Tarun Bevara

Beaverton, OR | tarunbevara10@gmail.com | +1 (989)-317-6067 | linkedin.com/in/tarunbevara10 | https://github.com/TarunBevara10

Summary

AI/ML Software Engineer with 3+ years of experience designing agentic and cloud-native ML applications on Azure and AWS. Delivered scalable AI/ML solutions for global clients including McDonald's and PepsiCo, optimizing model training, deployment, and inference pipelines using modern MLOps and cloud infrastructure.

Education

Central Michigan University, Master of Science in Computer Science, MI — Jan 2024 – Dec 2025
- **Coursework:** Artificial Intelligence, Machine Learning, Data Mining and Pattern Recognition, Analysis and Design of Algorithms, Applied Data Engineering, Cloud Computing
- GPA: 3.95/4.0

Gitam University, Bachelor of Technology in Computer Science, India — June 2018 – May 2022
- GPA: 8.71/10

Work Experience

SECURITY ANALYST - AI/ML, Central Michigan University, MI, USA — **Jun 2025 – Present**
- Researching a Hierarchical Multi-Agent RL framework to automate mitigation, investigation, and case creation across **Defender 365, Elastic, PowerShell, DFIR-IRIS, and Azure**.
- Built **FastAPI/Flask** microservices for agent coordination and Defender Graph API orchestration with **Azure Functions**, leveraging **Azure AI Foundry** for model lifecycle management and multi-agent orchestration experiments.
- Implemented MLOps workflows via **GitHub Actions, Docker, and Azure DevOps**, improving deployment speed by 45%
- Optimized inference and model serving using **NVIDIA Triton**, reducing latency 25% and enabling scalable multi-agent deployments on Azure Kubernetes Service (**AKS**).
- Monitored and investigated alerts in Microsoft 365 Defender using Elastic and **KQL** queries to correlate cloud telemetry, detect threats, and accelerate incident response.

SENIOR ANALYST – MACHINE LEARNING ENGINEER, Tiger Analytics, India — **June 2022 – Dec 2023**
- Built cloud-based **FastAPI** microservices on **AWS ECS** using Textract APIs to process 3K+ files, reducing manual review 80%.
- Integrated **React** frontends with backend inference via **REST, WebSockets, GraphQL**, enabling real-time predictions 60% faster.
- Optimized LLM fine-tuning on **AWS EC2** (p4d) with **CUDA-Docker**, achieving 66% faster training and retraining cycles.
- Improved text generation 30% using RLHF **(PPO)** and deployed scalable inference APIs on **SageMaker** Endpoints.
- Integrated **Qdrant DB** and **BM25 retrievers** for RAG search, boosting semantic precision 25% and cutting latency 60%.
- Deployed **QLoRA-quantized RAG models** on SageMaker, halving GPU usage and cutting cloud costs 40%.
- Deployed a CV model on **Azure AKS**, achieving 96% precision on 50K+ retail images from **Azure Blob** for compliance automation.
- Built **FastAPI** backend APIs integrated with a **React Native** UI, serving 1000's of requests concurrently with sub-200 ms latency.
- Integrated **PostgreSQL + Redis** for metadata caching, reducing query latency by 55% and improving throughput.
- Optimized on-device (android) inference using **ONNX Runtime and C++ bindings**, improving speed 30% and lowering memory 20%.
- Applied PCA + t-SNE for dimensionality reduction, preserving 98% variance and boosting real-time inference speed.
- Automated scalable MLOps CI/CD with **Terraform, Jenkins, AKS, Docker**, shortening release cycles 40%.

DATA ANALYST, Tiger Analytics, India — **Jan 2022 – May 2022**
- Supported 5+ client projects through advanced **predictive analytics** research, authoring 50+ technical tutorials to accelerate onboarding for new analysts and interns.
- Processed and visualized 100K+ financial records across 10 global companies using **Tableau** and **Alpha Vantage API**, improving analytical reporting speed by 25%.
- Deployed **Kafka** on **Amazon EC2** for real-time streaming and automated ETL pipelines via **AWS Glue + Glue Data Catalog**, cutting transformation time by 50%.
- Optimized large-scale data retrieval using **Amazon Athena** and **AWS S3**, enabling 20% faster query performance for key business insights.

DATA SCIENCE INTERN, SmartKnower, India — **Jan 2021 – Dec 2021**
- Preprocessed large datasets in **Python** and **Pandas**, handling outliers and categorical variables to achieve 20% faster training and improved data consistency.
- Applied **NLP** techniques (**TF**, **TF-IDF**, **Word2Vec**) and implemented Fusion-CNN, improving model F1-score to 92%
- Utilized **Scikit-learn**, **NLTK**, **Gensim**, and **Keras** to train supervised/unsupervised models (Logistic Regression, SVM, Naive Bayes, CNN) achieving 95% accuracy for NFR classification.
- Enhanced feature scaling with StandardScaler and MinMaxScaler, boosting model precision and efficiency by 15%.

