Kshitiz Regmi

Atlanta, Georgia, United States | kregmi3@student.gsu.edu | linkedin.com/in/kshitizregmi | github.com/kshitizregmi

Education

Georgia State University — Atlanta, GA
Master of Science (MS) in Computer Science — 04/2027

- GPA: 4.2/4.3
- Relevant Coursework: Distributed AI, Advanced Deep Learning, Digital Image Processing, Advanced Machine Learning, Introduction to Deep Learning

Experience

ALSER Lab, Georgia State University — Atlanta, GA
Graduate Research Assistant — 05/2026 – Present

- Developing a scalable framework for generating synthetic metagenomic FASTQ replicas for benchmarking analysis pipelines.
- Optimizing large-scale genomic data processing on Arctic Server HPC for reproducibility, scalability, and performance.

Georgia State University — Atlanta, GA
Graduate Teaching Assistant — 08/2025 – 04/2026

- Taught and supported Python programming, SQL, Pandas, NumPy, SciPy, Matplotlib, Seaborn, data analysis, visualization, and debugging.

Fusemachines Inc. — Kathmandu, Nepal
Machine Learning Engineer — 03/2020 – 08/2025

- Client: TIME Magazine. Designed the TIME AI platform ML architecture, owned design decisions and product roadmap, and communicated findings to stakeholders via clear reports and presentations.
- Led ML build and deployment of the TIME AI platform beta; increased user engagement by 9% and presented data-driven recommendations based on evaluation metrics.
- Built a production-grade conversational RAG system using embeddings-based retrieval, chunking, reranking, and a vector database; generated answers with citations.
- Implemented multi-turn conversation handling with context tracking and follow-up support using Datastore.
- Performed statistical data analysis on 200K articles to define RAG chunking strategy.
- Benchmarked semantic search retrieval engines by comparing Vertex AI Feature Store with OpenSearch, Pinecone, and FAISS; achieved 89% Recall@5 on 12K+ proprietary benchmark data.
- Built and deployed a shared embeddings and indexing service so RAG, semantic search, recommender system, and Document ChatAI reused the same pipeline; standardized workflows and reduced duplicate engineering.
- Developed Document ChatAI to help editors ask questions and find information across printed articles; collaborated with business teams to validate outputs.
- Designed, developed, and deployed DeepDive topic extraction and topic-based recommendations to help users explore topics and discover similar articles in real time.
- Built an unsupervised machine learning model for user segmentation using clustering on behavioral data; used NumPy, Pandas, scikit-learn, Plotly, and TensorFlow Embedding Projector to visualize embeddings and user similarity.
- Developed the email marketing engine achieving a 43% unique open rate; won the INMA Global Media Award 2023.
- Developed and deployed a recommendation engine that outperformed Google Vertex AI Recommendations and a retail AI baseline, achieving a 1.23% lift in click-through rate (CTR) in production A/B testing.
- Optimized the recommendation engine serving API to keep response time under 100ms while handling over 15 Million requests per month using autoscaling and load balancing.
- Deployed ML services to production on Cloud Run with monitoring for latency and errors, data validation checks, and scheduled retraining to handle data and concept drift.
- Led cross-team collaboration with Google Cloud engineers, data engineers, product leads, and stakeholders; contributed to product roadmaps and OKRs, and wrote architecture documentation.
- Built a multi-step multi-agent system on AWS Bedrock for fraud analytics that converts natural language to SQL.
- Built an LLM-powered agent that maps and transforms client data into a common data model using dbt SQL executed on Trino over Iceberg tables in a lakehouse.

Broadway Infosys — Kathmandu, Nepal
Data Scientist — 08/2022 – 09/2024

- Led Data Science and Machine Learning training for 500+ students, covering SQL, exploratory data analysis, data visualization, feature engineering, and model evaluation.

- Built MLOps project and ML pipelines using Git, Docker, and MLflow for experiment tracking and reproducibility.
- Mentored students on statistical hypothesis testing, hyperparameter tuning, and error analysis.
- Deployed machine learning models as REST APIs using FastAPI and built interactive frontends using Streamlit.

