Yi Xiang

elaine.yi.xiang@gmail.com | www.yixiangresearch.com

Senior Applied Scientist at AWS Bedrock developing representation-learning methods, contrastive-learning objectives, benchmarks, and evaluation frameworks for foundation-model systems. Core contributor to Amazon Titan Text Embeddings models and multimodal retrieval for Amazon Bedrock Knowledge Bases, with research spanning multilingual embedding models, contrastive learning, embedding fine-tuning and compression, multimodal retrieval, adaptive agent memory and stateful context management. ICLR 2025 Spotlight author and inventor on patents in representation learning and agentic AI, with a track record of translating model research into launched production capabilities.

EXPERIENCE

Amazon Web Services

Amazon Bedrock | Senior Applied Scientist | New York City | May 2023 – Present

- **Amazon Titan Text Embedding Models:** Led development of multilingual embedding capabilities covering 100+ languages; conducted contrastive-learning research and systematic ablation studies, achieving a 60% improvement over TTE v1 and outperforming evaluated OpenAI and Cohere multilingual embedding baselines.
- **Embedding Model Customization Platform:** Designed and developed an end-to-end platform for adapting embedding models from raw text, positive-pair, or triplet inputs, including synthetic training and evaluation data generation, fine-tuning, automated hyperparameter search, evaluation, and embedding-space visualization; improved retrieval performance by 25.7% on average across nine datasets.
- **Novel Contrastive Learning Objectives:** Developed a false-negative-aware contrastive loss for domain-specific embedding adaptation, delivering a 51% performance improvement for Amazon Bedrock Guardrails. Established experimentally validated fine-tuning configurations for production use.
- **Embedding Compression Research:** Developed post-training embedding compression methods through temperature-controlled contrastive training, enabling more storage-efficient retrieval representations while preserving model quality; published as an ICLR 2025 Spotlight.
- **Multimodal Retrieval:** Developed the modeling and evaluation approaches underlying multimodal retrieval over video and audio, including a text-auxiliary approach using scene summaries and audio transcripts and a parser-free approach using multimodal embeddings and multimodal reranking. Designed five benchmark datasets and conducted systematic ablation studies, achieving a 20% improvement on action-based datasets; the resulting capability launched as multimodal retrieval for Amazon Bedrock Knowledge Bases at re:Invent 2025.
- **MultiKB Agentic RAG:** Developed adaptive planning, context-management and evaluator-guided validation methods for an agent answering complex questions across structured and unstructured knowledge sources. Demonstrated through systematic ablations up to a 94% recall@5 improvement over plain RAG while reducing latency by 25% on multi-source QA benchmarks.
- **Adaptive Working Memory Management:** Developed a provenance-preserving working-memory approach for foundation-model agents that adaptively selects, compacts and curates task-relevant context while maintaining lossless source tracking, enabling efficient long-horizon agent operation without sacrificing traceability.
- **Scalable Experimentation and Evaluation:** Developed inference and evaluation capabilities supporting large-scale embedding and retrieval research, improving inference speed by 30-40% through bf16 optimization and dimension reduction and reducing evaluation runtime from four hours to 30 minutes.

AWS Machine Learning Solutions Lab | Applied Scientist | New York | Jan 2020 – May 2023

- **Applied Machine Learning Research:** Developed and deployed 10+ machine learning systems across healthcare, gaming, sports, finance and recruiting, generating $20M+ ARR and seven public customer references.
- **Selected Models:** Developed a transformer-based toxicity detection model for Epic Games; a three-tower recommendation model for Talent.com; a coordinate-regression deep neural network for catheter-tip localization in fluoroscopic images with Weill Cornell Medicine; and variational-autoencoder methods for anomaly detection.

Invesco Asset Management

Data Scientist | New York City | Mar 2018 – Jan 2020

Developed forecasting, recommendation, and customer-segmentation models using deep learning and NLP, and designed large-scale AWS/PySpark data pipelines that reduced reporting time from one week to two hours.

PUBLICATIONS

- Effective post-training embedding compression via temperature control in contrastive training, ICLR Spotlight, 2025
- Towards building a robust toxicity predictor, ACL, 2023
- A Coordinate-Regression-Based Deep Learning Model for Catheter Detection during Structural Heart Interventions, Applied Sciences, 2023

TECHNICAL WRITING AND TALKS

- Building an NLP-based job recommender at Talent.com with Amazon SageMaker, AWS AI Blogs, 2023
- Streamlining ETL data processing at Talent.com with Amazon SageMaker, AWS AI Blogs, 2023
- Build a Robust Text-based toxicity predictor, AWS ML Blog, 2022
- Deploy variational autoencoders for anomaly detection with TensorFlow Serving on Amazon SageMaker, AWS ML Blog, 2021
- "Automatic Prompt Optimization" Talk, AWS NLP Summit, 2023.
- "Robust and fast detection of toxic speech content via machine learning" Talk, ICML Expo Demonstration, 2022 | ACVC content moderation conference, 2022 | AMLC Tutorial, 2022 | AMLC content moderation workshop Oral Talk, 2022.

EDUCATION

Master of Mathematics in Finance, Columbia University | New York, NY | 2016 – 2018

B.Ec. in Financial Engineering, University of Science and Technology Beijing | Beijing, China | 2012 - 2016 | Academic Special Prize Scholarship, awarded to top 1%