Kapil Wanaskar

Cupertino, CA, USA   kapilw25@gmail.com   +16698379485   LinkedIn   Google Scholar

Work Experience

Amazon Web Services (AWS), *Software Dev Engineer - AI/ML* May 2024 – present | Cupertino, CA, USA

- Enabling MoE [Mixture of experts] on AWS - GPU/Accelerator [Trainium - TRN]
- Productionized scalable integration tests for CPU vs. TRN (activations, gradients, optimizer), enabling delta accuracy detection, reducing debug cycles by 20%, and improving model onboarding efficiency by 25% for DBRX, Llama4/5.
- Led implementation of NxD backend for Mixtral models using PyTorch+C++ extensions, optimizing CUDA kernels for parallel efficiency, reducing compile time by 10% and boosting throughput by 12%
- Created dashboards to compare MoE architectures (dropless/dropping), tracking MFUs, loss, peak memory, and weak scaling across 5+ PyTorch releases

Intuitive Surgical, *Software (Machine Learning) Engineer* May 2023 – May 2024 | Sunnyvale, CA, USA

- Developed FastAPI for VectorDB-based ML pipeline to detect real-time "Unknown" attacks, achieving 98% Precision
- Built PDF_GPT from Claude+Llama model and Code_GPT from CodeLlama model using AWS Bedrock and Langchain, improving security of company's data by 13%
- Leveraged Distributed Data Processing annotating 150+ GBs of Data in real time up to 99.9% accuracy.
- Using AWS SageMaker, hyper-parameter tuned 130,000+ combinations of Un-Supervised Attack detection model, identifying best Precision of 92% via MLflow

Vectorr.in, *Software Engineer* Mar 2018 – Jul 2022 | Mumbai, India

- Clustered 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%.
- Deployed Docker via CI/CD for automating deployment, achieving a 43% reduction in data overhead.

Research Publications

Multimodal Benchmarking and Recommendation of Text-to-Image Generation Models

- Evaluated 12+ (text-to-image) models (Stable Diffusion, CogView, FLUX, etc.) using DeepFashion 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 and ~15 point FID reduction

Evaluation of Local LLM models for shopping recommendation

- Benchmarked "Llama 3.1," "Gemma," "phi3.5," and "Qwen" on Answer Relevancy, Contextual accuracy, etc, reducing model selection time by 25%
- 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%

GPU based Performance Improvement of Reversible Dual Image Steganography

- Designed a Steganography Algorithm, intercepting encrypted message to withstand several attacks

Skills: C++ , CUDA programming

Prompt Recommendations for AI art, *IEEE AIKE, California, USA*

- 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

Surveillance Drone Cloud and Intelligence Service, *IEEE (MobileCloud), Greece*

- Proposed a surveillance drone cloud for efficient utilization of cloud computing and real-time data sharing

Detection of Cyber Security Threats using IOT Deep Learning

- Suggested TensorFlow deep neural system to classify stolen programming with source code literary theft

Real World Use of Deep Learning Models for Cyber Security in IoT Network

- Presented Deep reinforcement learning models for cyber security in IoT (Internet of Things) networks

Analyzing Effect of Workpiece Stiffness Variation on the Stability in Flank Milling of an Impeller Blade

- Scrutinized FFT (fast-fourier-transform) plots, chatter boundary plots, and stability region diagrams

Education

MS in Artificial Intelligence, Computer Engineering, CA, USA
San José State University

Master of Computer Integrated Manufacturing and Bachelor of Engineering, *Indian Institute of Technology (IIT) Bombay* Mumbai, India