Sakthi Ganesh Mahalingam

(602) 884-7518 | sakthiganesh1997@gmail.com | linkedin.com/in/msakthiganesh/ | github.com/msakthiganesh

EDUCATION

Arizona State University, Tempe – **Computer Science** | Master of Science — GPA: 3.9/4

Vellore Institute of Technology, India – **Electronics and Communication Engineering** | Bachelors — GPA: 3.3/4

SKILLS

Languages & DB	Python, SQL, MongoDB, Vector DB (FAISS, ChromaDB, Milvus), Web Development (HTML, CSS)
Libraries	PyTorch, TensorFlow, Transformers, DeepSpeed, Ray, LangChain, fast.ai, W&B, NLTK, Scikit-learn, spaCy, Xgboost, Pandas, NumPy, Postman, Git, Docker
Frameworks	FastAPI, Django, Flask, Apache AirFlow, React.js, Angular
Cloud Services	Azure (AZ-900 Certified), GCP (Vertex AI, Cloud Storage, BigQuery, Compute Engine)
Coursework	Statistical ML, NLP, Data Mining, Semantic Web Mining, Probabilistic Learning, Image Processing

PROFESSIONAL EXPERIENCE

Research Assistant – Machine Learning *– UDI, Arizona State University, Tempe* — *June 2023 – May 2024*

- Spearheaded research operations by architecting and designing a **Retrieval Augmented Generation chatbot** using LangChain, OpenAI, and local models, achieving an 89% reduction in research timelines.
- **Optimized resources by reducing GPU memory** requirements by 4x through efficient **fine-tuning of LLMs** (Llama2-7b, Falcon-7b) using DeepSpeed ZeRO-3 and Ray.
- Designed and **developed NoSQL MongoDB** for storing and retrieving previous conversations and Milvus for Vector DB and automated data retrieval, extraction, and ingestion processes using Apache Airflow.
- Developed **asynchronous APIs with FastAPI,** dockerized the app, and deployed on multi-GPU using vLLM and Ray.
- Evaluated RAG model performance using RAGAS and achieved over 95% scores in core metrics on 200+ pilot samples.

Senior Machine Learning Engineer *– Infosys R&D, Bangalore* — *August 2018 – July 2022*

- Identified critical data subsets and **designed a model-independent pipeline using PyTorch**, Snorkel, and Transformer and achieved #6 in the SuperGLUE Benchmark, a rigorous benchmark for natural language understanding tasks.
- **Enhanced RoBERTa-Large** (355M params) performance by converting complex logical tasks into critical task heads using Snorkel and PyTorch, achieving a score within a **4-point difference to billion parameter models** (PaLM – 540B).
- **Developed Infosys AI Platform SDK and APIs using Django**, allowing code interoperability between Vertex AI, Azure ML, and In-house ML and Deep Learning solutions, reducing pipeline development time and migration costs by 40%.
- **Created ETL pipelines** using SQL and Python to extract code databases and Git repositories and preprocessed the data.
- **Pre-trained PLBART and T5 models** on Infosys Git repositories **and fine-tuned** on CodeXGLUE, CodeSearchNet, and CodeGen **for Translation, Completion, and Generation tasks** using DeepSpeed - evaluated using BLUE and F1.
- Deployed code models using Nginx for load balancing, Flask, and Docker, handling 30,000+ API requests.
- Spearheaded research initiatives and implemented SOTA NLP papers by leading a team of ML engineers, significantly improving project advancements and team development.

PROJECTS

Augmenting Transformer Attention using Word Importance — **Github**

- Devised a **novel Word Importance based Attention Mechanism** (WIAM) for Transformer architecture using soft attention values between 0 and 1 based on word importance leading to 5% improved accuracy in complex logical reasoning tasks.
- Utilized normalized word importance scores from LIME along with the attention mask to **significantly enhance the initial training phase of the model** by focusing on critical tokens and debiasing the pre-trained model.

ACHIEVEMENTS

- Insta Award – Center for Emerging Technology Solutions, *Infosys Limited*
- Accelerated Early Career Program, *Infosys Limited*
- Finalist - Microsoft Convergence Hackathon, *Infosys Limited*

PUBLICATIONS

- Answer-Aware Question Generation from Tabular and Textual Data using T5, *International Journal of Emerging Technologies in Learning (iJET)*, 16(18), pp. 256-267. 2021.
- Unsupervised Convolutional Filter Learning for COVID-19 Classification, *Revue d'Intelligence Artificielle*, Vol. 35, No. 5, pp. 425-429. 2021.