**Sai Ram Chetla**

(216) 600-2746‬ | chetlasairamjp@gmail.com | linkedin.com/in/sai-ram-chetla/

<https://github.com/chetlasairam> | <https://chetlasairam.netlify.app/>

**SUMMARY**

*AI Engineer*

**5 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 – Present*

*Software 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 **PyTorch, TensorFlow,** **Llama** for its **cost efficiency** and **Hugging Face model** for token count.

*•* Upgraded **ES 1.7 term search index** to **ES 8.8** and implemented **Reciprocal Rank Fusion (RRF)** for **hybrid search,** optimizing **GPU**-accelerated query performance, utilized **DataDog** to observe the data updates.

*•* Developed an optimized relevant segment finding solution using **Python**, **SpaCy**, **FastEmbed and TensorFlow** leveraging **Azure Storage Class** **PVC** and **GPU instances** to reduce service pod initialization time by **40%**.

*•* Developed a full-stack search **UI** using **React.js** and **Node.js,** enhancing file search precision. Integrated **AngularJS** components for modular design and **PostgreSQL** for structured data storage.

*•* Developed and deployed advanced **Learn-to-Rank (LTR)** models utilizing multimodal datasets, including interviews, historical performance, and synthetic data, to enhance ranking accuracy and decision-making in **AI**-driven search and recommendation systems.

*•* Developed and optimized AI-driven search and recommendation systems using multimodal datasets, leveraging **Azure Machine Learning** for deployment and monitoring.

*•* Designed and implemented data pipelines for efficient indexing and retrieval of lease agreement data, utilizing **Azure Data Factory** and **Apache Spark** to process high-volume documents.

**Along consulting llc – Cleveland, Ohio, USA** *Sep 2022 – May 2024*

*Software Engineer*

*•* 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** integrating **PyTorch and GPU**-based acceleration for efficient computation model using **Hugging Face** and **LLama 3.1** to get humanized answers from segment.

*•* Partnered in an **AI system with GPT-4o, BERT** to automate the extraction of lease details from PDFs, utilizing **Pydantic functions** for accurate data validation and innovative prompt crafting for field-specific retrieval and **Langchain,** **OpenAI** for **NLP** task.

*•* Built scalable **ML pipelines** using **Azure ML** and **MLOps** principles, automating model training, evaluation, and deployment.

*•* Optimized and deployed **NLP** models in **Azure Cognitive Service**s, integrating with **Azure** Functions for scalable AI-powered applications.

*•* Deployed the backend infrastructure using **AWS (EC2, RDS)** and **Google Cloud Platform (GCP)** services for scalability. Optimized data storage with **Snowflake** and **PostgreSQL**, improving query performance.

**AMBIQUEST – Hyderabad, India** *Aug 2020 – Aug 2022*

*Software Engineer*

*•* Developed a **cross-platform** mobile application using **Flutter** for seamless experiences on **iOS** and **Android** devices. Integrated **machine learning** models using **TensorFlow Lite** for on-device **activity recognition** and personalized **workout recommendations**.

*•* Implemented **GPU-optimized** **computer vision** techniques with **OpenCV** to analyze and track user movements in real-time, and utilized **SQL databases** for efficient data storage and management, ensuring scalability and performance.

*•* Designed **RESTful APIs** using **FAST API** and integrated **PostgreSQL** for efficient data management in cloud-based services, incorporating **Snowflake** for efficient data storage and large-scale analytics using **Bash scripts** and **CI/CD pipelines** on **GitLab**.

*•* Ensured data security and user privacy by implementing **encryption standards** and robust **authentication protocols**.

**PROJECTS**

**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**, **OpenAIEmbedding** 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, AngularJS, React.js, JavaScript**.**

**AI tools: PyTorch, TensorFlow, GPT 4o,** BERT, **Llama, Open AI, Azure ML, LangChain , Hugging Face, AWS.**

**Database:**  MySQL, PostgreSQL, Snowflake database, **Vector databases** (Elastic Search, FAISS), **AWS, GCP, MangoDB**

**Others:** Machine Learning, **Deep Learning**, **FAST API**, **Kubernetes**, **Azure AI**, **Docker,GPU Acceleration, REST API**

**Big Data & Cloud:** Apache Spark, Hadoop, Azure ML, Azure Data Factory, Azure Cognitive Services, Azure Kubernetes Service (AKS)  
**MLOps:** MLflow, Kubernetes, CI/CD for ML, Azure ML Pipelines

**CERTIFICATIONS**

* **[Generative AI](https://www.udemy.com/course/generative-ai-with-google-bard-google-cloud/)** [For Beginners -](https://www.udemy.com/course/generative-ai-with-google-bard-google-cloud/) **[Google Gemini](https://www.udemy.com/course/generative-ai-with-google-bard-google-cloud/)** [&](https://www.udemy.com/course/generative-ai-with-google-bard-google-cloud/) **[Google Cloud](https://www.udemy.com/course/generative-ai-with-google-bard-google-cloud/)**
* Fundamentals of **Generative AI, Azure Open AI** service