Jiaxin Zhang

Senior Staff Research Scientist @ Salesforce AI Research | Agentic RL, Alignment Science & Reliable LLM Agents

jxzhangai@gmail.com 443-676-0871 Homepage LinkedIn GitHub Google Scholar Bay Area, CA

Highlights

- Research scientist working on reliable, honest, and calibrated long-horizon agents; my agenda turns uncertainty, confidence and consistency into a *first-class* training signals for RL/post-training, scalable evaluation, agent oversight and self-evolving.
- 50+ peer-reviewed papers (**30+** first/last-author, 4000+ citations, h-index 28); Area Chair for NeurIPS, ICLR, ACL, EMNLP, etc.
- Hands-on builder of environment and synthetic-data pipelines, agentic workflows, and distributed training/RL infrastructure, with large-scale systems experience up to **20,000+** GPUs and 10+ open-source projects (**3,000+** GitHub stars).
- Led research teams from idea to production systems for reliability, evaluation, and post-training, including deep research benchmarks over **1,000+** tasks, and platforms adopted by **1,600+** internal users recognized with a CTO Award (**top 1%**).

Technical & Research Focus

- **Post-training & RL:** agentic RL, SFT, RLHF/DPO/PPO/GRPO, on-policy distillation, reward modeling, synthetic feedback, calibration-aware training, environment design, scaling, evaluation & benchmark, and scalable RL infrastructure.
- **Alignment Science & Model Behavior:** honesty, calibration, uncertainty-aware supervision, hallucination detection & mitigation, factuality, reliability, robustness, safety, scalable oversight, and reward/evaluation signals.
- **Long-horizon Agents & Environments:** deep research and coding agent workflows, tool use, trajectory-level evaluation, failure analysis & attribution, simulation & benchmark design, evaluation infrastructure, and environment scaling.

Education

Ph.D. **Johns Hopkins University**, Computational Science (*Dean's Fellowship*) 2014–2018

Experience

Senior Staff Research Scientist, Salesforce AI Research 2025–Present

Leading a research team on calibrated post-training, agentic RL, and evaluation environments for reliable long-horizon agents.

Calibrated Post-training & Agentic Reinforcement Learning

- **Evolving Trainable Environments** – a self-evolving environment-scaling framework for agentic RL that automatically escalates task difficulty with agent capability, and synthesizes environments that generalize to real-world benchmarks.
- **Prospective Hindsight (PH)** – an agent-RL method that converts an agent's prediction–reality gaps into a training signal, aligning action-time self-belief with verifier/PRM outcomes; plugs into GRPO, on-policy distillation, and multi-turn RL.
- **CaOPD** – calibration-aware on-policy distillation for LLMs post-training that isolates privileged-context supervision as a source of overconfident model behavior and decouples capability learning from honest confidence calibration. [arXiv 26']
- **NuRL** – a synthetic-hint RL framework that breaks exploration stagnation on hard reasoning tasks by generating abstract guidance, pushing the reasoning frontier onto previously unsolved problems. [ICLR 26']

Agentic Alignment, Calibration, and Oversight

- **Agentic Confidence Calibration (ACC)** – a holistic trajectory-level confidence calibration across multi-step agent planning, moving beyond static token-level probabilities toward trajectory-level signals for agent oversight. [ICML 26']
- **Agentic Uncertainty Quantification (AUQ)** – propagates uncertainty through agent memory and uses uncertainty-aware reflection to recalibrate, targeting compounding error and irreversible failures in long-horizon execution. [arXiv 26']

Long-horizon Agents and Enterprise Environments

- Designed a Deep Research agent system with adaptive planning, multi-graph retrieval, and context engineering to synthesize long-form reports from heterogeneous enterprise sources, including CRM, Slack, and Web data. [ACL 26']
- Built production-style agent environments and benchmarks over **1,000+** real long-horizon tasks, evaluating evidence grounding, planning quality, reasoning depth, and failure modes for reliable agent training. [Salesforce Blog]

Senior Staff Research Scientist, Intuit AI Research 2022–2025

Founding research lead for reliable LLM systems, spanning LLM post-training, alignment, evaluation, and production deployment.

