Trisham Bharat Patil
AI Researcher and Engineer

Pune, MH, India. ⋄ +91-9175077283 ⋄ Email ⋄ LinkedIn ⋄ Website ⋄ GitHub: TrishamBP ⋄ Hugging Face: Trisham97, Quantbridge

SUMMARY

AI Engineer specializing in Large Language Models, Agentic AI, and AI Systems Engineering, with experience building scalable AI applications and production-grade inference and orchestration systems. Experienced across the LLM lifecycle, including model adaptation and post-training, RAG, agent memory, evaluation, inference optimization, and distributed AI workloads. Strong software engineering foundation with Python, TypeScript, cloud-native infrastructure, and containerized deployments, complemented by hands-on work with PyTorch, CUDA, GPU computing, distributed training, and high-performance LLM inference. Focused on translating advances in LLMs, agentic architectures, and AI infrastructure into reliable, efficient, and scalable production systems.

EDUCATION

Worcester Polytechnic Institute (WPI), Worcester, MA, USA. August 2019 – May 2021.
Master of Science **(MS)** in Mechanical Engineering

University of Pune (DY Patil Institute of Engg and Tech), Pune, MH, India. August 2015 – May 2019.
Bachelor of Engineering **(BE)** in Mechanical Engineering

TECHNICAL SKILLS

Languages & Core Runtimes: Python, TypeScript, JavaScript, Node.js (Async I/O), C++, Bash.
Data Engineering: High-Throughput ELT/ETL Pipelines, Streaming & Batch Data Ingestion, Schema Evolution (DLT), Analytical Warehousing (DuckDB, BigQuery), Apache Kafka, Paraquet File Format, Feature Engineering.
Backend & Distributed Infrastructure: Async I/O, Event Driven Design, REST, HTTP + SSE, STDIO, Express.JS, Koa.JS, Hapi, NestJS, Flask, FastAPI, Redis, Valkey, AWS SQS, AWS Elasticache, RabbitMQ, Celery, API Security (throttling/rate limiting), AWS IAM, Azure AD, Three.js, MongooseORM, SQLAlchemy ORM, Pydantic, Firebase.
Databases & Persistent Systems: PostgreSQL, MongoDB, AWS DynamoDB, Aurora Serverless.
Machine Learning and MLOps: Linear Regression, Logistic Classification, Random Forests, Gradient Boosting (XGBoost), KNN Classification, Ensemble Methods, Clustering Algo, ARIMA/SARIMAX, Prophet Forecasting, RNN Forecasting, Continous Monitoring, Drift Detection, Model Retraining, Performance Tracking.
Deep Learning: PyTorch, TensorFlow, JAX, Transfromer Models (BERT, T5, BART), Attention Mechanism (MHA, GQA, MLA), Horovod, CUDA, cuDNN, Mixed Precision (FP16/BF16 training), RNNs, LSTMs, GRUs, Generative Models (Diffusion, VAEs, GANs), Gradient Optimizers (AdamW, SGD), ONNX.
LLM Engineering: Hugging Face Transformers, SFT, DPO, GRPO, RLHF, PEFT, LoRA/QLoRA, Quantization, RAG, LLM Evaluation, vLLM, TensorRT-LLM, Distributed Training, FSDP, DeepSpeed, NCCL, CUDA, Triton, Inference Optimization.
AI Systems & Inference: C++, CUDA, ROCm, Triton Inference Server, TensorRT, TensorRT-LLM, Kubernetes, Docker, GPU Scheduling, Distributed Computing, Profiling, Mixed Precision.
AI Engg & Agentic AI: LangChain, LECL, LangGraph, LangFuse, LangSmith, CrewAI, GoogleA2A, OpenAI Agents SDK, Microsoft AutoGen, GraphRAG, Agentic RAG, Mem0AI, MCP, OpenClaw, NemoClaw, Open Shell, Hermes Agent, DeepAgents, Claude Managed Agents, NVIDIA Nemo Guardrails, Guardrails AI, AWS Bedrock Guardrails, Pinecone, Weaviate, S3 Vectors, Qdrant, PGVector, AWS Elasticsearch, DSPy, Prompt Engineering, Context Engineering, Loop Engineering, Graph Engineering.
Cloud Infrastructure & DevOps: Infrastructure as Code (Terraform), AWS (EC2, ECS Fargate, Lambda, SQS, RDS, Bedrock Runtime, SageMaker), GCP (Cloud Run, GCS), Docker, Kubernetes(k8s), CI/CD Pipelines (GitHub Actions).
Observability: Prometheus, Grafana, Jaeger Distributed Tracing, LangSmith LLM Tracing, DeepEval, Automated Testing (Pytest, Jest, Puppeteer E2E).

