Jasper Tan

JasperTan@utexas.edu | (908) 838-5638 | Personal Website | LinkedIn | GitHub

EDUCATION

The University of Texas at Austin May 2026
Master of Science in Electrical and Computer Engineering; Concentration: Machine Learning/Data Science GPA: 4.0/4.0

The University of Texas at Austin May 2025
Bachelor of Science in Electrical and Computer Engineering; Minor in Business Administration GPA: 3.85/4.0

Coursework: Reinforcement Learning, ML on Real-World Networks, Computer Vision, Generative Models in ML, Database Management, Data Science Laboratory, Software I & II, Computer Architecture, OS, Matrices, Probability, Algorithms

TECHNICAL SKILLS

Programming Languages: Python, SQL, C/C++, Java, Bash/Shell
Frameworks & Tools: PyTorch, SentenceTransformers, Diffusers, PEFT, Scikit-Learn, Weights & Biases, Streamlit
Technologies: LoRA/QLoRA, Snowflake, AWS (S3, EC2, SageMaker), Azure, Jenkins, Git, Linux, PostgreSQL, MongoDB, JIRA

EXPERIENCE

AMD - *ML Infrastructure Intern*; Austin, TX May 2025 – Present
- Engineering a scalable semantic search using OpenAI embeddings and rerankers to match documents with 200K+ tokens
- Prototyping contrastive learning, chunking, and fine-tuning strategies to improve long-form embeddings for hardware logs
- Reducing ETL latency by 97% on 1.6M-row datasets by optimizing Snowflake ingestion with Apache Arrow and Parquet files
- Automating log ingestion from MongoDB and deploying interactive Streamlit dashboards to enable AI-driven diagnostics

The University of Texas at Austin - *Machine Learning Researcher*; Austin, TX Aug 2023 – May 2025
- Advanced a novel human activity classification model incorporating feature fusion with real-time acoustic and inertial data
- Adapted a MobileNet V2 architecture via transfer learning for feature extraction and fine-tuning based on IMU data
- Deployed a lightweight computer vision model for object localization on edge devices, optimizing inference via LiteRT

Cvent - *Software Engineer Intern*; Tysons Corner, VA Jun 2024 – Aug 2024
- Designed date-time modals in JavaScript with optimized state management using Redux, React Hooks, and mutations
- Implemented real-time data visualization and logging with Datadog to monitor GraphQL queries to a PostgreSQL database

FirstParty - *Applied ML Intern*; New York, NY Jun 2023 – Jun 2024
- Developed Python pipelines in SageMaker using GPT embeddings to compute string similarities across unstructured data
- Leveraged cosine similarities and Levenshtein distance algorithms to generate confidence scores for data stored in S3
- Engineered ML-based data classification pipelines using natural language processing, achieving a 95% accuracy rate
- Employed object-oriented programming to design automated data ingestion apps for 100,000+ rows of web-scraped data

PROJECTS

Chatbot-Enhanced Recommender System May 2025
- Built a conversational recommender integrating LLMs with DeepFM and BERT encoders to improve inclusion@10 by ~20%
- Fine-tuned Gemma, Llama 2, and Mistral using LoRA on MovieLens 20M to simulate user queries and generate dialogue
- Trained a dual-encoder model (BERT + item embeddings) with triplet loss and VICReg to align conversations with items

Graph Reinforcement Learning for Semantic Segmentation Jan 2025
- Implemented graph-based reinforcement learning to improve semantic segmentation of 2D images and 3D point clouds
- Developed graph convolutional networks with dueling Deep Q-learning, optimizing node classification and navigation
- Designed large-scale graph environments handling 30,000+ nodes using k-nearest neighbor for 2D and 3D data

Fashion-Atlas May 2024
- Devised a garment re-identification application aimed at localizing clothes from images to give tailored recommendations
- Leveraged YOLOv8 to train a real-time object detection and classification neural network to crop and identify images
- Trained a CNN on a ResNet 50 architecture with a triplet loss function and cosine similarity to generate feature embeddings

Hindsight Experience Replay for Diffusion Models (HERD) May 2024
- Fine-tuned a text-to-image diffusion model (Stable Diffusion) using Reinforcement Learning to generate prompted images
- Built a distributed training pipeline using Transformer RL, image reward, and policy gradient methods (DDPO, DPOK, DDPG)

RationaLlama May 2024
- Fine-tuned an instruction-tuned Llama 2 using QLoRA to solve complex rational NLI tasks from the LogicQA dataset
- Employed 4-bit quantization with Bits and Bytes to minimize compute resources and achieved an 8% increase in accuracy