

“AI Scientists” with *Productive Friction*: Preserving Human Agency in AI-Mediated Research

Yun Huang

University of Illinois Urbana-Champaign
Champaign, Illinois, USA
yunhuang@illinois.edu

Abstract

Building on large language models (LLMs), “AI scientists” can increasingly automate scholarly workflows, from literature synthesis and hypothesis generation to experimentation, manuscript writing, and peer review. Drawing on three years of our work designing LLM-based systems for research ideation, multi-persona critique, structured deliberation, and collaboration support, we argue that a distinctive value of these systems lies in their ability to introduce *productive friction*: task-relevant pauses (temporal), choices (structural), debates (interactional), encounters with scholars (boundary), and engagement with contrary ideas (relational). Rather than treating friction solely as an inefficiency to eliminate, we show how carefully designed friction can create opportunities for researchers to exercise judgment, maintain ownership, assume responsibility, think critically, and remain open to alternative perspectives. We further propose a set of evaluation dimensions for assessing whether LLM-based research systems cultivate these human capacities, including agency, critical thinking, and collaboration. Finally, we invite the community to co-create a research agenda for designing future AI scientists that not only advance science but also empower the human researchers.

CCS Concepts

• Human-centered computing → Human computer interaction (HCI).

Keywords

AI scientists, productive friction, human agency, human-AI collaboration

ACM Reference Format:

Yun Huang. 2026. “AI Scientists” with *Productive Friction*: Preserving Human Agency in AI-Mediated Research. In *ACM AI Summit 2026 (AI Summit '26)*, August 31-September 02, 2026, Atlanta, GA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3806096.3844855>

1 Introduction

A growing number of LLM-based systems have been created to automate substantial portions of the research lifecycle: reviewing literature, generating ideas, running experiments, writing papers, and even simulating peer review. Although these systems differ in ambition and domain, they often share a goal, reducing the time and

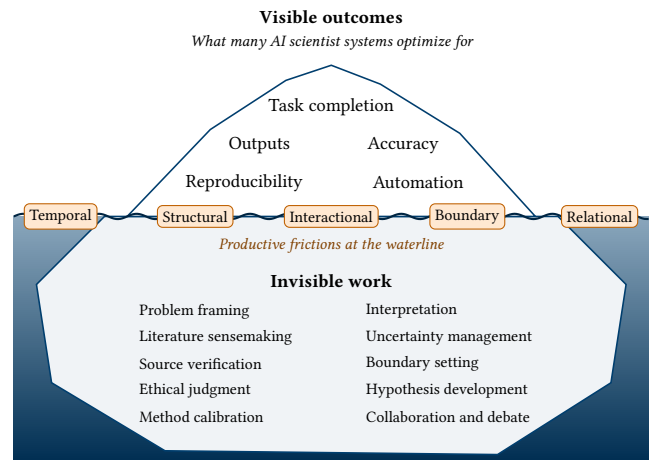


Figure 1: An iceberg model of AI-mediated research: above the waterline, the visible outcomes that AI systems optimize; at the waterline, deliberate *productive frictions*; below the waterline, the invisible work essential for trustworthy science. *Productive friction* is a designed interaction moment that requires human judgment before AI output is used. *Invisible work* is the judgment and coordination behind research outputs, such as framing problems, verifying sources, weighing alternatives, and negotiating with collaborators, that output metrics and automated pipelines do not capture.

effort research requires [1, 9, 10]. Because modern scholarly incentives (e.g., grant awards, tenure review, and citation metrics) heavily reward publication volume and rapid claims of novelty, researchers may be economically and reputationally incentivized to embrace *frictionless automation* so as to maximize throughput. This in turn can lead to paper flooding, shallow results, peer review collapse under AI-generated content, and the erosion of insightful scholarship [5]. Research communities suggest that science advances not only through efficiency and optimization but also through redundancy, divergent exploration, and failed attempts, which should be preserved as sources of scientific discovery [12].

Over the past three years, our research lab has built, evaluated, and piloted LLM-based systems to examine how to support researchers’ interpretation, direction-setting, and intellectual commitment when AI mediates their research [2, 6–8]. We ask which aspects of research AI can amplify, which practices should be protected, and how scholars can preserve epistemic diversity, creativity, and critical evaluation.



