

PUNEET SHIVHARE

AI Engineer · Multi-Agent Systems · LLM Infrastructure

+91-6394294541 | puneetshivhare85@gmail.com | linkedin.com/in/puneet-shivhare | github.com/puneet-shivhare

SUMMARY

AI Engineer with production experience architecting enterprise-scale autonomous agentic systems. Specialized in multi-agent orchestration, RAG pipelines, and LLM cost optimization. Reduced **human query resolution time by 80%** through agentic self-improvement and feedback loops, and achieved **60% LLM token cost reduction** in production.

EXPERIENCE

Bizacuity Solutions

AI Engineer

Jul 2025 – Present

Hyderabad

- Architected enterprise-grade **RAG** Engine with **AWS Bedrock** + **LangGraph**, integrating tools and third-party platforms like **Google Workspace** and **GitHub** for creating, updating, and managing issues with RBAC
- Built an agentic self-improving system on top of the RAG infrastructure that autonomously resolves user queries and surfaces relevant knowledge articles based on ticket context, reducing incoming support ticket volume and query resolution time by **80%**
- Designed multiple specialised agents and an orchestrated agentic system for Query Solver's evaluation and feedback loop that analyzes resolution outcomes and continuously refines its retrieval and generation strategy enabling autonomous improvement in ticket deflection rates by 60%
- Cut LLM inference costs by **60%** via structured A/B benchmarking across **GPT-5**, **Claude Sonnet**, **Haiku**, and **kimi 2.6** selecting optimal model per task by accuracy and cost-per-query matrix
- Built custom observability and tracing layer tracking per-tenant token usage, latency, and tool-call traces enabling real-time cost attribution across multi-tenant agent pipelines
- Designed autonomous sales intelligence system with **CRM** integration using tool-calling and **LangGraph** multi-agent workflows, cutting manual lead research time by **30%**
- Engineered near real-time ETL pipelines using **PySpark**, **Databricks**, and **AWS**; reduced storage overhead by **20%** and improved ingestion latency for downstream LLM applications

TECHNICAL SKILLS

Languages & Frameworks: Python, TypeScript, SQL, PySpark, **FastAPI**, Streamlit

AI & LLM: RAG, Multi-Agent Orchestration, Memory management, Tool Mapping, LangSmith, **LangChain**, **LangGraph**, Harness Engineering, MCP, Claude API, Hugging Face, Hermes, Machine Learning and Deep learning

Cloud & Infra: AWS (EC2, S3, Glue, Bedrock), Azure (blob, Foundry, container), GCP(Google chat api, google flow, Google Oauth, gemini api), Databricks, PostgreSQL, pgvector, Git, GitHub

Observability & Eval: LLM Tracing, A/B Model Benchmarking, Evaluation Loops, Custom Agent Skills, Pydantic Validation

PROJECTS

Autonomous Social Intelligence Agent | *LangGraph · Python · FastAPI · github*

- Built a multi-agent system using **LangGraph** that autonomously manages content creation, scheduling, publishing, and engagement across 7 platforms such as Instagram, Twitter/X, and Reddit without any manual intervention
- Engineered a trend intelligence agent that discovers top-performing reels and viral posts, generating contextual comments and interactions to drive organic reach; integrated a self-improving evaluation loop that scores engagement outcomes and autonomously refines agent prompts and content strategy over time

Agentic CRM | *TypeScript · FastAPI · PostgreSQL · pgvector · Hermes Agent · github*

- Built a Clay-inspired GTM-ready CRM with a fully agentic backbone integrated a containerized **Hermes** agent for autonomous task execution, persistent cross-session memory, and end-to-end lead enrichment workflows that operate continuously without human triggers
- Architected a hybrid memory layer using **PostgreSQL** and **pgvector** enabling the agent to retain and retrieve contextual lead history, interaction logs, and enrichment state; designed a **TypeScript** frontend and **FastAPI** backend supporting autonomous prospecting, lead scoring, and outreach pipelines end-to-end

EDUCATION

National Institute of Technology Karnataka, Surathkal (NITK)

2021 – 2025

B.Tech – Metallurgical & Materials Engineering (MME) with Minor in Economics · CGPA: 7.59