

Rohit Kumar

Research Engineer · LLMs & Generative Models

✉ rohitkumar.iisc@gmail.com  rohit.vision  [GitHub](#)  [LinkedIn](#)

Summary

I work on controllable and reliable generation, and on predicting inference performance and serving cost. At Armada AI, I am the sole owner of the Support RAG Agent and sole engineering owner of Almanac. M.Tech in AI, IISc Bangalore; co-first-author work on visual storytelling and learning from limited labels (CVPR 2026 Workshop, WACV 2025).

Skills

Retrieval & evaluation: LangGraph, Hybrid Search, BM25, Evals (RAGAS), Qdrant, Langfuse

Training & fine-tuning: Python, PyTorch, Supervised fine-tuning (SFT), LoRA, QLoRA

Serving & infrastructure: vLLM, Ray, Kubernetes, Docker, SLURM, AWQ, GPTQ

Vision & generative: Flux, ControlNet, CLIP, Novel-view Synthesis

Publications

AttriStory: Fine-Grained Attribute Realization for Visual Storytelling with Diffusion Models

Manogna Sreenivas*, Rohit Kumar*, Soma Biswas · CVPR 2026 Workshop

Inference-time AttriLoss raised StoryDiffusion VQAScore from 0.8363 to 0.8636 on the AttriStory benchmark.

[Paper](#) [Project page](#)

TACLE: Task and Class-aware Exemplar-free Semi-supervised Class Incremental Learning

Jayateja Kalla*, Rohit Kumar*, Soma Biswas · WACV 2025

92.35% average incremental accuracy on CIFAR-100 with 0.8% labels: +4.12 percentage points over SLCA+FT (ImageNet-pretrained ViT, 10 tasks).

[Paper](#) [Code](#) [Project page](#)

* Equal contribution.

Invited reviewer: NeurIPS, CVPR & ECCV 2026.

Experience

AI Engineer · Armada AI

Jun 2025 - Present

Trivandrum, India

Manager: Naveen Nair

- Sole owner of the [Support RAG Agent](#), built from an empty repository and deployed for Armada's internal Customer Support team
- Reduced Support Agent cost per query/run to approximately one-third of the initial graph on similar workloads by sharing reformulation and embedding across retrieval branches
- Built an [LLM inference planner](#) with spec-sheet predictions and optional GPU-profile calibration; used its predictions to choose serving configurations, then deployed them
- Sole engineering owner of [Almanac](#)'s local document-Q&A application, including model serving, ingestion, retrieval and SFT/LoRA/QLoRA fine-tuning
- Enabled incremental sparse-index updates and removal with online-learning BM25, addressing organizational vocabulary without full re-indexing
- Established retrieval-stage CRAG gating and end-to-end RAGAS evaluation against curated ground truth

Experience (continued)

Research Engineer · Avataar AI

Jul 2024 - Apr 2025

Bangalore, India

Managers: [Sohil Shah](#), [Shubham Goel](#)

- Developed diffusion-sampling changes for product shape and consistency and intrinsic-decomposition-based relighting for lighting consistency; both adopted into Avataar's production pipeline
- Fine-tuned sparse-view-conditioned Stable Diffusion models for novel-view synthesis, modifying image conditioning and running distributed training with SLURM

Teaching Assistant · IISc Bangalore

Aug 2023 - Apr 2024

Bangalore, India

- Teaching assistant for Digital Image Processing and Signal System in Practice
- Worked on continual learning (L2P, DualPrompt), self-supervised learning (MoCo, SimCLR), frequency-domain filtering and image features

Projects

LLM Inference Planning & Optimization

Case study

Engine-independent prediction of KV-cache capacity, latency, throughput and cost for dense, MoE and hybrid architectures. KV-cache capacity exact on 25/25 dense and MoE configurations. Spec-sheet-only predictions need no GPU access. With brief GPU profiling, dense mean error falls from 15.9% to 3.6% for TTFT (n=174) and 13.8% to 5.2% for batched TPOT (n=242). Errors are mean absolute percentages; profiled results are partly in-sample.

Document-model evaluation

Case study

Compared table-structure, OCR, layout and embedding models; built interchangeable pipeline components for model selection.

Education

M.Tech in Artificial Intelligence

2022 - 2024

Indian Institute of Science (IISc), Bangalore · CGPA: 8.0/10.0

Thesis advisor: [Prof. Soma Biswas](#)

B.Tech in Electrical Engineering

2018 - 2021

Bhagalpur College of Engineering, Bhagalpur · CGPA: 8.75/10.0

Advisor: [Prof. Arjun Kumar](#)

Diploma in Electrical Engineering

2015 - 2018

Government Polytechnic Muzaffarpur, Muzaffarpur

Achievements

GATE 2022: AIR 221 (Electrical), AIR 227 (Instrumentation).