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: 3.9/4.0

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: DeepRL, Grounded NLP, Systems in Generative AI, Reinforcement Learning, Generative Models, 3D Deep Learning, ML on Networks, Computer Vision, DB Management, OS, Probability and Stochastic Processes, Matrices, Algorithms

TECHNICAL SKILLS

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

EXPERIENCE

Synthefy - *Machine Learning Engineer*; Austin, TX Sep 2025 – Present
- Architected a foundational Patched Diffusion Transformer with In-Context Learning to synthesize physiological signals
- Optimized transformer throughput using FlashAttention, RoPE, and SwiGLU to efficiently process long-context sequences
- Reduced downstream prediction error by 12% via a TRSTR (train on real and synthetic, test on real) framework
- Engineered an ETL pipeline using Hydra and Pytorch Lightning to unify, normalize, and mask 20+ clinical datasets

AMD - *Machine Learning Infrastructure Intern*; Austin, TX May 2025 – Aug 2025
- Built a scalable semantic retrieval pipeline using OpenAI embeddings to match 100+ hardware logs via vector similarity
- Compressed 200k-token documents by 97% using regex, chunking, and fingerprinting to enable long-context embeddings
- Improved ETL speed by 45x on 1.6M-row datasets via optimized Snowflake ingestion with Parquet and Apache Arrow
- Automated log ingestion from MongoDB and deployed 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 Machine Learning Intern*; New York, NY Jun 2023 – Jun 2024
- Built embedding-based similarity pipelines in SageMaker using GPT embeddings and cosine similarity for structured data
- Designed NLP-based classification, achieving 95% accuracy via Levenshtein distance algorithms and alignment models

RESEARCH & PROJECTS

FluxServe - dLLM Inference Optimization Dec 2025
- Characterized LLaDA-8B architecture on A100s, proving compute-bound nature to justify aggressive dynamic batching
- Developed a trace-driven simulator to model diffusion transformer behavior, validating results against hardware execution
- Achieved 5.5x latency reduction by identifying an exploiting sub-linear compute scaling in diffusion-based LLMs

Chatbot-Enhanced Recommender System May 2025
- Built a conversational recommender system integrating DeepFM and BERT encoders to improve HitRate@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

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)