Jithin George

AI/ML Engineer | Backend Developer | Generative AI Specialist

www.linkedin.com/in/jithu010 | +919633473792| jithin.george.w@gmail.com

SUMMARY

Backend Developer with 2+ years of experience building AI-first backend systems using Python, FastAPI, and modern vector databases. Proven expertise in deploying GPT-powered RAG pipelines, scalable API services, and generative AI tools using LangChain, LlamaIndex, and cloud-native infrastructure. MSc (AI) graduate with solid ML/DL foundations and real-world enterprise deployment experience.

SKILLS & INTERESTS

Technical Skills:	Python, JavaScript (Node.js), FastAPI, LangChain, LlamaIndex, REST APIs, Microservices, SQL, MongoDB, Weaviate, Qdrant, Elasticsearch, Neo4j, Docker, Git, AWS, Azure, MLOps, CI/CD, Vector Databases, Redis, Semantic Search, Scalability Optimization.
AI/ML Expertise:	GPT-3.5, GPT-4, Prompt Engineering, ReAct, Chain-of-Thought, Retrieval-Augmented Generation (RAG), Embedding-Based Search, Machine Learning, Deep Learning, scikit-learn, TensorFlow, PyTorch, Model Monitoring, Model Evaluation, AIOps Pipelines, Ethical AI Practices, Proficient with NLP tools (spaCy, Hugging Face), recommendation engines, predictive models, and vector DBs (Pinecone, Qdrant).

WORK EXPERIENCE

Gapblue Software Labs **Ernakulam,Kerala**

Associate Consultant, *Mar 2024 - Present*

- Developed and deployed a production-grade Retrieval-Augmented Generation (RAG) system supporting 10,000+ queries per day, leveraging LangChain, LlamaIndex, and Azure OpenAI, which improved semantic search speed by 40% and enhanced retrieval accuracy by 35%.
- Designed and integrated vector databases (Qdrant, Weaviate, Pinecone) across enterprise systems, reducing data lookup latency by 3× and cutting average response time to under 300 ms.
- Led the architecture of Knowledge Explorer, a multilingual, glossary-aware GPT-4 solution integrating Azure OpenAI, Azure Translator, and Language Services, which improved internal information access by 60% and enhanced multilingual support across 5+ domains.
- Spearheaded RAG-based automation pipelines across four departments, cutting manual document processing time by 50% and contributing to an estimated annual savings of \$200,000.
- Developed an innovative RAG-based Q&A system that successfully secured a pivotal client contract valued at $500K - contributing to an impressive 22% increase in overall company revenue within six months.
- Optimized deployment processes for GPT-3.5, GPT-4, and GPT-4o models by enhancing inference time efficiency by 35%, which led to a significant reduction of API costs by 20%, saving the company approximately $250K annually.
- Delivered and maintained 3+ scalable AI applications in Agile teams, with 100% on-time deployment success and 30% client satisfaction growth over previous benchmarks.

Trainee Associate Applications Engineer, *Sep 2023 – Feb 2024*

- Built foundational expertise in backend development, Linux systems, and AI toolchains by contributing to full-stack solutions across Node.js, FastAPI, and React.
- Collaborated on LLM-powered backend modules, reducing GPT inference latency by 25% and enabling GPT-4 integration for internal RAG apps; also contributed to an external client-facing bot for patent, research, and innovation management, enabling faster semantic access to technical documentation.
- Assisted in front-end and back-end development, reducing bug resolution time by 40% before transitioning into Generative AI and NLP.

EDUCATION

MSc Computer Science (AI), CUSAT, Kochi **2021–2023**

Project: Forecasted AQI & heatwaves using Autoencoder–ARIMA + BiLSTM/DNN hybrid model.

BCA, Rajagiri College, Kochi **2018–2021**

Project: E-commerce web app with user auth & order management (HTML, JS, PHP, SQL).

ADDITIONAL INFORMATION

Certifications: Oracle AI Vector Search Professional, OCI AI Foundations Associate (2023).