PRAVEEN KUMAR

✉ inboxpraveen.17@gmail.com

Bangalore, India 📍

About

Principal AI Engineer specializing in scalable, production-grade LLM and speech intelligence systems. Turning complex AI into enterprise-ready solutions.

Objective

To lead the creation of scalable, high-impact AI systems that transform business operations through intelligent automation and real-world innovation. As a Principal AI Engineer, I aim to bridge strategy and engineering - driving technical excellence, mentoring teams, and turning complex AI concepts into simple, production-ready solutions.

Skills

Core Expertise

- End-to-end AI system architecture and deployment - spanning speech analytics, LLM-driven automation, RAG pipelines, OCR/IDP workflows, and multi-tenant SaaS systems.
- Specialized in production-grade AI engineering - transforming research-grade models into secure, scalable, and high-performance enterprise solutions.

Programming & Frameworks

- Languages: Python, SQL, Bash, JavaScript (React)
- Frameworks: FastAPI, Flask, PyTorch, Hugging Face Transformers, LangChain, SentenceTransformers, OpenAI SDK, Unsloth
- LLM Hosting: vLLM, Ollama, llama.cpp, Triton Inference Server
- Speech & Audio: FasterWhisper, WhisperX, pyannote.audio, TTS (Kokoro, Bark, XTTS)

Data & Model Engineering

- Expertise in LLM fine-tuning (LoRA/QLoRA), dataset generation, evaluation metrics, and prompt optimization.
- Experience with vector databases (Qdrant, FAISS, Chroma, PostgreSQL-Vector) and knowledge retrieval systems.
- OCR and layout-aware document parsing using PDFPlumber, Surya OCR, and PaddleOCR.

Infrastructure & DevOps

- Containerization: Docker, Docker Compose, NVIDIA Container Toolkit, MIG partitioning.
- MLOps: CI/CD (GitHub Actions, Jenkins), environment versioning (UV, Conda, Poetry).
- Orchestration & Scaling: Redis Queues, Celery, Load Balancing, Nginx, Multi-service Docker Networks.
- Deployment: AWS (EC2, S3, ECR), Azure, and On-prem GPU clusters (T4, A10, H100 MIG).

System Architecture & Backend

- Design of microservice-based AI backends with modular APIs, async processing, and event-driven pipelines.
- Database design with PostgreSQL, Redis, and TimescaleDB for scalable AI and analytics workloads.
- Authentication & Security: Vault, JWT, 2FA, role-based access control, secure secrets management.

Monitoring, Logging & Observability

- Fluent Bit, Loki, OpenSearch, Grafana for centralized log aggregation and performance dashboards.
- Experience building real-time dashboards for AI service health, latency metrics, and inference utilization.

Leadership & Delivery

- Proven track record in technical leadership, mentorship, and cross-functional collaboration.
- Successfully delivered 3 production AI products.
- Focused on aligning AI outcomes with business KPIs, compliance, and operational excellence.

Experience

TrellisSoft, Inc. — *Bengaluru, India*

Principal AI Engineer | *Jul 2025 – Present*

- Spearheading architecture and delivery of scalable AI ecosystems across speech, LLM, and RAG domains.
- Driving enterprise AI adoption through secure, high-throughput inference platforms (vLLM, Ollama, NVIDIA MIG).
- Leading multi-service orchestration, GPU optimization, and CI/CD pipelines for production-ready AI workloads.
- Partnering with leadership to align AI strategy with business outcomes and innovation roadmaps.

Senior Artificial Intelligence Engineer | *Jun 2023 – Jun 2025*

- Led design and deployment of multilingual speech intelligence, real-time agent assist, and QA automation platforms.
- Architected RAG-based knowledge systems and layout-aware OCR pipelines for insurance and finance automation.
- Fine-tuned and served domain-specific LLMs (Qwen, SmolLM, OLMo) via vLLM for high-performance inference.
- Mentored engineering teams, standardized AI design patterns, and earned multiple internal awards for innovation and mentorship.

