AI Agents for Discovery in the Wild

Abstract

The workshop focuses on the challenges practitioners face when taking promising autonomous discovery systems—such as AlphaEvolve, GEPA, and Claude Code— from the lab to the real world. We are interested in concrete problem domains and the issues that arise when moving beyond benchmark settings. What new difficulties emerge when we go from toy problems such as "circle packing" to complex real-world problems? By bridging the gap between laboratory research and real-world implementation, this workshop aims to chart a roadmap for the AI discovery engines capable of finding novel solutions in the wild.

1 Workshop Summary

Recent advances in LLM-driven optimization have established a highly effective class of AI discovery systems that autonomously generate, evaluate, and refine solutions. The promise of this AI-driven research is already yielding significant success stories. For example, AlphaEvolve [5] demonstrated that LLM-powered evolutionary search could discover novel algorithms and optimize large systems. Open-source frameworks like OpenEvolve [6], SkyDiscover [4], ShinkaEvolve [3], and GEPA [1] have rapidly replicated and extended these capabilities. Furthermore, systems like Glia [2] have shown that multi-agent LLM architectures can produce human-expert-level designs for complex distributed systems, while agentic coding tools like Claude Code and Cursor are empowering engineers to tackle optimization problems directly.

While these results are striking, they predominantly originate from somewhat well-defined benchmark problems featuring clean evaluation functions, known optima, and single objectives. Translating these promising techniques from laboratory settings to real-world deployment poses a unique set of challenges that the research community is only beginning to understand and address.

We aim to anchor the workshop around the following:

- **Problem Formulation and Evaluation**: Real-world problems rarely have clean, fast evaluation functions. How do we handle expensive simulations, noisy measurements, and multi-objective tradeoffs? How should problems be formulated so that LLM-driven search can effectively explore them?
- **Scaling, Cost, and Efficiency**: LLM-driven search requires substantial compute. What are the right tradeoffs between exploration depth, model cost, and solution quality? How do different frameworks and algorithms compare in practice?
- **Trust and Validation**: Benchmarks provide ground truth; real-world domains often do not. How do we validate AI-discovered solutions? When should practitioners trust an AI-generated optimization over a human-engineered one?
- **Human-AI Collaboration**: Domain expertise is critical for real-world problems. What is the right interface between human experts and discovery systems? How does human-in-the-loop guidance affect search quality and adoption?
- **Systems Infrastructure**: What infrastructure is needed to run discovery systems reliably and at scale? How do we handle failures, reproducibility, and integration with existing workflows?
- **Case Studies and Lessons Learned**: We actively solicit experience reports from practitioners who have deployed AI discovery systems on real problems, including both successes and failures.

This workshop is directly relevant to ACM CAIS, as AI discovery systems are inherently agentic. They operate as autonomous agents that reason about solution spaces, invoke tools (such as compilers, simulators, and evaluators), and iteratively refine their outputs. The workshop addresses core conference themes: the design and optimization of agentic systems, systems support for running them at scale, and practical experience deploying them in production environments targeted at solving real-world hard problems involving novel ideation and discovery.

2 Related Workshops

No prior ACM, ICML, ICLR, or NeurIPS workshop has focused specifically on the practical challenges of building and deploying AI agents for discovery and optimization in real-world settings. The closest related efforts include: the *AI with Recursive Self-Improvement* workshop (ICLR 2026), which focuses on the algorithmic foundations of self-improving AI systems and recursive optimization loops; *AI for Science* workshops at NeurIPS and ICML, which cover a broad range of AI applications in science, but do not focus on LLM-driven search optimization; and *ML for Systems* (MLSys, NeurIPS), which applies ML to systems problems, but does not address the broader range of real-world domains we target. In contrast, our workshop focuses on the design and deployment challenges for AI agents in discovery and optimization in real-world settings. The topic has grown rapidly over the past year, following the release of AlphaEvolve [5] (May 2025) and the subsequent emergence of several open-source frameworks, making this a timely venue for the community to convene.

3 Workshop Details

3.1 Scheduled Activities

We expect 80–120 participants. The workshop will run for a full day and combine invited talks, contributed presentations, poster sessions, a panel discussion, and a live demo session to encourage both technical depth and interactive discussion.

Morning		Afternoon	
08:30–08:45	Opening Remarks	1:30–2:00	Invited Talk 4
08:45–09:15	Invited Talk 1	2:00–2:30	Invited Talk 5
09:15–09:45	Invited Talk 2	2:30–3:15	Panel Discussion
09:45–10:15	Contributed Talks (3×15 min)	3:15–3:30	Coffee Break
10:15–10:30	Coffee Break	3:30–4:15	Poster Session 2
10:30–11:00	Invited Talk 3	4:15–4:45	Demo Session
11:00–11:45	Poster Session 1	4:45–5:00	Closing Remarks
11:45–1:30	Lunch		

Core workshop activities. The workshop will feature five invited talks from leading researchers working on AI-driven discovery systems and agents. Contributed talks will highlight selected papers presenting emerging ideas and early results. Two poster sessions will provide opportunities for detailed technical discussions and networking among participants.

