

Kapil Wanaskar

 San Francisco, CA, US  kapilw25@gmail.com  +16698379485  LinkedIn  Google Scholar

Work Experience

- First American Title, Senior ML Engineer** Apr 2025 – Present | Remote, CA, USA
- Engineered MultiAgent [WebSearch + Verification] workflow automating 250,000+ daily repetitive tasks, eliminating \$1.25M in manual processing costs (\$5 per task) while achieving 94% accuracy rate
 - Optimized model costs by 78% through knowledge distillation: Fine-tuned Qwen2.5-VL-7B to replace Gemini-2.5-Pro-API, deployed via Ollama achieving 2.3x faster inference speeds in production MultiAgent workflow [**Patent Pending**]
- Amazon Web Services (AWS), Applied ML Engineer** May 2024 – Mar 2025 | Cupertino, CA, USA
- Implemented GRPO, an advanced Reinforcement Learning (RL) fine-tuning strategy, to optimize Llama3.1 on structured information extraction from OCR-processed real estate documents; improved training efficiency by 40% over PPO while preserving accuracy across 100K+ samples
 - Optimized large-scale distributed model training across 100+ Trainium accelerators with int4 quantization, implementing efficient data parallelism that reduced inter-node communication overhead by 35% and memory usage by 60%
- Intuitive Surgical, Software Engineer - ML** May 2023 – May 2024 | Sunnyvale, CA, USA
- Built FastAPI + VectorDB-based inference pipeline to detect out-of-distribution robotic surgery logs; achieved 98% precision, enabling early-stage anomaly flagging by the clinical safety team
 - Supervised fine-tuned (SFT) Llama via PEFT (LoRA) on few-shot human-labeled feedback
 - Post-trained embedding encoders and re-indexed FAISS via Online Reinforcement Learning (RL) for similarity updates, improving security by 13%
 - HyperParameter tuned 130,000+ variations of unsupervised models on 150+ GB data using SageMaker + MLflow; achieved 92% precision and 99.9% accurate training inputs
- Vectorr.in, Software Engineer** Mar 2018 – Jul 2022 | Mumbai, India
- Engineered unsupervised customer segmentation system 10k+ (daily) visits stored in Snowflake database, surging customer satisfaction from 3.1 to 4.8.
 - Integrated Apache Kafka and Superset to segment real-time audience data for digital marketing while training Unsupervised models on AWS EC2, amplifying ROI by 23%.

Research Publications

- Multimodal Benchmarking and Recommendation of Text-to-Image Generation Models, IEEE CISOSE 2025**  2025
- Received the "BDS Best Student Paper" award
 - Evaluated 12+ (text-to-image) models (Stable Diffusion, CogView, FLUX, etc.) with ground truth from DeepFashion Multimodal dataset for alignment
 - Designed Weighted Score metric combining CLIP-Score, LPIPS, FID, MRR& Recall@3 via min-max normalization
 - Integrated metadata features and CLIP embeddings to align generated with ground truth image and prompt context
 - Metadata-augmented models (Flux, InContext LoRA) showed ~19% higher Weighted Score & ~15 point FID reduction
- How Robust Are Small Vision-Language Models? A Benchmark of Adversarial Attacks Across 16 VLMs and 17 Attack Types**  2025
- Designed a modular framework to benchmark 3B/4-bit VLMs (e.g., Qwen2.5-VL) under transfer and black-box attacks (PGD, CW, Pixel, SimBA), measuring performance degradation across 15+ adversarial methods
 - Built quantized inference pipeline and evaluation suite (VQA accuracy, SSIM-aware constraints); found CW-L ∞ and Pixel attacks degrade accuracy by up to 52.94%, while GeoDA surprisingly improved performance (+5.88%)
- Evaluation of Local LLM models for shopping recommendation**  2024
- Benchmarked LLMs (Llama 3.1, Gemma, phi3.5, Qwen) for e-commerce Q&A; reduced hallucinations by 30%, streamlining recommendation pipeline selection in shopping agents
 - Transitioned from RAG to FAISS embeddings to optimized vector retrieval, reducing hallucinations by 30%.
 - Evaluated LLMs using "Nemo-Mistral", fine-tuning benchmarking scripts to cut evaluation runtime by 40%
- Prompt Recommendations for AI art, IEEE AIKE, California, USA**  2023
- Extracted features of 5000 images via text embeddings and ensemble models
 - Proposed Graph-based evaluation of 3 recommendation Algorithms and Community Detection Algorithms, via analyzing absence of ratings or preference scores

Education

- MS in Artificial Intelligence, Computer Engineering,** San José State University CA, USA
- Master of Computer Integrated Manufacturing and Bachelor of Engineering, Indian Institute of Technology (IIT) Bombay** Mumbai, India