D Bippin Shekar Goud

9100863981 | bippinshekar@gmail.com | LinkedIn | GitHub

Experience

Founder & Lead Engineer Dec 2025 – Present

Leaps *Early-stage*

- Building an AI-native career execution platform that scores candidate fit, identifies skill gaps, and autonomously applies to high-match roles, replacing volume-based job search with fit-driven guided execution.
- Engineered autonomous application systems for LinkedIn (multi-step, turn-based, dynamically structured per company) and Ashby (single-form with company-specific open-ended fields requiring contextual candidate reasoning), two fundamentally different automation problems resolved under one architecture.
- Validated end-to-end automation in closed alpha across 400+ job applications on LinkedIn and Ashby, with waitlist-based access in place ahead of a funded beta rollout.

Independent AI Consultant Nov 2025 – Present

The AI Auditor

- Auditing and rebuilding AI systems for early-stage startups, identifying root causes of failure in production LLM pipelines, retrieval systems, and agentic architectures.
- *Pulsegen.io:* Rebuilt a stateless analytics bot into a full agentic TypeScript system over unstructured MongoDB, introducing Text-to-Mongo query generation and contextual memory, cutting latency 40% (80s to 40-50s).
- *Pulsegen.io:* Scaled an LLM-driven VoC engine processing 100K+ inputs/day, increasing throughput 33x and turning raw customer feedback into decision-ready product insights.

Founding AI Engineer Sept 2024 – Nov 2025

8bit.ai

- Built an enterprise Text-to-SQL engine from scratch across PostgreSQL, MySQL, and a data lake (Iceberg) spanning 250+ tables across multiple schemas, including a metadata jargon mapping layer that translated business terminology (e.g. "cumulative ARR by department") into actual column structures without users needing to know data team naming conventions.
- Fine-tuned Llama 70B (QLoRA) and OpenAI's fine-tunable model on 60,000+ curated business SQL pairs, raising SQL generation from 80% to 98% and execution success from 55% to 85%+, with pre/post guardrail and SQL correction layers forming a full eval pipeline for production reliability.
- Led development of a generative AI insights platform with interactive dashboards enabling executives to instantly probe trends and explore key metrics, driving a $500K+ lead pipeline and improving customer retention.
- Designed multi-modal orchestration routing across T2SQL, RAG, workflow generation, and web research, with Redis-backed memory retrieval replacing full history injection, extending conversation depth from 4 to 20+ turns before context exhaustion.
- Migrated orchestration from LangChain to LangGraph for HITL flows, then eliminated all framework dependencies in favor of direct SDK integrations with custom provider abstractions, removing black-box behavior entirely.

Projects

Askr | *Open Source* | GitHub

- Built a session orchestration daemon for Claude Code that monitors context and quota thresholds, auto-checkpoints via git, and keeps co-founders in sync across sessions, eliminating manual recovery and team drift during active development.

Technical Skills

Languages: Python, Go, TypeScript, SQL, Bash
Backend & Infrastructure: FastAPI, Redis, Docker, Kubernetes, REST APIs, GraphQL
Frontend: Next.js, React
AI / ML: LLMs, OpenAI, Anthropic, Gemini, Llama, HuggingFace
LLM Systems: RAG, Fine-tuning (QLoRA, LoRA), Prompt Engineering, Evaluation, Guardrails, HITL Architectures
Vector & Retrieval: FAISS, Pinecone, Weaviate, Hybrid Search, Embeddings
Agents & Orchestration: Multi-Agent Systems, LangGraph, Tool Calling
Databases: PostgreSQL, MySQL, MongoDB, Redis, Data Lakes (Iceberg)