Aswin Kumar J

Baltimore, MD | +1 667-802-9383 | aswinkj1@umbc.edu | LinkedIn | GitHub | Medium

Education

University of Maryland, Baltimore County — **Baltimore, MD**
MS in Information Systems *(specializing in Artificial Intelligence)* — Aug 2023 - May 2025 (expected)

Skills

Python3, Apache Spark, Tableau, ROS, Microservices, PyTorch, CUDA, TensorFlow, LLM, Pinecone, OpenCV, Transformer, Docker, Kafka, TypeScript, Next.js, Nuxt.js, React.js, HTML5, CSS3, Javascript, AWS EC2, S3, IAM, SageMaker, DynamoDB, Bedrock, ELB, Route 53, CloudWatch, Azure, GCP, API, REST, SOAP, GraphQL, gRPC, JSON, GIT, GitLab, SQL, MongoDB, Jupyter Notebook, Google Colab, VsCode, Shell, CI/CD, Linux.

Work Experience

University of Maryland, Baltimore County — **Baltimore, MD**
Research Assistant — Nov 2023 - Present

- Developed end-to-end navigation system of swarm robotics using Python3, and ROS1, integrating with Boston Dynamics SPOT robot and Clearpath's Husky
- Utilized OpenCV to capture and process images from video streams, optimizing image interpretation with a transformer-based GPT model
- Pioneered new navigation techniques by integrating SPOT SDK and Garmin GPS module, advancing robot positioning capabilities based on location data received from the web-based Dashboard
- Fine-tuned LLM3-codellama, instruct, and Molmo-7B-D models on edge devices to create AI-enabled navigation systems and chatbots, boosting the efficiency of autonomous robot controls by 40%
- Architected and developed a multimodal chatbot that can effectively respond to user's request with voice based on the visual videos that are visible from the robot's POV, trained it to give context-rich answers using RAG techniques and Langchain, voice model using gTTS/Kyutai-Moshika

University of Maryland, Baltimore County — **Baltimore, MD**
Full Stack Developer/ML Engineer — Dec 2024 - Jan 2025

- Developed and implemented the front-end for an educational web application using Nuxt.js through dynamic routing and state management, allowing for a smooth and interactive user experience with Data Science Caselets.
- Designed and integrated backend services for data intake and processing in MongoDB, which allows for predictive modeling of metacognitive abilities in problem-solving in data science.
- Deployed and optimized a full-stack application on AWS to achieve high availability and scalability, allowing over 200+ active users every day with high performance.
- Successfully integrated the Meta 3.2 3B Instruct model from AWS Bedrock, which allowed students to interact with a conversational LLM for real-time queries, enhancing the learning experience in data science and improving student engagement by 80% with the DSPS platform.
- Developed an adaptive student performance prediction model using regression and collaborative filtering algorithms, including sparse matrix factorization, to predict learner responses for unseen Caselets and integrated it into the existing Caselet platform, enabling personalized learning experiences and improving engagement by 95%.

Headstarter AI — **Remote**
SWE Fellow — Jul 2024 - Sep 2024

- Collaborated as a team of 4 in deploying a chatbot solution that improved Airlines' customer support efficiency by 50% using AWS Bedrock, Pinecone, Sagemaker, and multi-language support
- Enhanced the customer chatbot model with multi-language support, feedback mechanisms, and user authentication using Clerk, expanding its functionality to assist broader sections of customers
- Built a web scraper pipeline using BeautifulSoup that extracts and upserts data from Rate My Professor into a Pinecone vector index and integrated with a RAG system using LlamaIndex and OpenAI GPT-4o, this application provides up-to-date and context-relevant answers to user queries, enhancing the accuracy and relevance of responses
- Orchestrated the implementation of coding, design, and deployment best practices using EC2, fostering a collaborative team environment
- Built and deployed a feature-rich pantry tracker app in Vercel, enabling item addition, deletion, updating, and searching functionalities with Next.js, React.js, and Firebase

Vagus Technologies Inc **Trichy, India**
Senior Software Engineer Jul 2018 - Aug 2023

An in-house Applicant Tracking System (ATS) is a software application used by end clients and human resources professionals to streamline and automate the recruitment and hiring process

- Developed robust, scalable, and secured web applications for the applicant tracking system to ensure the continuity of business processes and client satisfaction
- Designed and configured caching database and backend applications, improving application efficiency by 30%
- Consumed APIs while utilizing Python requests to read numerous JSON reports and file automatic bugs for intermittent bugs during A/B testing & Static testing phase
- Designed and implemented RESTful APIs to facilitate data migration from legacy systems to AWS cloud services
- Utilized Docker for containerization, automating deployment processes on AWS EC2 and S3, and employed AWS Route 53 for scalable DNS management. This approach streamlined data transfer, reduced latency by 30%, and enhanced system scalability and reliability

Ahoy Systems Pvt Limited **New Delhi, India**
Technical Associate Jan 2017 - Jun 2018

An IoT-based solution for the modern lifestyle including Automatic ATM Security, pH analysis reports Indian Government from various Sugar plants in India.

- Designed and developed a security system based on BeagleBone Black using various sensor data, improving ATM security and reducing incidents by 30%
- Executed a system to generate alerts to 3rd party security and maintenance teams, ensuring timely responses and reducing downtime by 20%
- Programmed cron jobs to execute shell scripts for raising alerts on web portals, enhancing system reliability and reducing manual intervention

Google Summer of Codes'16 - Forced Alignment of words (RedHen Labs) Mar 2016 - Jul 2016

- Engineered a forced alignment system using Kaldi ASR, SRILM, and IRSTLM, optimized for HPC clusters, reducing alignment time by 60% for large-scale news video datasets
- Developed Python scripts to automate the alignment workflow, increasing processing speed by 75% and enabling the system to handle 500+ hours of video content daily
- Integrated Edinburgh Speech Tools for advanced phonetic analysis and feature extraction, significantly improving word-level alignment precision
- Collaborated with Red Hen Lab to integrate the system into their framework, resulting in a 40% increase in research output for multimodal communication studies

Projects

Volatility Forecasting using GARCH-LSTM Oct 2024 - Dec 2024

- Volatility prediction using GARCH-LSTM and SVR compared it against the Moment-1-Large model using Streamlit for representation.

AI-based Customer Chatbot for Delta Airlines Aug 2024 - Oct 2024

- An AI-powered chatbot for customer support using a custom RAG pipeline built with Llama 3.1 8b model, LangChain, Pinecone, and deployed on AWS Bedrock, EC2. Rich user engagement using Next.js, React, and Boto3.

Named-Entity Recognition using DistilBERT dataset Mar 2024 - May 2024

- A model that recognizes rare named entities using the DistilBERT and BERT Transformer.

Certifications

- **AWS Cloud Solution Architect (Associate)** Feb 2025
- **Neo4j Certified Professional** Oct 2024
- **AWS Educate Machine Learning Foundations** Oct 2024
- **AWS Certified Cloud Practitioner** Sep 2024
- **Oracle Cloud Infrastructure 2024 Generative AI Professional** Jul 2024

Publications

- AAAI 2025 Spring Symposium: Edge LLMs for Real-time Contextual Understanding for Robots