The **panel discussion** will bring together invited speakers and experts from academia and industry to discuss open challenges in deploying AI agents for discovery and optimization in real-world settings.

The **demo session** will feature live demonstrations of discovery systems (e.g., SkyDiscover, OpenEvolve, GEPA, Glia) applied to real problems. These demos will allow attendees to observe how these tools operate in practice and discuss practical issues such as deployment, evaluation, and integration into real-world workflows.

3.2 Procedure for Selecting Papers

We will solicit 4–6 page extended abstracts (excluding references) through an open call. Each submission will be reviewed by at least two program committee members and evaluated based on relevance, technical quality, novelty, and potential to stimulate discussion. We particularly encourage experience reports and case studies from real-world deployments of AI-driven discovery systems.

We expect to accept approximately 15–25 papers for poster presentation, with 2–4 selected for contributed oral talks.

Topics and Areas of Interest. The workshop focuses on AI systems that autonomously search, optimize, and discover solutions using LLMs, agents, and evolutionary methods:

- Agentic and Evolutionary Systems for Discovery
- LLM-Driven Search and Optimization
- Discovery Problems and Evaluation
- Real-World Deployment of Discovery Systems
- Infrastructure for Discovery Systems
- Multi-Agent Systems for Discovery
- Human-AI Collaborative Discovery
- Safety, Verifiability, and Trust in AI-Generated Solutions

3.3 Program Committee (preliminary)

We are assembling a program committee with expertise in AI-driven optimization, machine learning systems, agents, and scientific discovery, spanning academia and industry. This is a non-exhaustive preliminary list of potential PC members, intended to reflect the communities we plan to involve. The program committee will include the workshop organizers as well as additional faculty, postdoctoral researchers, PhD students, and industry researchers. Preliminary list: Alex Dimakis (UC Berkeley); Batu El (Stanford); Eric Liang (Databricks); Shu Liu (UC Berkeley); Rui Meng (Google); Melissa Pan (UC Berkeley).

3.4 Tentative Schedule for the Workshop

Paper submission deadline	Sun, April 12, 2026
Acceptance notification	Tue, May 5, 2026
Camera-ready deadline	Fri, May 15, 2026
Workshop day	Tue, May 26, 2026

3.5 Workshop Organizers

The following are the workshop organizers, alphabetically listed.

Shubham Agarwal is a PhD student in EECS at UC Berkeley, advised by Ion Stoica and Aditya Parameswaran. His research focuses on AI-driven systems and the reliability and efficiency of LLMs and agents. shubham3@berkeley.edu

Lakshya Agrawal is a PhD student in EECS at UC Berkeley, advised by Matei Zaharia and Dan Klein. His research focuses on LLM systems and evolution methods for optimizing prompts. lakshyaaagrawal@berkeley.edu

Mert Cemri is a PhD student in EECS at UC Berkeley, advised by Ion Stoica, Kannan Ramchandran, and Alex Dimakis. His research focuses on multi-agent LLM systems and efficient machine learning. cemri@berkeley.edu

Alex Dimakis is a Professor of EECS at UC Berkeley and an IEEE Fellow for contributions to distributed coding and learning. His research spans generative AI, information theory, and machine learning. He is also the co-founder of Bespoke Labs. alexdimakis@berkeley.edu

Batu El is a PhD student at Stanford University, advised by James Zou. His research focuses on machine learning systems and optimization. batuel@stanford.edu

Alex Krentsel is a PhD student in EECS at UC Berkeley, advised by Scott Shenker and Sylvia Ratnasamy. His research focuses on reliable and efficient large-scale control systems. akrentsel@berkeley.edu

Eric Liang leads the ML for systems team at Databricks. His past work includes systems for scalable machine learning and reinforcement learning. ekl@databricks.com

Shu Liu is a PhD student in EECS at UC Berkeley, advised by Ion Stoica. Her research spans LLMs, agents, data systems, and cloud computing. lshu@berkeley.edu

Rui Meng is a research scientist at Google Cloud AI Research. His work focuses on AI agents for discovery, representation learning, and language modeling. rmeng@google.com

Sylvia Ratnasamy is a Professor of EECS at UC Berkeley. Her research focuses on computer networking, distributed systems, and network infrastructure. She received the ACM Grace Murray Hopper Award. sylvia@cs.berkeley.edu

