Imama Shehzad

LinkedIn: linkedin.com/in/imamashehzad | GitHub: github.com/imcoza | Email: imamashehzad@gmail.com | Phone: +91 8923222242 | Location: India (Open to Remote)

EXPERIENCE

NLP Engineer June 2024 - Present

Gnani AI *Bangalore, India*

- Architected and developed production-ready Indic-language small language models (SLMs) and large language models (LLMs) with 4B and 12B parameters using pruning, model merging , and knowledge distillation achieving 8% and 15% improvement on IFEval benchmark, demonstrating expertise in model compression and optimization for production deployment.
- Implemented and compared parameter-efficient fine-tuning (PEFT) methods including Low-Rank Adaptation (LoRA), Infused Adapter by Inhibiting and Amplifying Inner Activations (IA3), Adaptive LoRA (AdaLoRA), and Quantized LoRA (QLoRA) using HuggingFace PEFT library and Transformer Reinforcement Learning (TRL); reduced prompt length by 48% and achieved 20% inference speed increase while maintaining model response generation accuracy through rank-64 adapters and 4-bit quantization, optimizing for cost-effective production inference.
- Enhanced Gemma 27B decoder-only transformer LLM with function calling and tool use capabilities using multi-stage supervised fine-tuning (SFT) with instruction-tuning for multi-turn conversation; achieved 15% accuracy improvement on Benchmark for Function Calling in LLMs (BFCL) with 45.99% vs 29.8% baseline, and improved multi-turn conversational accuracy to 14.25% vs 11.62% base performance, enabling production-ready conversational AI systems.
- Developed synthetic dataset preparation pipelines for English and Indic languages using LLM-based data generation and data augmentation techniques for preference dataset and multi turn conversation dataset.
- Fine-tuned compact transformer models ranging from 1B to 7B parameters as routers for query classification and routing using transfer learning and few-shot learning, optimizing latency and inference costs while maintaining quality through dynamic batching and model quantization.
- Built comprehensive machine learning testing framework with automated regression testing, A/B testing , model versioning (MLflow), and performance monitoring (latency, throughput, accuracy metrics) for production AI systems, ensuring reliability and scalability.
- Deployed production models using vLLM for high-throughput inference serving with optimized latency and throughput through continuous batching, PagedAttention, and tensor parallelism. Implemented MLOps best practices including model versioning and A/B testing protocols for production deployment, ensuring 99.9% uptime and scalable model serving infrastructure.

ML Engineer Intern March 2024 - May 2024

Meroku *Remote*

- Fine-tuned Stable Diffusion latent diffusion models for domain-specific image generation using LoRA-based fine-tuning, prompt engineering (negative prompts, style embeddings), and optimization techniques (gradient checkpointing, mixed precision training) for production deployment, demonstrating expertise in computer vision and generative AI.
- Architected production-grade RAG-powered LLM agent system using Haystack framework, integrating sentence transformers (multi-qa-MiniLM-L6-cos-v1), semantic search (cosine similarity, approximate nearest neighbor search), and document processing (chunking, metadata extraction) to reduce hallucinations and improve factual accuracy in conversational AI applications.

Projects

SIGMA-VL: Vision-Language Model (VLM) from Scratch — Python, PyTorch, Computer Vision, NLP GitHub

- 3B parameter VLM: SigLIP vision encoder (27-layer ViT), Gemma language model (18-layer decoder with GQA), multimodal fusion. Implemented RoPE, GQA, RMSNorm, SwiGLU, KV-Cache. QLoRA with 4-bit quantization: 75% memory reduction.
- Production FastAPI REST API with Docker, GPU support. Benchmarks: COCO Captioning (BLEU-4: 0.165, ROUGE-L: 0.428), VQAv2 (40.00% accuracy). KV-Cache optimization: 30% speedup.

NeuralSearch: Hybrid Search Engine with AI Query Intelligence — Python, FastAPI, Vespa, RAG GitHub

- Hybrid search: BM25 + semantic search using ensemble embeddings (MiniLM-L6-v2, MPNet-base, BGE-base). Query expansion via GPT-OSS-120B, MMR algorithm, LLM-based intent classification. Performance: 50-250ms latency, 10-50 requests/second.

ArticleForge: Production-Grade NLP/ML System with Comprehensive Evaluation — Python, NLP, Machine Learning, Feature Engineering GitHub

- Production-ready NLP/ML system: TF-IDF vectorization (scikit-learn, sparse matrices), TextRank extractive summarization (PageRank-inspired, $O(n^2)$ complexity), Sentence-BERT embeddings (all-MiniLM-L6-v2, 384-dim), semantic similarity (cosine similarity). Evaluation metrics: BLEU, ROUGE-L (F1: 0.35-0.45), METEOR, perplexity. Performance: Spearman correlation ¿0.80 (STS benchmark), 10,000 words/second processing throughput.

ReTox Hate Hunter: RLHF for DeToxified Text Generation — Python, RLHF, PyTorch

- RLHF-based toxicity mitigation: fine-tuned FLAN-T5 using Meta AI reward model, PPO algorithm with TRL. Result: 38.85% toxicity reduction while maintaining text quality.

Education

University of Allahabad Allahabad, India

Bachelor of Technology (Computer Science); GPA: 8.9/10.0 *2020 - 2024*

Technical Skills

Programming Languages: Python, C++, SQL, JavaScript

ML Frameworks & Libraries: PyTorch, TensorFlow, HuggingFace Transformers, Sentence-Transformers, Scikit-learn, XGBoost, LangChain, Haystack

Large Language Models: GPT, BERT, T5, FLAN-T5, Gemma, Llama, Fine-tuning, SFT, PEFT (LoRA, IA3, AdaLoRA, QLoRA), Instruction Tuning

NLP & Multimodal AI: Named Entity Recognition, Sentiment Analysis, Text Classification, Conversational AI, Multilingual NLP, Vision-Language Models (VLMs), Vision Transformers, Stable Diffusion

ML Techniques: RLHF, DPO, PPO, Knowledge Distillation, Model Pruning, Quantization, Model Compression, Transfer Learning

MLOps & Infrastructure: vLLM, Model Serving, MLOps, AWS (S3, EC2, SageMaker), Docker, Kubernetes, CI/CD, A/B Testing, Model Versioning

Search & Retrieval: RAG, Vector Embeddings, Semantic Search, BM25, Hybrid Search, Information Retrieval

Certifications

Generative AI with Large Language Models DeepLearning.AI and AWS

Deep Learning Specialization Stanford University and DeepLearning.AI