Sai Ram Chetla
(330) 910-9583 | chetlasairamjp@gmail.com | linkedin.com/in/sai-ram-chetla/
https://github.com/chetlasairam | https://chetlasairam.netlify.app/

SUMMARY

AI/ML Engineer

2+ years of professional work experience in the areas of Python Programming, Artificial Intelligence, Machine Learning. Proven ability to seamlessly **adapt new tools and technologies** while actively pursuing ongoing skill acquisition.

EDUCATION

Master of Science, Computer Science *August 2022 – December 2023*
Kent state university, Kent, Ohio, US - **GPA: 3.85**
Bachelor of Technology,Electronics and Communication *August 2017 - May 2021*
VNR Vignana Jyothi Institute of Engineering and Technology, Telangana, India. - **GPA: 3.78**

EXPERIENCE

MRI Software – Cleveland, Ohio, USA *May 2024 – Dec 2024*
AI/ML Engineer

- Partnered in the creation of **AI chatbot** (**RAG**) tailored for **real estate software**, specializing in lease agreement comprehension and analysis, Ensuring optimal functionality using **Llama** for its **cost efficiency** and **Hugging Face model** for token count, and **Azure Cognitive Services** for enhanced **natural language processing(NLP)** and text extraction capabilities.
- Upgraded **ES 1.7 term search index** to **ES 8.8** and implemented **Reciprocal Rank Fusion (RRF)** for **hybrid** search, while integrating **Apache Spark** for large-scale data processing to optimize indexing performance and scalability.
- Developed an optimized relevant segment finding solution using **SpaCy** and **FastEmbed**, leveraging **Azure Storage Class PVC** to reduce service pod initialization time by **40%**, and utilized **Azure Data Factory** to orchestrate and automate data workflows for efficient segment processing.
- Crafted a **React and Node.js-based search UI for Elasticsearch**, customizing the backend for advanced **vector search and hybrid search** capabilities, enhancing file search precision beyond standard term querying, and incorporating **ML pipelines** using **Azure ML** to fine-tune search relevance and ranking models.
- Created **Python** script to index millions of documents into **Elasticsearch**, extracting text from PDFs using **OCR** and **vectorizing** the content with **OpenAI**'s **text-embedding-3** model using **Hugging Face** and **LLama 3.1** to get answers from segments, while implementing **MLOps** practices to streamline model deployment, monitoring, and continuous integration on **Azure**.
- Partnered in an **AI system with GPT-4o** to automate the extraction of lease details from PDFs, utilizing **Pydantic functions** for data validation, innovative prompt crafting for field-specific retrieval, and **Langchain** with **OpenAI** for **RAG** operations, enhanced by **Azure Cognitive Services** for intelligent document understanding and **Apache Spark** for processing large datasets in parallel.
- Researched and evaluated **GPU**-accelerated inference frameworks such as **ONNX** Runtime for optimizing **LLM** pipeline performance, comparing execution times across **CPU** and **GPU** for segment extraction tasks, and identifying opportunities for model compression and hardware-aware deployment within real estate **AI systems**.

AMBIQUEST – Hyderabad, India
Software Engineer *Jan 2021 – Aug 2022*

- Developed a document classification system using traditional machine learning algorithms like **Naive Bayes** and **Support Vector Machines (SVM)** with **TensorFlow** for model training, incorporating feature extraction through **TF-IDF** and term frequency to categorize legal, financial, and operational documents.
- Built an end-to-end **machine learning** pipeline with **TensorFlow**, involving data prepossessing, **vectorization, model training**, and hyper parameter optimization, applying techniques like cross-validation to improve classification accuracy and efficiency.
- Designed **RESTful APIs** for secure communication between the mobile app and **cloud-based** services, and automated deployment and **continuous integration** processes using **Bash scripts** and **CI/CD pipelines** on **GitLab**.

PROJECTS

RATCH AI: Build a complete website which helps job applicants to help in their job application process. Used **FastAPI, React, Python, OpenAI API, Spacy, Pydantic models** and **LLama** to build the website.
Multiple AI ChatBot: Enables users to upload PDFs and ask questions, supporting multiple bots for simultaneous PDF queries. It leverages **LangChain, OpenAI, Hugging Face, RAG**, and various AI tools.
AI PDF Explorer: Extracts data from numerous PDFs in minutes. Users can query candidate experience and get results in a table. Utilizes **Llama** for **NLP**, **OpenAI Embedding** for relevance, **Hugging Face** for token counting, and **LangChain OpenAI** embeddings.
Kent Connects: Developed a **React.js** social site with **HTML**, **CSS**, **JavaScript**, **Firebase SDK** (**Authentication**, **Realtime Database**, **Firestore**, **Storage**), version control with **Git**, and deployed via **Netlify**/**Firebase CLI**.
Husk: Developed a **Flutter** chat app with **Firebase Cloud** , **Postman, ZegoCloud** for calls, and **Firestore** for real-time messaging.

SKILLS

Languages:	**Python**, C, C++, **Java**, HTML, JavaScript**.**
AI tools:	**GPT 4o, Llama, Open AI, Azure ML, Lang Chains, Hugging Face, AWS Sage maker.**
Database:	MySQL, Snowflake database, **Vector databases** (Elastic Search, FAISS), **AWS (cloud)**
Others:	Machine Learning, **Deep Learning**, **FAST API**, **Kubernetes**, **Azure AI**, **Docker, REST API**

CERTIFICATIONS

- **Generative AI** For Beginners - **Google Gemini** & **Google Cloud**
- Programming, **Data Structures** and Algorithms using **Python**
- Fundamentals of **Generative AI, Azure Open AI** service