Manuj Mehrotra,
Sr. NLP/LLM Engineer
8+ Year of Experience
+91-8957551977

https://github.com/MANUJMEHROTRA
https://www.linkedin.com/in/manujmehrotra
mehrotra.manuj7@gmail.com

8+ years of experience architecturing and deploying language models(LLM and SLM). Expertise in efficient training (FSDP,DDP), inference optimization (Quantization, Distillation), and full-stack ML infrastructure.

Roles and Responsibilities:

Publicis Sapient, Hybrid (Bengaluru), Senior NLP Engineer *March 2024-present*

LLM-to-SLM Migration for Intent Detection (Call Transcripts)

- Migrated a multi-intent detection system from LLM to a custom-trained BERT-based Small Language Model (SLM).
- Reduced inference latency and cost by ~35% with <2% accuracy drop, saving $1M annually in operational costs.

Introduced a Real time System (RTS) for inbound calls to call agents:

- Designed a real-time system using custom BERT models on live call utterances to suggest responses to agents via CRM.
- Decreased average turnaround & wait time by 10%, improving call center efficiency.

Call Transcript Analyzer pipeline:

- Optimized batch processing pipeline for 250K+ daily call transcripts using **semantic caching**.
- Boosted processing throughput while maintaining low latency and high accuracy of attribute extraction.

Prompt Versioning and Experiment Tracking

- Built a prompt management system using git-actions and cloud storage enabling version-control for prompt development during experimentation across evolving prompt templates.
- Streamlined experiment reproducibility and evaluation via custom pipelines integrated with inference metrics logging.

AB-InBev, Hybrid (Bengaluru), in CPG Domain *January 2022-March 2024*

Social Listening Platform (BrandWatch):

- Built a social listening engine clustering posts across Instagram, LinkedIn, Twitter, Telegram using **Approximate Nearest Neighbor (ANN).**
- Introduced **dynamic GPU utilization** by selectively running clustering on “clusters of interest,” reducing compute overhead on A100 GPUs without latency compromise.

Recommendation Engine:

- Created a recommender engine and deployed it over an end-to-end Machine learning pipeline over the Azure for a B2B platform in the beer industry in India and Africa, driving upsell and cross-sell opportunities.
- Contributed $5.2M to the company's net revenue in FY2022-23, with over 70% engagement.

AB Testing library:

- Developed an AB Testing library widely used within the organization, with over 400 downloads. The library offers a test and control framework for various experiments.
- Standardized the procedure and methodology for reporting numbers to the organization's FP&A and P&L.

Target Setting and Demand Forecast:

- Developed a ensembled-forecasting technique for setting monthly targets for beer distributors, reducing historical MAPE by 30% and saving over $1 million annually.

PharmaACE Analytics, Pune, Healthcare domain *January 2020-July 2021*

- Created a **predictive alters** for weekly schedule for BDR (Medical Representative), using insurance claims data thereby helping in uplifting the conversion rates per visit.
- Worked on creating a **forecast model** using to predict the monthly revenue of drug, and achieving less than 15% MAPE, which was later used for Resource planning and Budgeting of resources.

Tata Consultancy Services Ltd, Pune *January 2017- August 2019*

- Developed and deployed PMAS (Propensity Model at Scale) on Salesforce Cloud, a look-alike model for user profiling, to generate propensity scores for email campaigns
- Improved email campaign KPIs by implementing strategies in a 20% increase in response and conversion rates compared to rule-based campaigns. Predictive alerts aided in cold start and IP-warming situations.

Project: *December2023-March 2024*

From-Scratch Implementations | *Personal Research* | Jan 2025 -Present

- **Reproduced state-of-the-art architectures** including **Mistral 7B (MoE)**, **Gemma**, and **GPT-2** from first principles in PyTorch, adhering to original publication specifications to deepen understanding of transformer mechanics and MoE routing. (link)
- Implemented Multimodal Vision-Language Model (VLM) **PaliGemma (VLM)** with **grouped query attention** and **KV caching**, achieving a 4x improvement in inference speed for image-captioning tasks. (link)

Efficiency Benchmarking Study: BERT vs. MobileLLM | *Personal Research* | Jan 2025 -Present

- **Conducted a comparative analysis** of model compression techniques on a proprietary dataset of call transcripts. Systematically evaluated **Quantization Aware Training (QAT)**, **Post-Training Quantization (PTQ)**, distillation, and pruning.
- **Results: QAT (INT8) achieved a ~40% reduction in inference cost** with a negligible <2% accuracy drop, demonstrating the viability of SLMs for enterprise deployment.

Project:

Fine-tuning & Custom Training: BERT, GPT-2, Mistral-7B, Llama-7B, Phi-3B and VLMs like (Llama-Vision-11B, PaliGemma)
Training Techniques: Fine-Tuning (SFT), Parameter-Efficient Tuning (LoRA, Adapters, MoRA), Distillation, Quantization (QAT, GPTQ, AWQ), Pruning
Distributed Training: Distributed Data Parallel (DDP), Fully Sharded Data Parallel (FSDP), Torchrun, Ray, Deepspeed.
Applied ML & Data Science: Recommendation Engines, Propensity Models, Predictive Modeling, Demand Forecasting, AB-Testing, Social Listening & Clustering.
Frameworks & Tools: PyTorch, Hugging Face Ecosystem, TensorRT, ONNX, Scikit-learn
MLOps & Cloud: SQL, Azure ML, Docker, Git, MlFlow

Education: *August 2012- May 2016*

- B.Tech, Institute of Engineering and Technology (I.E.T.), Lucknow, in Electronic and Communications.

Teaching & Mentoring *August 2012- May 2016*

UpGrad – Data Science Instructor

- Delivered 20+ live classes (~20 students each) on Data Science/ML topics. feedback
- Achieved 4.7+/5 feedback score in 80% of sessions.

Volunteer Experience:

- Volunteered and mentored Corporate Social Responsibility (CSR) Empower initiative, is a training program with an objective to empower Security personnel and working staff of TCS with spoken Functional English.
- Mentored empower under privileged students with coding skill with an objective to empower them with working knowledge of coding under Robotex program