

KARAN TANWAR

tkaran.iitd@gmail.com • +91 95999 88036 • [linkedin.com/in/karantanwar](https://www.linkedin.com/in/karantanwar) • github.com/karan469 • Bengaluru, India

INTERESTS

LLM post-training and alignment — supervised fine-tuning, reward modelling and preference optimisation — and the systems work that makes large models cheap enough to serve inside a latency-bound request path. Four years of production ranking has handed me the same problems at higher throughput: conflicting objectives, learning from implicit and negative feedback, and inference budgets measured in milliseconds.

PUBLICATIONS & RESEARCH

TRUST: Transaction Risk via Unified Sequence and Topology

AAAI 2026

- Karan Tanwar et al. *Proceedings of the 40th AAAI Conference on Artificial Intelligence*. DOI: [10.1609/aaai.v40i47.41450](https://doi.org/10.1609/aaai.v40i47.41450).
- Unifies a transformer over transaction sequences with topological signal from a buyer–seller interaction graph to score delivery-refusal risk, and shows that the two views are complementary rather than redundant. Deployed platform-wide, it now scores every order placed on Meesho at checkout time.

Consistent and Uncertainty-Driven SLAM | M.Tech. Thesis, advised by Prof. Chetan Arora

IIT Delhi, 2022

- Treated per-pixel depth uncertainty as a first-class input to pseudo-RGBD SLAM rather than a post-hoc confidence score, and added a temporal-consistency objective so that predicted depth maps agree across a video sequence.
- Backpropagated an ORB-SLAM trajectory signal through gradSLAM to train the depth model without ground truth, cutting Absolute Trajectory Error **15%** in five epochs, with gains in both trajectory accuracy and reconstruction density over the D3VO, gradSLAM and DROID-SLAM baselines it was layered onto.

EXPERIENCE

Meesho — AI Engineer 2, Feed Ranking & Trust | Bengaluru

Apr 2024 – Present

- Own the home-feed multi-task ranker — a CGC/PLE architecture over a shared FT-Transformer encoder predicting pCTR, pCTCVR and conditional pCVR inside a strict serving budget — and rebuilt it for near-cold users with soft negatives, a learned gate that suppresses items resembling past skips, and an ESMM-style CTunCVR task over the unclicked space. **+5 pp NDCG@50** on orders.
- Built the Return-to-Origin risk model published as TRUST — a transformer over transaction sequences, behavioural features and graph embeddings — and compiled its serving path with TensorRT, cutting latency **400 ms → 20 ms** to allow synchronous scoring at checkout. **2% lower platform RTO, +1.6% NMV**.
- Built Meesho's platform-wide product-similarity model (FLAVA embeddings, hierarchically clustered and FAISS-indexed) to impute catalogue attributes — shipping weight, fabric, taxonomy fields — from matched products where supplier-declared data is unreliable, with Gemini filling the residual long tail. Weight accuracy **60% → 83%**; median shipping cost down **Rs 5.3 per catalogue across 500 sub-categories, +1.4% NMV**.
- Learned self-supervised session embeddings that surface supplier collusion and scripted bot traffic; at a **0.5%** session flag rate the detector holds **81% precision** against manual review and recovers **2.3×** the fraudulent GMV of the prior rules engine. Now a standing feature in Meesho's platform-wide trust and risk model.

Amazon — Software Development Engineer, Central Machine Learning | Bengaluru

Jul 2022 – Mar 2024

- Led LLM4AMZN, adapting generalist instruction-tuned LLMs to Amazon-domain corpora (AmaFLAN) through continued pre-training and instruction tuning, and beating FLAN baselines on accuracy and PR-AUC across a suite of zero- and few-shot catalogue understanding tasks.
- Tuned LoRA rank, adapter placement and schedule to match baseline quality at faster convergence, then ran the model as a pseudo-labeller over cold-start categories for Amazon Business and cut annotation demand **~30%**.
- Built the data and serving infrastructure around it: automated ingestion through AWS Comprehend, DynamoDB and a governed data lake, plus an ECS Fargate orchestration pushing **22,000+** daily Q&A updates into AWS Kendra, measurably improving the retrieval accuracy behind Amazon's Live Assistant.
- Worked with the Internal Search team to recalibrate per-attribute ranking weights in Elasticsearch on a recurring schedule, coupled with a learning-to-rank model over the same feature set, lifting internal search **NDCG by 10%**.

Amazon — Applied Scientist Intern, Search by Data | Remote

Feb – May 2021

- Built an automatic feature-selection method that, for catalogue datasets with 100+ attributes, recommends the top-*k* attributes maximising downstream PR-AUC/ROC-AUC, replacing both hand-picked feature sets and the brute-force cost of training over every available attribute in the catalogue.
- Used Vowpal Wabbit for fast training over high-cardinality text and categorical features, and for interpretable importance ranking of the attributes it selected.

