

Jiarui (Jerry) Yan

Jerry2@cs.cmu.edu | (206) 840-8166 | [linkedin.com/in/JerryY](https://www.linkedin.com/in/JerryY) | github.com/JerryYan123

EDUCATION

Carnegie Mellon University, School of Computer Science

Pittsburgh, PA

Master of Science in Computational Data Science GPA: 3.8/4.0

Expected 05/2027

Coursework: Large Language Models, Advanced NLP, Multimodal Machine Learning, Deep Learning, Cloud Computing, Distributed Systems

University of Washington, Information School

Seattle, WA

Bachelor of Science in Informatics: Data Science Major, Applied Mathematics Minor

09/2021 – 06/2025

GPA: 3.94/4.0 | Dean's List (all terms) | Magna Cum Laude (top 3.5%)

Coursework: Machine Learning, Natural Language Processing, Numerical Analysis, Probability I & II, Linear Algebra, Methods in Data Science

RESEARCH INTEREST

LLM pre-training data, post-training, reinforcement learning, agentic systems, agent planning

PUBLICATIONS & CONFERENCE PRESENTATIONS

[1] Aditya Kumar*, Zhihan Lei*, **Jiarui Yan***, Joshua W. Momo, Lauhitya Reddy, Cassandra A. Cohen, Arthur Kajiya, and William W. Cohen. "Learning to Construct Practical Agentic Systems." Under review.

[2] **Jiarui Yan**, Weiwei Sun, Sijie Li, Wenhan Li, Yiming Yang. "TraceML: An Empirical Analysis of Human-Agent Planning in Machine Learning Development." Accepted at the WAB workshop at **COLM 2026**; full paper under review.

[3] **Jiarui Yan***, Zichun Yu*, Shlok Sanghvi, Chenyan Xiong. "HTML2Text: A Unified Model-Based Refiner for Pre-training Data." Manuscript in preparation.

[4] Zhihan Lei, **Jiarui Yan**, Joshua Wisdom Momo, William W. Cohen. "Inducing Reasoning Primitives from Agent Traces." Under review.

[5] **Jiarui Yan**, Jane Tan, and Yong Tan. "Will the Fork Be with You? How Forks Drive User Contributions on Open-Source Platforms." 35th POMS Conference (UTD24), ISOM Track, 2025.

[6] Tingting Qi, Tianshu Wang, and **Jiarui Yan**. "The Spillover Effects of Different Monetary Incentive Levels on Health Experts' Free Knowledge Contribution Behavior." Internet Research, 31(6), 2143–2166.

*Equal contribution.

RESEARCH EXPERIENCE

RL Environment Scaling Laws (manuscript in preparation)

Palo Alto, CA

Research Assistant, Prof. Sanmi Koyejo (Stanford)

05/2026 – Present

- Developing **scaling laws** that predict how much **RL post-training** will improve a model from measurable properties of the training environment, rather than from compute alone.
- Designed metrics that make these environment properties controllable and cheap to measure, enabling prediction before a full training run is committed.
- Running controlled experiments that hold the RL algorithm fixed to isolate environment effects across math, coding, and tool-use tasks.

Pre-training Data Refinement (manuscript in preparation)

Pittsburgh, PA

Graduate Researcher, Prof. Chenyan Xiong (CMU)

01/2026 – Present

- Built an end-to-end model-based HTML-to-text refiner on **Qwen3-1.7B** with **32K-token** context that replaces the standard multi-stage heuristic pipeline (resiliparse, trafilatura, DCLM rules) in a single pass, lifting token retention from **40.5% to 54.7%** on DCLM CommonCrawl.
- Trained the refiner in two stages: **SFT** warmup on quality-merged labels from a cleaning pipeline, followed by **GRPO** reinforcement learning under a composite quality-and-faithfulness reward with BERTScore anti-drift; orchestrated with **vLLM** and **Verl** across multi-GPU clusters.
- Pretrained **400M** and **1.4B** language models on **8.2B / 28B** refined tokens and benchmarked against heuristic baselines, outperforming the best baseline on the 22-task **DCLM Core** zero-shot suite by **3-4%** average accuracy.

Agentic Systems with Pseudo-Tools

Pittsburgh, PA

Graduate Researcher, Prof. William W. Cohen (CMU, Google DeepMind)

12/2025 – 06/2026

- Built a Python framework on **Pydantic-AI** for modular agentic LLM systems centered on proposed **pseudo-tools** — recursive LLM calls under restricted context — with **LiteLLM** routing across **100+** model APIs and typed interfaces swappable between Python functions, LLM calls, **ReAct** sub-agents, and sandboxed code-gen backends.
- Conducted the first systematic engineering study of modular agents across **19** evaluation suites (BBH, FinQA, MedAgentBench, τ -Bench, RuleArena), providing concrete evidence that **static workflows are cheaper and more accurate than dynamic planners**: 0.64 vs 0.56 average correctness at half the cost on DeepSeek-V3.1.
- Designed three learning methods (pseudo-tool inducer, code distillation, orchestration hill-climbing) and applied **NSGA-II** evolutionary search across interface bindings and backbone-model assignments to recover cost-quality Pareto frontiers; the learned orchestrator reached **0.76** vs 0.55 for a human-engineered baseline.

