

Avinash Vikram Singh | AI/ML Engineer

| Founder | Full Stack Developer

Bengaluru, India

M +91 7052985015 • E avinashvs0052@gmail.com

W avinashvs-portfolio.vazmora.com • in avinash-vikram-s-40b263233
gh AVINASH0052

Summary

AI engineer and solo founder. I build production systems end to end, from retrieval architecture through to billing. Lead the engineering team behind PorsAI's enterprise AI orchestration platform. Shipped Vazmora, a live financial document intelligence SaaS running nine tools in production for over a year. Scored 1.11564 BPB inside 16MB on the OpenAI Parameter Golf challenge with a depth recurrent transformer. Focus is agentic systems, retrieval grounding and inference efficiency.

Experience

PorsAI Remote
Founding Engineer Dec 2025 to Present

- Lead the engineering team that built PorsAI Engage, an enterprise AI orchestration platform that runs agents, voice, chat, workflows and knowledge on one governed runtime.
- Shipped the agent marketplace (voice, chat, lead generation, credit operations, knowledge, SRE and SDLC agents) on top of workflow orchestration with human in the loop approvals, an LLM gateway and RAG.
- Engineered multi tenant backend architecture in Java Spring Boot with RBAC, audit logs, PII redaction and strict tenant data isolation across enterprise healthcare clients.
- Shipped multilingual voice agents and omnichannel chat across WhatsApp, Instagram and web. Fine tuned Llama models with the Hugging Face PEFT stack, with evaluation loops built on real client transcripts.
- Built a clinical AI diagnostic app for hair and skin analysis plus a predictive lead scoring model feeding sales routing.
- Work directly with clients on setup, configuration, validation and troubleshooting inside their live environments.

Indian Capital and Investment Remote
Python (AI) Developer Jun 2025 to Nov 2025

- Joined as founding engineer and stood up the ML environments and model deployment workflows on GCP and Vertex AI from scratch.
- Built Python pipelines for automated unstructured data extraction and integrated AWS cloud native tooling, improving infrastructure reliability 25%.

ARD Information Systems Noida, India
Data Scientist Nov 2023 to May 2025

- Built a Random Forest model ranking candidate profiles against job descriptions. Cut manual screening time 40%.
- Integrated the ranking engine into a live ATS over REST APIs with real time scoring under 30ms.
- Diagnosed and fixed a production latency regression by precomputing features and adding a Redis cache layer.
- Ran EDA across a 50GB+ candidate database and surfaced skill trends that lifted placement success 15%.

ARD Information Systems Noida, India
Data Scientist Intern Jul 2023 to Sep 2023

- Optimized SQL queries and refactored legacy Python aggregation scripts across portals, raising throughput 20%.
Built a quarterly hiring demand regression prototype at under 5% MAE.

Research and Competitions

OpenAI Parameter Golf Challenge Apr 2026
Depth Recurrent LLM, 10min_16mb Track

- Designed an 11 layer depth recurrent LLM with recurrence on layers 3 to 5, reaching 1.11564 BPB using 8H/4KV GQA, 4x MLP and U-Net skip connections.
- Built a compression pipeline with Hessian GPTQ SDClip, int6 matrices and int8 embeddings, then Brotli-11, to fit a hard 16MB budget.
- Added Score First SGD chunk test time training, MuonEq-R row normalized optimization and an SP8192 vocabulary.

Christ University

MSc Thesis, Uncertainty Calibration in Foundation Models

2026

- Benchmarked zero shot Amazon Chronos and Google TimesFM against locally trained LSTMs on NASA C-MAPSS engine degradation data.
- Found regime dependent uncertainty miscalibration. Coverage@80 swung from 38% to 61% across operating regimes while point accuracy stayed flat. Validated with Wilcoxon signed rank tests and Cohen's d.

Projects

Founder

vazmora.com

Vazmora, Financial Document Intelligence SaaS

2025 to Present

- Founded and shipped a production SaaS that auto extracts KPIs such as revenue, margins, EBITDA, EPS and P/E from annual reports, regulatory filings and earnings documents with no templates required.
- Nine tools in one platform: structured text extraction, OCR, table parsing, translation across 35+ languages, rich text editing, AI chat with page citations, financial intelligence, invoice extraction and quarterly comparison.
- Rebuilt the retrieval pipeline for citation grounding after the first version answered fluently without proof. Every answer now traces to an exact source page and a relevance threshold returns not found rather than guessing. Tested against adversarial queries with no true answer and cross page reasoning.
- Built credit based per page billing from scratch on Razorpay. Runs on Hetzner with FastAPI, Next.js, PostgreSQL, LangChain, Qdrant and Cloudflare R2, DNS, CDN and SSL.

Next.js 14, TypeScript, Prisma, Bolna

voicetable-alpha.vercel.app

Voice Table AI, Restaurant Operations

2025

- Automated voice confirmation on cash on delivery orders to cut return to origin rates, with webhooks for live order updates, transcript logging and an analytics dashboard.

Streamlit, LLMs, OCR, FAISS

pdftranslationextraction-avinash.streamlit.app

Document Processing and Translation Tool

2024

- Multilingual engine covering Hindi, Arabic and Cyrillic while preserving layout, tables and columns, with context aware image captioning and interactive question answering.

Skills

Languages: Python, TypeScript, JavaScript, SQL, Java, C++

ML and AI: PyTorch, TensorFlow, Scikit-learn, Transformers, Hugging Face PEFT, LoRA fine tuning, LangChain, LangGraph, RAG, Vertex AI, agentic frameworks, GPTQ quantization

Full Stack: FastAPI, Next.js (App Router), React, Node.js, Tailwind CSS, Java Spring Boot

Data and Infra: PostgreSQL, MongoDB, Redis, Qdrant, AWS, GCP, Hetzner, Cloudflare R2, Docker, Git

Tools: Claude Code, Codex, MCP, Pandas, NumPy, Tableau

Education

Christ University

Bengaluru

MSc Data Science, Reliance Foundation Postgraduate Scholar

2025 to 2027

Delhi University

New Delhi

BSc (Hons) Computer Science

2020 to 2023

Certifications: Machine Learning, Stanford on Coursera. Data Science using Python, University of Delhi.

Achievements

2025: **Reliance Foundation Postgraduate Scholarship**, recipient. Selected from a competitive national pool.

2025: Marketing Head, Expectation 25, Christ University. Ran departmental fest outreach to record participation.