Muhammad Qasim

Applied ML Engineer — Retrieval Systems & Search Infrastructure

kas-sim.github.io linkedin.com/in/kas-sim github.com/Kas-sim amkassim444@gmail.com

Technical Summary

Languages: Python (4 yrs), Java (2 yrs),C++ (1.5 yrs), SQL
ML / AI: PyTorch, TensorFlow, HuggingFace, BERT fine-tuning, embedding models, reranking
Retrieval Systems: RAG pipelines, hybrid search, dense vectors, cross-encoder rerankers
Systems: Linux, Docker, CI/CD (GitHub Actions), OCR pipelines, persistent storage
Algorithms: Tries, Inverted Index, HashMaps (O(1)), Graphs, Priority Queues, TimSort
Practice: LeetCode (35 Easy, 2 Medium)

Selected Projects

MQNotebook — Enterprise RAG Platform

Designed and implemented a production-grade RAG system to ingest unstructured enterprise data including scanned PDFs, flattened documents, PPTX speaker notes, and spreadsheets.

Built an OCR-first ingestion pipeline using pdf2image and Tesseract to recover text from image-only PDFs with page-level provenance.

Implemented hybrid retrieval (dense vectors + symbolic filters) followed by cross-encoder reranking to maximize context precision for LLM inference.

Solved Windows file-locking issues in persistent vector stores via session-isolated storage and lazy cleanup mechanisms.

Deployed on Streamlit Cloud with secure per-user API key injection and freemium access control.

DevShelf — Vertical Search Engine (Java)

Built a vertical search engine from first principles implementing tokenization, stop-word removal, and an inverted index using `HashMap<String, List<int>>`.

Implemented Trie-based autocomplete, graph-backed recommendations, priority-queue ranking, and deterministic merge-sort ordering.

Architected offline indexing and online querying paths to achieve constant-time metadata lookup and low-latency search.

Led a team of two developers, defined MVC architecture, enforced Git workflows, and authored full technical documentation.

BabyGPT — Generative Modeling Foundations

Implemented a character-level LSTM language model with a custom tokenizer to understand sequence modeling and next-token prediction.

Built a temperature-controlled sampling loop to manage creativity versus coherence during inference.

Experience

Arch Technologies (Remote)
Machine Learning Intern *Present*

Fine-tuned BERT models by replacing and retraining classification heads; ran experiments, evaluated metrics, and iterated on preprocessing pipelines.

Produced reproducible training and evaluation scripts and collaborated with engineers to integrate models into internal prototypes.

Education

Sukkur IBA University
Bachelor of Computer Science — AI Systems & Distributed Computing *Expected 2028*