Pavankumar Kushnure

+91 8739838392 | pavankumarkushnure@gmail.com | pavankushnure.website | GitHub | LinkedIn

Summary

Full-stack AI engineer (3+ yrs) shipping production systems with React, Node.js, Python/FastAPI, and LLMs on AWS/GCP. Built LLM gateways, guardrails, agentic workflows, and document extraction pipelines,sub 100ms hybrid search, enterprise CAD integrations, and Stripe billing owned end-to-end from design to deployment.

Technical Skills

GENAI & LLM: LLM API integration (Anthropic, OpenAI, Gemini, Groq) , RAG (Retrieval-Augmented Generation), agentic workflows, MCP (Model Context Protocol), Embeddings, Hybrid Search, Reranking, prompt engineering, LLM guardrails & evaluation, LangChain, LlamaIndex, LangGraph
Frontend: React.js, Next.js, Redux, Recoil, Micro-frontends
Backend: Node.js, Express.js, FastAPI, REST APIs, WebSockets, Webhooks, Pub/Sub Messaging
Databases: MongoDB, PostgreSQL, Algolia (Elasticsearch), Vector Databases (Pinecone, Weaviate,Qdrant)
Cloud & DevOps: AWS, GCP, Docker, CI/CD, Terraform, Nginx, Kubernetes
Tools & Services: Git, Postman, Swagger, Apache JMeter, Zapier, Stripe API, PM2

Experience

Bravent LLC Virginia, USA (Remote)
Lead AI Engineer *(Python, FastAPI, TypeScript, LLMs, MCP, Docker, GitHub Actions)* *Mar. 2026 - Present*

- Built an LLM API gateway unifying multiple model providers behind one policy-enforced endpoint, with multi-provider routing, per-app API keys, wallet-based billing, and token/cost governance (budget alerts, prompt-cache enforcement, context-budget truncation).
- Engineered a runtime guardrails engine enforcing 160+ security rails — PII detection & masking, DLP, prompt-injection and token-smuggling blocks, SSRF/egress allow-listing, and rate-limiting — plus a document PII pre-flight that extracts, scans, and masks uploads before they reach a model.
- Led development of a governed agentic-workflow platform: a visual agent builder, MCP-based tool integrations, and a governed runtime with subagent recursion under shared budget/depth caps, human-in-the-loop approval, and inbound triggers (webhooks, Slack, email, event polling).
- Built an LLM evaluation and observability stack (LLM-judge, exact-match, and trajectory scorers with golden-set regression gates, OpenTelemetry tracing, SIEM event export, and real-time alerting) and shipped it to production on cloud infrastructure via Dockerized GitHub Actions CI/CD.

Naya Studio New York, USA (Remote)
Software Engineer *(Reactjs, Nodejs, Expressjs, MongoDB, GCP)* *Oct. 2023 - May 2026*

- Designed and built a universal search system indexing 250K+ files (images, PDFs, videos, 3D models, links), delivering 50ms keyword search via Algolia and 150–200ms semantic retrieval via MongoDB Atlas Vector Search.
- Architected and implemented Stripe billing pipelines with automated webhooks, tiered subscription plans, and team-seat management, eliminating ∼20 hours/month of manual billing overhead.
- Integrated Onshape and Autodesk CAD rendering via OAuth 2.0, webhooks, and cloud translation APIs, enabling real-time 3D model sync and unlocking enterprise client adoption.
- Developed and shipped an Onshape App Store extension enabling direct project import/export between Onshape and Naya, reducing context-switching and increasing design-team adoption by 40%.
- Built a multi-select interaction system for batch editing and AI-powered asset generation (e.g., image/text/3d model generation via LLM APIs), reducing multi-asset workflow time by 50%.
- Re-architected the project dashboard into a drag-and-drop, card-based workspace with nested folder uploads, reducing manual asset setup time by 70%.

Kratin LLC Nagpur,India
Intern Technologist *(ASP .Net Core, C#, SQL Server,Internet Information Services (IIS))* *Jan 2023 - Aug 2023*

- Built and scaled a task management system handling 250K+ tasks, 3K users, and 100GB of files, which improved tracking and real-time collaboration.
- Optimized the top 10 high-traffic APIs by improving SQL queries and minimizing database calls, reducing average response time from 1.7s to 1.2s.

