VISHESH KUMAR

New York, USA | +1(631) 633-1439| Website **|** vishesh.kumar.1@stonybrook.edu **|** LinkedIn/vishesh71 **|**Github/vishesh711

EDUCATION

Masters of Science (Data Science), Stony Brook University, New York, USA **Aug 2023 – Present**

Bachelors of Technology (Computer Science), SRM Institute of Science and Technology**,** Chennai, IN **July 2019 - May 2023**

WORK EXPERIENCE

Gen - Artificial Intelligence Intern – A79.ai, Palo Alto, CA **Dec 2024 – Present**

- Engineered high-performance **AI workflows** with advanced retrieval, caching, and ETL pipelines, reducing response latency by **40%** and increasing system throughput by **35%** for real-time decision-making.
- Developed **AI architectures** with **Minikube-based deployments**, optimizing containerized workflows and worksheets by cutting deployment time by **30%**, while ensuring integrity of large-scale agents.
- Enhanced **Celery logging** and observability by customizing loggers and distributed tracing, improving debugging efficiency by **40%** and ensuring robust monitoring in asynchronous AI-driven systems.

Artificial Intelligence Intern – BSNL ALTTC Unit, Ghaziabad, IN **Aug 2021 – Mar 2022**

- Designed and deployed a next-generation **Smart Pole prototype**, leveraging **5G connectivity** and **AI-driven IoT architectures**, paving the way for scalable smart city applications and advanced R&D initiatives.
- Implemented state-of-the-art **image processing algorithms** for real-time anomaly detection and surveillance, achieving an industry-leading **91%** accuracy, significantly enhancing operational security.
- Optimized network performance using AI-powered edge computing, reducing latency by **30%**, improving bandwidth efficiency, and enabling seamless integration with large-scale IoT ecosystems.

PROJECTS

Multi-Agent Reasoning Model for Collaborative Intelligence **Jul 2024 - Dec 2024**

- Developed a multi-agent reasoning model, using probabilistic **graphical models**, hierarchical **role-based task allocation**, and **Swarm API** for Intelligent precise decision-making in dynamic, high-complexity scenarios.
- Built a high-performance pipeline using **OpenAI GPT** models, tiktoken for **token-optimized prompt caching** and parallel reasoning workflows, achieving **20%** higher accuracy and **35%** reduced computational latency.
- Orchestrated agents using JSON configurations, **reinforcement learning**, and multi-path refinement, validated on benchmarks with **colorama-enhanced interfaces** and robust logging.

Refined RAG: Precision Retrieval for Improved Responses **Aug 2024 - Dec 2024**

- Designed a pipeline to process PDFs using **PDFium**, extracting text and creating semantic chunks for efficient retrieval. Stored data in a **Qdrant vector database**, enabling high-performance similarity-based searches.
- Implemented a retriever combining similarity search, **FlashRank** for reranking, and a chain of **LLM-based filters** to refine query responses, by creating a local **gemma2:9b** model via **Ollama** on a local machine.
- Optimized performance with **Groq API** for fast LLM inference, achieving up to **300% improvement in response latency** and **50% higher concurrent request handling** and Built an interactive UI using **Streamlit**.

Ensemble Modeling Framework: R Package for Predictive Analytics **Feb 2023 – May 2023**

- Developed a **modular R package** implementing **ensemble algorithms** (Random Forest, Bagging, Lasso, Ridge, Elastic Net, Logistic Regression), optimized for high-performance regression and classification tasks.
- Built a preprocessing pipeline, integrating **correlation-based feature** screening, **data transformation,** and **dimensionality reduction**, improving model efficiency and compatibility with large datasets by **30%.**
- Streamlined usability with Roxygen2**-**based documentation**, CRAN**-compliant standards & interactive user vignettes, achieving a **20%** increase in prediction accuracy through techniques like **blending and stacking**.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, MATLAB, SAS

Frameworks and Libraries: Langchain, RAG, LLM, NumPy, Pandas, Pytorch, Scikit-learn, TensorFlow, Keras, OpenCV

Tools and Technologies: AWS (EC2, S3, Lambda), Docker, Apache Spark, GCS, Tableau, RStudio, Linux, Git, Minikube

Data Science: Time Series Forecasting, Sentiment Analysis, Machine Learning(LSTM, CNN, NLP), Gen-AI, Deep Learning

Statistical Analysis: Chi-square Analysis, Statistical Distributions, Investment Decisions, Statistical Inference