Weiran Yao

Phone: +1 (412) 613-1327 • Email: weirayao@gmail.com • Website: https://weirayao.github.io/ • LinkedIn: linkedin.com/in/weiranyao/

EMPLOYMENT

Senior Research Scientist, *Salesforce AI Research*, Palo Alto, CA — Jan 2023 – Current

Areas: *AI Agent, Multi-Agent System, Finetuning & Alignment, Data Pipeline, Prompt Optimization*

Tech Lead for LLM Agent Incubation Projects. Led a team of 3 research/applied scientists for design, development and deployment of prototype multi-agent AI systems for demonstrations to top executives, with end-to-end development (spearheaded both frontend and backend development).

- DigitalHQ: Multi-Agent Slack Workspace for AI Employees [Blog][Demo] — Jan 2024 – Current
- SEDA: Software Engineering Digital Assistant [Blog][Demo] — Apr 2024 – Current
- Salesforce WebAgent [Blog][Demo] — Aug 2023 – Oct 2023

LLM/SLM **Finetuning & Alignment**. Conducted post-training research to align long-context models to specialize in self-reflection of task executions. Contributed to Salesforce in-house xLAM-series agentic model development by aligning the model for function call in CRM production environment.

- Retroformer 236B – General Critic Model for Agentic Self-Reflection [Report][Code][Model][Data]
- Salesforce X Coder [Report][Paper][Code][Huggingface]
- Salesforce xLAM 1B | 7B | 8x7B – Large Action Model for Function Call [Report][Model][Blog]

Conducted research on **Synthetic Data Pipeline** for LLM function call. This pipeline enabled a 7B model to **outperform several gpt-4 models** for function call on Berkeley Function-Calling Leaderboard.

- APIGen: Automated Pipeline for Generating Function-Calling Datasets [Website][Paper][Data]
- AgentOhana: Unified Data and Training Pipeline for Effective Agent Learning [Paper][Code]

Prompt Engineering and Optimization. Conducted research to automatically optimize the system prompt of LLM agent towards multi-objectives, e.g., accuracy, consistency, latency, and applied it to product.

- Einstein Copilot Meta-Prompt Optimization. Latency metrics improved by 50%.
- PRAct: Optimizing Principled Reasoning and Acting of LLM Agent [Report][Code]

AI Interpretability. Conducted scalable sparse autoencoder research for extracting universal concepts across large models. Applied the approach for safe model alignment even with limited feedback.

- UniMind: Universal Concept Extracted Across Large Open Models [Paper][Code]
- Editing Arbitrary Propositions in LLMs without Subject Labels [Paper][Code]

Engineering Products. Developed Function Call and Structured Output API endpoints for Salesforce xLAM service based on vLLM inference backend. Developed automatic root cause analysis algorithms for Salesforce Database Throttles with scalable, real-time anomaly detection.

- OpenAI-Compatible Function Call + Structured Output API Endpoint
- AI for IT Operations: dbCPU Throttle Incident Categorization and Causation Analysis [Blog]

Ph.D. Researcher, *Carnegie Mellon University*, Pittsburgh, PA — Sep 2017 – Dec 2022

Areas: Fundamentals of *AI Interpretability*

My research focused on provable **AI Interpretablity** with sparse or disentangled autoencoders to identify concepts from videos and non-stationary time series. Some selected work below.

- Temporally Disentangled Representation Learning [Paper][Code]
- Learning Temporally Causal Latent Processes from General Temporal Data [Paper][Code]
- Prompt Learning with Optimal Transport for Vision-Language Models [Paper][Code]

OPEN-SOURCE SOFTWARE

- **AgentLite:** Lightweight Library for Building LLM Multi-Agent System (356 Stars)
- **CausalAI:** Scalable framework for Causal Analysis of Time Series and Tabular Data (246 Stars)
- **Merlion:** A Machine Learning Framework for Time Series Intelligence (3.3k Stars)

EDUCATION

Carnegie Mellon University, School of Computer Science, Pittsburgh, PA

- Ph.D. in Advanced Infrastructure Systems (Advisor: Kun Zhang) — Aug 2017 – Aug 2023
- M.S. in Machine Learning — Aug 2019 – May 2021

TECH STACK

Programming Language: Python, JavaScript, HTML/CSS, Bash, SQL

Tools and Frameworks: Git, LaTeX, PyTorch, Triton, Spark, Docker, Kubernetes, Streamlit, FastAPI

PUBLICATIONS

CONFERENCE AND JOURNAL PUBLICATIONS

[23] SWE-Committee: Combating Instability in Autonomous Software Engineers.

