

Junwei Quan

Phone: +86 15618378093 +1 4372208626 **Email:** junwei.quan@mail.utoronto.ca

LinkedIn: linkedin.com/in/jwquan0102 **GitHub:** github.com/junwei0102 **Google Scholar**

Education

- University of Toronto** **2026.09 - 2027.12** **MScAC in Applied Computing - Artificial Intelligence**
- **[Honors]** Ontario Vector Institute AI Master's Entrance Scholarship.
- University of Toronto** **2022.09 - 2026.06** **BSc: Computer Science Specialist (AI Focus), Data Science Specialist, Mathematics Minor** **CGPA: 3.91/4.00**
- **[Honors]** 2025 NSERC USRA, Canada's national Natural Sciences and Engineering Research Council research award; Dean's Nomination, Faculty of Arts & Science, University of Toronto.

Research Internships

Long-Horizon Planning for Offline Goal-Conditioned RL via Test-Time Graph Search **Research Assistant** **2025.03 - 2026.04**

- **[Algorithm Innovation]** Proposed **Test-Time Graph Search (TTGS)** to address value-error accumulation and one-shot policy execution failures in Offline Goal-Conditioned RL: sampled states from offline datasets to build weighted graphs, transformed pre-trained goal-conditioned value functions into estimated distance signals, and used Dijkstra search to generate executable subgoal sequences, enabling fixed policies to perform long-horizon task decomposition and planning without additional training.
- **[System Implementation]** Built an end-to-end experimental pipeline with CUDA-enabled JAX, OGBench, and MuJoCo; integrated GCIQL, HIQL, QRL and other base agents; wrapped offline training, graph construction, distance matrix computation, subgoal selection, whole-task evaluation, and WandB visualization; optimized JAX memory usage and intermediate-result caching for large-scale value evaluation, and managed cluster training/evaluation jobs with SLURM.
- **[Experimental Evaluation]** Designed comparisons between base policies and TTGS across graph sampling strategies, distance signals, and distance-penalty thresholds; systematically analyzed fixed-policy failure cases in maze/locomotion long-horizon tasks; validated lightweight test-time planning on OGBench, e.g., on pointmaze-giant-stitch-v0, HIQL success rate improved from 0.0% to 80.9% and GCIQL from 0.0% to 98.0%; co-first-authored paper published in **ICML 2026**.

rh-bench: LLM Agent Safety Benchmark for Reward Hacking Detection **Research Assistant** **2026.05 - Present**

- **[Problem Modeling]** Built **rh-bench**, a unified benchmark for reward hacking detection, targeting reward hacking, evaluation gaming, reward tampering, and deployment-time monitoring evasion in LLM agents; abstracted failures that satisfy proxy metrics while deviating from true user intent as a binary detection task, and organized them into six hacking taxonomies for systematic safety evaluation across scenarios and data sources.
- **[System Implementation]** Designed and implemented a unified data schema for rh-bench, standardizing multiple public reward hacking data sources into evaluable samples; preserved clean-vs-hack contrast pairs and multi-turn tool-use agentic trajectories to improve dataset traceability, scalability, and deployment relevance.
- **[Evaluation Pipeline]** Built a zero-shot **LLM-as-Judge** detection pipeline; used unified judging prompts, structured outputs, and risk probability scores to evaluate different off-the-shelf LLMs as reward hacking detectors; performed stratified analysis by behavior type, data source, sample construction method, and multi-turn tool usage. Experiments show the best model reaches about 93–94% overall accuracy and AUC-ROC above 0.96, but accuracy drops by 9–13 percentage points on agentic trajectories, indicating that multi-turn tool-use remains a deployment monitoring challenge.

MDM-Prime-v2: Subtoken Representation and Scalable Training for Diffusion Language Models **NSERC USRA Research Intern** **2025.05 - 2025.12**

- **[Algorithm Innovation]** Contributed to **MDM-Prime-v2** to address limited use of internal token structure and unclear subtoken granularity choices in masked diffusion language models under BPE tokenization: introduced binary encoding and index shuffling on top of MDM-Prime's partial-mask generation framework to increase subtoken information content, provide more stable design guidance for token granularity, and improve diffusion language model training and generation.
- **[System Implementation]** Built and maintained generative-model training and evaluation pipelines for multiple tokenizer/subtokenizer configurations, model scales, and large-scale training reproduction; managed training jobs on CUDA GPU clusters with SLURM, and used Docker to standardize dependencies, runtime environments, configuration files, logs, and checkpoint management, reducing experimental variance across machines and dependency versions.
- **[Experimental Evaluation]** Designed systematic comparisons across subtoken granularity, encoding methods, index-shuffling strategies, hyperparameters, and training stability, supporting 170M ablations, 1.1B-scale training, and generation-quality evaluation; two core techniques improved 170M average accuracy on 8 zero-shot commonsense reasoning tasks from 40.02% to 43.76% (+3.74 points), while the 1.1B model reached 49.42% after 540B tokens and reduced generation perplexity from SMDM's 77.24 to 53.06 at 1024 sampling steps; paper submitted to **EMNLP 2026**.

Publications and Submissions

- **TTGS**, co-first author, **ICML 2026**; **MDM-Prime-v2**, third author, **EMNLP 2026 submitted**; **rh-bench**, co-first author, **AAAI 2026 planned submission**.
- Additional ICSE 2025 Demo and MO2RE 2025 software engineering papers.

Skills

- **AI Training Systems:** PyTorch, JAX, CUDA, SLURM, Docker, WandB; GPU cluster training, experiment environment packaging, memory optimization, checkpoint/log management, and reproducible experimental pipeline construction.
- **RL / Robotics:** Offline RL, Goal-Conditioned RL, Planning, Control, MuJoCo / OGBench / D4RL, World Model, VLA, IL.
- **LLM / Generative Models / Agent Safety:** Transformer, discrete diffusion language models, LLM Agent, LLM-as-Judge, Reward Model, RLHF / DPO / PPO / GRPO basics, reward hacking.