OYA INC Kathmandu, Nepal
Artificial Intelligence Engineer Intern 04/2019 – 09/2019

- Built an ETL pipeline processing 80K+ user-item records to generate model-ready signals for a collaborative filtering recommendation engine.
- Implemented data preprocessing, data cleaning, and model training and deployment pipelines.
- Deployed real-time inference APIs using Flask, moving the system from prototype to production.

Projects

Federated Conformal Prediction for Mitigating Hallucinations in Federated Fine-Tuned Vision-Language Models via Abstention | Python, Fine-tuning, GenAI

- Proposed **Fed-BLEND**, a novel federated conformal prediction method that bins calibration data by predictive entropy and shrinks local per-bin quantiles toward a global anchor using a sample-size-aware weight.
- Deployed on a federated LoRA fine-tuned Qwen2.5-VL-3B model across five non-IID clients; reduced hallucination rate from 13.05% to 3.79% (71% relative drop) while keeping useful-answer rate at 76.63% and Precision@Commit at 95.29%.

Time Series Prediction with GCTAF and Supervised Contrastive Learning | Python, PyTorch, Time-Series, Attention

- Built an attention-based model in PyTorch to forecast 60-step delta-based magnetic field trajectories across 8 SHARP features, outperforming a persistence baseline.
- Separately trained a flare risk classifier using Supervised Contrastive Learning with missing-aware input triplets and focal loss to handle class imbalance and irregular temporal gaps.
- Applied two-stage representation learning — contrastive pretraining followed by linear probing + fine-tuning — achieving TSS value of 0.75 on flare classification.

Multi-model Semantic Image Search Engine
Text to Image, Image to Image, and Text+Image to Image Search, open source

- Developed a semantic multimodel image search engine on Myntra Fashion Product Dataset

Certifications

- Introduction to AI and Machine Learning on Google Cloud | Google Cloud
- Machine Learning Engineering for Production (MLOps) | Coursera
- TensorFlow Developer Certification | Google
- Sequences, Time Series and Prediction | deeplearning.ai
- NLP in TensorFlow | deeplearning.ai
- AI for Medical Prognosis | deeplearning.ai
- Data Science in Stratified Healthcare and Precision Medicine | University of Edinburgh

Mentorship & Service

TreeHacks 2026 | Stanford University, Stanford, CA

- Mentored hackathon teams on agent development, MCP server setup, MVP scoping, and deployment best practices.

Georgia State Undergraduate Research Conference Judge | Georgia State University

- Evaluated undergraduate research presentations using a standardized rubric and provided constructive feedback to student researchers.

NFTE StartUp Tech Showcase Judge | Tucker Middle School, DeKalb County, GA

- Judged student startup pitches and provided feedback on business ideas, creativity, and presentation quality.

Skills

- **GenAI, LLMs, and Agentic AI:** Model Context Protocol (MCP), Agentic System Design, Multi-Agent Systems, Tool Calling, LLM Fine-Tuning, LoRA, PEFT, RAG Pipelines, LLM Evaluation, LangChain, Hugging Face, Gemini, AWS Bedrock
- **Machine Learning and Deep Learning:** PyTorch, TensorFlow, Scikit-Learn, Transformers, CNN, LSTM, Attention Mechanisms, Recommender Systems, Time-Series Forecasting, Model Optimization
- **MLOps, Cloud, and Serving:** GCP, Vertex AI, Cloud Run, BigQuery, AWS, SageMaker, EC2, MLflow, Docker, REST APIs, FastAPI, CI/CD, GitHub Actions, Model Drift Monitoring, Data Drift Monitoring

- **Vector Search and Retrieval:** Vertex AI Feature Store, AWS OpenSearch, Qdrant, Pinecone, FAISS, Semantic Search
- **Data Science and Data Engineering:** Python, SQL, PostgreSQL, Pandas, NumPy, dbt, Apache Iceberg, Trino, Exploratory Data Analysis, A/B Testing, Statistical Analysis