SHUBHAM S. DESHMUKH

Linkedin: https://www.linkedin.com/in/shubhamdeshmukh619 | Ph: +91 8446236895 | Email: shubhamdeshmukh40@yahoo.in

SUMMARY

AI/ML professional with 7+ years of experience in semiconductor and healthcare AI solution development specializing in transformer-based LLM development, performance and scalability engineering, and distributed multicore inference. Expertise in memory-aware optimization, throughput optimization, latency and bandwidth maximization and predictive pipelines. Published researcher with open-source contributions to AI frameworks, adept at delivering scalable, high-performance AI solutions.

EDUCATION

Master of Engineering, Data Science and Engineering — Bits Pilani, Expected: Sep 2025
Bachelor of Engineering, Computer Engineering — Savitribai Phule Pune University, 2017
Diploma, Computer Engineering — MSBTE, 2014

WORK EXPERIENCE

Samsung Semiconductor India Research | MLE - Staff Engineer | Bengaluru — Jun 2021 - Present

- Alleviated the DDR memory bottleneck in CPU-based Transformer inference by integrating CXL, lifting throughput from a **1×** baseline to **1.99× at 51.3%** CXL memory utilization and up to **2.5× at 90.2%** utilization.
- Engineered the CMM-D mechanism, increasing memory-bandwidth utilization by **60%** and capacity by **20%** at peak.
- Validated a **3× IPM** uplift across **95-core (5×19)** and **144-core (6×24)** deployments versus DDR-only setups.
- Implemented a custom memory-categorization allocator in the PyTorch embedding constructor, directing allocations to pre-reserved DDR and CXL pools cutting allocation overhead by **30%** and memory fragmentation by **40%**.
- Designed and developed ML-driven DRAM failure prediction via chip telemetry, extending prediction lead time by **20%**.

Buddhimed Technologies | NLP Engineer | Bengaluru — Aug 2020 - June 2021

- Trained and fine-tuned **5** OCR-enabled clinical classification models on **100 K+** scanned discharge reports and radiology reports, reducing manual chart-review time by **40%** and processing all records through a secure OCR pipeline for sensitive hospital data.
- Designed and deployed an ML-driven document classification system for discharge summaries and lab reports, automating extraction of clinical entities (diagnoses, medications, lab values) and increasing structured EMR data coverage by **50%.**

Concentrix Global analytics | Data Analyst | Pune — May 2019 - Aug 2020

- Processed over 1 million customer interactions per month, delivering analytics-driven insights that boosted NPS by **12%**.
- These insights informed targeted strategies to capture customer expectations, preferences, aversions, and intents, while analysing ideas, suggestions, and issues, resulting in a **25%** increase in targeted campaign engagement.

MyNalanda Solutions and Services | Data Analyst | Pune — Feb 2018 - Jan 2019

- Developed ML-driven models to integrate and interpret data from 5 stakeholder groups capturing diverse skillsets from academic divisions and other units with **92%** predictive accuracy and a **30%** reduction in reconciliation time.

IIM Lucknow | Remote Intern | Pune — Nov 2017 - Jan 2018

- Solving data science problems and business case studies, including Harvard Business School cases.

PUBLICATIONS

1. Optimizing Transformer Based Inference Efficiency Using CMM-D - **First author** — 21st IEEE INDICON - 2024
2. Scaling Transformer Inference through CPU and Memory Optimization (TBP)- **First author** — 28th IEEE SNPD - 2024
3. Suspect Recommendation for Memory Devices - **Primary author** — 4th IEEE WINTECHCON - 2023
4. Twitter Outburst detection - **First author** — IJSER - 2017

OPEN SOURCE AND ACHIEVEMENTS

- Contributor of opensource Scalable Memory Development Kit for heterogeneous memory.
  - https://github.com/OpenMPDK/SMDK/blob/main/CONTRIBUTOR.md
- Runner-up at IEEE Wintechcon 2024 and Best Paper award in Techcon 2024 at SSIR
- Received Citation for the Work Quality
- National level Gold Medalist in 9th National Choi Kwang Do Championship

TECHNICAL SKILLS

- **Languages & Frameworks:** Python, PyTorch ; **LLM & Attention Mechanisms:** GPT, JAX-int, Transformers, KV-cache Optimization, Multi-Head/MLA/Auxiliary-Loss-Free Load Balancing ; **Architectures & Compilation:** MoE, XLA, QAT, Sparse Activation, Conditional Computation ; **Performance and Optimization:** DualPipe, Distributed Multicore Inference, Dynamic Batching, RAG (HNSW, FAISS), Quantization (FP32, BF16, INT8), PEFT (LoRA/Multi-LoRA), ZeRO.