Aswin Kumar Janakiraman

Baltimore, MD | aswinkj1@umbc.edu | LinkedIn | GitHub | Medium

Education

University of Maryland, Baltimore County **Baltimore, MD**
MS in Information Systems - 3.9/4.0 Aug 2023 - May 2025
Relevant Courseworks: Deep learning with PyTorch, Information Extraction, Data mining & Analytics using Tableau, Machine learning, Multi-Cloud Computing and Resource Management using Python & Terraform

Skills

Languages: Python, GoLang, JavaScript, TypeScript
Databases: SQL Server, DynamoDB, PostgreSQL, MongoDB, PL/SQL, Oracle Database 23ai, Pinecone
Frameworks: Flask, Django, FastAPI, GraphQL, React.js, Redux, Next.js, Nuxt.js, Microservices, Kafka
Technologies: PySpark, ROS, JSON, Markdown
Cloud: AWS EC2, S3, IAM, SageMaker, Bedrock, Lambda, Amplify, Kinesis, Athena, ECS, GCP Vertex AI Studio
AI/ML: PyTorch, TensorFlow, HuggingFace, Transformers, LLM, LangChain and LangGraph
Agents: MCP, ACP, Google A2A, OpenClaw, Hermes and Autoresearch
Tools: Git, VSCode, Docker, Jenkins, Terraform, OAuth2, Jupyter Notebook, Tableau, OpenCV, PyLint, JsLint
Testing Frameworks: Unit testing and A/B Testing

Work Experience

Shock, Trauma and Anesthesiology Research Center (Prime-AI lab), UMB **Baltimore, MD**
Senior AI/ML Engineer Nov 2025 - Present

- Engineered an Agent-powered AIS scoring pipeline using MCP, Python, and MongoDB, achieving 83% expert-validated accuracy across 20 years of Maryland Trauma data, estimated to save $40M in healthcare costs
- Optimized local Ollama-based LLM inference by tuning data sampling strategies via Terraform and agent tool-calling intervals, improving model throughput by 40%
- Designed a VLM-based Surgical Assistant using RAG on image-text data, automating surgical documentation across 5 high-critical procedures and reducing manual documentation effort by 60%
- Collaborated with 10+ clinicians and surgeons to propose and validate AI-driven workflows, accelerating research-to-deployment cycles by 35%
- Built a Human-in-the-Loop medical review panel using LangExtract to detect documentation errors in Trauma patient records, improving record accuracy by 75% across reviewed cases
- Presented monthly literature reviews on Agentic AI architectures and model development trends to a research team, contributing to a publication submission for JMIR and 2+ notable AI conferences

Center for Real-time Distributed Sensing and Autonomy, UMBC **Baltimore, MD**
Graduate Research Assistant Nov 2023 - May 2025

- Engineered an end-to-end navigation system or the fleet of Boston Dynamics SPOT and Clearpath's Husky using Python and ROS, enhancing the cross-platform navigation system by 75%
- Developed voice-based Llava-1.5b VQA model using RAG and Prompt Engineering, improving image interpretation accuracy by 90%
- Optimized the VQA model's voice throughput utilizing advanced NLP and voice models and reduced the processing time to 1.25 seconds/response
- Architected and developed a multi-agent workflow using LangGraph, and MCP to execute high-level natural language voice commands and improving robot mission success by 85%
- Presented monthly literature reviews on Agentic AI architectures and model development trends to a research team, contributing to a publication submission for JMIR and 2+ notable AI conferences

Informatics for Human Flourishing, UMBC **Baltimore, MD**
ML Engineer Dec 2024 - Jan 2025

- Developed an educational web application using Nuxt.js, improving the click-through rate with A/B Testing by 90%
- Migrated the legacy application into AWS EC2, integrated with MongoDB to boost transaction rate by 80%
- Integrated Meta 3.2 3B Instruct model from AWS Bedrock with the front-end for in-built chatbot experience and increased user engagement with the application by 85%

- Pipelined a caselet recommendation model from AWS SageMaker with the application and improved user engagement by 95%

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

- Built business prototypes using Python, FastAPI and React, reducing the time-to-market by 70%, brought $2M+ in revenue
- Implemented Memcache for efficient retrieval of data from PostgreSQL, reducing the latency to 50ms as measured by reduced API response times during peak loads
- Designed RESTful APIs for data migration from on-premise to AWS cloud, reducing migration time by 65%
- Utilized Docker to contain and deploy applications in AWS, reducing manual overhead by 70%

Internships

Headstarter AI **Remote, US**
Software Engineering Fellow Jul 2024 - Aug 2024

- Developed and deployed a chatbot using React, Next.js, and Python on AWS EC2 integrated with Bedrock
- Enhanced chatbot model with multi-language support using AWS Bedrock and user authentication with Clerk, expanding assistance to broader customer sections
- Improved the chatbot with RAG pipeline using LlamaIndex and OpenAI GPT-4, knowledge base using BeautifulSoup, and enhancing groundedness by 92%
- Improved web interface design and model prompts via A/B testing, resulting in an 80% retention rate

Google Summer of Codes ’16 - Forced Alignment of words (RedHen Labs) **Remote, US**
Summer Intern 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

- **Kaggle BirdCLEF+ 2025: Audio-based species recognition ||** Developed an audio-based species identification model (birds, amphibians, mammals, insects) from the Middle Magdalena Valley of Colombia, using NFNet and SERESNext models with a 0.880 F1 score
- **RefactorAI ||** A smart code refactoring application using TypeScript, Terraform, REST API, AWS, and transformer || Designed and implemented the core logic using the CodeBERT transformer model deployed in AWS Sagemaker to analyze and translate the code with 85% of semantic equivalence || Designed a scalable infrastructure with Terraform to manage AWS Serverless services for seamless code snippet management and reducing the manual overhead by 70%
- **NSF HDR ML Hackathon - Sea Level Anomaly Detection Challenge ||** Deployed an anomaly detection model within AWS Sagemaker, trained on 20 years of real-time sea level data across 25 coastal regions of the US, with a precision of 89%

Publications & Achievements

- SIGCSE 2025 - CASTCurate: An Agentic System to Accelerate the Collection and Annotation of Data-Driven Stories
- AAAI 2025 Spring Symposium: Edge LLMs for Real-time Contextual Understanding for Robots
- Won a Bronze medal in a Kaggle competition - BirdCLEF+ 2025 by securing 192/1825 position worldwide

Certificates

- AWS Solution Architect (Associate), AWS Educate Machine Learning, GCP Machine Learning Engineer, Oracle Cloud Infrastructure 2024 Generative AI Professional, OpenCV University - PyTorch and TensorFlow Bootcamp, Certified Neo4j Developer