HARSH PUROHIT

+91 8815428116 ◇ Pune, Maharashtra
er.harshpurohit@gmail.com ◇ linkedin.com/in/harshpurohit13

SUMMARY

AI-Backend Engineer with hands-on experience designing and delivering production-grade Generative AI and Agentic AI systems across healthcare and enterprise domains. Currently serving notice period and holding an offer. Specialized in LangChain/LangGraph-based multi-agent workflows, FastAPI backend services, and end-to-end AI feature ownership. Led delivery of systems with 90+ branching decision paths, low-latency APIs, and real-world client usage, while working closely with product, frontend, and clinical stakeholders

CORE SKILLS

Programming & Backend Engineering: Python, FastAPI, Git, API versioning, Environment-based configuration management

Generative AI & Agentic Systems: LangChain, LangGraph, Prompt Engineering, Multi-agent orchestration, Decision-tree driven workflows, Retrieval-Augmented Generation (RAG), LLM APIs (OpenAI or equivalent)

Cloud & Deployment: AWS Bedrock, Amazon ECS, AWS Fargate, Amazon RDS, Amazon S3, Docker, containerized API deployments, HIPAA-based deployments

Machine Learning: Classical Machine Learning (scikit-learn), Deep Learning (CNNs), NLP fundamentals

Leadership & Delivery: Project-level technical leadership, client-facing requirement discussions, task planning, code reviews, mentoring, cross-team collaboration

PROFESSIONAL EXPERIENCE

AI Backend Engineer (Project Lead) July 2024 – Present
64 Squares LLP, Pune, Maharashtra

- Owned architecture, development, and delivery of multiple production Generative AI workflows across healthcare and enterprise use cases, used by **internal teams and client stakeholders**.
- Designed and implemented Agentic AI workflows using LangChain, combining deterministic logic with LLM-driven reasoning to support **90+ branching decision paths**.
- Built scalable backend services using FastAPI, achieving **sub-second to low-second response latency** for most conversational and retrieval workflows.
- Reduced manual intake and review effort for client teams by automating clinical and document analysis workflows using AI-driven systems.
- Led a project team of **4–6 engineers**, handling task planning, technical reviews, and delivery coordination across parallel workstreams.
- Acted as the primary technical interface for clients, translating business and clinical requirements into scalable AI system designs and iterative releases.
- Implemented vector-based retrieval pipelines handling **hundreds of documents per deployment** with improved response relevance and contextual accuracy.
- Ensured domain-safe AI responses and guardrails, particularly for healthcare-related workflows.

TECHNICAL PROJECTS

Virtual Pain Clinic (VPC) – Agentic Healthcare AI Platform
Role: Technical Lead / AI Engineer

- Led the design and implementation of a healthcare-focused Agentic AI platform simulating structured clinical intake and assessment workflows.
- Architected a multi-agent system comprising:
  - **Screening Agent** for patient eligibility validation using structured clinical logic.
  - **Intake Agent** implementing complex branching flows across ∼9 clinical modules and ∼90–100 adaptive questions.
- Designed dynamic decision logic to adapt question paths based on user responses, closely mirroring real-world clinical triage processes.
- Integrated a 3D pain mapping flow, enabling anatomical pain selection and persistence of spatial metadata for downstream reasoning.
- Implemented clinical-safe rejection and referral flows to ensure appropriate physician guidance when screening criteria were not met.
- Integrated LLM workflows using **AWS Bedrock** as the primary model interface for agent reasoning and RAG pipelines.
- Used **Amazon RDS** for structured application data and workflow state management across AI-driven intake and document systems.
- Designed structured AI outputs mapping responses to lay diagnosis summaries and ICD-10 descriptions without exposing raw medical codes.
- Delivered conversational and assessment workflows with **sub-second to low-second response latency** under normal usage.
- Supported **dozens of concurrent user sessions** during internal and client-side testing.
- Reduced manual intake effort for clinical teams by **standardizing and automating assessment flows**.
- Owned backend implementation end-to-end and coordinated delivery with frontend, product, and clinical stakeholders.

Tech Stack: Python, FastAPI, LangChain, LLM APIs, JSON-based decision trees, Vector Search

ViziSmart – Voice-Enabled AI Assistant
Role: AI Engineer

- Contributed to a voice-driven AI assistant for intelligent property management use cases.
- Integrated LLM-powered conversational logic with backend services to support natural language interactions.
- Implemented intent handling and response structuring for voice-based workflows.

Tech Stack: Python, LLM APIs, Backend Services

EDUCATION

Bachelor of Technology, Chhatrapati Shivaji Institute of Technology 2020 – 2024
Relevant Coursework: Computer Science

High School (12th Grade), Vishwadeep Senior Secondary School 2020
Board: CBSE