Projects

MCPOne | Live | *TypeScript, Next.js, PostgreSQL, Drizzle, gRPC, MCP, LLM APIs (Gemini/Anthropic)*

- Built a policy-enforced LLM governance runtime for regulated federal (FedRAMP/IL4/IL5) workloads over REST, gRPC, and MCP, screening every request through inline PII redaction and prompt-injection and content-safety scans before it reaches a model, with an append-only hash-chained audit ledger in PostgreSQL.
- Designed a guardrail enforcement engine mapping a 1,248-rule catalog to real runtime mechanisms (36 gateway gates, 42 per-request runtime gates, 192 infrastructure attestations), plus an autonomous cron agent that adversarially re-tests live production guardrails hourly across six harm categories.
- Shipped the multi-tenant Next.js/React SaaS with a 5-tier LLM router (BYOK → credits → local Gemma via Ollama → managed Gemini → canned fallback) streaming over SSE, next-auth, and Stripe metered billing; deployed via GitHub Actions CI (build → typecheck → test → Trivy scan) with Drizzle migrations and PM2 reloads.

ProposalIQ | Live | *TypeScript, Python, Qdrant, Anthropic/OpenAI/Gemini, Redis, Docker, OCI*

- Engineered a production RAG pipeline over Qdrant that fuses dense embeddings (text-embedding-3-large, 3072-dim) with BM25 sparse retrieval via Reciprocal Rank Fusion (k=60); chose hybrid over dense-only after semantic search kept missing the exact identifiers (clause numbers, CLINs) that federal solicitations turn on.
- Orchestrated 3 LLM providers (Claude, GPT-4o, Gemini) behind a unified layer with per-model token and cost accounting, and integrated agentic opportunity sourcing from SAM.gov via MCP tools driven by a Claude tool-use loop, alongside the HigherGov REST feed.
- Built an LLM evaluation subsystem for A/B model judgments and weighted rubric scoring (writing quality, completeness, accuracy), surfacing win-rate and tie analytics to compare models and prompts on measured results instead of intuition.
- Shipped the multi-tenant platform end to end: React SPA, Express API (162 endpoints, 33 data models), and a Python retrieval service, with SSE-streamed AI drafting, real-time collaborative DOCX editing over Socket.IO, 442 automated tests, and CI/CD to Oracle Cloud (GitHub Actions → pm2 → Nginx, Dockerized OnlyOffice).

Doc Extractor | Live Demo | Code | *Python, FastAPI, Claude API, Mistral OCR, PyMuPDF, Pydantic*

- Built an AI document-extraction engine that turns messy invoice PDFs into clean, validated JSON, attaching a per-field confidence score and source citation to every field and auto-flagging low-confidence fields for human review.
- Engineered hybrid parse-then-LLM routing (PyMuPDF for born-digital PDFs, Mistral OCR for scans) into Claude structured outputs, with a multi-signal per-field confidence engine, model-cascade escalation on low confidence, and schema self-heal retries.
- Validated extraction on a 50-invoice synthetic eval set with per-field precision/recall/F1 and a golden-set CI regression test.
- Shipped a FastAPI async-job backend and a drag-and-drop, two-pane reviewer UI (PDF viewer + editable field ledger with click-to-cite), deployed on Railway and Vercel.

FeatureGate | Code | *Node.js, Express.js, React, MongoDB, Redis, Docker, Terraform, GCP*

- Built an open-source, self-hosted feature flag platform supporting segment-based targeting, a rule evaluation engine with consistent hashing for percentage rollouts, and real-time flag propagation via SSE in under 500ms.
- Designed and published a Node.js SDK on npm with in-memory flag caching and local evaluation (sub-1ms per check), eliminating per-request network calls and providing a LaunchDarkly-compatible developer experience.
- Containerized with multi-stage Docker builds and deployed to GCP Cloud Run via a 4-stage GitHub Actions CI/CD pipeline (lint → test → build → deploy) with Terraform-provisioned infrastructure including Memorystore (Redis) and Cloud Run auto-scaling.

Education

Shri Guru Gobind Singhji Institute of Engineering and Technology,Nanded — June 2023
*Bachelor of Technology in **Computer Science and Engineering*** — *CGPA 9.02*

Jawahar Navodaya Vidyalaya Wardha — Sep. 2011 – May 2018
Class XII(PCM) & Class X — *94% & CGPA 10*