

MOHAMMAD USAMA ANSER

Full Stack Software Engineer

+92-321-7011-024 | work.usama.anser@gmail.com | [linkedin.com/in/usama-anser-a34406176](https://www.linkedin.com/in/usama-anser-a34406176) | Lahore, Pakistan

SUMMARY

Software Engineer with nearly 5 years building production systems, focused on LLM infrastructure and AI-powered applications. Depth in Python, FastAPI, and agent tooling; shipping tool-calling runtimes, MCP servers, RAG pipelines, and LLM-driven data workflows used by engineering teams in production. Backed by 4 years of production web engineering across Rails, React, Postgres and AWS, which is what makes integrating AI into existing systems, not just greenfield demos, straightforward.

EXPERIENCE

Full Stack Software Engineer

PromptLayer | USA - Remote | 01/2026 - Present

PromptLayer is a prompt engineering and LLM observability platform that AI teams use to manage, version, evaluate, and monitor prompts in production

- Shipped 10+ major features across a Python/FastAPI + React/TypeScript stack, owning end to end from Postgres schema to frontend UI, used by 100+ teams.
- Built the Tool Registry from scratch; versioned, reusable tool definitions with an editor UI, public API, RBAC permissions, and response-schema enforcement.
- Engineered the agentic auto-execution loop: a multi-provider engine driving LLM-to-tool round-trips across streaming and non-streaming paths in the playground, evaluations, and workflows, supporting 10+ LLM providers with per-call OpenTelemetry (OTEL) trace spans.
- Added AI-mocked agent runs; deterministic LLM-generated tool results that let users test full agent loops without hitting live endpoints, cutting agent iteration time by 50% and eliminating external calls from test runs.
- Developed agent infrastructure for PromptLayer's in-app AI assistant: a Slack integration with OAuth account linking, JSON response schemas for its 60+ tools, and structured outputs with charts rendered as images.
- Added first-party support for provider-native tools (web search, code execution, file search, bash/text-editor tools) across Anthropic, OpenAI, Google, Vertex AI, and Amazon Bedrock.
- Built AI-generated trace summaries and per-call spans for agent execution loops, cutting time-to-diagnose a failed run by 70%.
- Built PromptLayer's official MCP server exposing 60+ API tools to Claude, Cursor, and other MCP clients.
- Automated weekly release-notes generation by dog-fooding PromptLayer's own workflow product, replacing a 3-hour manual process.

Senior Software Engineer

Devsinc | USA - Remote | 02/2024 - 01/2026

Devsinc is an IT services and consulting company building scalable web applications for startups and enterprises across the US, Europe, and the Middle East

- Built an LLM-powered lead engagement pipeline end to end: GA4-instrumented contact forms captured user responses, which moved via SFTP into Celery workers for async processing, with AWS Glue running ETL before the normalized data reached the model layer.
 - Designed the generation stage: enriched behavioural data with location and profile signals from Analyze360, fed structured context to an LLM that produced personalized outreach emails, and persisted processed records to BigQuery for downstream analysis.
 - Orchestrated the async pipeline with Google Cloud DAGs, handling multi-stage processing across SFTP ingestion, transformation, model inference, and warehouse load.
 - Engineered product integrations with the Microsoft ecosystem (e.g., Office 365, Azure) via Microsoft Graph API, boosting customer engagement by 30%.
 - Implemented Trailblazer in an application for modular and reusable code, boosting code maintainability and readability.
 - Collaborated with cross-functional teams, including product managers, designers, and QA engineers, to ensure the delivery of high-quality software.
-

Software Engineer

Devsinc | USA - Remote | 11/2021 - 02/2024

Devsinc is an IT services and consulting company building scalable web applications for startups and enterprises across the US, Europe, and the Middle East

- Designed and executed a Python based web application incorporating advanced features like user authentication, authorization, and real-time WebSockets, boosting user engagement by 35%.
- Engineered a React application with RxJS for real-time data streaming and state management, boosting user experience.
- Developed and executed a distributed caching system using Redis, improving response times of a Rails application by 45%.
- Integrated Elasticsearch into a Python based application to enable advanced search functionality and improve data retrieval time by 10%.
- Implemented new features and enhancements, and optimized existing codebase, leading to a 30% improvement in application performance.

SKILLS

AI Agents	Airflow	Anthropic API	AWS	AWS Glue	BigQuery	Celery	CI/CD	Django
Docker	EC2	Elasticsearch	Embeddings	ETL Pipelines	FastAPI	Fine-tuning		
Google Analytics 4	Google Cloud Platform	JavaScript	Kubernetes	LangChain	LangSmith			
LLM Evaluation	LLM Observability	MCP	Next.js	OpenAI API	pgvector	PostgreSQL		
Prompt Engineering	Prompt Management	Pytest	Python	RAG	Rails	React	Redis	
REST	Ruby	S3	TypeScript					

PROJECTS

Lead Engagement Platform

LLM-driven lead engagement and personalization platform. Captured user feedback through GA4-instrumented forms, processed responses through an async pipeline, and generated personalized outreach emails via LLM using enriched behavioural and location data.

Stack: Python, Celery, AWS Glue, Google Cloud, BigQuery, SFTP, LLM APIs, GA4

Meal Suite

Meal Suite is a platform programmed to help seniors track their diet and medication routines. It features an interactive reporting module that generates daily, weekly, and monthly reports for seamless sharing with doctors, guardians, and physicians, fostering improved health management

Stack: Python, Django, React, PostgreSQL, Celery, AWS

CRM365

CRM365 is a customer-centric insights company, providing customer relationship management software and applications focused on sales, customer service, marketing automation, e-commerce, analytics, and application development

Stack: Python, React, PostgreSQL, Microsoft Graph API

Book Pedia

Book Pedia is an AI-based application that provides book readers with in-depth chapter summaries and answers to questions about their favorite books

Stack: Python, FastAPI, React, PostgreSQL, pgvector, OpenAI API, RAG, Embeddings

EDUCATION

Bachelor's in Computer Science

FAST National University of Computer & Emerging Sciences

09/2017 - 05/2021 | Lahore, Pakistan

INTERESTS

Passion: Passionate about AI development, focusing on positive societal impact and data privacy.

Hobbies: Enjoys writing, gaming, reading books, and exploring new foods.