

KARTIK SHENOY

(323) 986-9136 | kartikshenoy.com | kartik.shenoy@gmail.com | [LinkedIn](#) | [Github](#) | [Google Scholar](#) | San Jose, CA

SUMMARY

Applied ML researcher and engineer with 5+ years of experience in post-training LLMs (SFT, GRPO, PPO), Agentic AI, multimodal learning (vision-language models, cross-modal attention), and large-scale ranking, retrieval and document understanding production systems using deep learning, traditional ML modeling at scale. Published researcher (IEEE, Semantic Web Journal, Alexa Prize semifinalist) with 2 filed patents on multimodal embeddings and document understanding. Strong collaborations across product, engineering and stakeholder teams to drive innovation. Skilled in experimental design, machine learning infrastructure, mentorship to convert business requirements to scalable production systems.

SKILLS

LLMs & Post-Training: RLHF, GRPO, SFT, QLoRA/PEFT, TRL, Agentic AI orchestration, LLM evaluation, RAG, reward modeling

Deep Learning: PyTorch, Transformers (ViT, DeBERTa, LayoutLM, BART), multi-modal training, FSDP, mixed-precision, multi-task learning, knowledge graphs, Computer Vision (YOLO, ConvNeXt), Reinforcement Learning, Triton, Cuda programming, JAX

Retrieval & Ranking: Learning-to-Rank, ANN/Vector Search (OpenSearch), RAG, candidate generation, contrastive, Siamese learning

Infrastructure: Python, SQL, AWS (SageMaker, EC2, S3, Athena), Docker, Kafka, Git, MySQL, DynamoDB, distributed training

EXPERIENCE

Senior Machine Learning Engineer | Bill.com, San Jose, CA

Feb 2023 – Present

- Achieved $F1=0.90$ on document-invoice boundary detection (+50% over production LSTM baseline) by post-training **Qwen 2.5** (1.5B-7B) with **GRPO + QLoRA**; conducted systematic ablations across reward policy design, **LoRA rank**, and **SFT** baseline ($F1=0.82$) to validate the RL approach.
- Developed a **cross-modal ViT-DeBERTa (image+text) fusion** model for the same boundary detection task, training with **FSDP** at scale ($F1=0.86$); comparative analysis across architectures informed the team's decision to shadow launch the GRPO approach.
- Spearheaded and shipped a production **entity search system** over 27M **multimodal (logo+text)** invoice embeddings achieving **91% top-1 accuracy at 280ms P95 latency** by implementing a pointwise **Learning-to-Rank** pipeline with XGBoost Ranker with custom negative sampling and **OpenSearch ANN** (*Patent filed on 1/31/2025: Multimodal Embedding System for Business Recommendations.*)
- Engineered **multi-task training** document understanding achieving 89% classification & 76% field extraction accuracy by jointly training Siamese network-based classification + NER heads on a shared **LayoutLM** (2D attention transformer) encoder.
- Designed **logo extraction (object detection)** pipeline achieving 0.90 mAP (**YOLOv8**) and 87% similarity accuracy (**ConvNeXt**) to automate entity deduplication across millions of invoice logos.
- Shipped production **LLM agent** orchestration: retrieval, tool-selection, context engineering, prompt engineering and grounding mechanisms (to prevent hallucinations) layers using **LangChain**, **N8N**, **LLM evaluation** and monitoring framework using **Braintrust**, **DataDog** for several agents including **customer support, payment delivery through IVR + voice phone calls and invoice understanding**.
- Owned end-to-end retrieval pipeline design (query/doc representation, negative sampling, A/B experimentation). Mentored multiple engineers.

Research Assistant | Information Sciences Institute, USC (Dr. Jon May)

Jan 2022 – Dec 2022

- Improved **multi-agent RL Diplomacy bot** win-rate by 5% by engineering natural-language DAIDE message generation and rule-based strategic reasoning on top of targeted changes to the **reinforcement learning reward modeling policy** in DipNet.

Research Assistant | Centre of Knowledge Graphs, USC (Dr. Filip Ilievski)

Feb 2021 – Jan 2022

- Boosted **graph embedding** quality by +10.6% **Spearman correlation** (0.66 to 0.73 on WordSim353) by retrofitting node representations with **BERT** embeddings + structural features from **Wikidata**, **Probase**, **DBPedia**. Published in Semantic Web Journal.
- Built automated quality-assessment pipeline over **1.1B Wikidata statements**, identifying low-quality links in the knowledge graph on the basis of deleted, deprecated and constraint violation signals.

Software Developer | Barclays Global Service Centre, Pune, India

Jul 2018 – Dec 2020

- Devised a real-time **fraud detection** system on streaming transactions at 20ms avg latency (ROC-AUC 0.7) using ensemble models over **Kafka + PySpark**, reducing false-positive escalations.

EDUCATION

M.S. Computer Science (Concentration in AI), Honors | University of Southern California | GPA: 4.0/4.0

Dec 2022

B.Tech Computer Engineering | University of Mumbai, India | GPA: 9.56/10

Jun 2018

PUBLICATIONS & PATENTS

- 2 US Patents Filed:** Layout-aware field extraction (19/410,970) & Multimodal embedding system (19/043,191) - 2025
- "A study of concept similarity in Wikidata", Semantic Web Journal, October 2023. [[Paper Link](#)]
- "Does Wikidata Support Analogical Reasoning?" Iberoamerican Conf. on Knowledge Graphs, 2021. [[Paper Link](#)]
- "Viola: A Topic Agnostic Generate-and-Rank Dialogue System", Alexa Prize 2020. [[Paper Link](#)]
- "Real-time Indian Sign Language (ISL) Translation", IEEE ICCNT, October 2018. [[Paper Link](#)] [[Github Link](#)]

PROJECTS

- RAG Health Assistant for Los Angeles General patients:** Built an end-to-end pipeline with ChromaDB, LangChain, FastAPI, OpenAI APIs, PHI guardrails to promote HPV vaccination. Poster Presentation at USC ShowCAIS 2026. [[Demo Link](#)].
- Alexa Prize Semifinalist (2021):** Built generate-and-rank dialogue system (DialoGPT, SNIPS NLU intent classifier, FSMs, BERT, AI safety guardrails).
- Retrieval-Augmented QA:** Fine-tuned BART-large achieving 92.6% top-1k retrieval accuracy, 43.9% EM accuracy on Natural Questions dataset.