PRASHANT ARORA

GENERATIVE AI ENGINEER (LLM EVALS, AGENTS & RAG)

prashantarora998@gmail.com | +91-8630831390 | Delhi NCR, India · Open to relocate · Immediate joiner
linkedin.com/in/prashant-arora-3063b5155 | github.com/neyoxdrago | neyoxdrago.com

SUMMARY

Generative AI Engineer with 5+ years in data science/ML and 17 months owning LLM evaluation, prompt architecture, and backend AI for SlidesAI (14M+ downloads). Built release-gating evals, hybrid RAG, multi-agent workflows, and an MCP server for Codex and Claude. Led a team of five at Accenture; shipped an Azure OpenAI HR copilot.

TECHNICAL SKILLS

Generative AI & Agents: Agentic AI, Multi-Agent Systems, MCP, Prompt Engineering, Function Calling, Structured Outputs, Prompt Caching, Guardrails

RAG & Retrieval: RAG, Hybrid Search, Semantic Search, Vector Database, FAISS, Chroma, Embeddings, Reranking, Chunking, Multimodal RAG

Evaluation & Observability: LLM Evaluation, promptfoo, LangSmith, LLM Tracing, Monitoring, A/B Testing, CI/CD, LLMOps

Engineering & Models: Python, TypeScript, FastAPI, LangChain, Docker, Cloudflare Workers, Cloudflare Agents SDK, Hugging Face, OpenAI, Claude (Anthropic), Azure OpenAI, GCP Vertex AI (Gemini), OpenRouter

Data & Machine Learning: pandas, Spark (PySpark), scikit-learn, PyTorch, TensorFlow, MLflow, MLOps, SQL, NLP

PROFESSIONAL EXPERIENCE

Zesterv LLC
Generative AI Engineer — **Jun 2025 - Oct 2026**

- Built LLM evals (500-case benchmark plus 100-200 production samples per run) in promptfoo and LangSmith, using LLM-as-judge and rule checks for hallucination and groundedness; every metric must clear its threshold.
- Gated every prompt and model change with a CI/CD regression gate that compares variants side by side in staging; it blocked a newer model that beat public benchmarks but scored worse on SlidesAI cases.
- Owned backend AI architecture for SlidesAI (FastAPI REST APIs, serverless Cloudflare Workers) with task-based model routing across Claude, Gemini, and small models.
- Built and launched the SlidesAI MCP server, used by Codex and Claude agents (about 100 requests a day) to generate full branded decks.
- Cut deck generation time from about 2 minutes to 80 seconds with parallel LLM calls, response streaming, and prompt changes that raised the prompt cache hit rate from 30% to 86%.
- Built hybrid RAG for SlidesAI (BM25 + semantic search on Upstash Vector, Cohere reranking, multimodal over text and images); retrieval quality checked in development and testing before release.
- Built LLM observability in LangSmith and Latitude: TTFT, tokens-per-second, and token-usage dashboards, plus tracing fixes that cut trace volume from 60K to 20K a month.
- Built multi-agent orchestration on the Cloudflare Agents SDK: Durable Objects chat memory, capped retries that feed failures back to the model, and a retry-worthiness check.
- Ran a head-to-head OCR test on about 150 documents; chose Datalab's Chandra API after the self-hosted open-source model missed critical extractions.

Accenture Solutions Pvt. Ltd. (formerly BRIDGEi2i)
Decision Science Practitioner Sr. Analyst — **Mar 2021 - May 2025**

- Built the Text-to-SQL layer and led deployment of an Azure OpenAI HR copilot (GPT-3.5, GPT-4) with function calling and RAG over FAISS and Chroma, serving 500+ HR stakeholder queries a week.
- Led a team of five on client analytics and GenAI projects, owning client communication and requirements.
- Trained and deployed ML and deep learning models (PyTorch, scikit-learn) with MLflow on Azure ML, using drift alerts to keep production stable.
- Built parallel batch pipelines in pandas and PySpark across heterogeneous data sources, reducing run time through query optimization and parallel processing.
- Delivered NLP pipelines on employee survey comments and Power BI dashboards for executive workforce and people analytics reporting.

EDUCATION

B.Sc. Computer Science (Honors) — **2017 - 2020**
Keshav Mahavidyalaya, University of Delhi
Algorithms, data structures, databases, and software engineering.