Kexun Zhang, Weiran Yao, Yihao Feng.
ArXiv Preprint 2024.

[22] **APIGen: Automated Pipeline for Generating Verifiable and Diverse Function-Calling Datasets.**

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, Caiming Xiong.
ArXiv Preprint 2024.

[21] AgentOhana: Design Unified Data and Training Pipeline for Effective Agent Learning.

Jianguo Zhang, Tian Lan, Rithesh Murthy, Zhiwei Liu, Weiran Yao, Juntao Tan, Thai Hoang, Liangwei Yang, Yihao Feng, Zuxin Liu, Tulika Awalgaonkar, Juan Carlos Niebles, Silvio Savarese, Shelby Heinecke, Huan Wang, Caiming Xiong.
ArXiv Preprint 2024.

[20] AgentLite: A Lightweight Library for Building and Advancing Task-Oriented LLM Agent System.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Liangwei Yang, Zuxin Liu, Juntao Tan, Prafulla K Choubey, Tian Lan, Jason Wu, Huan Wang, Shelby Heinecke, Caiming Xiong, Silvio Savarese.
ArXiv Preprint 2024.

[19] CaRiNG: Learning Temporal Causal Representation under Non-Invertible Generation Process.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, Kun Zhang.
International Conference on Machine Learning (ICML) 2024.

[18] Causal Layering via Conditional Entropy.

Itai Feigenbaum, Devansh Arpit, Huan Wang, Shelby Heinecke, Juan Carlos Niebles, Weiran Yao, Caiming Xiong, Silvio Savarese.
Causal Learning and Reasoning (CLeaR) 2024.

[17] Editing Arbitrary Propositions in LLMs without Subject Labels.

Itai Feigenbaum, Devansh Arpit, Huan Wang, Shelby Heinecke, Juan Carlos Niebles, Weiran Yao, Caiming Xiong, Silvio Savarese.
ArXiv Preprint 2024.

[16] DRDT: Dynamic Reflection with Divergent Thinking for LLM-based Sequential Recommendation.

Yu Wang, Zhiwei Liu, Jianguo Zhang, Weiran Yao, Shelby Heinecke, Philip S Yu. *ArXiv Preprint 2023.*

[15] Temporally Disentangled Representation Learning under Unknown Nonstationarity.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, Kun Zhang.
Advances in Neural Information Processing Systems (NeurIPS), 2023.

[14] **Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization**.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, Silvio Savarese.
International Conference on Learning Representations (ICLR) 2024. ***(Spotlight Presentation)***.

[13] BoLAA: Benchmarking and Orchestrating LLM-Augmented Autonomous Agents.

Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, Silvio Savarese.
International Conference on Learning Representations (ICLR) 2024.

[12] Rex: Rapid Exploration and Exploitation for AI Agents.

Rithesh Murthy, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Le Xue, Weiran Yao, Yihao Feng, Zeyuan Chen, Akash Gokul, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, Silvio Savarese.
International Conference on Learning Representations (ICLR) 2024.

[11] On the Unlikelihood of D-Separation.

Itai Feigenbaum, Huan Wang, Shelby Heinecke, Juan Carlos Niebles, Weiran Yao, Caiming Xiong, Devansh Arpit.
The International Conference on Probabilistic Graphical Models (PGM) 2024.

[10] Salesforce CausalAI Library: A Fast and Scalable Framework for Causal Analysis of Time Series and Tabular Data.

Devansh Arpit, Matthew Fernandez, Itai Feigenbaum, Weiran Yao, Chenghao Liu, Wenzhuo Yang, Paul Josel, Shelby Heinecke, Eric Hu, Huan Wang, Stephen Hoi, Caiming Xiong, Kun Zhang, Juan Carlos Niebles.
ArXiv Preprint, 2023.

[9] Non-Parametric State-Space Models: Identifiability, Estimation and Forecasting.

Chenghao Liu, Weiran Yao, Steven Hoi, Kun Zhang.
International Conference on Learning Representations (ICLR) 2023.

[8] Temporally Disentangled Representation Learning.

Weiran Yao, Guangyi Chen, Kun Zhang.
Advances in Neural Information Processing Systems (NeurIPS), 2022.

[7] **Prompt Learning with Optimal Transport for Vision-Language Models.**

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, Kun Zhang.
International Conference on Learning Representations (ICLR) 2023. ***(Spotlight Presentation)***.

