

PRINCE SINGH

Website: princesingh.xyz

Github: github.com/mrprince88

Linkedin: linkedin.com/in/mrprince88

Email: mrprincesingh88@gmail.com

Mobile: +91 8077978214

SUMMARY

Software engineer with 4 years building production systems, now focused on LLM applications: retrieval pipelines, tool calling assistants, and the evals that keep them measurable. Owns systems end to end, from retrieval design through latency, inference cost, and quality regressions.

EXPERIENCE

Fareportal

Feb 2025 – Present

Sr Software Engineer

Gurugram, Hybrid

- Architected an **LLM-powered travel assistant** embedded in the booking flow: tool-calling over live fare and inventory APIs, multi-turn conversation state, streamed responses, and guardrails gating any booking-mutating action. Serves **5k+ sessions/day** at **0.2s p95**.
- Built the **evaluation harness** for the assistant and RAG system: a golden set of **200+ labeled queries** (20 user journeys, 10+ queries each) scored for groundedness and answer relevance via LLM-as-judge, with retrieval precision and refusal rate tracked per release, caught **12 quality regressions** before they shipped.
- Designed a **RAG knowledge system** over **2k support documents** using semantic chunking, hybrid BM25 + dense retrieval on **pgvector** with cross-encoder reranking. Raised retrieval precision from **70% to 92%**.
- Reduced inference spend by **30%** through model routing (small models for classification and summarization, frontier models for generation), semantic caching, and prompt compression, without measurable quality loss on the eval set.
- Shipped a **call analysis pipeline** processing **5k+ daily recordings** through transcription, sentiment, and compliance scoring, routing low-confidence outputs to human review; improved QA throughput by **25%**.
- Led delivery of a performance management platform for **4k+ employees** across product, design, and HR stakeholders; raised review completion by **35%**.
- Tech Stack: **Python, FastAPI, LangChain, Vercel AI SDK, pgvector, PostgreSQL, Next.js, React, Node.js**.

Lodus

Feb 2024 – Jan 2025

Software Engineer

USA, Remote

- Designed a **hierarchical RBAC** system with inherited role scopes and per-resource policy checks, modelling the permission graph with product and security for **200 users**.
- Integrated **OpenSearch** for full-text search over a large document corpus, tuning analyzers and index mappings; **80%** faster than the prior SQL-based search.
- Cut API p95 latency by **35%** by moving synchronous work into **Celery** background jobs, and shipped a real-time feed serving **20k users**.
- Tech Stack: **Python, Flask, Celery, PostgreSQL, MongoDB, OpenSearch, Next.js, Kubernetes, Helm**.

Appleute

Sept 2022 – Jan 2024

Software Engineer (Mar 2023 – Jan 2024)

Germany, Remote

- Introduced frontend standards adopted by **6 engineers** across **5+ projects**, reducing UI defects and review churn.
- Owned end-to-end delivery of 2 client projects supporting **10k+ users**, including Heyfam, a Next.js gift catalog scoring **96/100** on Lighthouse; improved API performance **30%** by restructuring data fetching against Nest.js services.
- Tech Stack: **Next.js, React, Nest.js, Jest, MongoDB**.

Software Engineer Intern (Sept 2022 – Mar 2023)

- Led full-stack development of a social mobile app (Android, iOS) supporting **25k+ users** at **99.5%** uptime, plus a CMS for **15+ internal users** and automated health checks on **AWS Lambda**.

TECHNICAL SKILLS

AI Engineering: RAG (chunking, hybrid retrieval, reranking), agents & tool-calling, structured outputs, evals (golden sets, LLM-as-judge), tracing & observability, semantic caching, model routing

LLM Tooling: LangChain, Vercel AI SDK, OpenAI & Anthropic APIs, Langfuse, pgvector, OpenSearch

Languages: Python, TypeScript, JavaScript, Java, Go

Web: FastAPI, Node.js, Nest.js, Flask, Celery, React, Next.js, React Native

Data & Infra: PostgreSQL, pgvector, MongoDB, MySQL, OpenSearch, Docker, Kubernetes, AWS, GitHub Actions

EDUCATION

JSS Academy of Technical Education Noida

B.Tech in Computer Science and Engineering