This work is licensed under a Creative Commons Attribution 4.0 International License. *AI Summit '26, Atlanta, GA, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2638-5/2026/08

<https://doi.org/10.1145/3806096.3844855>

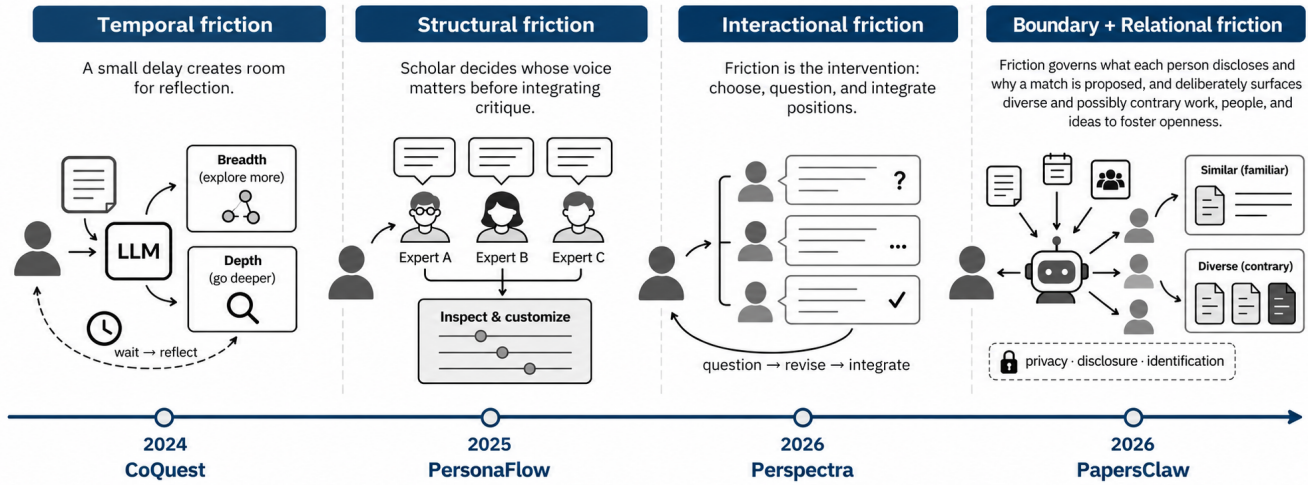


Figure 2: The evolution of *productive friction* across our LLM-based systems. These designs shift the role of AI from supporting individual ideation toward fostering critical engagement, preserving human agency, and enabling serendipitous scholarly connections and foster openness to contrary ideas.

Drawing on this systems research, we propose *productive friction* as a design principle for AI-mediated research. As illustrated in Figure 1, *productive friction* sits between the visible outcomes that many industry systems optimize for and the invisible work that is essential to science. *Productive friction* creates an intentional moment of pause, choice, disagreement, boundary-setting, or encounter with unfamiliar work that blocks premature automation and requires human judgment before AI outputs are accepted, integrated, or acted on. Drawing on empirical findings from our systems, we discuss five types of productive friction: *temporal*, *structural*, *interactional*, *boundary*, and *relational*. These *productive frictions* differ from arbitrary difficulty because they are tied to the human agency the system is meant to preserve, much as scaffolds that problematize a task make it harder in the short term but more productive for research and learning [13]. Each type can be designed into *dyadic* interaction, where a researcher works directly with AI agents, or into *polyadic* interaction, where AI mediates exchanges among researchers [14]. We then outline a research agenda for designing and evaluating AI scientists that preserve human agency in meaningful scholarship.

2 Types of *Productive Friction*

Our LLM systems were designed for and evaluated primarily with HCI scholars, because HCI is a revealing stress test for “AI scientist” discourse. HCI research is highly interdisciplinary: systems researchers often combine technical prototyping with theories from psychology and the social sciences, identify conflicts among stakeholders, apply mixed-method evaluation, and develop situated interpretations of human behavior. Research ideation therefore involves deliberating about what matters, to whom, under which values, and with what consequences, rather than only generating plausible hypotheses [6, 8].

Figure 2 presents a timeline of our systems research. We began the development of our first system in early 2023, when ChatGPT had captured wide attention and using LLMs to summarize or synthesize papers was becoming the norm. Agentic frameworks soon became popular, moving from single agents to multi-agent integration. In retrospect, we gradually shifted our attention from testing model capabilities to empowering individual researchers’ agency, and then to fostering interdisciplinary collaboration. Below, we discuss each case study in more detail.