Long-Horizon ML Agent Planning

Pittsburgh, PA

Graduate Researcher, Prof. Yiming Yang (CMU)

12/2025 – 06/2026

- Introduced a large-scale dataset of paired human and agent trajectories for long-horizon ML development, covering **4,672** Kaggle runs across **134** MLE-bench competitions and **151K** code snapshots, designed to expose planning failures that outcome-based benchmarks cannot detect.
- Reconstructed development trajectories through a 4-stage pipeline (state ingestion, lineage DAG construction, topological pruning, schema normalization) and fine-tuned **Qwen3-1.7B** labelers via distillation from a **GPT-5-mini** teacher over 136 state tags and 85 action labels, validated against human annotations.
- Identified a structural planning gap in which agents compress multiple development phases into a static loop; designed an agent harness that shifted this behavior and improved leaderboard score on most of 7 evaluated competitions within a 12-hour budget. Released on **HuggingFace** under CC-BY-4.0.

PRIOR RESEARCH EXPERIENCE

High-Dimensional Learning for Multi-Sourced Data

Seattle, WA

Research Assistant, Prof. Wanning Chen (UW)

11/2023 – 12/2024

- Deployed a transfer-learning **matrix completion** framework for multi-source high-dimensional data, introducing proxy representations to assist prediction on sparse target domains together with **SVD** low-rank approximation.
- Conducted experiments on synthetic and NetEase real-world data, aggregating **2M+** users and content into 200×70 interaction matrices for Like and Click behaviors; tuned hyperparameters by cross-validation on Frobenius norms, reducing RMSE from **0.6 to 0.35**.

Behavioral Analysis of Open-Source Software Contributors

Seattle, WA

Research Assistant, Prof. Jane Tan (SMU), Prof. Yong Tan (UW)

11/2023 – 12/2024

- Built a large-scale GitHub and Hugging Face repository mining pipeline tracking developer activity and fork interactions, aggregating create, commit, and fork events into time-series panel data.
- Empirically showed that different fork types have distinct effects on contribution activity, with no observable cross-repository spillover.

Retrieval-Augmented Generation Question Answering System

Seattle, WA

Student Researcher, Prof. Ranjay Krishna (UW)

01/2024 – 06/2024

- Built an end-to-end multimodal **RAG** system over course materials, extracting **2M+** tokens from lecture videos and slides via PyDub, **VOSK** speech-to-text, and PyTesseract **OCR**, then embedding documents with **BERT** for dense retrieval.
- Integrated **LLaMA-2 (13B)** with a **Pinecone** vector database, achieving **82%** precision and **77%** recall on 200+ evaluation queries and outperforming a baseline LLM.

INDUSTRY EXPERIENCE

Google LLC

Sunnyvale, CA

Software Engineer Intern

05/2026 – 08/2026

- Built an **LLM agent** (NUAC) automating Workspace **SRE access authorization**, parsing natural-language peer-approval signals in Buganizer tickets into structured, policy-compliant endorsements.
- Hardened the agent with deterministic guardrails, metadata-grounded requester/approver checks, **prompt-injection** risk scoring, and immutable-history validation.
- Built the evaluation stack: a golden dataset of human-annotated and synthetic adversarial cases with daily automated monitoring, reaching **93.75% F1**; shipped **23 CLs** across API, agent, and eval infrastructure.

Quantum Crypto

Beijing, China

Research Intern

01/2025 – 08/2025

- Developed an LLM agent on **Qwen2.5-14B** with a **RAG** pipeline (query rewrite, retrieval, reranking, prompt engineering), embedding internal corpora with **sentence-transformers**; built the supporting transaction-monitoring backend on FastAPI and PostgreSQL.
- Researched **futures-spot hedging** and **funding-rate arbitrage** strategies for cryptocurrencies.

Elile AI

Seattle, WA

Machine Learning Engineer Intern

05/2024 – 08/2024

- Built an **Airflow** ETL pipeline into AWS RDS for sensor ingestion and trained **XGBoost** models for predictive maintenance on **4K+** EV charging stations, achieving 7-day-ahead failure prediction.
- Containerized ML services with **Docker** and CI/CD, cutting deployment latency by **70%**.

Huawei Technologies

Beijing, China

Data & Algorithm Engineer Intern

06/2023 – 09/2023

- Architected a distributed geospatial pipeline for **10M+** Points of Interest (Spark / Hadoop / HDFS + Elasticsearch + Kafka streaming) and enhanced aggregation via AOI spatial features.
- Trained **BERT**-based embedding models (PyTorch / NLTK) for duplicate detection, achieving **0.89** precision and **0.76** F1 on a ~100K validation set.

TEACHING

INFO 498: Text Mining & Analytics (Spring 2023, Fall 2023) Prof. Lucy Lu Wang

INFO 330: Databases & Data Modeling (Spring 2024) Prof. Lucy Lu Wang

- Designed assignments and evaluated projects covering corpus analysis, TF-IDF retrieval, text classification, and information extraction; supported students on NLTK / scikit-learn / spaCy pipelines.
- Led weekly SQL recitations and group project development for 35 students; gave individualized feedback on relational data

modeling and query design.

SKILLS

Programming Languages: Python, C++, Java, TypeScript, JavaScript, SQL, R, Shell

ML / LLM Frameworks: PyTorch, TensorFlow, vLLM, Verl, Pydantic-AI, LiteLLM, LangChain, Hugging Face, CUDA

Big Data: Apache Spark, Hadoop, HDFS, Hive, Kafka, Parquet, NumPy, Pandas, Tableau

Cloud / DevOps: AWS, Google Cloud Platform, Docker, Kubernetes, Airflow, Git, CI/CD