

Rahul Bhatia

Senior Applied Research Scientist | Machine Learning & AI Leader — Generative AI, Personalization & Applied Research at Scale

Hyderabad, India | rbhatia46@gmail.com | +91 89898 34570 | linkedin.com/in/rbhatia46 | github.com/rbhatia46

SUMMARY

Senior AI/ML leader with 6+ years architecting and scaling production machine learning systems — spanning generative AI, retrieval, personalization, and applied research — for global enterprises. Track record of shipping systems that moved measurable business outcomes (revenue, cost, accuracy) at ServiceNow, Fidelity Investments, CRED, and Rakuten. Systems delivered span 200M+ daily production requests, multi-million-dollar cost savings, and double-digit revenue and efficiency gains. Researcher, published at ACL 2026.

SKILLS

- **Languages:** Python, Scala, SQL, Java
- **Machine Learning & GenAI:** LLMs, RAG & GraphRAG, Multi-Agent & Agentic AI Systems, Multimodal Retrieval (ColPali, VLMs), Embeddings & Vector Search, Transformers, Diffusion Models, Contrastive Learning, Model Fine-Tuning & Evaluation, Recommendation Systems & Personalization
- **Infrastructure & Data Systems:** NVIDIA Triton Inference Server, TensorRT, Model Quantization, Distributed Training & Inference, GPU Clusters, PySpark, Hive, Snorkel

EXPERIENCE

Senior Applied Research Scientist

ServiceNow

Sep 2024 – Present

Hyderabad, India

- **Multi-Agent Search Pipeline with Tool-Calling & Source Arbitration:** Architected and led development of a multi-agent enterprise search system — a Planner agent decomposing queries across specialized Retriever agents (dense bi-encoder, domain-tuned BM25, structured metadata, ColPali multimodal index) and a Synthesizer agent performing confidence-weighted arbitration and contradiction detection — using A2A inter-agent protocols with shared context state. Improved multi-hop answer accuracy 31% and cut irrelevant results 26% over a single-retriever baseline.
- **Production AI Systems at Scale:** Directed the productionization of embedding, reranker, and multimodal models on NVIDIA Triton Inference Server with dynamic batching, TensorRT optimization, and FP16 quantization — sustaining 20M+ daily embedding and 200M+ daily reranking requests across enterprise tenants at sub-50ms / sub-200ms p99 latency, and accelerating research-to-production deployment cycles.
- **Multimodal RAG for Enterprise Document Q&A:** Led the design and deployment of a multimodal RAG pipeline for enterprise document question-answering using ColPali-based visual retrieval and VLMs, extending Q&A coverage to images and tables across enterprise document collections.
- **Next-Generation Multimodal Search — ColPali Tiling Extension:** Designed a novel tiling-based extension to ColPali for multimodal document retrieval, improving retrieval performance 17% over the out-of-the-box baseline while reducing storage footprint and inference latency through quantization and PCA-based compression of patch vectors.
- **Embedding Fine-Tuning Framework — Dynamic Hard-Negative Mining:** Built a unified hard-negative mining pipeline spanning multiple negative sources with teacher-model filtering, curriculum-based scheduling, and per-epoch refresh, enabling bi-encoder fine-tuning to state-of-the-art performance — 87% hit@5 and 76% recall@5 on an in-house multilingual corpus.

Data Scientist, L5

Fidelity Investments

Dec 2022 – Sep 2024

Bengaluru, India

- **CustomerDNA — Generative AI for Client Servicing:** Conceived and led the launch of a generative AI product synthesizing a customer's full-year interaction history across 5 channels into a single briefing for support reps — adopted enterprise-wide to elevate personalized client servicing.
- **LLM-Based Call Summarization at Scale:** Spearheaded an end-to-end call summarization system fine-tuned for customer support calls, generating 2–3 line summaries at GPT-4-level quality for roughly 1/250th the cost; scaled on a 4-node GPU cluster with quantization to score 20M+ calls per year.
- **GraphRAG for Financial Fundamental Analysis:** Built a GraphRAG-enhanced conversational system analyzing S&P 500 earnings reports and financial market news, modeling indirect dependencies between financial entities to improve retrieval over a standard RAG baseline.
- **CRAFT — Cost-Reduced Architecture for Fine-Tuning LLMs:** Devised a novel, cost-reduced architecture for fine-tuning text-to-text transformers using generative pre-training, enabling efficient, economical fine-tuning of large language models in the absence of labeled data.

- **Multi-Label Intent Classification & Generative Banner Design:** Built a few-shot contrastive-learning classifier tagging helpdesk calls across 500 unique intents, and led a team of interns fine-tuning a Stable Diffusion pipeline for marketing banner generation that outperformed DALL-E-2 on in-house evaluation metrics.

Data Scientist

CRED

Nov 2021 – Dec 2022

Bengaluru, India

- **Address Ranking & Enrichment Pipeline:** Built an address ranking and enrichment pipeline ingesting 15K new addresses/day from SMS data, lifting collections efficiency 10% and the whitelisting funnel 4–5% — saving approximately \$1.34M/month for the lending business.
- **Bad Address Classification:** Built an address classification model to identify undeliverable addresses at 96% accuracy, improving collections operational efficiency by 5%.
- **Address Deduplication:** Built a character-embedding-based address deduplication model with hierarchical clustering (94% accuracy) to eliminate duplicate field visits during collections.

Associate Data Scientist

Rakuten

Jan 2020 – Nov 2021

Bengaluru, India

- **Large-Scale Customer Personalization:** Built geo-personalization models from terabyte-scale GPS trajectories to create Geo User Personas, driving a 32% revenue increase and 25% CAC reduction in a business-unit campaign.
- **Spatio-Temporal Clustering & Transport-Mode Classification:** Built large-scale nearest-POI and transport-mode classification models, processing a monthly batch of 300M points across 100K users in under 5 minutes using PySpark.
- **Trajectory Representation Learning & Purchase Prediction:** Built an encoder-decoder model for offline trajectory embeddings and a purchase-probability classifier (AUC 0.87) predicting next offline transactions for a retail chain.

Data Science Intern

InnovAccer

May 2019 – Jul 2019

Noida, India

- **Appointment No-Show Prediction:** Built a Random Forest classifier (AUC-ROC 0.85) with LIME-based interpretability to predict patient no-shows — a problem costing the US healthcare industry over \$150B/year — delivering an estimated \$50K in savings for a business client.

PUBLICATIONS

- Gupta, V., **Bhatia, R.**, Varshney, V., Naik, M. “RESOLVENOW at UZH Shared Task 2026: Rule-Based Type Classification with LLM-Driven Multi-Label Tagging for UN Resolutions.” 13th Workshop on Argument Mining and Reasoning (ArgMining), ACL 2026.

EDUCATION

B.Tech, Computer Science

LNM Institute of Information Technology

Jul 2016 – Jul 2020

Jaipur, India

OTHER INVOLVEMENT

- **Google Summer of Code 2019 — Student Mentor:** Mentored 2 GSoC students as a long-time open-source contributor to Public Lab.
- **Silicon Valley Fellow:** Selected (0.39% acceptance) for a state-sponsored startup exposure trip; pitched an AI product to investors at Silicon Valley Bank, California.
- **Conference Speaker:** PyCon Malaysia 2019 & PyCon Indonesia 2020 — talks on interpretable ML and categorical encoding techniques.