# USMAN MEHBOOB MOHAMMED

+1 602-748-6222 • musmanme@asu.edu • linkedin.com/in/mohammedusmanmehboob
github.com/usmanmohammed2306 • Portfolio

## SUMMARY

Software engineer with 2 years 9 months of enterprise experience and current ASU research in LLM training, agents and retrieval. Seeking US Applied AI engineering roles after May 2027; open to relocation.

## EDUCATION

**Masters in Computer Science** Graduating May 2027
Arizona State University, Tempe, AZ GPA: 4.00/4.00
B.Tech., GITAM University, India | May 2020
Relevant coursework: Topics in Natural Language Processing, Image Processing and Analysis, Semantic Web Mining, Advances in Robot Learning, Agentic AI, Knowledge Representation and Reasoning, Information Assurance and Security, Statistical Machine Learning, Operational Deep Learning

## TECHNICAL SKILLS

**Languages:** Python, Java, SQL, TypeScript, JavaScript
**LLM Training:** PyTorch, VERL, GRPO, SFT, LoRA, DPO-style objectives, FSDP, FlashAttention-2, vLLM, Slurm, Weights & Biases
**Retrieval and Agents:** RAG, FAISS, BM25, Sentence Transformers, scikit-learn, agent memory and evaluation
**Software Systems:** FastAPI, React, Spring Boot, REST APIs, SQLAlchemy, PostgreSQL, SQLite, Docker, AWS, Git, pytest, Playwright

## RESEARCH EXPERIENCE

**Arizona State University, Tempe, AZ: Research Intern** 05/2026 - Present

- Optimized Qwen3-4B GRPO training on four A100 GPUs, raising model FLOPs utilization from 9% to 21% by correcting loss scaling, aggregation and EOS masking across five comparable configurations.
- Designed and implemented mistake-specific supervision and sequence-level DPO-style optimization for GRPO.
- Resolved an FSDP collective deadlock caused by empty micro-batches by synchronizing optimizer behavior across ranks; added training diagnostics.
- Published 6,413 GSM8K preference pairs with verified numeric answers and traceable negatives using Qwen2.5-3B and vLLM.
- Separately fine-tuned Qwen3.5-4B with rank-16 LoRA on 25,627 reasoning examples.

## PROFESSIONAL EXPERIENCE

**Arizona State University, Tempe, AZ: Information Systems Grader** 10/2026 - Present

- Grade coursework and resolve student questions in a 180-student Machine Learning in Business course.

**SunTec Business Solutions, India/UAE: Software Engineer** 03/2022 - 11/2023

- Migrated legacy components into 4+ Java/Spring Boot microservices deployed with Docker on AWS in an eight-person agile team; received the CTO Spot Award.
- Reduced generation time by 33% for workflows covering 10M+ customer pricing reports using parallel processing, application caching and SQL optimization.
- Mentored 5+ engineers in Java, system design and deployment; led code reviews and coordinated releases with product and QA.

**SunTec Business Solutions, India: Associate Software Engineer** 03/2021 - 02/2022

- Reduced manual data entry by over 80% with Excel-to-UI automation; delivered 10+ features and resolved 20+ production defects.
- Refactored core Java data-processing workflows, reducing execution time by 50%.

## SELECTED AI AND ENGINEERING PROJECTS

**TabRL: Two-Stage Text-to-Table Generation**

PyTorch, LoRA, FSDP, FlashAttention-2 | Team code | Schema model | Table model

- Led development of a two-stage Llama 3.1 8B pipeline converting unstructured text into typed JSON schemas and schema-conditioned tables; implemented and trained both LoRA stages.
- Built training across 2K-8K-token contexts, checkpoint merging and inference validation; published two checkpoints and produced 120 schema predictions and 120 table outputs for error analysis.

**MultiplexRAG: Query-Adaptive Retrieval**

Python, FAISS, BM25, Sentence Transformers, scikit-learn | Code

- Built a learned router selecting BM25, dense FAISS search or reciprocal-rank fusion, with weakly supervised logistic regression and a confidence-based fallback.
- Increased MRR@10 from 0.5110 to 0.5270 on 324 NFCorpus validation queries versus the strongest fixed retriever; implemented query-level splits, ranking-metric checks and ablations.

**PRAXIS: Procedural Memory for Tool-Using Agents**

Python, BM25, TF-IDF, vLLM, tau-bench, BFCL | Code

- Built inference-time procedural memory that retrieves reusable tool guidance from prior agent trajectories without model fine-tuning.
- Added memory validation, PII redaction, mutation checks and three-layer train/test isolation; retrieved up to five memory cards within a 2,400-character context budget.

**OwnStack: Governed Agent Workflows**

Python, FastAPI, React, TypeScript, PostgreSQL, Docker | Local prototype

- Built tenant-scoped software-change workflows with exact approvals, durable two-attempt repair and explicit human adoption of generated proposals.
- Implemented external validation for a narrow PDF format, blocked an early-exit false pass and rejected stale worker results; checked recovery with synthetic candidates and providers.

**IntentCart: Evidence-Grounded Workstation Planning**

Python, FastAPI, SQLite, JavaScript, constrained optimization | Local prototype

- Built a planner with source-bound specifications, explicit unknowns, conflict diagnosis, minimal-change repair and approval-bound sandbox checkout.
- In a fixed 384-goal synthetic evaluation, recovered all 73 minimum-cost feasible baskets using 143 fixture lookups versus 1,920 for full-census inspection.

**AutoFrame: Vision-Language Camera Control**

Python, FastAPI, Gemini, Android, ESP8266 | Team code

- Developed a team photography prototype integrating vision-language reasoning, an Android camera server and a three-axis hardware rig for iterative framing.
- Integrated retrieval memory from 167 attempts across 29 objectives; evaluated 28 autonomous sessions with proposal validation, stopping conditions and execution traces.

**Relief Atlas: Consent-Gated Task Planning**

TypeScript, React, Fastify, PostgreSQL/PGlite, MCP | Local prototype

- Implemented organizer-scoped draft resumption, deadline clarification and review-only MCP tools for meal, visit and ride planning; exact approval gates request creation and queued outreach.