Skills

Programming Languages: Python, C++, C#, JavaScript, SQL, .NET, CUDA
Web Dev : FastAPI, Flask, REST, GraphQL, WebSockets, OpenAPI, ReactJS, React Native, Angular
Databases: PostgreSQL, MySQL, Qdrant, Redis, Azure SQL, Elastic Stack (Elasticsearch, Kibana, Logstash)
Visualization Tools: Tableau, Plotly, Seaborn, Matplotlib, Pandas, NumPy, Grafana

Deployment tools: Jenkins pipeline, Docker, Kubernetes, AWS Sagemaker
AI/ML Tools: PyTorch, TensorFlow, Keras, Scikit-learn, LangChain, Hugging Face Transformers, vLLM, Ollama, Optuna, ONNX Runtime, Nvidia Triton
Cloud & DevOps: Microsoft Azure, Azure AI Foundry, Azure OpenAI, Azure Functions, AWS (SageMaker, Lambda, ECS), Docker, Kubernetes, Jenkins, GitHub Actions
Other: Other: PowerShell, Terraform, App Dynamics, Grafana, CI/CD Automation

Projects

AIOS – AI Operating System (*Paper Implementation*)

- Built a multi-agent AI framework using Python, PyTorch, LangChain, vLLM, and Hugging Face Transformers, integrating 5+ LLM backends (OpenAI, Anthropic, Groq, etc.).
- Implemented RAG + FAISS-based retrieval and distributed orchestration for 10K+ daily interactions with reduced latency and higher reasoning accuracy.
- Designed an OS-like scheduler for agent process management, memory sharing, and asynchronous task execution—enabling modular plug-and-play agent services and resource optimization.
- Benchmarked agents on GAIA, HumanEval, SWE-Bench with Scikit-learn metrics for reproducible multi-environment testing.

Database Agent with Microsoft Semantic Kernel and Azure OpenAI

- Built an SQL query generator agent from natural language using Semantic Kernel and Azure OpenAI for natural language database querying (C#, .NET, SQL).
- Automated schema learning and contextual query generation with Kernel Memory connectors.
- Designed robust query security: relevancy filters, custom QA, and safe execution controls.
- Deployed scalable, multi-modal agent service with REST API and Docker for fast, versatile integration.
- Enabled natural language querying for 1M+ database rows with agent, achieving 90%+ accuracy on benchmark datasets.

Intelligent Multi-Modal Fusion System for Urban Autonomous Vehicle Navigation

- Re-implemented TransFuser, a Transformer-CNN fusion model combining RGB + LiDAR data via ResNet + PointNet encoders for end-to-end driving control.
- Built and trained models in PyTorch + CARLA, integrating ROS + OpenCV for synchronized multi-sensor pipelines.
- Applied Adam optimizer and mixed-precision GPU training to improve convergence and reduce training time.
- Visualized learned attention and driving behavior using TensorBoard and Matplotlib for interpretability.

Graph-Based RAG Autonomous Agent

- Implemented a controllable RAG pipeline using LangChain, FAISS, and LlamaIndex for retrieval-augmented generation with structured control.
- Designed graph-based reasoning modules enabling dynamic retrieval, reduced hallucination, and interpretable responses.
- Deployed FastAPI + Streamlit for real-time demos, integrating PDF/document loaders and vector-storage workflows.
- Achieved high relevance and interpretability across diverse reasoning tasks with minimal debugging overhead.
- Improved answer relevance by 25% and reduced hallucinations by 35% in real-world RAG tasks, validated across 500+ knowledge queries.

Research Work

Pit tag impact on Lamprey movement using Computer Vision in low light **Jan 2025 – May 2025**

- Processed 10K+ hours of underwater footage using OpenCV denoising and temporal smoothing to reduce noise and motion blur.
- Fine-tuned a YOLOv11 (CSPDarknet-RepVGG) model with SPP modules for small, elongated lamprey detection in low-light, turbid water.
- Used MOG2 background subtraction with temporal filters to isolate true biological motion from reflections and air bubbles.
- Integrated BoTSORT tracking with CVAT annotations, optimizing thresholds to reduce ID switches and align detections with PIT timestamps.

Creating fashion design sketches using Conditional GAN **Jun 2025 – Present**

- Currently working on AI powered fashion design system using Condtional GANs and increasing accuracy.

Certifications

- Microsoft AI & ML Engineering - Professional (2025)
- Microsoft SQL server - Professional
- IBM RAG and Agentic AI - Professional (2025)
- IBM Deep Learning with PyTorch, Keras and TensorFlow - Professional (2025)
- Advanced Machine Learning on Google Cloud - Professional (2025)
- Google Cloud AI Infrastructure - Professional (2025)
- Oracle Cloud Database Service - Professional (2025)
- OCI Multicloud Architect Professional (2025)
- ISRO - Geoprocessing Using Python (2019)