Somayeh Yarahmadi

San Jose, CA | somayeh.yarahmadi@gmail.com | linkedin/syarahmadi | github.com/syarahmadi

Applied Scientist specializing in large language model adaptation, multimodal learning, and retrieval-augmented reasoning

SUMMARY

Applied Scientist (Ph.D., Electrical Engineering) with 7+ years bridging AI research and deployment. Expertise in large language models, multimodal learning, retrieval architectures, and representation learning. Known for rigorous experimentation and model diagnostics—identifying, explaining, and fixing real-world failure modes in complex ML systems. Proven ability to turn research insights into deployable, high-reliability AI that scales across multi-million-document and multimodal datasets.

EXPERIENCE

Senior Applied Scientist, Siemens *Apr 2024 - Present (Remote)*

Lead applied research scientist for retrieval-augmented and multimodal LLM systems used in technical knowledge search and automation. Owned model experimentation, evaluation protocols, and translation of research findings into production systems across business units.

- **Retrieval and Reasoning:** Raised Recall@50 by 12% on multi-document queries through hard-negative mining and adaptive context window tuning. Extended retrieval to semi-structured data (tables and configuration graphs) using graph propagation, cutting off-topic generations by one-third. *Failure mode:* Early retrievers overfit to common query templates; implemented stratified sampling and loss reweighting to recover long-tail precision.
- **Multimodal Learning:** Built and aligned BLIP-2 and EVA-CLIP encoders on 60K manual–diagram pairs, improving diagram search nDCG@20 by 9 percentage points. *Failure mode:* Initial fine-tuning on high-resolution diagrams caused embedding collapse; stabilized training via staged loss weighting and balanced-resolution batches.
- **Evaluation and Deployment:** Ran controlled studies on prompt design and context sensitivity, reducing unsupported claims by 35%. Benchmarked open-weight models (Llama-3-70B, Mistral-7B, Falcon-180B) for retrieval-conditioned fine-tuning, documenting scaling and context compression behavior. *Failure mode:* Context extensions beyond 16K tokens degraded factuality; introduced adaptive context windows and chunk overlap heuristics to maintain recall. Deployed optimized inference with vLLM paged attention caching and FP8 quantization, doubling throughput within latency SLOs. Established reproducible evaluation combining automatic RAG metrics (context precision, groundedness) with expert review for drift monitoring.

Data Scientist II, Power Engineers *Jan 2022 - Mar 2024 (Remote)*

Developed transformer-based forecasting models for grid reliability and outage-risk prediction. Worked cross-functionally to integrate ML forecasts into real-time planning and operations.

- **Forecasting and Modeling:** Improved early-alert accuracy (MAPE -15%) through temporal transformer architectures and feature fusion of weather, topology, and event data. *Failure mode:* Initial models overfit to rare extreme-weather events, generating unstable alerts; applied loss regularization and synthetic oversampling to stabilize predictions and cut false alarms by 20%.
- **Uncertainty and Automation:** Introduced quantile regression and conformal prediction for uncertainty calibration, producing actionable reliability bands. Automated retraining and validation pipelines on AWS to preserve stability under seasonal drift. Authored reports on temporal drift and adaptive fine-tuning that shaped enterprise forecasting standards.

Research Scientist, Power and Energy Center (PEC) *Sep 2018 - Dec 2021*

Conducted ML research on signal processing, simulation modeling, and hybrid data–physics methods for grid reliability. Collaborated with utilities and academic labs on interpretability and robustness.

- Developed interpretable CNN–Bayesian hybrid models for grid event classification; validated simulation-trained models on field telemetry with strong transfer performance. *Failure mode:* Early hybrid models diverged under sensor noise; implemented noise-aware priors and variance regularization to achieve stable convergence.
- Built simulation-to-field validation pipelines, improving fault detection precision and reducing false positives. Published IEEE papers on interpretable ML and hybrid simulation–ML modeling.

EDUCATION

Ph.D., Electrical Engineering — **Virginia Tech** Focus: Signal Processing, Robust Machine Learning, Computer Vision, Simulation Modeling

TECHNICAL AND RESEARCH SKILLS

- **Modeling and Algorithms:** Large Language Models (instruction tuning, retrieval adaptation, long-context scaling), Diffusion and Generative Models, Vision–Language Modeling, Graph Neural Networks, Temporal Transformers, Bayesian and Ensemble Methods, Active Learning, Self-Supervised and Contrastive Representation Learning.
- **Applied Research Methods:** Hypothesis-driven experimentation, Ablation and Sensitivity Studies, Evaluation Design, Statistical Significance Testing, Error Taxonomy Development, Uncertainty Quantification, Explainability and Model Interpretation, Fairness and Bias Diagnostics.
- **Retrieval and Reasoning Systems:** Retrieval-Augmented Generation (RAG), Dense/Sparse/Multi-Vector Retrieval, Cross-Encoder Reranking, GraphRAG, Context Optimization, Hard-Negative Mining, Long-Context Tokenization, Information Grounding Metrics.
- **Multimodal and Representation Learning:** Vision–Language Alignment (CLIP, BLIP, EVA, ColPali), Contrastive Embedding Spaces, Multimodal Pretraining and Finetuning, Feature Fusion for Text, Tables, and Images, Evaluation of Cross-Modal Consistency.
- **Evaluation and Experimentation Infrastructure:** RAG-Triad, Ragas, Human–LLM Evaluation Pipelines, A/B Testing, Continuous Benchmarking, Drift and Regression Analysis, MLflow, Weights and Biases, Model Governance Frameworks.
- **Systems and Deployment:** PyTorch, Hugging Face Transformers, vLLM, TensorRT-LLM, Mixed-Precision Training, Quantization (FP16/FP8), Scalable Serving (AWS SageMaker, EC2, S3), Docker, Linux.
- **Programming and Tools:** Python, C++, SQL, Git, Pandas, NumPy, SciPy, Matplotlib; data curation and annotation pipeline design.
- **Research and Collaboration:** Experiment Design Documentation, Reproducibility Practices, Mentorship of Junior Scientists, Cross-Functional Communication, Technical Writing for Research and Product Alignment.