Agam Pandey

Roorkee,India | Mail | Website | LinkedIn | GitHub | Google Scholar

Education

Indian Institute of Technology, Roorkee *August 2022 – Present*
BTech Civil Engineering; Fundamentals of ML, IEC-03 AI Techniques, CSL-537 DL.

GD Goenka Public School, Lucknow *April 2019 – May 2021*
Senior Secondary Education 95.2%; Physics, Mathematics and Computer Science.

City Montessori School, Lucknow *April 2008 – May 2019*
Secondary School Certificate 96.6%; Physics, Mathematics and Computer Science.

Experience

Applied AI Researcher | Mem0 *Remote, SF | Jan 26–Present*
- Building long-context memory for agents; retrieval, temporal reasoning, evals & benchmarking, and continual learning.

AI Researcher | Koustuv Sinha, Meta FAIR Lab *Roorkee, India | Aug 25–Present*
- Working on multimodal video–language causal reasoning benchmarks and foundation models (InstructBLIP, PerceptionLM, Qwen7B); official collaboration with the Department of Data Science and AI and Meta AI.

AI Intern | Swiggy, CEO's Office *Bangalore, KA | May '25–Aug '25*
- Built multi-agent conversational recommender for Swiggy Scenes using LangGraph, Llama-2, multimodal (Text and Vision) GraphRAG embeddings, and retrieval-based ranking models.
- Designed LLM backend on Firebase,Cloud Functions, integrating Apify-Snowflake data pipelines for lead analytics and intent-to-action workflows. CRM with Cimba AI, React, and Node.js, to automate lead scoring and assignment.

AI Engineer Intern | Akki AI *Remote, Delhi | Dec 24–Apr 25*
- Copilot for startups: Built multi-agent conversational and CUA system using CrewAI, LangGraph, and HybridRAG with startup specific personas and dynamic memory (Trends/Core values).
- Designed RLHF-inspired evaluation protocols with preference scoring, feedback logging, and deployed scalable agent clusters on AWS EC2 (FastAPI, Celery) for distributed workloads.

AI Engineer Intern | EzAIx Inc. *Remote, US | Mar 25–Apr 25*
- Built multi-agent onboarding assistant: OpenAI CUA, LangGraph, GPT-4 Vision, and RAG for Microsoft Suites.
- Selenium for autonomous UI control and real-time context understanding via vision, dialog, and task agents.
- FastAPI, Celery and MongoDB Atlas microservices for async orchestration and vector-based multi-tenant retrieval.

Data Science Intern | Mindshift Analytics *Remote | LoR Aug 24–Oct 24*
- Built scalable algorithms analyzing fuel efficiency and vehicle dynamics from GIS telemetry across mining fleets.
- RDP trajectory simplification, DBSCAN clustering, pipelines for anomaly detection, performance insights

Machine Learning Intern | Strabag–Efkon India *Gurgaon | GitHub | Jun –Jul 24*
- Fleet Optimization algorithm tool with custom ML pipelines and Linear Programming (PuLP) for reduced response time of emergency vehicles with double coverage of accident locations, implemented on Hyderabad-ORR, ensuring 8 & 10-min primary & secondary time, 100 % coverage & 80% reliability.
- Built modular scripts in Python which is now being scaled by the company on other Indian National Highways.

Research and Publications

- Agam Pandey *"Before It Persists: Write-Time Defense for Multimodal Agent Memory"* ICML SCALE Workshop 2026
- Agam Pandey, et al. *"CroPA++: Exposing Vulnerabilities in Vision Language Models and Enhancing Adversarial Transferability of Cross-Prompt Attacks"* NeurIPS Reliable ML Workshop [OpenReview]
- Agam Pandey, et al. *"Revisiting CroPA: A Reproducibility Study and Enhancements for Cross-Prompt Adversarial Transferability in Vision-Language Models."* MLRC TMLR [OpenReview] Best Paper Award
- *"Zero-Shot Vision Language Reasoning via Dual-layer Scene Graph Chain of Thoughts"* AAAI-26 Student Abstract Paper
- *"TransPatch: Learning Universal Adversarial Patch for ViT–CNN Cross-Architecture Transfer in Semantic Segmentation"* AAAI-26 Student Abstract, Paper
- Agam Pandey, et al. *"StegGNN: Learning Graphical Representation for Image Steganography."* Submitted to ICCV 2025 (CV4DC Workshop). [Manuscript rejected; under revision for resubmission] Submission

Projects

Hybrid IR RAG Recommendation system with Langchain Planner *GitHub | App | HF API*
- Designed and evaluated a production-ready hybrid recommendation/IR system combining BM25, BGE dense retrieval, weighted RRF fusion, and a fine-tuned MiniLM cross-encoder, improving val recall@10 from 0.08 (BM25) to 0.46.
- Built a structured query planner with Pydantic json and LLM-guided rewrites (Qwen-2.5B with schema constraints), enabling intent/skill/constraint injection before retrieval, 2× recall@10 gains over retrieval-only baselines.

- Ablation studies on candidate depth, fusion weights, and rerank depth (train/val 52/13), identifying top-200 RRF candidates, rerank@10 as the optimal quality–latency trade-off for CPU-only deployment.

