

BILWASIVA BASU MALLICK

+1 (614) 405-1679 basumallick.2@osu.edu <https://linkedin.com/in/bilwasiva/> [Github](#) [Website](#)

Summary

- Accomplished ML Scientist with **9.5+** years of experience in architecting **39** enterprise-scale AI/LLM systems.
- Proven leader with AI/ML expertise in Generative AI, Large Language model, Agentic AI, Transformers, NLP, NLU, Vision, Recommendation, Deep Learning; proficiency in building, mentoring Engineering & Research teams.
- Bridging frontier AI research with business strategy via **T25 STEM MBA** to scale technology driven products.
- Specialization:** GenAI, Multimodal AI, LLMs, Distributed Training, AI Infra, MLOps, Reinforcement Learning

Education

The Ohio State University

Columbus, Ohio

Master of Business Administration (MBA) - STEM MBA (Major- Tech/AI)

Aug 2025 – Apr 2027

Award: Fisher MBA Leadership Fellow, Graduate Assistant — Score: GRE- 336, GMAT FE-685 — **GPA- 3.9/4**

BITS Pilani

Rajasthan, India

Master of Engineering (ME) — Bachelor of Engineering (BE) in Computer Science CS Dept Rank - 1st — CGPA:9.3/10

Work Experience

KLA

May 2026 – Aug 2026

AI Scientist

Ann Arbor, Michigan US

- Spearheaded AURORA, autonomous **10-agent** AI system converting bug reports/feature requests into production-ready pull-requests in **6 minutes** with self-correcting review loop, delivering **1k PRs** on org-wide GitHub for **19** repositories.
- Built **50ms** latency Agentic RAG microservices with BM25, MPNet, semantic re-ranking, Gemma-4 routing at **\$0 cost**.
- Designed production grade label based RBAC protecting **1.5M+ nodes** for 15 Neo4j Databases configuring 500+ custom roles, 250 automated tests (100% pass) at bitmap O(1) time at infinite scale reducing credential leak exposure by **92%**.
- Championed **AI adoption** publishing **29** reusable Claude Code skills for citizen-developers, security, CI/CD pipelines.

Amazon

Feb 2021 – Aug 2025

Principal Applied Scientist

Bangalore, India

- Pioneered GenAI CX ecosystem in Amazon Pay with Meta Llama 3.1 (8B) for intent classification, Hermes 3 (Llama 3.1 405B) chatbot, Qwen2.5-Coder (32B) & Mistral-Nemo (12B) for tool calling to automate queries for **70M customers**.
- Directed cross-functional org of 15 scientists, MLEs across Ads, Prime Video to build Rec-Rainbow, a 0-shot, LLM based personalization agent for session-aware recommendation, generating a **53% dwell time increase** & **44% lift in CTR**.
- Accelerated task execution of multi-hop workflows in AWS Bedrock by **3X** using LangGraph for stateful control, CrewAI for multi-agent coordination, Anthropic MCP for standardized tool integration, with parallelized execution across steps.
- Deployed Retrieval Augmented Generation (RAG) model for AWS Documentation org to streamline automated response for real-time queries, integrating Pinecone vectorDB, Longformer & Flann-T5, obtaining 87% BertScore & **0.91 MRR**.
- Innovated "HELIUM-PRISM", a robust lightweight multi-modal model, to accelerate product mapping & ranking in Kindle, Prime Video & promote de-duplication in digital catalog for **84 external competitors** across **3 continents**.
- Submitted patent for CV model with **98%** precision, **99%** recall in unmatched & **99%** precision in matched decision, outperforming baseline (VGG, ResNet, Inception, MobileNet, EfficientNet) with **9X speed** & **14.5X better memory**.
- Optimised model execution time by **250%** using EMR, enabling DCCS Org perform **0.3 Million** batch inferences daily.
- Augmented quarterly Kindle purchase by 15%** with enriched catalogue, optimal price matching, personalization.
- Implemented hierarchical clustering model with ResNet, Graph Transformer for layout agnostic promotion-monitoring to identify relevant product/promotion entity from webpage with metadata & obtained **99.6% recall**, **98.7% precision**.
- Evolved product-matching model Sherlock using hybrid of BERT & ELECTRA transformers, few-shot learning on text attributes for 6 GLs (ebook, music, video, podcast, app, games) & reduced onboarding time from **6 months** to **3 weeks**.
- Engineered real-time tag-generation with BART for catalogue search intelligence with **93%** precision@10 & **0.85 MRR**.
- Created recommendation architecture for MiniTV with NeuMF, BERT4Rec & attained **0.87 NDCG** & **0.89 mAP**.
- Integrated a multi-agent pipeline (Mixtral-8x7B chatbot) with code-generation, Function Calling tools on AWS Lex with **31% increase in customer query resolution speed**, automating **38k** customer support tickets quarterly globally.
- Mentored **51+ engineers** from **7 teams**, established ML vision of DA org & organized biweekly ML learning sessions.
- Contributed **45 business ideas** to chalk out **2024 Science roadmap** & **top 12** got approved by Org VP for 2024.