Post-training & Knowledge Adaptation

- Led post-training of domain-specialized (financial) LLMs with SFT, RLHF, DPO, PPO, and GRPO, targeting expert-level reasoning and reliable model behavior.
- **IMFL** – an interactive human-in-the-loop multi-fidelity learning framework that combines sparse expert supervision with weak labels to reduce adaptation cost under scarce high-quality feedback. [NeurIPS 23'][Blog]
- **Synthetic Knowledge Ingestion (SKI)** – combines RAG, SFT, and continual pre-training (CPT) to update LLMs with changing knowledge and improve factual correctness in dynamic knowledge environments. [EMNLP 24'] [Blog]

LLM Honesty, Hallucination & Uncertainty

- **SAC**^{3} – semantic-aware cross-check consistency method for reliable black-box hallucination detection; in production it cut refusal rate by 65% and saved 19k hours/year over internal GenAI Platform. [EMNLP 23'] [Blog]
- **SPUQ** – perturbation-based uncertainty quantification (UQ) for confidence and consistency evaluation; deployed in production safety/evaluation systems used by 1,600+ internal users. [EACL 24']
- **GAME** – gradient-guided attention editing for contextual hallucination detection and mitigation. [NAACL 25']

Reliable Enterprise Agents, Evaluations, and Optimization

- **PhaseEvo** – automatic evolutionary prompt-optimization system; deployed at scale (2,000+ developers), improving task KPIs by 10-30% and cutting completion time up to 480x. [ACL 25']
- **Divide-Conquer-Reasoning (DCR)** – multi-agent framework for evaluating and improving LLM consistency over complex logic chains [EMNLP 24'], with out-of-knowledge query detection for reliable RAG. [EMNLP 24']
- Advanced multimodal factuality and evaluation methods, including InterleavedEval, R2I-Bench [EMNLP 25'], modality-specialized synergizers [ICLR 25'], and statistical factuality guarantees for large vision-language models. [EMNLP 25']

Staff Research Scientist, Computer Science and Mathematics Division, Oak Ridge National Laboratory 2018–2022
Led research on distributed large-scale machine learning and AI for Science; served as PI/team lead on DOE-funded projects.

Large-scale Training and Distributed RL Infrastructure

- Architected synchronous distributed training on ORNL leadership-class supercomputers (Summit and Frontier), reaching near-linear scaling on **20,000+** NVIDIA V100 GPUs and accelerating ImageNet training (ResNet50) to **10** minutes.
- Scaled distributed RL to **1,000+** CPUs, building high-throughput infrastructure for robust RL training and evaluation.

Generative Modeling and Experimental Design for Scientific Discovery

- Developed generative models (normalizing flows, energy-based models, GANs, and diffusion models) to accelerate scientific simulation (DFT, MD, FEM) and experimental design. [NeurIPS 19'] [AISTATS 21'] [UAI 21'] [AAAI 23']

Selected Publications

Full list: Google Scholar. 50+ peer-reviewed papers, 30+ first/last author, 4000+ citations, h-index 28

Agentic RL, Post-Training & Reasoning

- **[NeurIPS 2026 in review] Jiaxin Zhang**, Xiangyu Peng, Qinglin Chen, Yu Li, Hiroaki Hayashi and Chien-Sheng Wu. Prospective Hindsight: Self-Calibrating Reinforcement Learning via Prediction–Reality Gaps.
- **[COLM 2026 in review] Jiaxin Zhang**, Xiangyu Peng, Qinglin Chen, Qinyuan Ye, Caiming Xiong, and Chien-Sheng Wu. The Illusion of Certainty: Decoupling Capability and Calibration in On-Policy Distillation.
- **[ICLR 2026]** Justin Chih-Yao Chen, Xiangyu Peng, Prafulla Kumar Choubey, Kung-Hsiang Huang, **Jiaxin Zhang**, Mohit Bansal, and Chien-Sheng Wu. Nudging the Boundaries of LLM Reasoning.
- **[EMNLP 2025] Jiaxin Zhang**. Confidence-Aware Reasoning: Optimizing Self-Guided Thinking Trajectories in Large Reasoning Models.
- **[EMNLP 2024] Jiaxin Zhang**, Wendi Cui, Yiran Huang, Kamalika Das, and Sricharan Kumar. Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models.
- **[NeurIPS 2023] Jiaxin Zhang**, Zhuohang Li, Kamalika Das, and Sricharan Kumar. Interactive Multi-Fidelity Learning for Cost-Effective Adaptation of Language Models with Sparse Human Supervision.

Alignment, Honesty, Calibration & Uncertainty

- **[ACL 2026] Jiaxin Zhang**, Wendi Cui, Zhuohang Li, Lifu Huang, Bradley Malin, Caiming Xiong, and Chien-Sheng Wu. From Passive Metric to Active Signal: The Evolving Role of Uncertainty Quantification in Large Language Models.
- **[EMNLP 2025]** Zhuohang Li, Chao Yan, Nicholas J. Jackson, Wendi Cui, Bo Li, **Jiaxin Zhang**, and Bradley Malin. Towards Statistical Factuality Guarantee for Large Vision-Language Models.
- **[NAACL 2025]** Yu Wang, Kamalika Das, Xiang Gao, Wendi Cui, Peng Li and **Jiaxin Zhang**. Gradient-guided Attention Map

Editing: Towards Efficient Contextual Hallucination Mitigation.

