SHREYAS PACHPUTE

Surat, Gujarat | +91 7228903007 | shreyaspachpute1107@gmail.com | Portfolio | LinkedIn | GitHub

SUMMARY

AI Architect with 2+ years of production experience architecting end-to-end GenAI systems, custom LLM deployments, and intelligent automation platforms that drive measurable business outcomes. Expertise in bridging enterprise software (.NET) and cutting-edge AI, optimizing inference pipelines across multi-GPU infrastructure, and leading cross-functional teams. Proven ability to translate business requirements into scalable, production-ready AI solutions—with demonstrated impact on lead generation, operational efficiency, and cost optimization. AWS Machine Learning Specialty Certified.

PROFESSIONAL EXPERIENCE

AI Architect | Commercient LLC *Jan 2024 – Present* | Remote

Role & Responsibilities:

Partner directly with C-level leadership and cross-functional teams to design and deploy production-grade AI systems that solve core business challenges. Own end-to-end architecture, from requirements translation to deployment optimization. Lead technical strategy for AI integration across the platform. Mentor junior engineers.

Key Contributions & Business Impact:

- **Sales Intelligence Automation:** Architected and deployed conversational AI agent on company website, replacing third-party bot (Drift). Directly generated qualified lead capture increase while achieving $12,000 annual cost savings. Engineered HubSpot API integration for seamless lead routing and CRM synchronization.
- **Enterprise AI Platform:** Designed and scaled no-code ingestion and deployment pipeline enabling customers to autonomously deploy domain-specific chatbots from unstructured data sources. System now serves 10+ enterprise clients with zero-touch onboarding, reducing customer setup time by 80% and support burden for internal teams.
- **Support Automation & Intelligence:** Fine-tuned custom LLMs on 10+ years of historical support ticket data and integrated into production support workflows. Achieved 60% reduction in manual triage effort and improved response consistency, enabling support team to focus on high-complexity issues while automating routine resolutions.
- **SQL & Integration Automation Engine:** Developed intelligent SQL view generator leveraging LLM pattern recognition to automatically template ERP-to-CRM data connectors. Eliminated 80% of manual integration coding previously required for onboarding, dramatically accelerating customer time-to-value and reducing engineering support overhead.
- **Natural Language Integration Builder (LangGraph):** Engineered conversational sync agent allowing non-technical users to set up complex ERP-to-CRM integrations using natural language. Reduced engineering support requirements by 70% for standard connector configurations, enabling customers to self-serve and decreasing deployment cycle time.
- **Production LLM Inference Optimization:** Engineered custom multi-GPU LLM inference infrastructure using vLLM with advanced batching, memory optimization, and parameter tuning. Achieved 150+ tokens/second throughput under concurrent load across distributed GPU servers, enabling real-time AI capabilities at scale with cost-efficient resource utilization.
- **Team Leadership & Development:** Mentored junior engineers in production ML systems, GenAI architectures, and cloud deployment best practices. Designed and delivered internal training programs on LLM deployment, RAG systems, and multi-agent orchestration—directly improving team capability and reducing onboarding time for new hires.

Technologies & Stack:

Python, C# / .NET Core, Langchain, LangGraph, vLLM, HuggingFace Transformers, RAG Architectures, Vector Database (Pinecone), MySQL, Multi-GPU Orchestration

CORE COMPETENCIES

AI & Machine Learning:

- Large Language Models (LLMs)
- Model Inference Optimization
- Retrieval-Augmented Generation (RAG)
- vLLM
- Fine-Tuning
- Generative AI
- AI Agent Orchestration
- Multi-Agent Systems
- LangGraph
- Prompt Engineering
- LangChain
- CrewAI

Software Engineering:

- Microservices Architecture
- API Design & Integration
- Python
- FastAPI
- .NET Core
- C#
- ASP.NET
- System Design

Cloud & Infrastructure:

- AWS
- Multi-GPU Deployment & Orchestration
- Docker
- MLOps
- Kubernetes
- CI/CD Pipelines

Data Engineering & Databases:

- PostgreSQL
- MySQL
- Vector Databases
- Data Pipelines
- SQL Optimization
- ETL
- Data Warehousing
- Analytics

Specializations:

- Enterprise Software Integration
- ERP/CRM Systems
- Production LLM Deployment
- Model Optimization
- Cross-Platform Development (.NET + Python)
- Team Leadership & Mentoring

CERTIFICATIONS & ACHIEVEMENTS

- **AWS Certified Machine Learning – Specialty (ML-C01) —** (*View Credential*)
- **Amazon ML Summer School 2023** (*Certificate*) — Selected participant; advanced ML concepts training directly from Amazon scientists.

EDUCATION

Bachelor of Engineering - Computer Science

Sarvajanik College of Engineering & Technology, Surat, Gujarat

Graduation: 2020 – 2024 | CGPA 8.3/10

CAREER FOCUS & INTERESTS

Passionate about architecting intelligent automation systems that solve real-world business problems. Focused on bridging the gap between enterprise software and modern GenAI—helping organizations leverage AI to accelerate growth, reduce costs, and empower teams. Continuously exploring advanced LLM architectures, multi-agent orchestration patterns, and production ML optimization strategies. Seeking remote-first roles with forward-thinking companies building the next generation of AI-native products