Freecharge (Axis Bank) — Software Development Engineer Intern | Gurugram

Jun – Jul 2021

- Built a filesystem abstraction over Amazon S3 — list, delete and pre-signed URL issuance behind one API — so two Axis Bank services could manage stored assets without proxying object bytes through application servers. Cached paths and access metadata in MongoDB to serve **25K** requests a day, cutting server load **20%**.

EDUCATION

Indian Institute of Technology Delhi | Dual Degree (B.Tech. + M.Tech.), Computer Science 2017 – 2022

- M.Tech. CPI **9.67/10**. Coursework: Machine Learning, Computer Vision, Artificial Intelligence, Information Retrieval, Parallel & Distributed Programming, Probability & Stochastic Processes, Linear Algebra & Differential Equations.

TECHNICAL SKILLS

LLMs & post-training: Supervised fine-tuning, reward modelling, preference optimisation (DPO, GRPO), parameter-efficient adaptation (LoRA), domain-adaptive continued pre-training, pseudo-labelling and distillation; instruction-tuned model evaluation.

Ranking & retrieval: Learning-to-rank; multi-task learning (CGC/PLE, gradient balancing); CTR/CVR/CTCVR modelling (ESMM family); negative-signal and implicit-feedback modelling; embedding-based retrieval; exposure and position debiasing (IPW/IPS, PAL, ESCM²); multimodal and sequential representation learning.

Inference & training systems: Low-latency serving under strict SLAs; TensorRT and ONNX graph optimisation; mixed-precision numerics (FP32/BF16/FP8); distributed training (PyTorch DDP, AMP); throughput and tail-latency profiling, batching policy, precision calibration.

Languages: Python, C++, Java, SQL, Scala (PySpark/Spark SQL).

Frameworks & tooling: PyTorch, Hugging Face Transformers / PEFT / TRL, scikit-learn, NumPy, Detectron2, Vowpal Wabbit; PySpark on Databricks (GCP), MLflow, Docker, Git; AWS (EC2, ECS Fargate, S3, DynamoDB, Kendra, Comprehend), Linux and shell tooling.

SELECTED EARLIER PROJECTS

Spotify Million Playlist Dataset Challenge | Information Retrieval, Prof. Srikanta Bedathur 2021

- Playlist continuation from a seed set of tracks, comparing neural collaborative filtering against an effective nearest-neighbour baseline; NDCG@500 of 0.215 and 0.332 respectively.

Vector-Space Retrieval with Wildcard and Entity Queries | Information Retrieval, Prof. Srikanta Bedathur 2020

- End-to-end tf-idf retrieval system supporting prefix wildcards (mit*) and typed entity queries (L:new* → documents mentioning New York). **Highest NDCG among 76 participants**, and cut index construction time 17% by replacing linear list scans with a hash-based posting index.

YT-Connect: Synchronised Video Watching | Independent project 2020

- Real-time co-watching platform for YouTube and Vimeo built on Node.js socket programming: an event-driven channel between browser and server keeps playback state consistent across clients in a room, with independent rooms, live chat and shared ambient-lighting control.

3D Object Augmentation with ArUco Markers | Computer Vision, Prof. Chetan Arora 2019

- Recovered camera intrinsics from multiple marker views, then used marker pose estimation to solve for the homography and augment 3D objects into the scene; extended into a two-marker AR ping-pong game driven by real-time pose estimation across both marker planes.

POSITION OF RESPONSIBILITY

Mentor, Multimodal Modelling Group | Meesho 2025 – Present

- Mentor two junior data scientists on the multimodal team — scoping their problems, reviewing model design and code, and setting the evaluation bar before anything reaches an online experiment.

Teaching Assistant, Introduction to Machine Learning (Prof. Rahul Garg) | IIT Delhi Jul – Dec 2021

- Authored and graded the course assignments, hosted as Kaggle competitions, and ran the weekly doubt-clearing sessions for the cohort throughout the semester.

Technical Coordinator, Rendezvous | IIT Delhi Jul – Oct 2019

- Led backend engineering for IIT Delhi's cultural festival website, supervising 12 activity heads, and shipped its authentication, admin and sponsor portals on AWS.

ACHIEVEMENTS

- **World rank 989** of 96,000 registered contestants, [Google Code Jam 2020](#) (Qualification Round).
- Top **0.5%** nationally in JEE Mains 2017 and top **2%** in JEE Advanced 2017, from a pool of over 1.1 million candidates.