Ion Stoica is a Professor of EECS at UC Berkeley, holding the Xu Bao Chancellor Chair. He co-founded Databricks and Anyscale and is an ACM Fellow and member of the National Academy of Engineering. His recent work focuses on AI-driven systems research. istoica@berkeley.edu

Matei Zaharia is an Associate Professor of EECS at UC Berkeley and co-founder and CTO of Databricks. He created Apache Spark and received the ACM Doctoral Dissertation Award and the Presidential Early Career Award. matei@berkeley.edu

3.6 Efforts for Diversity and Inclusion

Our organizing team includes researchers from academia and industry, spanning different career stages and research backgrounds. We will distribute the call for papers broadly across multiple research communities, including machine learning, systems, and AI for science, to encourage participation from diverse institutions and domains.

We will actively encourage submissions from early-career researchers, researchers at underrepresented institutions, and members of underrepresented groups in computing. In addition, we aim to ensure diversity among invited speakers, panelists, and the program committee in terms of research areas, institutional representation, career stage, and demographic background.

3.7 Workshop Speakers and Panelists

The following researchers have confirmed or been invited as speakers and panelists, spanning academia and industry, and representing a range of perspectives on AI-driven discovery.

Aditya Akella (UT Austin, confirmed) is a Regents Chair Professor of Computer Science at the University of Texas at Austin and Director of the LDOS and InfraAI initiatives. His research focuses on cloud and Internet infrastructure, and recently on AI-driven systems for automated systems management and optimization.

Mohammad Alizadeh (MIT / Glia, confirmed) is the NEC Professor of Software Science and Engineering at MIT EECS and a principal investigator in CSAIL. He is a co-founder of Glia, an AI system for automated systems design and optimization, and his research focuses on AI-driven systems, networking, and large-scale infrastructure.

Azalia Mirhoseini (Stanford / Ricursive Intelligence, confirmed) is an Assistant Professor of Computer Science at Stanford. Her research focuses on capable, reliable, and efficient AI systems for high-impact real-world problems, including systems and chip design.

Graham Neubig (CMU / OpenHands, confirmed) is an Associate Professor at Carnegie Mellon University's Language Technologies Institute and Chief Scientist at OpenHands. His research spans NLP, code generation, and AI agents for software development.

Alexander Novikov (Google DeepMind, invited) is a researcher at Google DeepMind and one of the developers of AlphaEvolve, an LLM-based system for algorithm discovery and optimization.

Joey Gonzalez (UC Berkeley, confirmed) is a Professor in EECS at UC Berkeley, and a co-director and founding member of the Sky Computing Lab and RISE Lab. His group works on AI and data systems, including projects such as Gorilla, Chatbot Arena, vLLM, and SGLang.

Aditya Parameswaran (UC Berkeley, confirmed) is an Associate Professor of Computer Science at UC Berkeley and co-director of the EPIC Data Lab. His research spans data systems, interactive data analysis, and document-centric AI systems, including DocETL.

James Zou (Stanford, confirmed) is an Associate Professor of Biomedical Data Science and, by courtesy, of Computer Science and Electrical Engineering at Stanford. His research focuses on reliable, human-compatible, and statistically rigorous AI, especially in health and science.

References

[1] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning, 2026.

[2] Pouya Hamadanian, Pantea Karimi, Arash Nasr-Esfahany, Kimia Noorbakhsh, Joseph Chandler, Ali ParandehGheibi, Mohammad Alizadeh, and Hari Balakrishnan. Glia: A human-inspired ai for automated systems design and optimization, 2025.

[3] Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. Shinkaevolve: Towards open-ended and sample-efficient program evolution, 2025.

[4] Shu Liu, Mert Cemri, Shubham Agarwal, Alexander Krentsel, Ashwin Naren, Qiuyang Mang, Zhifei Li, Akshat Gupta, Monishwaran Maheswaran, Audrey Cheng, Melissa Pan, Ethan Boneh, Kannan Ramchandran, Koushik Sen, Alexandros G. Dimakis, Matei Zaharia, and Ion Stoica. Skydiscover: A flexible framework for ai-driven scientific and algorithmic discovery, 2026.

[5] Alexander Novikov, Ngân V~u, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, S Shirobokov, Borislav M Kozlovskii, Francisco J R Ruiz, Abbas Mehrabian, M P Kumar, Abigail See, Swarat Chaudhuri, George Holland, A Davies, Sebastian Nowozin, Pushmeet Kohli, Matej Balog, and Google Deepmind. AlphaEvolve: A coding agent for scientific and algorithmic discovery. June 2025.

[6] Asankhaya Sharma. Openevolve: an open-source evolutionary coding agent, 2025.