Our first exploration was a single-LLM system called CoQuest [6], designed to support literature-grounded research-question generation. We conducted a formative study to understand how HCI researchers develop novel research questions (RQs), and then examined how researchers used a mind map to navigate LLM-generated RQs. Through experimental studies, we found that breadth-first and depth-first modes shaped users’ perceived creativity differently. For example, the breadth-first mode made users “feel” more creative by the end of the study session, but when rating individual RQs generated by the LLM, users perceived those suggested in the depth-first mode as more creative. Interestingly, “active waiting” moments caused by AI processing delays correlated with a stronger sense of control: while waiting, users actively explored other areas of the map and interacted more with the system. We call this *temporal friction*; a small delay appeared to make room for reflection.

Building on this single-LLM design, we expanded the map-based exploration with multiple LLM personas. PersonaFlow [8] allowed users to customize the “domain experts” they wanted so that they could receive multiple perspectives. Users thus decided whose perspectives mattered before integrating critique, rather than passively accepting the default AI perspective. By layering personas, they also reshaped the forms of expertise they drew upon. We therefore view this design as *structural friction*. To our surprise, multiple

personas improved users’ perceived helpfulness and relevance without increasing their perceived cognitive load. Allowing users to customize personas also increased their sense of control and their recall of generated ideas.

Subsequently, another multi-agent system, Perspectra [7], extended structural friction by letting multiple agents engage in forum-style discussion. Users selected which agents to engage with, examined arguments through threaded discussions, and revised their proposals in response to targeted critiques. Analyzing users’ questions through the lens of argumentation frameworks, we found that, compared with a group-chat baseline, users showed more frequent and deeper critical-thinking behavior and made more proposal revisions. Users’ critical-thinking activities (e.g., application, analysis, inference, and evaluation) increased significantly in response to agents’ counterarguments challenging prior claims. This *interactional friction* proved to be an effective intervention: users shifted from directly copying AI responses toward treating them as positions to question, select, and integrate into their own proposals.

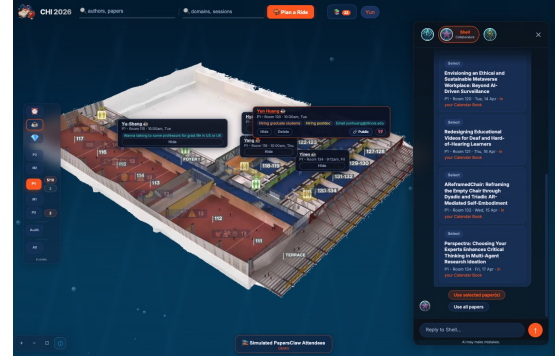
These three case studies focus on AI that supports individual cognition, where humans engage with AI-generated ideas, perspectives, and critiques. In late 2024, we began building chatbot apps for CSCW, CHI, and several other ACM conferences. Early versions focused on recommending papers, but the app has since evolved to facilitate intellectually productive encounters between researchers.

Conventional recommender systems often emphasize relevance, compatibility, shared interests, and user preferences. PapersClaw [2] pursues a different design objective: recommending intellectually meaningful differences, not only similarities. For example, two researchers might use different theoretical frameworks, methodologies, or disciplinary vocabularies while investigating related underlying phenomena. AI could identify these connections, explain why the differences matter, and encourage researchers to explore them. The goal is curiosity and potentially productive intellectual exchange, not disagreement for its own sake. PapersClaw’s potential distinction lies in deliberately designing explainable intellectual friction as a mechanism for collaborative discovery.

PapersClaw extends this research direction into AI-mediated human-to-human interaction, where AI can help researchers discover, interpret, and act upon intellectual differences. Because the AI stands between researchers, friction can govern what each person discloses and why a match is proposed. Ethical polyadic agents must be designed around social boundaries such as privacy, disclosure, and identification [14]; we call the friction that enforces these boundaries *boundary friction* (Figure 3a). PapersClaw’s agent also adds *relational friction* (Figure 3b): alongside the similar items that recommenders tend to favor [4], it deliberately surfaces papers, people, and possibly contrary findings outside the user’s usual themes to create productive distance. This friction targets openness to contrary evidence, perspective-taking across paradigms, and diverse ties.

3 Research Agenda

As a pilot, PapersClaw has not yet been formally evaluated. We envision such AI systems moving beyond recommending similar people and relevant information toward facilitating intellectually productive encounters across disciplinary boundaries. By revealing latent



(a) **Boundary friction introduces a physical-digital dimension.** Attendee cards displayed on the venue floor plan allow researchers to selectively disclose their locations ([Public] vs. [Hide]). When two researchers are physically near each other, an LLM-based agent surfaces unexpected intellectual connections that may prompt in-person interactions.



