Ashwin Sateesh Kumar

ashwinsateesh5@gmail.com • LinkedIn • GitHub • Portfolio • (925)-445-6494 • Dallas, TX

TECHNICAL SKILLS

- **Programming Languages**: Python, R, SQL
- **Frameworks:** PyTorch, TensorFlow, Scikit-Learn, LangChain, Hugging Face
- **Data & ML Libraries:** NumPy, Pandas, SciPy, OpenCV, NLTK
- **Tools & Platforms:** AWS, GCP, Docker, FastAPI, Git, Bash, CUDA, FAISS, Vector Databases, UI Integration (TypeScript, HTML)
- **Core Skills:** ML System Design, Model Optimization, NLP, Computer Vision, Statistical Modeling, CI/CD

WORK EXPERIENCE

AI Software Engineer, Assistant Vice President – *Citibank* Dallas, TX | December 2024 – Present

- Developed a ReAct-based multi-agent vulnerability finder with an explicit reasoning-action loop and stateful tool orchestration, integrating a dependency-aware deep code analyzer for source-to-sink reasoning and knowledge-base validation
- Extended the system with a user-driven vulnerability report generation agent, reducing penetration testing lead times by over 50%
- Productionized a multi-tool security platform with custom built MCP server, exposing in-house ML-driven tools like architecture diagram and cloud log analyzers to automate threat assessment for over 8000 security analysts
- Architected a microservices-based RAG system with async chat history management, PostgreSQL vector storage, and security guardrails against prompt injection and adversarial attacks, enabling reliable session recovery, consistent context, and fast access
- Optimized the RAG pipeline using dual-layer memory with BM25 + cross-encoder reranking, RAGAS-based chunk retrieval evaluation, and prompt caching; reduced redundant vector lookups and stabilized LLM grounding via relevance-based context pruning

Machine Learning Engineer – *Abecedarian* Boston, MA | January 2024 – December 2024

- Engineered a scalable Yoga Assistant by fine-tuning GPT-3.5 with parallel queue-based continual learning and automated web scraping for real-time updates, integrating Chainlit UI and perplexity-based benchmarking to achieve 2x reduction in model latency
- Enhanced multimodal capabilities by fine-tuning Stable Diffusion and distilling it into a lightweight Latent Diffusion Model for yoga posture visualization, reducing the inference time by 50% and validating image-text alignment using CLIP-based evaluation
- Led the design and development of an Agricultural Policy Recommendation System integrating multimodal ML models – Vision Transformers + PCA + Xgboost for remote sensing, LSTM for forecasting, and BERT for news analysis to deliver real-time insights
- Enabled data-driven decision-making for stakeholders by orchestrating LangChain workflows and a custom fine-tuning pipeline to transform multivariate forecasts into actionable policy recommendations, deployed via a cloud-based Streamlit interface

Research Assistant – *Northeastern University* Boston, MA | July 2023 – November 2023

- Uncovered distinct visual and linguistic patterns in social media communication by training a custom disentangled multimodal β-VAE (U-Net for images and BERT for text) and analyzing learned interpretable latent factors from 400k posts
- Achieved fine-grained control over multimodal generation by manipulating disentangled latent factors; deployed the system on GCP with bfloat16 quantization, improving inference performance by ~30%

Machine Learning Engineer Intern, R&D - *Signify (Phillips Lighting*) Boston, MA | June 2022 – December 2022

- Developed a hyperspectral vision pipeline using infrared and RGB imagery for plant anomaly detection, driving real-time closed-loop lighting corrections that sustained healthy growth across commercial-scale grow environments
- Optimized household lighting automation across 18 homes by forecasting adaptive lighting scenes using SARIMAX and Xgboost, improving energy efficiency by achieving 97% prediction accuracy
- Implemented a vision-based user re-identification system for personalized home automation, enabling reliable identity matching across camera views by adapting an omni-scale feature learning architecture with mean average precision (mAP) of 0.95

Trainee Software Engineer, Machine Learning - *KPIT Technologies Ltd* Bengaluru, India | July 2019 – November 2020

- Improved large-scale autonomous driving perception by enhancing annotation quality of 1M+ images using transfer learning and human-in-loop (HIL) workflows
- Devised a U-Net-based semantic segmentation model for traffic scenes understanding, achieving 0.89 IoU in spatial detection tasks

PROJECTS

HealthBot: LLM-based Healthcare Assistant Python | PyTorch | Hugging Face | LoRA | RLHF

- Engineered a healthcare chatbot achieving a F1 score of 0.96 for disease classification using Bi-RNNs and a BERT-enhanced Named Entity Recognition system; fine-tuned GPT-2 with LoRA and reinforcement learning (RLHF) to deliver precise semantic responses

EDUCATION

Northeastern University Boston, MA | December 2023

Master of Science in Data Science ***(GPA: 3.8/4)***

- ***Courses:*** *Algorithms, Machine Learning, Data Mining, Database Management Systems, Deep Learning*

PES University Bengaluru, India | July 2019

Bachelor of Engineering in Electronics and Communications

- ***Courses:*** *Data Structures (OOPs), Linear Algebra, Artificial Neural Networks, Pattern Recognition, Image Processing, Signal Processing*