Artificial Intelligence Engineer | *Aug 2020 – Jun 2023*

- Built NLP, ASR, and CV models including chatbots, sentiment analysis, and document recognition systems.
- Designed semi-supervised data pipelines and automated retraining workflows for large-scale deployments.
- Recognized as “Brainiac of the Quarter (2021)” and “Employee of the Quarter (2022/2023)” for technical excellence and delivery impact.

Projects

Nexus360: Real-Time AI Voice Agent

Role: Principal Architect & AI Lead | Tech: vLLM, FasterWhisper, XTTS, LangChain, React, FastAPI, Redis, PostgreSQL, Docker, NVIDIA MIG

- Designed and deployed a real-time AI voice agent capable of conducting human-like outbound sales and support conversations with full compliance and dynamic call flow logic.
- Integrated speech-to-text (FasterWhisper), LLM orchestration (vLLM + LangChain), and TTS (XTTS/Kokoro) for low-latency voice interactions.
- Developed persona-driven dialogue engine supporting 150+ customer types and multilingual conversations.
- Implemented real-time sentiment tracking, compliance tagging, and AI-guided recommendations for agents.
- Scaled to serve 50+ concurrent real-time calls per node using Redis streaming and GPU MIG partitioning.
- Delivered enterprise-grade dashboards and analytics via the *NexusAssist* interface for agent performance monitoring and insights.

LossLens: AI-Powered Insurance Document Intelligence Platform

Role: AI System Architect | Tech: PDFPlumber, Surya OCR, PaddleOCR, Qwen 2.5 (LoRA), vLLM, FastAPI, React, PostgreSQL, CI/CD, AWS EC2

- Built a multi-tenant AI document processing suite automating loss-run and claims extraction for insurance carriers like Travelers, KeyRisk, and CCMSI.
- Engineered layout-aware OCR + LLM pipelines for field-level data extraction, validation, and correction workflows.
- Designed RAG-based QA evaluation and summarization layer for claim verification and compliance audits.
- Implemented custom field-mapping UI allowing tenant-specific schema control while maintaining centralized model IP.
- Achieved 98% field accuracy and 60% faster processing through optimized model orchestration and GPU-based parallelism.
- Integrated secure access control, audit logs, and 2FA for enterprise-grade data handling and compliance.

PulseAI360: AI Call Quality & Performance Intelligence Platform

Role: Principal AI Engineer | Tech: FasterWhisper, WhisperX, pyannote.audio, vLLM, PostgreSQL, Docker, Grafana, Redis, Flask, H100 MIG

- Architected an end-to-end call ingestion, transcription, and QA analytics system processing 60+ recordings per minute.
- Integrated speaker diarization, sentiment/emotion detection, and LLM-based coaching insights for QA and compliance.
- Delivered timestamped, speaker-labeled transcripts with evaluation metrics and real-time dashboards for supervisors.
- Deployed on GPU-accelerated infrastructure with Redis streaming, CI/CD automation, and centralized monitoring via Loki & Grafana.
- Enabled business leaders to view call trends, agent performance, and compliance insights through a unified analytics layer.

Education

MVJ College of Engineering

Aug 2016 – July 2020

- Specialization in “Computer Science and Engineering”.
- Attained an 8.0 GPA on a scale of 10 GPA.

Achievements

- GitHub Sponsored by Recall.AI for Open Source Contributions
- “Excellence Appreciation” award in July 2025 for product contributions and leaderships.
- “Mentorship and Coaching Award for the year” in Jan 2024 for providing impactful mentorship.
- “Excellent Project Deliverables” award in 2023 for Employee Training & OCR System.

Habit/Interest

- Editor and Writer of Analytics Vidya Publication on Medium.
- Playing football & FIFA. Huge Fan of ‘Lionel Andres Messi’.