

PRASAD RATNAKANT SAWANT

AI Engineer

Email: prasadsawant1618@gmail.com | Phone: +91 7820923701 | India
[Portfolio](#) | [GitHub](#) | [LinkedIn](#)

PROFESSIONAL SUMMARY

Applied AI Engineer with nearly 3 years of professional experience across AI engineering and data analytics, leading product-oriented Generative AI, RAG, agentic workflows, OCR-enabled automation and enterprise data solutions. Hands-on with Python, FastAPI, LangGraph, MCP, Gemini, PostgreSQL/pgvector, MSSQL, tool calling, evaluation, multi-tenant applications and Linux/Docker deployment. Owns solution design through full-stack delivery, enterprise integration, validation and iterative improvement with business teams.

TECHNICAL SKILLS

- AI Agents & LLMs:** LangGraph, Agno ADK, Gemini, locally hosted LLMs, prompt engineering, tool/function calling, human-in-the-loop workflows, multi-agent applications, LangSmith tracing
- RAG & MCP:** MCP servers using stdio and Streamable HTTP, document ingestion, OCR, chunking, embeddings, semantic retrieval, reranking, citations, response grounding, RAG evaluation, token-usage tracking, ChromaDB, PostgreSQL/pgvector
- Backend, Data & Integrations:** Python, FastAPI, PostgreSQL, MSSQL, SQLAlchemy, Alembic, REST/JSON APIs, Quasar/Vue, Office.js, SQL, ETL, QuickBooks, Oracle NetSuite, Maconomy, SFTP and external API integrations
- Production & ML:** Docker, Linux, Nginx, Unicorn, GitHub Actions CI/CD, MLflow, structured logging and monitoring, latency/failure analysis, PaddleOCR, YOLO, CPU model quantization

PROFESSIONAL EXPERIENCE

AI Engineer | VSKY Solutions Pvt. Ltd.

August 2025 – Present

- Architecture & Integration:** Build modular FastAPI and REST/JSON services with PostgreSQL or MSSQL data layers, Quasar interfaces, Office.js add-ins and embeddable assistant widgets for enterprise workflows.
- Security & AI Quality:** Apply human validation, ERP authentication and role controls, tenant and document-group isolation, encrypted credentials and server-side secrets; evaluate prompts, retrieval, reranking and grounded responses with citations, token tracking, latency and failure analysis.
- Production Delivery:** Deploy and operate AI services on Linux using Nginx, Unicorn, Docker and GitHub Actions CI/CD, using MLflow where applicable and supporting releases with structured logging and operational monitoring.
- Technical & Product Leadership:** Lead end-to-end solution design and small-team delivery across backend services, user interfaces, enterprise integrations and release validation, using stakeholder feedback to prioritize production improvements.
- AI Recruitment Automation:** Led an Outlook add-in using Office.js, FastAPI, Quasar, PostgreSQL, Gemini, PDF extraction and PaddleOCR. Generated recruiter-ready resume summaries, extracted candidate data into a human-validated form, read live ERP master data, and registered approved records and files through ERP authentication and role controls. HR feedback reported a reduction from about **20 minutes to no more than 45 seconds** per candidate.
- Production RAG Assistants:** Delivered two production assistants for application user guidance and SOP queries, covering documentation, ingestion, embeddings, retrieval, reranking, grounded-answer evaluation and custom token-usage tracking. Also built an embeddable JavaScript widget that frontend teams can mount with a script tag.
- Enterprise Data Analyst Agent:** Built a production MSSQL agent with Agno ADK over governed database views and a business knowledge corpus covering terminology, schemas, column purpose and logic. The agent generates SQL, validates query safety in code, executes validated queries, analyzes results and answers users from retrieved data.
- ERP Expense Automation:** Led a Gemini and custom YOLO workflow that extracts receipt data and creates structured ERP records. Applied model quantization for CPU inference, helped reduce manual expense-approval time by **approximately 60%**, and contributed as AI team lead to the project's SOC 1 certification.
- Statement-of-Work Automation:** Built an analysis and ERP data-entry solution using Python, FastAPI, PostgreSQL, LangGraph and Gemini. Added validation and tool-driven actions, reducing manual review and data-entry effort by **approximately 50%**, while leading a small team across solution design and full-stack delivery.
- SDLC Automation:** Designed a human-approved agent that converts feature requests into requirements, technical plans and executable work orders, with codebase analysis, timesheet support and release-note generation; reduced manual SDLC effort by **approximately 85%**.

- Built Python- and SQL-based ETL, data-processing and reporting solutions integrating QuickBooks, Oracle NetSuite, Maconomy, SFTP systems and external APIs.
- Cleaned, transformed and validated enterprise data, and automated recurring processing and reporting workflows using SQL, Power BI and SSRS.
- Contributed to OCR-based invoice processing and Microsoft Teams automation for operational workflows.

KEY MLOPS & TECHNICAL HIGHLIGHTS

Open RAG MCP — Python, FastAPI, MCP, PostgreSQL/pgvector, ChromaDB, Quasar

GitHub | Project Page | Documentation | Live Demo

- Built and deployed an open-source private-knowledge platform with user-isolated workspaces, document groups, ingestion, OCR, chunking, embeddings, semantic search, reranking, grounded answers, citations and revocable access keys.
- Implemented provider model configuration, isolated document groups, pasted-text and file ingestion, processing-state tracking and a Search Bench for result-count, reranking, diversity, chunk metadata and relevance-score validation.
- Published configurable AI agents with instructions, allowed origins, conversation history and citation controls, supported by a playground for testing streamed answers and inspecting sources.
- Exposed validated group knowledge through an embeddable Web SDK, server-side streaming API and bearer-authenticated MCP tools, with BYOK configuration, one-time secret display and group-scoped retrieval.

Multi-Tenant Training Portal — FastAPI, PostgreSQL, SQLAlchemy, Alembic, Quasar, Vue, TypeScript

GitHub | Project Page | Live Demo

- Built and deployed a tenant-isolated employee-training platform with role-governed workspaces for platform administrators, tenant administrators, HR, trainers and trainees; included external authentication, encrypted integration credentials and private media delivery.
- Implemented the learning cycle from employee onboarding and trainer-to-trainee mapping through versioned subjects, rich lesson content, configurable knowledge checks, HR approval, assignments, due dates, completion tracking and performance reporting.
- Added tenant-controlled AI authoring for outlines, lesson content and tests from objectives and source material, while retaining trainer review and HR approval before learners receive content.
- Separated a Quasar/Vue/TypeScript frontend, role-aware FastAPI services and PostgreSQL state; kept integration secrets server-side, encrypted credentials at rest and served private media through authenticated short-lived access.

EDUCATION

M.Sc. in Mathematics — SRM College, Kudal, University of Mumbai

2021 – 2023

B.Sc. in Mathematics — SRM College, Kudal, University of Mumbai

2019 – 2021