Oracle

Jan 2019 – Jan 2021

Senior Applied Scientist

Bengaluru, India

- Established vision of Multimodal Large Language Model & crafted 4-quarter roadmap with **25% faster** time-to-delivery.

- Owned 9 Document Understanding models **leading a 16-member team** to deliver customized solution for **42 clients**.
- Spearheaded the operational efficiency drive of OCI by designing org level Knowledge Graphs to automate KTLO oncall triage maintenance, reduce manual efforts & conserve bandwidth of 50+ support engineers in mission-critical projects.
- Contributed to leading project milestones & critical phases of ML life-cycle development across 4 Cerner Health verticals (Relation, Extraction, Assertion, NER) impacting **27k customers globally** & enhancing user retention rate by **18%**.
- Executed 4-quarter strategic roadmap of next-gen automation platform across 5 business units with **25% faster** launch.

Rakuten R&D

Aug 2018 – Jan 2019

Research Scientist

Bangalore, India

- Developed Object Detection Model using YOLOv3 to detect & localize antenna pole structure from video feed data and enhanced business Key Performance Indicator (KPI)- team Revenue growth of **2.4X** & headcount reduction of **90%**.
- Obtained model **confidence score of 0.84** along with IOU (Intersection over Union) of 0.72 & deployed model in GCP.

Cisco Systems

Jan 2018 – July 2018

Software Development Engineer

Bangalore, India

- Succeeded in building a hybrid model using Gaussian Naive Bayes, Gradient Boosting, Random Forest Regression, KNN to predict N/W parameters in Cisco's Webex stream with normalized RMSE of 0.33- **improving availability by 78%**.
- Built Deep Reinforcement Learning model for fault tolerance in Cisco's Smart Router with confidence score of **87.86%**.

Technical Skills

Machine Learning: Classification, Regression, Ensemble (Random Forest, XGBoost, LightGBM, CatBoost), A/B testing, Multi-Armed bandit, Recommender system, Ranking, Probabilistic Graph Models, Reinforcement Learning, Causal inference

Deep Learning: CNN, LSTM, GAN, State Space Models (Mamba), Multimodal Fusion (CLIP, SigLIP), Latent Diffusion Models, Flow Matching, Video Generation (Sora), Graph Neural Networks (GNN/GAT), ASR (Whisper-v3), RLHF (PPO), Direct Policy Optimization (DPO/KTO), Knowledge Distillation, Liquid Neural Networks (LNNs), Autoencoders.

Functional ML Skills: Statistics, Probability, Regularization, Feature Engineering, Hyperparameter Tuning, Optimization

Computer Vision: Object Detection Segmentation (YOLOv3, SSD, Fast/Faster/Mask R-CNN, UNet, EfficientDet, Vision Transformer (ViT)), Pretrained (VGG16/19, InceptionV3, ResNet50, InceptionResNetV2, MobileNetV2, EfficientNetB7)

Python libraries: pandas, numpy, scipy, scikit-learn, statsmodels, optuna, gensim, spacy, Ragas, Polars, Unsloth, Pydantic

ML Tools: PyTorch, TensorFlow, Keras, PySpark, OpenCV, NLTK, JAX, Dask, HuggingFace, LlamaIndex, W&B, vLLM

Natural Language Processing: Word2Vec, fastText; Transformers: BERT, RoBERTa, BART, XLNet, DistilBERT, T5

