



Syed Zain Gillani

AI Engineer · GenAI · RAG · LLM Applications

Open to relocation to the UAE — available immediately

CONTACT

@ zain@codingeagles.dev

+92 325 1736468

zain.codingeagles.dev

Muzaffarabad, Azad Kashmir,
Pakistan

CORE SKILLS

AI / Generative AI

RAG Systems AI Agents
Prompt Engineering
Context Engineering
Fine-tuning (LoRA) Embeddings
Vector Databases MCP
LangChain Claude SDK
Google AI SDK Model Deployment

Automation

n8n Zapier
Workflow Automation

Web & Backend

HTML5 CSS3 JavaScript
Python FastAPI REST APIs
Docker Linux

Data & Networking

PostgreSQL MySQL pgvector
TCP/IP HTTP/HTTPS DNS

LANGUAGES

English Conversational
Urdu Native
Punjabi Native
Hindko Native

EDUCATION

BS Computer Science

In progress — 2nd year
Boys Degree College, Muzaffarabad,
Azad Kashmir, Pakistan

NATIONALITY

Pakistani

PROFESSIONAL SUMMARY

Junior AI Engineer with 2+ years of software experience and a focused ~2 years building production generative-AI systems. I design and ship **RAG pipelines**, autonomous **AI agents**, and secure, low-latency **LLM applications** using the Anthropic Claude and Google AI SDKs. Comfortable across the full workflow — prompt and context engineering, embeddings and vector search, fine-tuning, and deploying custom models on Linux servers with hardened, low-latency APIs. Currently building CodingEagles' internal AI platform while completing my Computer Science degree.

EXPERIENCE

AI Engineer — CodingEagles (Agency) & Freelance

2024 — Present

- Designed and shipped **CodingEagles' internal AI platform**, unifying RAG search, AI agents and workflow automation into one system used across the agency's day-to-day operations.
- Built a **custom RAG pipeline** over the agency's internal documentation and infrastructure — chunking, embeddings and a vector store (pgvector/Qdrant) with reranking — cutting internal information lookup from minutes to seconds.
- Developed **AI agents for local restaurants** (ordering, menu Q&A and customer support) using function calling and the **Model Context Protocol (MCP)** to connect live business data.
- Applied **prompt and context engineering** to reduce token usage and response latency, and added **prompt-injection defenses** and input validation for secure, production-grade deployments.
- Fine-tuned** open LLMs (LoRA) for domain-specific tasks and **deployed custom models on Linux servers** behind FastAPI, with Docker, monitoring and rate-limited secure APIs.
- Automated recurring business workflows with **n8n and Zapier**, integrating AI steps into email, CRM and messaging pipelines.

Web Developer — CodingEagles (Agency) & Freelance

2022 — 2024

- Built **responsive websites and web apps** for clients using HTML5, CSS3 and modern JavaScript, focused on clean, accessible and fast-loading interfaces.
- Delivered projects end-to-end — from layout and styling to REST API integration and deployment — establishing the software foundation that carried into AI engineering.

KEY PROJECTS

CodingEagles Internal AI Platform **Deployed**

Internal AI system powering agency-wide search, agents and automation from a single interface.

Stack: Claude & Google AI SDKs · Python / FastAPI · vector DB · n8n · Docker on Linux

Infrastructure RAG Pipeline **RAG**

Retrieval-augmented assistant over CodingEagles' internal docs and infrastructure for instant, sourced answers.

Stack: embeddings · pgvector / Qdrant · hybrid retrieval + reranking · LangChain

Restaurant AI Agent **Agents**

Conversational AI agent for local restaurants handling menu Q&A, orders and customer support.

Stack: LLM function calling · MCP · secure API integration · low-latency serving

For more projects and my full portfolio, visit zain.codingeagles.dev