Seyed Ali Hosseini

Senior AI Engineer & Tech Lead

Enterprise AI, LLMs, OCR & RAG

Tehran, Iran
+98 939 315 0876
Ali79HM@yahoo.com
linkedin.com/in/ali79hm
github.com/ali79hm

Professional Summary

Senior AI Engineer and Tech Lead with 4+ years of experience architecting and scaling enterprise AI systems. Progressed from AI Engineer to Technical Lead at Roshan, overseeing two core teams: Alefba (Persian OCR) and LLM Engineering. Proven expertise in applied machine learning, deep learning, LLM inference optimization, and backend development. Adept at bridging the gap between high-level technical strategy and hands-on implementation, with a strong track record of mentoring engineers, driving architecture decisions, and delivering robust AI solutions.

Technical Skills

LLMs & AI Systems	Inference Optimization, Model Deployment, Quantization, Benchmarking, Embedding Systems, RAG, Real-time Voice AI
ML & Computer Vision	Computer Vision, OCR, Deep Learning, Unsupervised Learning, Document Layout Analysis
Frameworks & Libraries	PyTorch, Transformers, LangChain, vLLM, llama.cpp, OpenVINO, TensorRT, DeepLabV3
Backend & Infrastructure	Python, FastAPI, Django, Flask, LiveKit, Docker, Kubernetes, Triton Inference Server, Linux, Git
Data & Databases	SQL, Elasticsearch, NumPy, Pandas, JavaScript

Professional Experience

Technical Lead

March 2025 – Present

Roshan

Tehran, Iran

- **Technical Strategy & Roadmap:** Defined the long-term architectural vision for the Alefba OCR and LLM teams, effectively translating high-level product requirements into actionable engineering tasks and sprint milestones.
- **Team Mentorship & Code Quality:** Mentored a team of AI engineers, driving engineering excellence through rigorous code reviews, pair programming, and establishing standardized coding practices.
- **Engineering Leadership & CI/CD:** Led the transition and standardization of development workflows for the "Fahm" project, defining branching strategies, release cycles (dev/beta/stable), and CI/CD pipelines to minimize production bugs.
- **LLM Architecture & Deployment:** Architected LLM inference pipelines across diverse hardware architectures utilizing vLLM, llama.cpp, and OpenVINO. Implemented load balancing across models to optimize speed, accuracy, and resource allocation.
- **Enterprise RAG Systems:** Engineered a unified data extraction pipeline for processing diverse document types (Word, Excel, HTML, PDF, Video, Audio) into structured formats for organizational chat systems. Evaluated and benchmarked embedding models to identify the optimal balance of latency and retrieval accuracy.
- **Real-time AI Integration:** Developed and deployed real-time voice conversational AI applications using LiveKit.

AI Engineer

July 2022 – March 2025

Roshan

Tehran, Iran

- **Advanced OCR Development:** Spearheaded the development of the "Alefba" Persian OCR system. Improved handwriting recognition, Arabic diacritics support, and robustness against noise and complex layouts using Attention mechanisms and synthetic data generation.
- **Advanced Training Techniques:** Implemented unsupervised representation learning for segmentation via contrastive and reconstruction methods. Leveraged Teacher-Student pseudo-labeling with large open-source models to continuously bootstrap and expand training datasets.
- **Model Optimization & Training:** Maximized GPU utilization and accelerated training speeds by implementing TAR dataloaders, dynamic learning rates (OneCycleLR, ReduceLROnPlateau), Label Smoothing, and mixed augmentation strategies (random padding, cropping).
- **Document Layout Analysis:** Enhanced structural extraction by replacing heuristic methods with DeepLabV3 segmentation, transitioning towards polygon-based detection, and developing advanced logic for dynamic newspaper column parsing and table extraction.

- **Algorithmic Problem Solving:** Resolved LSTM memory limitations for processing long/continuous text lines by engineering a fixed-chunking and overlapping algorithm combined with CTC loss.
- **Production Scalability:** Improved system resilience by redesigning the document queuing architecture to support fair dynamic resource allocation for long vs. short documents, and implemented Triton inference server health-checks to prevent request failures during crashes.
- **Inference Acceleration:** Accelerated inference speeds using TensorRT and implemented inference on video files via frame-by-frame extraction.

Projects

KDE Jalali Calendar Widget JavaScript, QML, Python

- Developed an open-source KDE Plasma widget for the Jalali (Persian) calendar featuring Google Calendar integration, built utilizing JavaScript, QML, and Python.

Education

B.Sc. in Computer Engineering 2019 – 2023
Shahed University Tehran, Iran

- **GPA:** 17.31 / 20