- **[EMNLP 2024]** Wendi Cui, Zhuohang Li, Damien Lopez, Kamalika Das, Bradley Malin, Sricharan Kumar, and **Jiaxin Zhang**. Divide-Conquer-Reasoning for Consistency Evaluation and Automatic Improvement of Large Language Models.
- **[EMNLP 2023] Jiaxin Zhang**, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC^3: Reliable Hallucination Detection in Black-Box Language Models via Semantic-Aware Cross-Check Consistency.

Long-horizon Agents and Environments

- **[ICML 2026] Jiaxin Zhang**, Caiming Xiong, and Chien-Sheng Wu. Agentic Confidence Calibration.
- **[COLM 2026 in review] Jiaxin Zhang**, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. Agentic Uncertainty Quantification.
- **[ACL 2026]** Prafulla Kumar Choubey, Kung-Hsiang Huang, Pranav Narayanan Venkit, **Jiaxin Zhang**, Vaibhav Vats, Yu Li, Xiangyu Peng, and Chien-Sheng Wu. Don't Stop Early: Scalable Enterprise Deep Research with Controlled Information Flow and Evidence-Aware Termination.

Multimodal Architecture, Reasoning, Alignment & Evaluation

- **[ICML 2026]** Ying Shen, Zhiyang Xu, Jiuhai Chen, Shizhe Diao, **Jiaxin Zhang**, Yuguang Yao, Joy Rimchala, Ismini Lourentzou, Lifu Huang. Latte-flow: Layerwise timestep-expert flow-based transformer.
- **[ICLR 2025]** Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, **Jiaxin Zhang**, Qifan Wang, Yu Cheng, and Lifu Huang. Modality-Specialized Synergizers for Interleaved Vision-Language Generalists.

Mentorship & Advising

Mentored 10+ PhD interns, students, and junior researchers; several now at leading AI labs (Google DeepMind, NVIDIA):

- **Student & Researcher**: Zhuohang Li (Vanderbilt → Google DeepMind), Ankita Sinha (Intuit → NVIDIA), Wendi Cui (Intuit)
- **Ph.D. Intern**: Yu Wang (UCSB → Amazon), Justin Chih-Yao Chen (UNC), Haoyi Qiu (UCLA), Jixuan Chen (UCSD)

Professional Activities and Services

Area Chair and Program Committee

- Area Chair: NeurIPS, ICLR, ACL, EMNLP and NAACL 2024–Present
- Reviewer: NeurIPS, ICML, ICLR, ACL, EMNLP, NAACL, TMLR, JMLR, CVPR, ICCV, ECCV, AAAI, AISTATS, KDD 2020–Present

Invited Talks (2025–Present)

- "LLM hallucination, and reliable long-horizon agents" – Google Research, Meta Superintelligence Labs, EPFL, UCSD

Research Project Funding

- Principal Investigator (PI) or co-PI on 7 Department of Energy (DOE) ASCR/ORNL AI projects (6.4M+ total) 2018-2022

Honors and Awards

- **Intuit CTO Award (Top 1% Performance, Company-wide)**, Intuit 2024
- **Intuit A2D Innovation Award (Top 1%, Team Lead)**, Intuit 2024, 2025
- **Promising Early-Career Researcher Award**, Oak Ridge National Laboratory, US Department of Energy 2020
- **Chinese Outstanding Students Abroad Award**, Ministry of Education of the P.R. China 2019
- **China National Scholarship**, Ministry of Education of the P.R. China 2009, 2012

Technical Skills

- **LLM Training & Evaluations**: agentic RL, SFT, RLHF, DPO/PPO/GRPO, on-policy distillation, reward/evaluator design, synthetic data & simulation, environment scaling & evaluation, agent benchmarks, calibration, honesty & factuality evaluation
- **Infrastructure:** Python, PyTorch, Megatron-LM, verl, slime, vLLM, SGLang, distributed training/evaluation, HPC scaling.

Selected Open-Source Software & Resources

- SAC^3: Reliable Hallucination Detection in Black-Box LLMs (100+ paper citations)
- UQpy: Uncertainty Quantification with Python (340+ stars, maintained since 2018)
- Intuit AI Research Open Source GitHub Organization (Founder and Owner)
- Curated community resources: Awesome-LLM-Uncertainty-Reliability-Robustness (800+), Awesome-LLM-RAG (1.3k+)