Uma Maheshwar Amanchi

♦ Mountain View, CA ♦ uma.amanchi23@gmail.com ♦ (646) 240-8724 ♦ linkedin.com/in/umamaheshwar-amanchi

WORK EXPERIENCE

ArchKey Solutions (Contract) St. Louis, MO
Software Engineer – AI/ML Jun 2024 – Present

- Designed and deployed an agent-based invoice processing system using **GPT-4o (Azure OpenAI)**, automating extraction and classification of unstructured invoice data, and accelerating procurement decision-making.
- Integrated **Azure SQL** for structured storage and **Azure AI** Search for semantic retrieval of part details to build an end-to-end **MLOps** pipeline reducing invoice processing costs by over $1.5M annually and enabling real-time semantic lookup for frequently purchased parts.
- Built a production-grade conversational AI system using **GPT-4o** over enterprise finance databases, with Redis caching, Cosmos DB for persistence, and Entra-based **OAuth/OIDC** authentication to secure user access.
- Implemented **MLOps** lifecycle using Azure ML, DevOps, and **AI Hub** with support for CI/CD pipelines, enabling scalable deployment, experimentation tracking, and model versioning across projects.
- Mitigated hallucinations in **LLM** responses by integrating Bing Search agents for grounding and developed a materials classifier using **Azure Language Studio** to enhance extraction accuracy.

Gainwell Technologies (Contract) McLean, VA
Software Engineer Nov 2023 – Jun 2024

- Finetuned a **Multilingual DeBERTa** model to auto-moderate user-generated content in healthcare support portals, ensuring compliance with community standards and reducing manual review time.
- Replaced legacy **BERT** model with optimized variant, boosting content moderation automation rate by **15.81%** and auto-approving 500k entries resulting in **$100,000** savings in data labelling efforts.
- Curated the training dataset for content moderation using the **GPT-4** and increased the training dataset quality and diversity by leveraging the self-instruct methodology.
- Trained a **Mixtral-8x7B Mixture of Experts(MoE)** model using **Low-Rank Adaptation (LoRA)** for content moderation and deployed a **4-bit Quantized Model** into production.

Eitacies Inc (Contract) Santa Clara, CA
Software Engineer Nov 2022 – Nov 2023

- Developed the **Natural Language Understanding (NLU)** module financial services chatbot by fine-tuning a BERT Transformer model on domain-specific data, enhancing **intent classification** accuracy by 10% over the previous **BI-LSTM** approach.
- Extended English-based chatbot to **Multilingual** settings by collecting in-domain data for French, German, and Spanish languages and finetuned the multilingual BERT checkpoint **using PyTorch framework, Hugging Face library**.
- Implemented a **T5 Transformer** model to summarize lengthy customer service tickets, achieving a **BLEU** score of 0.7, thereby improving response efficiency for support teams.

KSU - College of Public Health Kent, OH
Machine Learning Intern Mar 2022 – Aug 2022

- Designed an internal plagiarism detection system using **Retrieval Augmented Generation(RAG)** to detect similar copies of assignments from previous years.
- Developed system primarily uses the **Dense Passage Retrieval (DPR)** as the retriever and **BART** as a generator.
- Enhanced the overall system by adding a **DeBERTa-based Re-Ranker** and improved the overall recall@1 by 10 points.

Accenture Hyderabad, India
Application Development Associate Jan 2021 – Jul 2021

- Created an **XG-Boost** classifier to predict the category of incoming tickets and obtained a strong **Recall** of **0.95**.
- Designed a **Linear Regression** model to estimate the total time required to close a ticket and achieved **R2** score of **0.87.**
- Leveraged **K-Means Clustering** and its cluster concepts to identify the incoming pattern of unknown tickets and alerted the on-call team to quickly resolve the rapidly increasing clusters.

PROJECTS

Enhancing Multilingual Capabilities of Large Language Models – Generative AI

- Curated the high-quality multilingual fine-tuning dataset by translating the alpaca dataset into 11 languages.
- Improved the **Mistral-7B LLM** multilingual capabilities by **20%** through fine-tuning on translated Alpaca dataset
- Optimized the Multilingual prompts performance using **Prompt Engineering** and **Chain-of-Thought (CoT) Prompting** by **10%.**

Course Enrolment Dropout Prediction - ML

- Curated a dataset by merging data from **10 Data Sources** where each data source has more than **15 Million Records**.
- Implemented data cleaning, normalization, and feature engineering and fitted the data with the **XG-Boost** classifier.
- Conducted **Grid-Search** to obtain optimal hyper-parameters which gave a strong recall of 0.95 and F1-score of 0.91

Multilingual Question Answering - NLP

- Constructed a language-agnostic **Dense Passage Retriever (DPR)** by leveraging **LABSE – Multilingual Sentence Transformer** and obtained a gain of 10 points in recall compared to the language-aware XLM-RoBERTa model.
- Engineered a retriever-reader architecture for document-grounded QA by utilizing the **Cross-Encoder XLM-RoBERTa** model for passage re-ranking and **Multilingual BART** for text generation.

EDUCATION

Kent State University Kent, OH
Master of Science in Computer Science Dec 2022
Selected Coursework: Machine Learning, Deep Learning, Natural Language Processing, Large Language Models

SKILLS

- **Programming Languages:** Python, Java, SQL, C++, JavaScript,
- **AI & ML:** Pytorch, Hugging Face, LangChain, Spacy, TensorFlow, Pandas, Numpy, Scikit-learn, Matplotlib, Git, AWS, Azure, Real Time Image Processing, Computer Vision Solutions