Large Language Models (LLM): Grok-3, Falcon 3, AWS Nova, Nemotron-4, Cohere Command R+, Perplexity Sonar, SearchGPT, Jamba, Mixture of Experts (MoE), Multimodal Vision (GPT4o, Llama4, Llava-v1.6.), AI Safety (TruthfulQA, HELM, MMLU), Small Language Models (Phi-4, Gemma 3, Qwen 2.5/3), Open-Weight (Mixtral 8x22B, Mistral Large 3)

Frontier & Reasoning Model: OpenAI (o3-pro, o4-mini, GPT-5 (Sol, Luna), GPT-6 Astra), Anthropic Claude (Sonnet 5, Opus 4.8, Fable 5), Google Gemini (Flash/Pro), Meta Llama (Llama 4 Scout/Maverick, Llama 3.3), DeepSeek (R1, v4)

GenAI: Fine Tuning: SFT, IFT, PEFT (LORA, QLORA, DoRA), RLAIIF, RLVR, GRPO, ORPO, SimPO; Agentic RAG, GraphRAG, RAG, test-time compute scaling, prompt engineering (Chain of Thought, Tree of Thought, Few-shot learning).

Google Cloud Platform (GCP): VertexAI, Feature Store, Model Registry, BigQueryML, GKE, Pretrained AI API, TPU

Multi-agent Agentic AI: LangGraph, Langchain, CrewAI, Autogen, AutoGPT, Phidata, Smolagents, Semantic Kernel

Distributed Training Infrastructure: FlashAttention-3, PyTorch FSDP, DeepSpeed ZeRO 1/2/3, Megatron-LM

Retrieval: Hybrid Search (Dense, BM25, HNSW); ColBERT, ColPali, Cross-Encoder Rerankers (BGE, Cohere), Self-RAG

AI Infra & Inference Engines: SGLang, TensorRT-LLM, TensorRT, NVIDIA Triton Inference Server, BentoML, ONNX runtime, TorchServe, Ray Serve, Ollama, quantization (AWQ, GPTQ, FP8/FP4 kernel), PagedAttention, GQA/MQA

VectorDB & AI Platform: Qdrant, Weaviate, Pinecone, ChromaDB, pgvector, Elasticsearch, Cursor, Copilot, Horovod

MLOps, DevOps & Observability: MLflow, Airflow, TensorBoard, Kubeflow, AutoML, FastAPI; Docker, Kubernetes, Helm, Microservices, Git, Bazel, Node.js, Bash, Terraform, Jenkins, Prometheus, Grafana, Splunk, Datadog, ELK stack.

Hardware Compute: NVIDIA H100/B200 GPU Cluster Management, CUDA acceleration, Red-teaming infrastructure

Programming Languages: Python, C++, Java, Scala, C, R, JavaScript, TypeScript, React, Ruby, Rust, Golang, Kotlin

Database: SQL, PostgreSQL, NoSQL (MongoDB, Cassandra, HBase, Redis), Snowflake, Databricks, BigQuery

Distributed System: Delta Lake, Iceberg, Flink, Spark, Kafka, Redpanda, Hadoop, HDFS, YARN, Trino, Presto, DuckDB

AWS: Bedrock, SageMaker JumpStart, Inferentia2, Trainium, Athena, S3, Lambda, EC2, EKS, API Gateway, ElastiCache

Awards & Achievements

- Received **Amazon Employee of the Quarter** award 4 times across Prime Video, Music, Kindle orgs globally. **2024**
- Achieved the **2nd Prize** in ProdCon in entire India for pitching the prototype of my Think Big idea. **Nov, 2022**
- Awarded with the **First Prize** for my winning solution & implementation of tech idea- "Auto Generation of Video from Text" in **Amazon Hackathon'22** out of **300+ employees** & won **Think Big Award** for a business idea **June 2022**
- Won the **Star Performer Award, WOW Award & Budding Star Award** recognition in entire Org. **2021-2023**
- Felicitated with Swami Vivekananda Memorial Rashtriya Purashkar- **State Governor Award** for my Board results.
- Secured the **1st position out of 1 Million students** in **12th Board** exam with **99.2% [Rank- 1]** in entire WB state
- Secured the **4th position out of 1 Million students** in **10th Board** exam with **96.7% [Rank- 4]** in entire WB state
- Received KVPY scholarship & Performed within **top 1%** in National Astronomy, Physics, Chemistry, Maths Olympiads.