HASSAN PASHA

PRINCIPAL AI ENGINEER & HEAD OF AI · LLM + VISION SYSTEMS · INFERENCE

Chicago, IL | +1 (312) 721-7019 | hassanpasha.pe@gmail.com | hassanpasha.com

PROFILE

Principal-level AI engineer and technical leader who both sets AI strategy and ships the models himself. Nine years spanning large-scale training-data engineering, LLM fine-tuning and serving, and computer-vision research — with founding/CTO-level ownership of teams, architecture, and delivery. Deep hands-on range across inference optimization (**TensorRT-LLM, SGLang, vLLM**), **CUDA / H100** throughput tuning, **LoRA** adapter infrastructure, retrieval-augmented generation, and autonomous agent systems, owning the full path from dataset design and distributed training to low-latency production on cloud GPU fleets. Built long-context reasoning-over-code datasets, an 8.3M-node full-text scholarly graph (530.7M edges), and real-time voice and vision pipelines now running in production — and founded **Sequence**, a computer-vision & AI-agent platform acquired in a **$5M exit**. Available for principal / fractional-lead AI engagements.

SELECTED RESEARCH & SYSTEMS

Sequence — Computer Vision & AI-Agent Platform **(Founder · $5M Exit)** — Acquired · CV · Pose · AI Agents · AWS

- Built real-time **threat- and activity-detection** pipelines: human pose estimation, keypoint extraction, skeleton normalization, and **temporal transformers** (MMAction2) for behavioral event classification with feedback-loop learning.
- Curated and versioned human-activity datasets (SPHAR, LIRIS), handling ragged-vs-padded tensor batching, normalization, and pooling strategies for multi-frame reasoning.
- Orchestrated multi-stage **AI-agent workflows** over **FastAPI** services with event-driven automation, Dockerized model serving, and scalable inference on **AWS**.

DBSA-27B — Domain-Adapted Reasoning Model — Long-context reasoning · LoRA · TensorRT-LLM

- Led continued pretraining and instruction tuning of a **27B-parameter** reasoning model on curated long-context **reasoning-over-code** corpora, extending effective context for multi-file program understanding.
- Designed a **LoRA adapter infrastructure** for cheap per-domain specialization — hot-swappable adapters served from a shared base to cut fine-tuning cost and GPU footprint by an order of magnitude.
- Compiled and quantized the model with **TensorRT-LLM** and served via **SGLang / vLLM**, achieving high-throughput, low-latency inference on **H100** GPUs with paged KV-cache and continuous batching.

Full-Text Scholarly Knowledge Graph & RAG Corpus — OpenAlex spine · 8.3M nodes · 530.7M edges

- On the **OpenAlex** 492M-work spine, built a full-text-linked scholarly knowledge graph — extracting and parsing complete papers from **arXiv** (via Tralics) and **PMC, bioRxiv, medRxiv** (via JATS) — yielding **~8.3M full-text nodes** and **530.7M edges**.
- Modeled the edge set with non-text bridge nodes and trailing nodes anchored to full-text nodes, giving graph-aware **RAG** a dense, grounded structure over real scientific text.
- Engineered the ingestion, dedup, embedding, and vector-index pipeline; tuned hybrid dense/sparse retrieval and reranking to feed grounded context into downstream reasoning models.

LiveKit Real-Time Voice AI — Streaming ASR/TTS · Low-latency agents

- Delivered a production **LiveKit** voice-AI agent with streaming speech-to-text, LLM reasoning, and text-to-speech in a sub-second turn loop, deployed on autoscaling GPU infrastructure.

CORE COMPETENCIES

LLM & Training: Fine-tuning, LoRA/PEFT, RLHF-style tuning, long-context, distributed training

GPU & Systems: CUDA, multi-GPU optimization (H100/H200/A100/B200/B300), Triton

Computer Vision: Pose estimation, action recognition, temporal transformers, MMAction2

Cloud & MLOps: AWS (SageMaker, Lambda, ECS), Docker, CI/CD, model serving, IaC

Inference & Serving: TensorRT-LLM, SGLang, vLLM, quantization, KV-cache, continuous batching

Retrieval & Agents: RAG, vector search, embeddings, graph retrieval, AI agents, tool use

Frameworks: PyTorch, TensorFlow, Transformers, FastAPI, Django

Languages: Python, C++, C, CUDA, SQL, Julia

PROFESSIONAL EXPERIENCE

Founding Head of AI → Principal AI Engineer · ***Aldea (first 5, acquired by SubQ)*** — 2025 – Present

- Founding team member and A-class shareholder; led a team driving AI research and large-scale **long-context training-data** engineering — built the **DBSA-27B** reasoning model and **reasoning-over-code** datasets.
- Built the **LoRA** fine-tuning and adapter-serving infrastructure and the **TensorRT-LLM / SGLang / vLLM** inference stack, tuning throughput and latency for production serving.
- Built an **8.3M-node full-text scholarly graph** (530.7M edges) on the **OpenAlex** spine, powering graph-aware **RAG** over parsed full text from arXiv, PMC, bioRxiv, and medRxiv.
- Evaluated GPU fleets across **H100, H200, A100, B200, and B300**; ran pricing, vendor management, and technical due diligence to help close a **512-GPU cluster** deal.

Principal Solution Architect — AI / Cloud · ***CVS Health, Boston MA*** — May 2022 – 2025

- Architected secure, serverless AI systems on AWS (Lambda, Step Functions, API Gateway, SageMaker), building scalable microservices and ML inference endpoints.
- Built ETL and feature pipelines feeding ML models; designed automated security measures that reduced non-compliant service creation by **70%**.
- Led a team of 7, drove code standards and CI/CD, and defined an ML approach to future security remediation.

Lead ML Engineer · ***NOCD, Chicago IL*** — Dec 2020 – Jun 2021

- Built ML-backed matching algorithms (Python, Flask, React) and HIPAA-aware data architecture on PostgreSQL, DynamoDB, and Redis for a revenue-critical patient application.

Lead AI Developer → CTO → Advisor · ***Evolution Business, Chicago IL*** — Aug 2018 – Jan 2021

- Built and trained **computer-vision** and deep-learning models (TensorFlow) for a healthcare SaaS product; owned data curation, model accuracy, and distributed training/serving infrastructure.
- Stood up ETL batch pipelines (Airflow, Databricks) and predictive models that increased lead generation.

Lead Engineer · ***Advance Technology Services, Schaumburg IL*** — Jan 2017 – Aug 2018

- Developed custom **NLP** algorithms (tokenization, bag-of-words over millions of scraped RFP PDFs) that helped close **$6M** in won RFPs; cut response time **50%** with improved datasets and ETL flow.

EDUCATION

Doctoral Research — Theory of Mind & Machine Cognition · ***University of Illinois Chicago*** — Ph.D. research (not completed)

Computational models of belief, intent, and reasoning about other minds — the capabilities now central to modern LLMs.

B.S. in Computer Science · ***University of Illinois Chicago*** — 2017