

Pushkar Raj

Generative AI | ChatBot | RAG

pushkarraj6@gmail.com | 8527589417 | Noida,UttarPradesh | LinkedIn
Github | HackerRank | StackOverflow

PROFILE

Results-driven **Senior GenAI Engineer** with **7+ years** of experience building scalable **LLM, GenAI, and AI/ML solutions**. Proven track record of delivering production-ready AI systems that improve performance, reduce latency, and drive measurable business impact across cloud platforms.

PROFESSIONAL EXPERIENCE

Senior Machine Learning Engineer, *Siemens (Brightly Software)* 09/2025 – Present | Noida

- **Architected a scalable multi-agent RAG system** using **LangGraph, LangChain, Azure OpenAI, and FastAPI**, improving workflow modularity by **40%**.
- **Optimized semantic retrieval** with **OpenSearch Serverless**, metadata filtering, and dynamic **top-k** search, improving response relevance by **25%**.
- **Built a persistent memory framework** using **LangMem, LangGraph, DynamoDB, and Snowflake**, enabling **99.9%** reliable cross-session conversations.
- **Improved API performance** through asynchronous processing, caching, and concurrent execution, reducing latency by **45%** and increasing throughput by **3×**.
- **Enhanced production reliability** with automated **CI/CD**, structured logging, and comprehensive testing, reducing deployment failures by **60%** while maintaining **90%+** core test coverage.

Senior Machine Learning Engineer, *Neuralgo.ai* 03/2024 – 09/2025 | Noida

- Spearheaded the implementation of **advanced GenAI and large language models** to enhance customer interaction platforms.
- Developed solutions using llms for generating **SQL queries** and analyzing survey data to **create insightful** text summaries.
- Designed and deployed scalable AI-driven solutions, improving overall system performance.
- Enabled 80 + hrs./week in time savings through automation and intelligent query handling.
- Focused on secure, real-time, multi-modal data handling aligned with Responsible AI principles with red teaming score of greater than 95%.

Senior Data Scientist, *Ascendum Solutions* 06/2023 – 02/2024 | Gurgaon

- Introduced the use of **Document AI** models in the organization.
- Collaborated with multiple teams in order to protect the **Data Privacy** of the customer.
- Used **open-source AI** models for the development of the projects.

Data Scientist, *Appnox Technologies Pvt. Ltd* 12/2022 – 05/2023 | Noida

- Successfully **lowered the cost of production** by self-hosting models on the cloud.
- Implemented **cost-effective strategies** in order to generate data for model **finetuning and training.**
- Build **scalable** endpoints of different models with **minimum downtime** in production.
- **Researched** through various published **journals** to adapt new technologies in the organisation.
- Utilized **free** and **open-source** models, and finetuned them for various **downstream tasks.**

Project Engineer, *IIT BOMBAY* 06/2017 – 07/2021 | Mumbai

- Developed and deployed various **AI and ML-based** solutions in various domains including Defence, Weather, and Robotics.
- Oversaw the development of machine learning pipelines, ensuring **scalability, efficiency, and robustness**.

PROJECTS

NextGen CPGRAMs

- Architected and developed the **Nextgen CPGRAMS** platform using a **Neo4j Graph Database** and **FastAPI** backend, creating a dynamic Knowledge Graph of over 50,000 interconnected nodes (departments, officials, locations) for precise grievance mapping.
- Implemented an AI-powered routing engine leveraging **Graph RAG** to analyze complaint semantics, **slashing misrouting errors by over 85%** compared to the previous keyword-based system (from ~40% error rate to <6%).
- Successfully processed and routed over **15,000 citizen complaints**, accelerating the average resolution lifecycle **by 60%**—reducing turnaround time from an average of 30 days down to just 5 days.

Chatbot Agent (SA based client), *neuralgo.ai Team Size 4*
- Developed a **GenAI** Text2SQL agent using **Llama-Index** and an **LLM** for real-time queries on **Snowflake**.
- Architected a chatbot agent with **LangChain**, utilizing **RAG**/**Graph RAG** and custom **embedding models**.
- Improved model performance through **fine-tuning**, strategic **tokenization**, and rigorous **model evaluation**.
- Delivered a production-grade solution on **AWS** with robust **guardrails** and integrated **CRUD APIs**.