ByeLabs: Local Multi-Agent Document Processor | HiLabs AIQuest Hackathon *GitHub | Finals*

- An end-to-end automated roster system converting diverse raw mails into structured Excel exports and analytics.
- Fine-tuned SLM (Qwen3-4B-Instruct) with GRPO and LoRA adapters for robust schema-aware data extraction, normalization, and validation from noisy unstructured documents.
- A fully local processing pipeline with observability, versioning, rollback, and production-grade monitoring (Prometheus, OpenTelemetry, Grafana), ensuring data quality, auditability, and compliance without reliance on third-party APIs.

DL Playground| Data Science Group, IIT Roorkee *GitHub*

- Built interactive drag-and-drop deep learning architecture playground with real-time tensor inference and validation.
- Implemented automatic shape propagation across Conv2D, Pooling, Transpose, Add, Concat, Flatten, and Linear layers.
- Developed compiler-style error detection with visual debugging for tensor dimensions and PyTorch code generation.

FeedCode | Data Science Group, IIT Roorkee *GitHub Repo*

- Built FeedCode, LLM-driven system infers user's coding style from solved problems for personalized code feedback.
- Designed a dual-LLM pipeline for coding-style reasoning and snippet evaluation to generate context-aware feedback.
- Lightweight Flask–MongoDB backend integrating ML workflows for real-time style analysis and code comparison.

Knowledge Graph Embeddings for GraphRAG LLMs | Data Science Group *GitHub | Gitbook*

- Quantitively analyzed CLIP/USE's image & text embeddings with UMAP,t-SNE and PCA for dimension reduction.
- Integrated knowledge graph embeddings to enhance contextual understanding for RAG-LLMs.
- Developed deployed pipeline to create Knowledge Graphs for images and text within a same vector space.

Multimodal Agentic GraphRAG: Zomato Nugget | Data Science Group *GitHub*

- Developed RAH Chatbot, a Gen AI HybridRAG system with multimodal chunking and a MongoDB data lake.
- Weaviate (vector+BM25), Neo4j (graph) with a LangChain ReAct agent (TinyLlama) for contextual responses.
- Deployed application with Docker + FastAPI with LangFuse observability, enabling scalable end-to-end analytics.

Boosting Ensemble Predictive Modeling for Match Outcomes| Amex Superbowl'24 *GitHub*

- Feature engineered batsman/bowler/match statistics temporal windows, generating high-signal team-strength features.
- Trained XGBoost/LightGBM/CatBoost with GridCV tuning, constraints, and leakage-safe temporal validation.
- Built a boosted ensemble inference pipeline with class probabilities and per-model feature attribution.

Spatiotemporal Attention for Trajectory, Conflict Zone Prediction with PET *GitHub*

- Developed and trained SDT-ATT model for highway trajectory forecasting using real-world lane-change datasets.
- Modeled vehicle interactions with social tensors, temporal attention for 20-frame history and 30-frame prediction.
- Optimized PyTorch pipeline for applications, enabling conflict zone prediction and multi-agent collision avoidance.

Tech Enhanced AI Interview Platform| Techshila | STC IITR *AI Interview Platform*

- Developed ML model for interview question generation, using Mistral 7B LLM fine-tuned with Q-LoRA adapters.
- Integrated Faster Whisper for Automatic Speech Recognition & CTranslate2 for automatic speech-text conversion.
- Deployed the model on HuggingFace, with a frontend for inputs and a Flask API backend for audio processing.

Skills

Programming Languages: Python, SQL, PyTorch, DSPy **Machine Learning/NLP:** scikit-learn, Hugging Face, spaCy, NLTK, Pandas, NumPy, LangChain, LlamaIndex, Opentelemetry, CrewAI, mem0, W&B, MLFlow, vLLM. **Tools:** Docker, Kubernetes, Git/GitHub, AWS, GCP, Firebase **API Frameworks:** Django, FastAPI, Flask **Databases:** MySQL, Neo4j, MongoDB.

Service and Leadership

Conference Reviewer | ICML, ICLR, AAAI *Oct 2025 – Present*

- Reviewer for International Conference on Machine Learning (ICML), May 2026–Present.
- Reviewer for International Conference on Learning Representations (ICLR), Feb 2026–Present.
- Program Committee Member for AAAI; reviewed research submissions for Student Abstract and Poster Program 26.

Joint Secretary | Data Science Group, IITR *May 2024 – Present*

- DSG, a undergrad run group works on innovative open source projects in the domain of Machine/Deep Learning.
- Organized hackathons and industry collaboration with WWT and Drexel University for research projects.

Joint-Secretary |Athletics IITR, Institute Sports Council *July 2024 – Present*

- Co-Led the institute team of 50+ athletes with the Secretary, fostering a strong track & field culture. Run Club.
- Coordinated training sessions, events like Annual Athletics Meet and Fitness Camp for campus residents.
- Inter Hostel'24, Institute Trophy'24: Team Gold (Captain), Gold (5000m), Silver (800m, 4x400m Relay);