SHUBHAM GAUR

+91 9814070637 | shubhamgaur22091999@gmail.com | LinkedIn |
Website | GitHub

Software Engineer - AI/ML

Software Engineer with 4+ years of experience in building production systems, transitioning into Applied AI/ML Research. Hands-on experience in designing and evaluating applied ML systems under real-world constraints, including a local Retrieval Augmented Generation (RAG) pipeline, achieving 87% answer accuracy with citation grounding, and a multilingual OCR+NLP evaluation system, reducing defect escape rate from 15% to 3%. Strong foundation in Python, system design and evaluation, with a growing focus on applied machine learning and data-driven experimentation. Seeking AI Residency opportunities to deepen research and large-scale ML capabilities and contribute to real-world ML product development.

SKILLS

- **AI/ML:** RAG Architecture, Vector Databases (ChromaDB/ FAISS), GenAI (Prompt Engineering), Natural Language Processing (NLP), Embeddings (Sentence Transformers), LangChain, OCR (Teserract, PaddleOCR), ML Evaluation Metrics, A/B Testing, Failure Analysis
- **Languages:** Python, C#, SQL
- **ML/ Data Libraries:** Hugging face Transformers, Scikit-learn, Pandas, Numpy, Matplotlib, TensorFlow
- **Software Engineering:** Agile/Scrum, CI/CD (Azure Devops), Data Structures & Algorithms, Distributed Computing, Git, OOPs, Rest APIs, System Architecture
- **Soft Skills:** Analytical Problem-Solving, Cross-Functional Collaboration, Mentorship & Team Development, Stakeholder Communication

PROJECTS

Local RAG System for Private Document Q&A

Personal Project | 2025

- Designed and implemented an end-to-end Retrieval-Augmented Generation (RAG) system for answering queries over private documents without external API dependencies.
- Built full pipeline: document ingestion, token-based chunking with overlap, embedding generation, vector retrieval and LLM-conditioned response generation.
- Formulated and tested hypotheses on how chunk size, overlap and embedding choice affect retrieval recall and adopted a hybrid chunking strategy (tokens + overlap), improving answer accuracy by ~35%.
- Evaluated answer accuracy using manually curated Q/A pairs with citation correctness as the acceptance criterion, and performed systematic failure analysis across partial retrieval, semantic drift, and citation mismatch to further improve accuracy.

ML-Powered Multilingual Evaluation System

HCLTech | 2024-25

- Designed and deployed a production ML-based multilingual pipeline to automate translation quality assessment across 12+ languages.
- Developed end-to-end system: screenshot capture -> OCR extraction -> Unicode normalization ->

embedding-based similarity scoring -> confidence-based decisioning.

- Conducted comparative evaluation of OCR engines across 8 languages, analyzing error distributions under varying script complexity and image noise.
- Reduced translation failure rate by 25%, catching critical errors missed by manual review.

Real-Time Context-Aware AI Assistant

Early-Stage Applied ML Research

Focus: Exploring design and ML tradeoffs for real-time, context-aware AI assistants under mobile constraints

- Investigating architectures for a lightweight AI assistant capable of responding to on-screen context in real- time (chat-head style interaction)
- Experimenting with screen content extraction techniques and compared on-device vs cloud-based inference tradeoffs, evaluating model capability, privacy and compute limits.
- Analyzing privacy implications of screen-level context awareness and exploring strategies for sensitive data handling.

Learning Objectives:

- Real-time inference constraints and latency budgeting
- Model compression and optimization concepts for resource-constrained environments.
- Early-stage evaluation strategies for interactive AI systems.

WORK EXPERIENCE

Senior Engineer - Systems & Automation

HCLTech (formerly Becton Dickinson) | 08/2023 - Present

- Transitioned from traditional automation into ML-assisted evaluation systems using OCR and NLP for multilingual content validation, reducing production quality failure by 70%.
- Built an ML-based localization testing solution, integrating Optical Character Recognition (OCR) for screenshot text extraction and multilingual validation, significantly reducing defect leakage and improving translation accuracy.
- Conducted systematic experiments to benchmark OCR engines and embedding models through precision/recall driven experiments, applying classical ML concepts such as similarity metrics and threshold optimization to support production readiness.
- Developed quality monitoring dashboards and integrated automation into Azure DevOps CI/CD pipelines, enabling centralized reporting of automation health, ML evaluation metrics and release readiness.
- Mentored junior engineers on Python and ML evaluation practices, conducted code reviews, and enforced Git- based version control best practices.

Engineer - Automation and Data Validation

Coforge | 03/2021 - 07/2023

- Built and maintained data validation pipelines for automobile-related enterprise systems using SQL, ensuring data integrity across upstream and downstream services.
- Designed C#-based automation workflows incorporating Excel-driven data inputs and Extent Reports-based execution dashboards to enhance visibility and reduce manual testing effort.
- Performed API verification for automobile system services, including financial and billing modules, ensuring request/response correctness and schema adherence.

- Validated complex multi-step data transformation, related to pricing, invoicing and financial calculations and identifying data flow issues across systems.
- Conducted root cause analysis on failures in distributed systems, collaborating with development teams to identify data and integration issues.

EDUCATION

Bachelor's in Computer Science Engineering | Chandigarh University, Gharuan
Session: 2017-2021 | 7.86 CGPA

12th (CBSE) | Kendriya Vidyalaya, Chandigarh
Session: 2016-2017 | 87%

10th (CBSE) | Kendriya Vidyalaya, Chandigarh
Session: 2014-2015 | 10 CGPA

CERTIFICATIONS

Core Python Certification
01/2026 - 06/2026 | ATL Education Foundation (6-week program)

ADDITIONAL

x

Publicly documenting my AI/ML learning process and applied projects on LinkedIn, translating technical concepts into practical, real-world insights.

Research Interest: Real-time context-aware AI systems that operate under mobile and on-device constraints. Interested in learning useful representation from on-screen context, understanding latency, accuracy and privacy tradeoffs, and evaluating interactive AI assistants in resource limited settings.