Cheng Chang

chc012@alumni.stanford.edu | chc012.github.io | linkedin.com/in/chc012 | github.com/chc012 | Google Scholar

Education

Stanford University — Stanford, CA, USA
MS in Computational and Mathematical Engineering — Sep 2022 – Jun 2024

University of California San Diego — La Jolla, CA, USA
BS in Cognitive Science; BA in Philosophy — Sep 2018 – Jun 2022

Experience

Uniphore (acquired Orby AI) — Palo Alto, CA, USA
Senior AI Scientist — Feb 2026 – Present

- **AI Research:** Drive research on software-engineering agent evaluation and LLM-as-a-judge meta-evaluation.
- **Software Engineering:** Lead development of a production-grade system that evaluates AI agents end-to-end using an LLM-as-a-judge approach.

Senior Software Engineer — Aug 2025 – Jan 2026

- **AI Research:** Developed coding agents for search automation; created an open-source GUI automation benchmark and trained web navigation models.
- **Software Engineering:** Built services to help users select small language models (SLMs) for custom enterprise use cases.

Orby AI — Mountain View, CA, USA
Software Engineer — Jul 2024 – Aug 2025

- **Multimodal Model Research:** Trained generative models and built web automation agents achieving state-of-the-art results on public benchmarks.
- **Software Engineering:** Integrated LLMs into backend automation tools for code generation and web interaction.

Software Engineer Intern — Apr 2024 – Jun 2024

- Built a distributed web UI data crawler for foundation model training.

SALT Lab, Stanford University — Stanford, CA, USA
Graduate Student Researcher — Oct 2023 – Mar 2024

- Developed CARE 2.0 (LLM-based counselor training platform); implemented UI, backend APIs, and database.

GrubMarket, Inc. — San Francisco, CA, USA
Data Science Intern — Jun 2023 – Aug 2023

- Designed and implemented recommender systems for fresh produce.

Selected Publications

AppSci-Bench: Can LLMs Build AI Agents and Make Them Better? *Under review.*

Agent Behavior as Code: Efficient and Robust LLM Agents with Programmatic Specifications. *Under review.*

PADMÉ: Preference Alignment Data Synthesis for Meta-Evaluation of LM Agent Evaluators. *(To appear) NeurIPS TAE Workshop 2026.*

WARC-Bench: Web Archive Based Benchmark for GUI Subtask Executions. *ICLR 2026.*

[WebDreamer] Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents. *TMLR 2025.*

[UGround] Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents. *ICLR 2025.*

Tech Stacks

Languages: Python, C/C++, TypeScript/JavaScript, R, SQL, HTML, CSS, TeX/LaTeX

ML & LLM: PyTorch, Hugging Face, LiteLLM, OpenAI, Anthropic, Fireworks AI

Infrastructure & Tools: AWS, Kubernetes/Docker, Ray, Databricks, Linux, Git, FastAPI, Flask