Yi Xiang
elaine.yi.xiang@gmail.com

Senior Applied Scientist at AWS Bedrock building production-grade ML systems across representation learning, multimodal RAG, and agentic AI. Core contributor to Amazon Titan Text Embedding, multimodal RAG (launched at Re:Invent 2025), and agentic AI systems including Multi-KB RAG Agents and Agent Memory Systems. Previously led and delivered 10+ enterprise AI solutions at AWS ML Solutions Lab, driving $20M+ revenue and 7 public customer references. Publishes at ICLR, speaks at ICML and AMLC, writes for the AWS Machine Learning Blog, and holds patents in representation learning and agentic AI systems.

EXPERIENCE

Amazon Web Services

Amazon Bedrock | Senior Applied Scientist | May 2023 – Present

core ML & research contributions

- Model Training: Key contributor to Amazon Titan Text Embedding (TTE) models. Led development of multilingual capabilities covering 100+ languages for TTE v2. Conducted systematic research on contrastive learning and designed comprehensive ablation studies. Achieved state-of-the-art results outperforming OpenAI and Cohere multilingual models with a 60% improvement over TTE v1. TTE is the most widely adopted text embedding model by AWS customers.
- MultiKB Agentic RAG: Led the design and implementation of a MultiKB RAG agent capable of answering complex queries across multiple vector stores with both structured and unstructured data. Achieved up to 94% improvement in recall@5 over plain RAG on QA benchmarks through advanced context management, dynamic query planning, and self-evaluation features, validated by comprehensive ablation studies.
- Embedding Fine-tuning: Led the fine-tuning development for Titan Text Embedding Models, resulting in a 25.7% average performance improvement across nine retrieval datasets. Established production-ready parameter specifications through rigorous experimentations. Innovated a false negative–aware contrastive loss that delivered a 51% performance improvement for Amazon Bedrock Guardrail.
- Video RAG: Developed video/audio RAG capabilities using text-auxiliary and parser-free multimodal embedding, leading to a 20% performance gain on action-based datasets. Enhanced system with Universal Multi-modal Reranker integration and creation of five benchmark datasets.
- Model Infrastructure: Redesigned a unified inference package with enhanced scalability, added dimension reduction, and achieved a 30–40% speed improvement through bf16 precision optimization. Streamlined the evaluation pipeline, reducing processing time from 4 hours to 30 minutes.

Leadership & Scope

- Technical Leadership: Owned technical direction for complex machine learning systems, defining end-to-end architecture, modeling strategy, and implementation approach from ambiguous problem statements through production readiness. Led discovery and design phases to translate high-level business objectives into scalable, extensible technical roadmaps, making principled trade-offs grounded in empirical evaluation and system constraints.
- Stakeholder & Execution Leadership: Drove execution across cross-functional teams by aligning researchers, engineers, and leadership around clear technical priorities and delivery milestones. Maintained alignment via regular progress updates, executive briefings, and technical deep dives, ensuring timely delivery of production-quality systems.

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

- ML Model Development: Led, designed, and built production machine-learning systems for AWS enterprise customers in healthcare, gaming, sports, and finance industries, generating over $20M in annual recurring revenue (ARR) with seven public customer references.
- Representative deployments included: a transformer-based toxicity detection model for Epic Games; a triple-tower deep pointwise recommender for Talent.com; a DNN for cardiovascular catheter tip detection at Weill Cornell Medicine; a variational autoencoder for anomaly detection; a named-entity recognition (NER) model for insurance filings; and a golf goal-prediction model for the PGA Tour.

Invesco Asset Management

Data Scientist | March 2018 – Jan 2020

- Developed machine learning solutions for sales and marketing enablement using deep learning and NLP, including fund flow forecasting (XGBoost, LSTM), product recommendation, and customer segmentation.
- Applied time-series signal processing techniques to detect inflow, outflow, and reallocation patterns across firms, products, territories, and financial advisers. Utilized Redshift and PySpark on EMR for data processing.
- Designed and implemented large-scale ETL processing for petabytes of data and data quality pipelines with AWS and Airflow, reducing reporting time from one week to two hours.

PUBLICATIONS

- Effective post-training embedding compression via temperature control in contrastive training, ICLR, 2025[Link]
- Building an NLP-based job recommender at Talent.com with Amazon SageMaker, AWS AI Blogs, 2023 [Link]
- Towards building a robust toxicity predictor, ACL, 2023[Link]
- A Coordinate-Regression-Based Deep Learning Model for Catheter Detection during Structural Heart Interventions, Applied Sciences, 2023 [Link]
- Streamlining ETL data processing at Talent.com with Amazon SageMaker, AWS AI Blogs, 2023 [Link]
- Build a Robust Text-based toxicity predictor, AWS ML Blog, 2022 [Link]
- Deploy variational autoencoders for anomaly detection with TensorFlow Serving on Amazon SageMaker, AWS ML Blog, 2021 [Link]

TALKS

- “Automatic Prompt Optimization” AWS NLP Summit, 2023.
- “Robust and fast detection of toxic speech content via machine learning” ICML Expo Demonstration, 2022. [Link] | ACVC content moderation conference, 2022 | AMLC Tutorial, 2022 | AMLC content moderation workshop Oral Talk, 2022.

EDUCATION

Master of Mathematics in Finance, Columbia University | 2016–2018
B.Ec. in Financial Engineering, University of Science and Technology Beijing | 2012-2016

OTHERS

- Dedicated Community Builder: Founded the first virtual coffee chat program in the organization in May 2020 during the pandemic, with over 50 individuals participating globally and more than 100 coffee chats scheduled.
- Public speaking: Passionate about communicating complex technical ideas to diverse audiences. Delivered multiple tech talks at conferences such as ICML and AMLC, with audiences exceeding 600 attendees.