(b) **Relational friction: The Seahorse agent introduces productive intellectual distance by recommending papers beyond users’ usual research domains (+ Outside your domains).**

Figure 3: PapersClaw [2] screenshots featuring CHI 2026 papers illustrate two forms of AI-mediated productive friction: (a) *boundary friction* and (b) *relational friction*.

connections, explaining meaningful differences, and supporting situated human interaction, such systems may enable researchers to generate questions and possibilities that neither participant would have developed independently.

We invite the research community to co-create a design and evaluation agenda centered on human agency, intellectual distance, divergent exploration, and interpersonal negotiation. Informed by the presented work, we identify the following design implications, which can be integrated and further evaluated in a long-term agenda:

- (1) **Require judgment before elaboration.** Instead of giving an answer right away, systems may require researchers

to specify a perspective, objective, stakeholder, or framing before generation continues.

- (2) **Make disagreement inspectable.** Many AI chatbots give users two options and leave them to compare. Beyond presenting options side by side, systems should expose contestation among agents and alternative framings rather than converging too quickly on a single “best” output.
- (3) **Preserve authorship through cognitive forcing designs.** Most AI chatbots offer a “copy” button; interfaces should instead help users translate critique into their own arguments and discourage direct copy-paste adoption.
- (4) **Balance intellectual boundaries with serendipitous discovery.** AI-mediated collaboration should protect privacy and intellectual autonomy through selective disclosure, transparent recommendation logic, and consent-aware matching, while enabling unexpected scholarly connections.
- (5) **Shift from convergent teamwork to divergent collaboration.** Traditional teamwork research often examines how teams coordinate, manage conflict, and accomplish shared objectives. However, organizational behavior research also recognizes the creative value of dissent, disagreement, and divergent thinking. AI-mediated research may intentionally stimulate divergent collaboration.

Through the proposed *productive frictions*, certain scholarly work can be deliberately protected, because delegating it would erode the human accountability that makes research trustworthy. To evaluate whether productive frictions protect this invisible work, we propose the following dimensions, each defined by what it examines and how it could be measured:

- **Epistemic distance:** How different are the researchers an encounter connects? Measures: differences in theories, methods, assumptions, and domains.
- **Divergence generation:** How does the interaction open up new possibilities? Measures: new questions, alternative framings, hypotheses, and conceptual combinations.
- **Perspective transformation:** How does a researcher’s understanding change? Measures: revised assumptions, re-framed problems, and perspective-taking.
- **Friction navigation:** Can researchers engage constructively with disagreement? Measures: clarification, constructive challenge, argument integration, and openness.
- **Boundary agency:** How do researchers control what AI shares about them and their unpublished work? Measures: consent, selective disclosure, refusal, and perceived autonomy.
- **Emergent intellectual value:** What new knowledge comes from the unexpected encounter? Measures: novelty, significance, and originality of jointly developed ideas.
- **Longitudinal development:** How to support collaborative capabilities grow over time? Measures: sustained cross-disciplinary engagement and intellectual exploration.

Another important insight is that *productive friction* should not be treated as a binary condition. Different forms of friction may

have varying effects depending on their intensity, timing, and participants’ readiness to engage. We may explore several design variables:

- *Friction intensity:* How strongly does the encounter challenge existing assumptions?
- *Friction velocity:* How quickly are intellectual differences introduced, interpreted, and acted upon?
- *Friction dimensionality:* How many intellectual dimensions are involved, such as theory, methodology, values, and disciplinary assumptions?
- *Friction timing:* When is the intervention introduced, and is there an immediate opportunity to act?
- *Friction adaptability:* Can the system adjust its interventions to researchers’ interests, capabilities, and willingness to engage?
- *Friction agency:* Can participants control, modify, reject, or disengage from the proposed encounter?

We may design experiments to test if productive outcomes depend on an appropriate balance between intellectual challenge and participants’ ability to navigate that challenge.

Scope and limitations. A CSCW ’26 Special Interest Group [3] discusses related questions in social computing through five themes: knowledge building and research inquiry; team formation and collaboration; peer review and community building; interdisciplinary research development and evaluation, including “research taste”; and authorship, ownership, and governance. Several of these questions can be understood through the lens of productive friction. For example, designing systems that “encourage reflection, skepticism, and intellectual ownership” relates to temporal, structural, and interactional friction. Helping researchers discover complementary expertise while avoiding recommendations that reinforce existing academic hierarchies relates to relational and boundary friction.