[6] **Distribution-aware Goal Prediction and Model-based Planning for Safe Autonomous Driving.**

Jonathan Francis*, Bingqing Chen*, Weiran Yao*, Eric Nyberg, Jean Oh.
International Conference on Machine Learning (ICML) 2022. Workshop on Safe Learning for Autonomous Driving ***(Best Paper Award)***.

[5] Partial Disentanglement for Domain Adaptation.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, Kun Zhang.
International Conference on Machine Learning (ICML) 2022.

[4] Learning Temporally Causal Latent Processes from General Temporal Data.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, Kun Zhang.
International Conference on Learning Representations (ICLR) 2022.

[3] Data Driven Safety Risk Prediction of Lithium Ion Battery.

Yikai Jia, Jiani Li, Chunhao Yuan, Xiang Gao, Weiran Yao, Minwoo Lee, Jun Xu.
Advanced Energy Materials 2021.

[2] From Twitter to traffic predictor: Next-day morning traffic prediction using social media data.

Weiran Yao, Sean Qian.
Transportation Research Part C: Emerging Technologies 2021.

[1] Learning a Distributed Control Scheme for Demand Flexibility in Thermostatically Controlled Loads.

Bingqing Chen, Weiran Yao, Jonathan Francis, Mario Bergés.
IEEE SmartGridCommm. 2020.

PATENTS

[7] Systems And Methods For Language Agent Optimization", US Patent 18,498,257.

[6] Systems And Methods For Orchestrating LLM-Augmented Autonomous Agents, US Patent 18,494,393.

[5] Systems And Methods For Building AI Agents For Language Models, US Patent 63,555,382.

[4] Systems And Methods For A Unified Training Framework Of Large Language Models, US Patent 18,658,899.

[3] Systems And Methods For Editing A Large Language Model, US Patent 18,428,530.

[2] Systems And Methods For A Unified Training Framework Of Large Language Models, US Patent 18,658,899.

[1] Distributed Control for Demand Flexibility in Thermostatically Controlled Loads, US Patent 12,027,858. Patent Granted.

PRESS COVERAGE

[6] **VentureBeat**. "Salesforce proves less is more: xLAM-1B 'Tiny Giant' beats bigger AI Models."

[5] **The Stack**. "On-device agentic AI is here!"

[4] **MarkTechPost**. "Salesforce Research Introduces AgentOhana: A Comprehensive Agent Data Collection and Training Pipeline for Large Language Model."

[3] **MarkTechPost**. "AgentLite by Salesforce AI Research: Transforming LLM Agent Development with an Open-Source, Lightweight, Task-Oriented Library for Enhanced Innovation."

[2] **MarkTechPost**. "Salesforce AI Researchers Introduce the Evolution of LLM-Augmented Autonomous Agents and the Innovative BOLAA Strategy."

[1] **MarkTechPost**. "Meet Retroformer: An Elegant AI Framework for Iteratively Improving Large Language Agents by Learning a Plug-in Retrospective Model."

INDUSTRY TALKS

[3] DigitalHQ: Collaborative Slack Workspace for Digital AI Employees, invited talk at *Salesforce Dreamforce*, Sep 2024, San Francisco.

[2] PRAct: Optimizing Principled Reasoning and Acting of LLM Agent, invited talk at *Databricks Data + AI Summit*, Jun 2024, San Francisco.

[1] Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization, invited talk at *Moveworks*, Sep 2023, Mountain View.

BLOGS

[2] Meet Merlion: An End-to-End Easy-to-Use Machine Learning Library for Time Series Applications. Salesforce AI Research.

[1] CausalAI: Answering Causality Questions Using Observational Data. Salesforce AI Research.

MENTORING EXPERIENCE

Summer Interns @ Salesforce AI Research

- Kexun Zhang, Ph.D. student at Carnegie Mellon University, Language Technology Institute.
- Mengdi Wang, Ph.D. student at Carnegie Mellon University, Language Technology Institute.

Ph.D. Students @ Carnegie Mellon University

- Lingjing Kong, Ph.D. student at Carnegie Mellon University Machine Learning Department.
- Xiangchen Song, Ph.D. student at Carnegie Mellon University Machine Learning Department.
- Zemian Ke, Ph.D. student at Carnegie Mellon University Mobility Data Analytics Center.

INTERNSHIPS

Research Intern, *Salesforce AI Research*, Palo Alto, CA — May 2022 – Aug 2022

Host: Juan Carlos Niebles, Caiming Xiong

Proposed TDRL, a provable temporally disentangled autoencoder approach for extracting video concepts. Paper published at *NeurIPS 2023*.