HUILIN TAI

(646) 866-9171 | Houston, TX | ht62@rice.edu
www.linkedin.com/in/huilintai | haleyyy2001.github.io

EDUCATION

Rice University — Houston, TX, United States
Ph.D. in Computer Science (Incoming) — Aug 2026 – Present
Columbia University — New York, United States
Master of Science, Computer Science (Thesis-Based) (GPA 4.0) — Sep 2024 - Dec 2025
McGill University — Montreal, QC, Canada
Bachelor of Science, Statistics and Computer Science (GPA 3.86) — Sep 2020 - May 2024

RESEARCH INTEREST

My research focuses on efficient and interpretable large language models, with an emphasis on long-context inference, scalable deployment, and reliable evaluation. I study computationally efficient generation techniques such as speculative decoding, cache and memory optimization, and streaming inference, as well as representation diagnostics for layer probing and robustness under distribution shift. I am also interested in multimodal language models and their applications to biomedical and clinical data, where reliability, interpretability, and efficient integration of multimodal signals are critical for real-world deployment.

PUBLICATIONS

[1]Huilin Tai. "Cross-Species Antimicrobial Resistance Prediction from Genomic Foundation Models." Master's Thesis, Columbia University, 2026. arXiv:2603.11141.
[2] Huilin Tai, Qian Li, Jingtao Wang, Jiahui Tan, Ryann Lang, Basil J. Petrof, Jun Ding. "CellSexID: Sex-Based Computational Tracking of Cellular Origins in Chimeric Models." Cell Reports Methods (2025).
[3] Mingxiao Huo, Jiayi Zhang, Hewei Wang, Jinfeng Xu, Zheyu Chen, **Huilin Tai**, Ian Yijun Chen. "Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding." TTODLer Workshop at ICML 2025 (2025).
[4] Pengliang Ji, Chuyang Xiao, **Huilin Tai**, Mingxiao Huo. "T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation." CVPR 2024 Workshop (2024).
[5] Adam M.R. Groh, Nina Caporicci-Dinucci, Brianna Lu, Maxime Bigotte, Elia Afanasiev, Joshua Gertsvolf, Dale J. Hatrock, Victoria Mamane, Sienna Drake, **Huilin Tai**, Jun Ding, Alyson Fournier, Catherine Larochelle, Jo Anne Stratton. "Ependymal cells undergo an astrocyte-like gliosis in response to chronic and acute neuroinflammation." Journal of Neurochemistry (2024).
[6] Xiaorong Guo, **Huilin Tai**, Xiaoqing Li, Peng Liu, Jin Liu, Shan Yu. "SPARC is a novelty prognostic biomarker for ovarian cancer and associated with immune signatures and drug response." Clinical and Experimental Obstetrics & Gynecology (2024).
[7] Dehai Wu, Congyi Zhang, Guanqun Liao, Kaiming Leng, Bowen Dong, Yang Yu, **Huilin Tai**, Lining Huang, Feng Luo, Bin Zhang, Tiexiang Zhan, Qiuhui Hu, Sheng Tai. "Targeting uridine-cytidine kinase 2 induced cell cycle arrest through dual mechanism and could improve the immune response of hepatocellular carcinoma." Cellular & Molecular Biology Letters, 105 (2022).

RESEARCH EXPERIENCE

Graduate Researcher — New York, United States
Columbia University (Supervised by Professor Mohammed AlQuraishi) — Aug 2024 - Dec 2025

Genome LLM Development & Evaluation

- Long-context LLM runtime (Evo, ~1B parameters): Built a multi-GPU distributed decoding runtime with block-streaming I/O, mixed precision, and MLflow experiment tracking. Reduced peak memory ~5× through block compression, fp16 precision, and HDF5 streaming, enabling inference on 5M+ token contexts.
- Interpretable genome embeddings: Modeled genome token embeddings as multivariate time-series signals and applied PCA + MiniRocket to capture local mechanistic patterns across millions of tokens. Integrated FAISS-based k-NN probes to identify prototype neighbors and quantify distribution shift, improving interpretability in long-context settings.
- Layer diagnostics for stable feature extraction: Performed representation diagnostics across 32 transformer layers (isotropy, effective rank, activation scale, attention sinks). Identified a stability boundary around Layers 10–11 and established a standardized layer-selection protocol for downstream tasks.

- OOD benchmarking for AMR prediction: Designed group-aware evaluation splits (leave-species-out and clustered splits), demonstrating that random cross-validation substantially overestimates model performance. Released a leakage-resistant benchmarking framework with reusable evaluation scripts.

Visiting Scholar New York, United States
The Feinstein Institutes May 2024 - Aug 2025

Spatial Transcriptomics Toolkit — Foundation Model Evaluation & Immune Profiling

- Built zero-training spatial transcriptomics pipeline integrating frozen pathology foundation models (Virchow2 ViT embeddings, PLIP text-image retrieval) with dual immune scoring methods (ssGSEA + mean-Z) to quantify morphology-expression coupling, achieving >0.7 inter-method agreement on 33-gene immune panel
- Benchmarked spatial clustering algorithms (SpaGCN, Leiden, BANKSY, COVET) for liver immune microenvironment segmentation, implementing automated confidence flagging and interactive visualizations with 0.85 silhouette score for immune niche delineation

Research Assistant
Carnegie Mellon University Aug 2024 - Present

T2VBench: Benchmarking Temporal Dynamics for Text-to-Video Generation

- Contributed to the computational backbone of the evaluation pipeline, integrating VLMs (LLaVA-1.5, InstructBLIP) with automated metrics (CLIPScore, BLIPScore, VQAScore) to benchmark temporal coherence.
- Ran large-scale experiments by processing 1,600+ prompts and 5,000+ videos, enabling correlation analysis between model outputs and human preference scores.

Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding

- Validated tree-based speculative decoding on LLaVA-1.5 with 3.28× end-to-end speedup at matched accuracy across diverse VLM benchmarks; implemented acceptance-threshold tuning and fallback logic.
- Contributed to pipeline optimization and evaluation scripts for draft vs. target models, ensuring reproducible performance testing.

Research Assistant Montreal, Canada
Mila - Quebec AI Institute (Supervised by Professor Jun Ding) Dec 2022 - May 2024

CellSexID: Ensemble Learning Framework for Biological Classification

- Developed CellSexID, a Python/CLI package for cell sex prediction across human and mouse samples (including transplant models), integrating multiple classifiers (XGBoost, SVM, Random Forest, Logistic Regression) as an ensemble method with cross-validation achieving 96% AUROC/AUPRC
- Open-sourced CellSexID with modular data loaders, containerized workflows, and comprehensive tutorials. Released Python/CLI APIs enabling reproducible quality control, demultiplexing, and donor-recipient tracking for research applications

Large-Scale Multi-Institutional Data Integration and Analysis

- Collaborated with clinical teams across three institutions (Lady Davis Institute, CHUM, MUHC) developing data integration methodology for 20TB+ datasets. Implemented standardized protocols and statistical frameworks with automated quality control to address batch effects
- Applied unsupervised learning (PCA, UMAP) to single-cell data, validating clinical research hypotheses about cellular subpopulations.

Deep Learning Model for Spatial Gene Expression Prediction

- Designed a Variational Autoencoder with Stochastic Variational Inference in PyTorch, enabling scalable learning for spatial transcriptomics data across diverse tissue types.
- Engineered custom Evidence Lower Bound and shift sigmoid transformation functions, enhancing model convergence and interpretability of latent space representations.

Research Assistant Montreal, Canada
McGill University (Supervised by Professor Hamed Hatami) Sep 2023 - Apr 2024

Randomized Group-Testing Algorithm Design for Hamming Distance Communication Protocol

- Devised a protocol based on a group testing algorithm to estimate Hamming distance between two n-bit strings and reduced upper bound of communication complexity from O(log n) to O(log log n).

- Adopted adaptive testing algorithm to improve algorithm efficiency, focusing on randomized algorithms for Hamming distance estimation.

Theoretical Bounds on Excess-Error Replicability in Agnostic Learning

- Investigated excess-error dependent replicability in agnostic learning to identify conditions that exhibit excess-error dependency.
- Designed an algorithm for covering hypothesis classes with finite VC dimensions, achieving accurate approximation of real error rates using empirical data.

Research Assistant Montreal, Canada
McGill University (Supervised by Professor Anmar Khadra) Nov 2022 - Jan 2024

Computational Modeling of Nanoparticle Binding Dynamics

- Implemented probabilistic models using Markov Chain Monte Carlo (MCMC) simulations in MATLAB to model nanoparticle binding dynamics, achieving strong experimental alignment through parameter optimization and serial engagement modeling
- Developed comprehensive mathematical framework employing Poisson and Rayleigh distributions to investigate 20+ selective binding phenomena, creating MATLAB-driven computational pipeline with randomization algorithms for binding capacity calculation and surface distribution probability visualization

WORK EXPERIENCE

Statistician Intern Remote
Harbin Medical University May 2021 - May 2022

- Performed survival analysis implementing statistical modeling in R, developing interactive nomograms for multi-year prognosis prediction and conducting Kaplan-Meier analysis with log-rank validation across hepatocellular carcinoma and ovarian cancer studies
- Implemented Gene Set Enrichment Analysis (GSEA) framework processing tumor risk stratification data, systematically identifying and ranking pathway enrichment patterns within KEGG and Hallmark databases using R-based analytical methods

Data Scientist Intern Shanghai, China
Yooden Technology Oct 2020 - Feb 2021

Real-Time Energy Optimization Pipeline for Data Center Robotics

- Engineered a streaming data pipeline for energy-saving robots in data centers, ingesting multi-sensor streams with SQL-based storage, ETL transformations, and anomaly detection logic, supporting real-time monitoring of power consumption.
- Contributed to interactive dashboards and automated load-balancing alerts using Python (Pandas, NumPy, Matplotlib); integrated scheduling workflows to enable proactive energy optimization and improved system reliability.

TEACHING EXPERIENCE

Graduate Research Mentor Remote
King Abdullah University of Science and Technology May 2024 - August 2024

- Led weekly briefings on 10+ recent papers in video editing and diffusion-model disentanglement, synthesizing object-level disentanglement methods and evaluation protocols for group discussion.
- Designed generative model architecture combining state-of-the-art video editing frameworks with Beta-VAE for disentangled representation learning, implementing custom loss functions and training protocols in PyTorch.

Teaching Assistant Montreal, Canada
McGill University Sep 2022 - May 2024

- Supported instruction across four courses (Math235 Algebra, Math240 Discrete Math, Math308 Statistical Learning, Math356 Honors Probability), collaborating with faculties to deliver consistent educational experience through grading, office hours, and student mentoring.
- Led statistical learning discussions in Direct Reading Program, guiding students through Bayesian inference and Generalized Linear Models applications, with emphasis on survival analysis methodology and practical data analysis challenges.

HONORS

4th Place — VideoVista Video Understanding & Reasoning Evaluation Competition (HITSZ-TMG) Jan. 2026

Mackey-Glass Research Bursary, Issued by McGill Faculty of Medicine Apr. 2023
Hugh Brock Scholarship, Issued by McGill University Sep. 2020