These discussions raise a broader design question: as AI systems automate increasingly complex scholarly activities, which forms of intellectual work should remain under human control, and who should decide? Our framework approaches this question by distinguishing friction that supports critical thinking, intellectual ownership, and accountability from friction that merely introduces inefficiency. Rather than eliminating friction indiscriminately, AI systems should preserve opportunities for researchers to exercise judgment and engage with intellectual differences. However, determining which forms of friction are productive requires consideration of researchers’ goals, capabilities, and values, rather than relying solely on technical optimization.

The value of *productive friction* also depends on disciplinary practices and task characteristics. In formal sciences where proofs can be machine-checked, friction matters less for verification than for choosing conjectures and understanding why a proof holds; errors caught might include counterexamples. In the social sciences, as in HCI, friction concerns whose values and context an analysis represents and which stakeholders it omits. Friction has boundary conditions: it is hard to justify where errors are inexpensive to correct and outputs are reliably verifiable. Thus, *productive friction* should be calibrated to the nature of the task, the consequences of errors, and participants’ capacity to engage.

Acknowledgments

I thank my current and former doctoral students in the Social and Artificial Intelligence Technologies (SALT) Lab at the University of Illinois Urbana-Champaign, along with our many collaborators, for inspiring the ideas discussed in this paper. This material is based upon work supported by the National Science Foundation under Grant No. 2119589. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Additionally, cloud computing resources for this project were obtained through CloudBank [11], which is supported by the National Science Foundation under Award No. 1925001.

References

- [1] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, et al. 2026. Accelerating Scientific Discovery with Co-Scientist. *Nature* 655, 8122 (2026), 487–496. doi:10.1038/s41586-026-10644-y
- [2] Yun Huang. 2026. PapersClaw.fun: Enhancing Academic Conference Experiences. <https://papersclaw.fun/>. Accessed March 30, 2026.
- [3] Yun Huang, Noshir S. Contractor, Tai-Quan Peng, Dashun Wang, Cuihua Shen, Manling Li, Yiren Liu, Hangyue Zhang, and Sangho Suh. 2026. Rethinking Social Computing Research in the Age of AI: Agency, Collaboration, and Governance. Special Interest Group, CSCW '26. CSCW '26 program item. <https://programs.sigchi.org/cscw/2026/program/content/268614>
- [4] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A Survey of Serendipity in Recommender Systems. *Knowledge-Based Systems* 111 (2016), 180–192. doi:10.1016/j.knosys.2016.08.014
- [5] Jianghao Lin, Rong Shan, Jiachen Zhu, Yunjia Xi, Yong Yu, and Weinan Zhang. 2025. Stop DDoS Attacking the Research Community with AI-Generated Survey Papers. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Position Paper Track*. doi:10.52202/085713-2700
- [6] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How AI Processing Delays Foster Creativity: Exploring Research Question Co-Creation with an LLM-Based Agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–25. doi:10.1145/3613904.3642698
- [7] Yiren Liu, Viraj Nischal Shah, Sangho Suh, Pao Siangliulue, Tal August, and Yun Huang. 2026. Perspectra: Choosing Your Experts Enhances Critical Thinking in Multi-Agent Research Ideation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–27. doi:10.1145/3772318.3791560
- [8] Yiren Liu, Pranav Sharma, Mehul Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 506–534. doi:10.1145/3715336.3735789
- [9] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv preprint arXiv:2408.06292. <https://arxiv.org/abs/2408.06292>
- [10] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. Towards End-to-End Automation of AI Research. *Nature* 651, 8107 (2026), 914–919. doi:10.1038/s41586-026-10265-5
- [11] Michael Norman, Vince Kellen, Shava Smullen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, Rob Fatland, Sarah Stone, Amanda Tan, Katherine Yelick, Eric Van Dusen, and James Mitchell. 2021. CloudBank: Managed Services to Simplify Cloud Access for Computer Science Research and Education. In *Practice and Experience in Advanced Research Computing (PEARC '21)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3437359.3465586
- [12] Julio M. Ottino and Brian Uzzi. 2026. Progression without Progress. *Science* 392, 6800 (2026), 791. doi:10.1126/science.aeh8945
- [13] Brian J. Reiser. 2004. Scaffolding Complex Learning: The Mechanisms of Structuring and Problematising Student Work. *Journal of the Learning Sciences* 13, 3 (2004), 273–304. doi:10.1207/s15327809jls1303_2
- [14] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3491102.3501855