Text2Sql Engine(U.K based client), *neuralgo.ai Team Size 3*
- Architected a **multi-agent system** where specialized LLM agents collaborate to generate complex SQL queries and deliver insightful data summaries.
- Engineered a high-performance **Graph RAG** by converting the client's relational database into a knowledge graph with clearly defined entities and relationships.
- Implemented multi-level **parallelism** (asyncio/multithreading) for data fetching, reducing query execution latency from databases by over 40%.
- Conducted rigorous system validation, using **ROUGE** scores for summary quality, execution accuracy for SQL, and extensive Human Evaluation to ensure model **faithfulness and relevance.**

Lead Management dashboard(US based client), *neuralgo.ai Team Size 4*
- Created an AI-assisted dashboard to generate automated responses for user queries with **GPT-4o.**
- Utilized **AWS eventBridge** to schedule the response.
- Integrated **Nylas** service to handle Client inbox, outbox, and **Webhooks** for notifications.
- **Created and deployed** APIs to handle various features in the AI-driven inbound **sales agents** for **high-volume** inbound sales businesses.

Digital Examiner(DEX), *Ascendum Solutions. Team Size 5*
- Achieved an **accuracy of 92%** with an **SVM** for the classification of itemized billings.
- Experimented with **Gen AI** models to classify **Itemized bills** from other documents.
- Fintuned models like **Lilt, LayoutLMV3, and Donut** for **document question answering**.
- Inference from **TAPAS and TAPEX-based** models for structured table understanding.

Automation of Large Language Models(LLM), *Appnox Technologies Pvt. Ltd. Team Size 5*
- Built **EC2 GPU clusters** for ultra-fast **LLM** Inference, and finetuning using ray clusters.
- Used auto-scaling to **manage head and worker nodes** for multi-node inferencing of **LLM**.
- Applied **pipeline parallelism** and **sharding** to generate faster inference from the LLMs.
- Deployed open-source **LLM** models like **OPT-175B** and **BLOOM-175B** on **4 x AWS p3.16xlarge** instances with a total of 512GB GPU clusters.

Yamak.ai, *Appnox Technologies Pvt. Ltd.,Team Size 10*
- **Prompt Engineering** GPT-3 API to **generate content** for different use cases of the platform.
- Data Mining and **Web Scraping** to finetune Causal Language Models like **GPT-2, OPT, and BLOOM** for Down-Stream Tasks.
- Built a Paraphrasing model after **finetuning** OPT and Bloom using the PyTorch framework.
- Built Causal Languages Models like Bloom, GPT-2, GPT-3, and OPT for **downstream tasks** like Paraphrasing, Summarization, and Question Answering.
- Used model **sharding** and **device mapping** for multi-GPU and multi-Node inference.
- Deployed models in **AWS** and wrote a custom python script to **CloudWatch** to monitor **GPU usage.**

SKILLS AND TOOLS

Programming Language: Python, C, SQL.
Framework: Pytorch, FastAPI, Langchain, llama-index,LangGraph
Library: Scikit-Learn, Numpy, Pandas, Selenium, Flask, Pyspark, Scipy, NLTK, Torch, Hugging Face, Ray, Alpa.
Visualization: Matplotlib, Seaborn, Wandb.
Algorithms: Machine learning, Deep Learning, Computer Vision.
Cloud platforms: AWS, Heroku, Azure.
Operating system: Linux, Windows.
Version Control: Git, GitHub, DVC, Bitbucket.
Database: MongoDB, My SQL, Postgres SQL, Snowflake
MLOPS: Docker, GitHub Actions, Circleci, Kubernetes, Flask API, MLflow.

EDUCATION

Bachelor of Technology, *A. P. J. Abdul Kalam Technical University* 2017 | Ghaziabad, India
Electronics and Communication Engineering

LANGUAGES

English Hindi