Chih-Hsuan (Bella) Yang

LLM Agents, Agentic Evaluation and Preference Learning

bellayang@anl.gov | chihhsuan-yang.github.io

GitHub | Hugging Face | Google Scholar | AgentsSci | LinkedIn

Postdoctoral researcher at Argonne National Laboratory studying multi-agent LLM systems, agentic evaluation, and AI for Science. Research combines protocol harnesses, trajectory-level diagnostics, and controlled memory studies with prior hands-on preference-model training and DPO/PPO experience.

Research Contributions and Accepted Publications

Precise but Uncoupled: Reviewer Precision Does Not Guarantee Critique Uptake in Multi-Agent Math Reasoning

First author. NeurIPS Main Track, accepted. 2026

Designed controlled analyses separating reviewer error detection, a solver's response to critique, and verified answer repair. Identified why accurate review alone need not improve final outcomes, informing agent feedback design.

Paper and project | Code | Data

LLMs Can Predict Failure Risk, But Struggle to Predict Which Collaboration Protocol Pays Off: Cost-Aware Protocol Routing Across Reasoning Tasks

First author. Findings of EMNLP, published. 2026

Compared four reasoning protocols on 4,181 math problems, with checks across scientific reasoning tasks and two solver families. Evaluated supervised routing, held-out generalization, confidence calibration, and inference quality-cost trade-offs; distinguished failure prediction from protocol-specific value.

Paper | Code | Data and router

BioTrove: A Large Curated Image Dataset Enabling AI for Biodiversity

First author. NeurIPS Datasets and Benchmarks, Spotlight. 2024

Co-led a scientific data effort yielding 161.9 million research-grade images across approximately 366,600 species, with vision-language models and zero-shot biodiversity benchmarks.

MolGen-Transformer: A Molecule Language Model for the Generation and Latent Space Exploration of Organic Molecules

First author. Computational Materials Science 269, 114549. 2026

Developed transformer-based molecular generation and controlled latent-space exploration for molecular and materials discovery.

Open Evaluation Resources

Scientific-agent protocol traces: a matched matrix of 6,653 problems, two actor models, and four protocols (53,224 trajectories), with structured outcomes and provenance. Supports protocol comparison and trace-level diagnostics; follow each source dataset's usage restrictions.

AIMS @ COLM 2026 workshop | Dataset

Current Research Directions

Study how role contracts and evaluator feedback affect answer recovery, and measure task-conditioned memory utility through paired with/without-memory evaluations using a frozen reader and controlled context.

Research and Industry Experience

Postdoctoral Researcher, Argonne National Laboratory Jul 2025 - Present

- Build LLM-agent evaluation harnesses for matched protocol comparisons, structured trajectories, and role-level token/call accounting; analyze critique uptake and verified answer repair.
- Study supervised protocol routing, calibration, and task-conditioned memory with held-out tests and controlled ablations; release reproducible research code and evaluation datasets.

AuroraGPT Post-Training Working Group participant; mentor four students.

Visiting Scientist, Corteva Agriscience Sep 2024 - Jan 2025

Trained preference models from scratch using real experimental outcomes to ground rewards. Adapted multiple language models with Direct Preference Optimization (DPO) and Proximal Policy Optimization (PPO); incorporated scientist preferences when model judgments were uncertain.

AI Resident, X, Google LLC Sep - Dec 2023

Developed multi-agent LLM frameworks representing stakeholders with different interests. Investigated information flow, coordination, communication, and bidding mechanisms.

AI Graduate Researcher, Iowa State University Jan 2020 - Jun 2025

Researched molecular language modeling, multi-fidelity materials design, neural solvers for partial differential equations, and multimodal biodiversity models; co-led the BioTrove data effort.

Research Methods and Implementation

Agentic evaluation: protocol orchestration, trace-level failure analysis, memory ablations, supervised routing, confidence calibration, and inference quality-cost trade-offs.

Post-training: experiment-grounded preference/reward modeling, reinforcement learning from human feedback (RLHF), DPO, PPO, and uncertainty-triggered scientist feedback.

Implementation: Python, PyTorch, scikit-learn, NumPy/pandas, Hugging Face, pytest, Git; reproducible Linux/HPC experiments and structured trace/data workflows.

Selected Manuscripts Under Review

Wrong but Useful: Trajectory Value Beyond Answer Correctness in Multi-Agent Messages
First author. Submitted to AAAI 2027, under review; public preprint. 2026

Beyond Division of Labor: How Roles Shape Feedback and Recovery in Multi-Agent Systems
First and corresponding author. AAMAS 2027, under review. 2026

What Is This Memory Useful For? Measuring Task-Conditioned Utility in Long-Term Agent Memory
First and corresponding author. AAMAS 2027, under review. 2026

Education

Ph.D., Mechanical Engineering, Iowa State University 2025
Co-major in Computer Engineering; minor in Applied Mathematics.

M.S., Computer Science, Iowa State University 2025

B.S., Bioenvironmental Systems Engineering, National Taiwan University 2019