

Divpreet Singh

Mohali, India

divpreet1012@gmail.com | +91-9914426127

linkedin.com/in/divpreet-singh12 | github.com/Divpreetai

Summary

AI Engineer with 6+ years of experience building scalable backend systems and production-grade AI/LLM pipelines. Experienced in RAG-based applications, entity resolution systems, AI-driven data processing, and LLM orchestration workflows. Strong background in designing scalable, secure, and maintainable AI-integrated systems for financial and enterprise environments.

Skills

Programming: Python, SQL, JavaScript (working knowledge)

Backend: FastAPI, Flask, REST APIs, Microservices

AI / ML: LLMs, RAG Pipelines, NLP, Sentence Transformers, Hugging Face, Embeddings, Semantic Search, Vector Search, FAISS, Prompt Engineering

RAG & LLM Application Design: Function Calling, Structured Outputs, Tool Calling, Re-ranking, Chunking Strategies, Grounding

AI Engineering & Evaluation: Agentic Workflows, LLM Orchestration, AI Evaluation Frameworks, Tool-Assisted Reasoning

AI Security: Prompt Injection Mitigation, Secure LLM Applications, Data Privacy, PII Handling

AI Engineering Tools: Cursor, GitHub Copilot, AI-assisted development workflows

DevOps & MLOps: CI/CD, Docker, Kubernetes, Model Deployment, Monitoring

Search & Data: Elasticsearch, FAISS, PostgreSQL, MySQL, SQLite

Processing: Ray, batch processing, large-scale data handling

Cloud & Infra: AWS, GCP, Linux

Other: Performance tuning, API security, logging, debugging

Experience

iAcuity Fintech — Software Engineer (AI/Backend)

Oct 2023 – Present | Remote

- Designed and implemented AI-driven entity extraction and resolution pipelines for financial transaction data
- Built scalable systems to process and normalize 300K–1.8M+ records, improving accuracy and consistency
- Implemented parallel processing using Ray, reducing execution time for batch pipelines
- Integrated LLM-based components into backend workflows for financial query answering
- Developed and optimized Elasticsearch pipelines for high-volume data
- Built data normalization pipelines for noisy financial narration inputs
- Improved system performance (30–35%) through query optimization and async execution
- Developed backend services using FastAPI and deployed via Docker
- Worked on resource optimization for CPU/memory-intensive workloads
- Used GCP for deploying GPU-enabled Docker environments for ML/AI pipelines
- Worked on RAG-based AI workflows, retrieval systems, and LLM orchestration pipelines for enterprise use cases
- Contributed to solution design discussions, architecture reviews, and implementation planning
- Created technical documentation including API specifications, workflow diagrams, operational guides, and implementation documents

Freako Fresh — Software Developer

Jan 2022 – Sep 2023 | Chandigarh

- Built backend services and APIs for analytics workflows
- Designed ETL pipelines for large datasets
- Improved API performance using indexing and async execution
- Delivered features end-to-end in Agile environment

Advantal Technologies — Software Developer

Jun 2019 – Nov 2020 | Gurugram

- Developed backend APIs using Python and SQL
 - Implemented authentication and validation workflows
 - Assisted in CI/CD and production issue resolution
-

Projects

Entity Resolution & AI Data Processing Platform

- Built AI-integrated entity extraction and resolution platform for financial datasets
- Developed pipelines for normalization, grouping, and deduplication
- Optimized workflows handling 300K–1.8M+ records
- Used embeddings, semantic similarity, and Ray for scalable matching

Tech Stack: Python, FastAPI, Elasticsearch, PostgreSQL, Docker, Ray, Hugging Face

AI-Powered Financial Question Answering System

- Developed RAG-based financial QA system using LLM workflows
- Worked on chunking, embeddings, and semantic search
- Implemented retrieval augmentation, contextual ranking, and orchestration workflows to improve answer quality and grounding
- Implemented retrieval evaluation, response grounding, and answer validation to improve factual accuracy
- Integrated LLM pipelines into backend services
- Explored multi-step orchestration and tool-assisted reasoning

Tech Stack: Python, FastAPI, RAG, LangChain, LlamaIndex, FAISS, PostgreSQL

Agent Evaluation & AI Workflow Platform

- Designed evaluation workflows for LLM correctness, grounding, and hallucination detection
- Built automated pipelines for prompt evaluation, response scoring, and testing

- Developed validation workflows for comparing prompts, models, and AI-generated outputs
- Explored agent-oriented execution and evaluation frameworks

Tech Stack: Python, FastAPI, OpenAI API, PostgreSQL, LangSmith

Production-Grade Multi-Agent Voice Support System

- Built a multi-agent voice-based customer support platform with speech-to-text, retrieval, reasoning, and text-to-speech agents
- Developed agent orchestration workflows for customer query handling and contextual response generation
- Integrated retrieval-based knowledge access and LLM-powered response generation
- Designed modular architecture supporting scalable deployment and future agent extensions

Tech Stack: Python, FastAPI, Whisper, LangChain, RAG, Vector Database, LLMs, Docker

OCR & Intelligent Document Processing Pipeline

- Built OCR workflows for PAN, Aadhaar, and financial documents
- Improved extraction from noisy and scanned inputs
- Developed preprocessing and structured parsing logic
- Document classification, layout-aware extraction, table parsing, and validation workflows

Tech Stack: Python, OpenCV, Tesseract, Pandas, Layout Parsing, Table Extraction

AI Infrastructure & Resource Governance System

- Worked on optimizing CPU/memory-intensive AI workloads
- Developed resource allocation and monitoring workflows
- Supported scalable execution of backend processing systems

Tech Stack: Python, Docker, Kubernetes, AWS

Education

M.E. (Computer Science & Engineering – Cyber Security) – Panjab University
B.E. (Information Technology) – Panjab University

Additional Information

- Experience working on production-scale AI systems from data processing to deployment
- Comfortable working in distributed teams and cross-functional environments
- Experience with RAG systems, AI evaluation workflows, and agentic AI architectures
- Interested in building scalable AI-integrated systems and intelligent automation platforms