Prasoon Kumar

Machine Learning Engineer | Generative AI | NLP & Speech | M.S. in Computer Science
San Francisco (Open to relocate) | prasoonkumar.23702@gmail.com | 925-997-8644 | LinkedIn

SUMMARY

Machine Learning Engineer with **4+ years of experience** building production-ready **AI systems across NLP, speech, and vision applications**. Skilled in generative **AI workflows, LLM fine-tuning, RAG systems, and speech processing**. Proven success reducing manual workload, **improving model precision, and accelerating deployment timelines**. Highly collaborative and adaptable to fast-paced startup environments with a bias toward shipping elegant, maintainable code.

EXPERIENCE

ML Engineer — *Sprouts AI*
Oct 2023 – Mar 2024

- Built LLM-powered job description generator using ChatGPT-4 and Llama 2, reducing manual recruiter effort by 60%.
- Developed **RAG systems leveraging Hugging Face models and Elasticsearch** to power real-time document retrieval.
- Integrated NLP services into production via Docker + GCP Cloud Run with **end-to-end CI/CD and observability.**

ML Engineer — *Automation Anywhere*
Mar 2024 – Sep 2024

- Fine-tuned BERT-based text classifiers and built summarization pipelines for enterprise automation use cases.
- Designed **real-time telemetry ingestion using Kafka + Spark** Streaming, processing 2M+ events/day.
- Maintained retraining pipelines and unit/integration testing in CI/CD with GitHub Actions.

MLOps Intern — *Vimaan Robotics Inc*
Feb 2025 – Present

- Built real-time vision inference services **deployed via AWS EKS & Lambda** with **sub-100ms latency.**
- Automated ML deployments with **Docker, Kubernetes, and FastAPI**; monitored performance with Prometheus.
- Collaborated with cross-functional engineers to troubleshoot production edge cases and ensure uptime.

Graduate Research Assistant — *University of Texas at San Antonio*
Aug 2021 – Oct 2023

- Researched domain adaptation and **out-of-distribution detection for robust NLP/vision systems**.
- Developed 3D scene understanding pipeline (YOLOv5, U-Net) and contributed to dataset augmentation workflows.
- Mentored undergraduate researchers and contributed to reproducible model evaluation scripts.

EDUCATION

MS in Computer Science — University of Texas, San Antonio
May 2023 — GPA: 3.8/4.0
B.Tech in Computer Science — IEM, Kolkata, India
May 2021 — GPA: 8.6/10

SKILLS

- **Languages & Tools:** Python, SQL, Bash, Git, Docker, Kubernetes, FastAPI, GitHub Actions
- **ML/NLP:** PyTorch, TensorFlow, Hugging Face, ChatGPT-4, BERT, Llama 2, RAG, speech-to-text
- **Data & Infra:** PostgreSQL, MySQL, Spark, Kafka, Redis, Firestore, MongoDB
- **Deployment**: GCP (Cloud Run, Pub/Sub), AWS (Lambda, S3, SageMaker), CI/CD, Prometheus, Grafana

PROJECTS

SmartRAG — Retrieval-Augmented Generation System
Built a document retrieval system using Elasticsearch and LLMs to power real-time context-aware Q&A for internal HR tools.
Speech Summarizer — Voice-to-Insight Pipeline
Built prototype for meeting summarization by converting audio to text using Whisper and summarizing with OpenAI APIs.

CERTIFICATIONS

- **Generative AI with Large Language Models — DeepLearning.AI & AWS**
- **Prompt Engineering for Developers**
- **MLOps Specialization — DeepLearning.AI**