EXPERIENCE

Cloudangles (On-Site) Hyderabad, Telangana, India.
Senior Innovation Engineer February 2026 – Present.
People Case Management Automation Platform:

- Led the end-to-end development of an AI-assisted Legal Document Intelligence platform for Centrica UK's HR, Employee Relations, and Legal teams. Partnered closely with stakeholders to understand grievance workflows, translate business requirements into scalable technical solutions, and independently owned the architecture, backend, frontend, AWS infrastructure, CI/CD pipelines, and production deployment from concept through release.
- Designed and evolved a scalable document processing platform capable of handling large case files, emails, chat messages, and supporting evidence. Evaluated multiple LLMs through structured experimentation using synthetic datasets, balancing response quality, latency, and operational cost before implementing specialized AI workflows for timeline reconstruction, policy analysis, document classification, sentiment analysis, draft generation, and human-in-the-loop review.

- Architected a distributed cloud-native solution using FastAPI, Next.js, AWS Lambda, SQS, ECS Fargate, S3, EventBridge, Docker, GitHub Actions, and Terraform. Refactored a monolithic workflow into an event-driven architecture to overcome execution time limits, improve reliability, enable concurrent processing, and support secure deployments across development, testing, pre-production, and production environments following enterprise architecture and security reviews.
- Improved AI accuracy and long-term usability by implementing iterative evaluation pipelines, advanced retrieval strategies, agent-based reasoning workflows, persistent semantic memory, and custom evaluation metrics. Increased response quality through continuous experimentation, retrieval optimization, and prompt refinement while ensuring every AI-generated output remained reviewable.

Gaius Networks (Remote) New York, New York, USA.
Senior Software Engineer. October 2023 – January 2026.
ParseTalent & Flipped.ai — Bootstrapped AI Platform

- Architected a cost-efficient CV intelligence platform by evolving CPU-based NLP extraction pipelines into scalable LLM-assisted workflows; benchmarked open-source models, implemented advanced RAG techniques (BM25, hybrid retrieval, reranking, cross-encoders, MMR), and optimized inference using vLLM and PagedAttention to overcome KV-cache bottlenecks and support concurrent enterprise workloads on AWS T4 infrastructure.
- Designed a distributed AI processing platform using FastAPI, PostgreSQL, Celery, RabbitMQ, and AWS Lambda, decoupling ingestion, extraction, retrieval, and inference services to execute thousands of parallel document-processing workflows while maintaining fault tolerance and operational scalability.
- Led end-to-end product engineering and commercialization, delivering a production AI platform with ATS integrations (Greenhouse, Zoho Recruit), authentication, billing, candidate search, matching, analytics, and recruiter-facing workflows supporting real-world hiring operations.

Enterprise AI Platform — Saudi Arabian Client (Leannode).

- Spearheaded development of an Arabic CV intelligence system by researching OCR failures, building custom extraction pipelines, curating labeled datasets, and fine-tuning PyTorch models; combined traditional NLP and LLM-based reasoning to increase parsing accuracy from ~80% to 95%+ across multilingual recruitment workflows leading to a $3M licensing deal.

TechGigs LLP (On-Site) Pune, MH, India.
Full Stack Developer. December 2022 – September 2023.

- Led the end-to-end delivery of multiple production platforms across e-commerce, food delivery, healthcare, and fitness domains, architecting high-availability Node.js/PostgreSQL/MongoDB systems and making pragmatic monolith-versus-microservice trade-offs to balance scalability, operational simplicity, and delivery velocity.
- Engineered end-to-end secure infrastructure and administration layers by deploying optimized async Node.js runtimes on AWS EC2, establishing RBAC systems via JWT/OAuth 2.0, and building Next.js control planes; instituted TDD frameworks using Jest and Puppeteer to guarantee zero-downtime deployments under production traffic.

RESEARCH ARTICLES & PUBLICATIONS

- **Domain-Adaptive Multi-Task NLP for Financial and Geopolitical Intelligence Extraction** — Multi-task learning, domain adaptation, NLP, information extraction, financial and geopolitical intelligence. **2026** (Link)
- **From Multi-Head Attention to Multi-Head Latent Attention: An Engineering Analysis of Attention Mechanisms for Large Language Model Inference** — Transformer architectures, MHA, MLA, KV-cache optimization, memory efficiency, and LLM inference. **2026** (Link)
- **A Scalable Memory Lifecycle Architecture for Agentic Legal Document Intelligence Systems** — Agentic AI, LLM memory architectures, RAG, semantic retrieval, memory compression, and lifecycle management. **2026** (Link)

HACKATHONS AND ACHIEVEMENTS

AMD Developer Hackathon — ACT II, Track 1 | July 2026: Ranked 34th out of ~20,000 participants (Top 0.2%); built a token-efficient autonomous AI agent achieving ~22× token reduction while maintaining 100% accuracy on the test evaluation. (GitHub)

CERTIFICATIONS

- **Data Engineering with Python and AI —** freeCodeCamp, 2026.
- **Natural Language Processing Specialization** — DeepLearning.AI / Coursera, 2024.
- **AI Engineering & Deep Learning** — Udemy: LLM Engineering, RAG, QLoRA, Agents; Agentic AI & MCP; Production LLM/Agent Deployment; PyTorch & TensorFlow Deep Learning, 2025–2026.