using AWS Lambda, enabling 1,000+ concurrent Lambdas and bulk processing of 100,000+ CVs with low latency and intelligent routing. - Engineered robust observability, autoscaling, and fault-tolerant infrastructure to sustain 99.9% uptime, ensuring smooth background execution of AI workflows and consistent support for 20+ concurrent users across all services.	New York, New York, USA. October 2023 – August 2025.
---	---

- Designed and developed the entire backend architecture using Node.js, Express.js, and MongoDB, enabling rapid development in just 3 months to meet investor deadlines — handling 10,000–15,000 concurrent requests with secure, low-latency APIs.
- Built a responsive Next.js frontend integrated with backend APIs, leveraging TypeScript and SEO optimization to deliver real-time candidate search, skill matching, and analytics dashboards — directly helping secure \$50K in seed funding.

Enterprise Project – Saudi Arabian Client – Leannode.

- Led design and deployment of a multilingual AI-powered HR platform for a Saudi Arabian enterprise client, localized for Arabic language and culture, directly contributing to a \$3M licensing deal.
- Developed and optimized AI modules including CV parsing, job description analysis, candidate matching, and sentiment analysis, driving widespread adoption across enterprise clients.
- Built a bilingual OCR pipeline using OpenCV, EAST, and Tesseract with Arabic language support, boosting CV parsing accuracy from ~80% to 95%+ through multi-engine bounding box detection.
- Integrated Hugging Face OPUS-MT transformers for Arabic–English CV translation with batch processing and JSON-configurable workflows, reducing manual recruiter review time by 60%.
- Benchmarked Qwen, Mistral, and LLaMA LLMs on NVIDIA L4 GPUs, applying CUDA optimizations and designing zero-/few-shot prompts for structured parsing at scale.
- Designed the API and orchestration layer with FastAPI, exposing RESTful endpoints for OCR and AI services, integrating with frontend, and decoupling routing, auth, and rate limiting from GPU inference for fault-tolerant microservices.
- Optimized LLM inference by moving from Hugging Face AutoTokenizer/pipeline on Qwen- 7B models to Berkeley’s vLLM, enabling continuous batching and one-server-per-GPU deployment; ran three replicas on Google Cloud Run with routing and queuing.
- Stabilized and structured extracted data by migrating to Amazon RDS for PostgreSQL, implementing normalized schemas with JSONB, transactional writes, and indexing to support analytical queries and enterprise integrations.
- After stabilizing the extraction schema, migrated entities to Amazon RDS for PostgreSQL, designing normalized tables with JSONB where appropriate, transactional writes, and targeted indexing to support analytical SQL, auditability, and reliable integrations.
- Interfaced directly with client stakeholders to gather evolving requirements, managed sprints in Notion, and followed Agile/Scrum practices to ensure smooth delivery cycles, high client satisfaction, and fast iteration.

TechGigs LLP (On-Site)
Full Stack Developer.

Pune, MH, India.
December 2022 – September 2023.

- Developed backend infrastructure by translating client-provided Figma designs into REST APIs for mobile and web applications, creating database schemas from scratch tailored to each client’s needs. Leveraged Node.js, Express.js, and MongoDB to handle 50,000+ daily API requests with 99.9% uptime, reducing downtime by 75% and improving deployment speed by 60%.
- Designed and integrated multi-database architecture utilizing MongoDB, PostgreSQL, and MySQL, resulting in 65% faster query response times, managing 10TB+ of data, and reducing data retrieval costs by 30% while maintaining 99.99% data consistency.
- Implemented comprehensive security protocols, including JWT authentication and OAuth 2.0, while proactively mitigating vulnerabilities such as DDoS attacks and SSL injection, ensuring application security and data protection.
- Implemented comprehensive testing strategies using Jest, Mocha, Puppeteer, and Chai frameworks, following Test Driven Development (TDD) to achieve high code coverage and maintain rigorous quality assurance standards.
- Developed and designed an Admin Panel using React.js or Next.js, enabling admins/clients to control functions within their main application. Used Material-UI (MUI) and Bootstrap to create responsive interfaces.
- Integrated D3.js and Chart.js for data visualization inside the Admin Panel, providing a comprehensive overview to track sales, costs, and other metrics. Additionally, created and integrated Admin-side APIs to support seamless dashboard interaction.

PROJECTS

- **Drug Information RAG System:** Built an FDA-powered Retrieval-Augmented Generation (RAG) system using GPT-4o-mini, LangChain, and ChromaDB. Enabled multi-query drug search, semantic retrieval, and structured LLM responses with Gradio UI, delivering real-time, source-attributed answers on drug safety, interactions, and regulatory status.
- **LoRA and QLoRA Fine Tuning:** Developed an end-to-end data and model fine-tuning pipeline by preprocessing and converting Amazon product review datasets into structured training formats. Fine-tuned Llama 4-7B Mini and GPT-4.1 Mini using LoRA and QLoRA techniques, optimizing model performance and efficiency for scalable downstream AI applications.
- **Football Match Analysis using Computer Vision:** Developed a sophisticated football match analysis system leveraging YOLO and OpenCV for player tracking and event detection. Implemented deep learning models in PyTorch to automate the analysis of soccer games from video footage.
- **MRI-Based Alzheimer’s Classification:** Classified OASIS MRI data into NC, MCI, and AD categories using CNNs, AlexNet, and Vision Transformer architectures, achieving high accuracy in early Alzheimer’s detection through advanced deep learning and image preprocessing techniques.
- **Distributed MLP Training with PyTorch and Horovod on MNIST:** Implemented data-parallel distributed mini-batch training of a Multi-Layer Perceptron on MNIST using PyTorch DDP and Horovod. Designed parallel data pipelines, optimized synchronization overhead, and integrated CUDA acceleration. Utilized Hadoop concepts, cuML, and NCCL backend to enhance scalability and performance across GPUs.
- **Movie Reservation System Backend:** Designed and developed a complete movie reservation system backend, implementing RESTful APIs using FastAPI for user authentication, seat booking, scheduling, and ticketing. Built a robust database schema using PostgreSQL, optimized SQL queries, and ensured secure data handling, covering all essential responsibilities of a backend engineer.