Jithendra Jagannatha Kagathi

+1 (612) 559 1897 | jagan028@umn.edu | Minneapolis, MN | LinkedIn | GitHub

NLP Engineer | Data Scientist | ML Engineer | Computational Linguist

Accomplished Data Science Master's student specializing in multimodal Generative AI and Large Language Models with production-ready implementation experience. Proven expertise in enterprise-grade ML solutions across healthcare diagnostics and conversational AI, delivering measurable business outcomes through RAG systems, model fine-tuning, and evaluation frameworks.

Education

University of Minnesota Twin Cities *September 2023- May 2025*
Master of Science, Data Science *3.6/4 CGPA*
Relevant Coursework: Natural Language Processing (NLP), Intro to Linguistics, Advanced Machine Learning, Principles of Data Base Systems, Multivariate Statistics, Time Series Analysis, Intro to Parallel Computing, AI for Health Care, Intro to Virtual Reality

JSS Science & Technology University *August 2017- June 2021*
Bachelor of Engineering, Electronics & Communication Engineering *8.86/10 CPGA*
Relevant Coursework: Artificial Intelligence, Digital Signal Processing, Intro to C, Data Structures and Algorithms, Linear Algebra, Advanced & Multivariable Calculus, Probability & Stochastic Processes, Fourier Series Integration Transforms & Applications

Experience

University of Minnesota *November 2023- Present*
Graduate Research Assistant, Dept. Of Health Informatics *Minneapolis, MN*

- Veterinary LLM: Developed an enterprise-grade diagnostic tool using **Llama 2** and **GPT-4**, achieving 89% accuracy and reducing diagnosis time by 5x. Processed 10,000+ veterinary cases using advanced **Prompt Engineering** and **RAG** with thorough empirical analysis and production-ready code that required minimal revision.
- DialogTree Framework: Contributed to the open-source framework **DialogTree**, implementing scalable enterprise architecture that simplified dialog creation for 4+ **Chatbot** projects through XML-directed management and LLM integration with comprehensive documentation and test suites.
- Hennepin County Healthcare Chatbot: Deployed a multimodal **Bilingual Chatbot** (English and Spanish) using the DialogTree framework and **LLAMA 3.1 70B** model with sub-5ms latency. Enhanced search efficiency by 35% for organ transplant services, collecting 50+ user feedback data points for continuous model enhancement and performance evaluation.
- PASCAL Smoking Cessation Project: Built and deployed a production-ready **React**-based chat UI for counselors, collecting 1,000+ conversational data points to fine tune LLMs with rigorous evaluation metrics, improving interaction quality by 20% and developing automated annotation processes for continuous model improvement.
- MNStar Health Assessor Training: Developing an enterprise training system using **RAG** with **Elasticsearch** indexing and **Kibana** visualizations, reducing training time by 60% and improving knowledge retention by 25% through analysing 15,000+ interactions with continuous model refinement.

Traderware *June 2024- August 2024*
AI/ML Intern *Remote*

- Developed and optimized a production-ready **RAG pipeline** for an enterprise trading application, enhancing query processing efficiency by 50% through the integration of advanced **Agentic Systems** (**AutoGen, LangChain, LlamaIndex**) and implementing **RAG Fusion** based query processing with over 100,000 chunks of SEC Form 10-K and 10-Q documents, with comprehensive empirical analysis of retrieval performance.
- Improved retrieval accuracy and system performance by 40% through the implementation of **GraphRAG** with **Neo4j**, leveraging **Multimodal** data processing for 50,000+ textual and visual data points, achieving 80% document retrieval accuracy using industry-standard evaluation metrics including cosine similarity and advanced reranking methods with thorough documentation.

Cohere *July 2024- August 2024*
Team Lead, Expedition AYA *Remote*

- Led development of production-ready **Indic LLM** SamskruthaLLM using **Llama-3.1** and **Gemma-2**, implementing various **PEFT and Quantization** strategies with modular, reusable code to achieve 40% improvement in translation quality across 11,000+ training pairs and published enterprise-grade model in **Hugging Face Hub**.
- Engineered scalable evaluation framework using **BLEU/ROUGE** metrics alongside custom **Linguistic Heuristics**, developing new evaluation paradigms resulting in 30% more accurate model assessment for Sanskrit-specific language tasks with comprehensive performance documentation.
- Designed efficient **Tokenization** and **Data Preprocessing** pipeline handling complex **Sanskrit Morphology**, addressing data quality issues and reducing **Inference** time by 25% while maintaining 95% accuracy on downstream tasks with continuous performance monitoring.

Accenture *August 2021- August 2023*
ML Engineer & Application Developer *Bengaluru, India*

- Developed an enterprise ML platform with automated pipelines on Azure cloud, reducing model deployment time by 40% through streamlined preprocessing, leveraging **Azure ML** for GPU-accelerated **BERT** and **XGBoost** model training with continuous performance monitoring and optimization.

Skills & Certifications

Development & Deployment: Docker, Kubernetes, Git, CI/CD, Production-ready code development
Cloud Technologies: AWS (ML services, S3, EC2), Azure, GCP
ML/AI Frameworks: PyTorch, Transformers, TensorFlow, JAX, Hugging Face, LangChain, LlamaIndex, AutoGen
Evaluation & Monitoring: MLflow, Weights & Biases, Custom evaluation metrics (BLEU/ROUGE)
Data Processing: SQL, PySpark, NLTK, Spacy, Data preprocessing pipelines
Languages: Python (primary), R, C, C++, Java, HTML
Certifications: AWS Certified Machine Learning Speciality, Azure Fundamentals (AZ-900)