From AI-as-Shortcut to AI-as-Partner: Student-AI Collaborative Inquiry for Health-Agent Education

Zhihan Jiang
Columbia University
New York, NY, USA
zj2445@cumc.columbia.edu

Xuhai Xu
Columbia University
New York, NY, USA
xx2489@columbia.edu

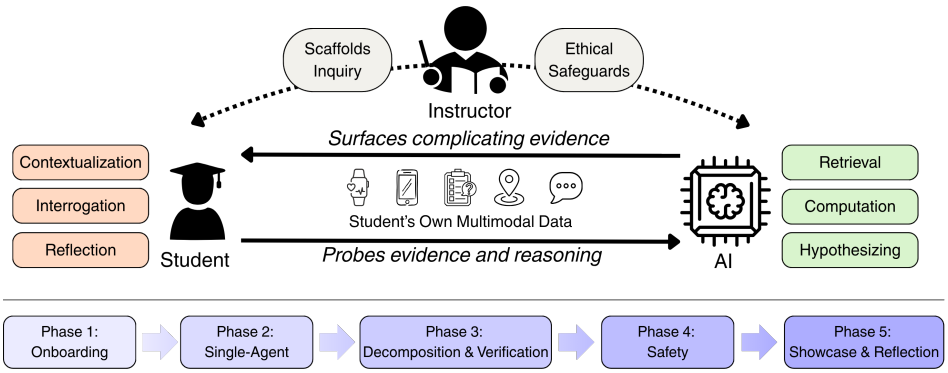


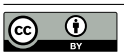
Figure 1: Student-AI Collaborative Inquiry (SACI). A student–AI pair co-investigates wearable-sensor data through mutual interrogation across five phases, with instructor scaffolding and ethical safeguards.

Abstract

Generative AI is in classrooms whether educators planned for it or not; the key pedagogical question is no longer whether students will use it, but what learning relationship its use creates. We use health-agent education over wearable and ubiquitous-sensing data as a stress test: productive AI use here requires students to interpret messy, longitudinal, personally meaningful data and challenge the agent's evidence rather than accept its output as a shortcut. We present **Student-AI Collaborative Inquiry (SACI)**, in which a student and an AI agent co-investigate the student's multimodal sensor data. SACI combines (1) personal data as the shared object of inquiry, (2) mutual interrogation between student and agent as the central pedagogical mechanism, and (3) responsible-use scaffolds for refusal, red-teaming, and research governance. We illustrate SACI through a short-format conference tutorial and a planned term-long course, and outline adoption choices, limitations, and an assessment plan for educators.

CCS Concepts

• **Social and professional topics** → **Computing education**; • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → **Interactive learning environments**; **Health informatics**.



UbiComp Companion '26, Shanghai, China

ACM ISBN 979-8-4007-2533-3/2026/10
https://doi.org/10.1145/3798063.3837303

Keywords

Education, Teaching Innovation, Large Language Models, LLM Agents, Wearable Sensing, Personal Health Informatics, Human-AI Collaboration, Student-AI Collaboration, Collaborative Inquiry

ACM Reference Format:
Zhihan Jiang and Xuhai Xu. 2026. From AI-as-Shortcut to AI-as-Partner: Student-AI Collaborative Inquiry for Health-Agent Education. In *Companion of the 2026 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '26), October 11–15, 2026, Shanghai, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3798063.3837303

1 Introduction

Institutions have responded to generative AI in the classroom with bans, detectors, and syllabus carveouts [4]; we ask instead not how to prevent AI use but how to shape it. Students may treat AI as a shortcut or engage it as a thinking partner; the pedagogical question is which stance instruction cultivates. Health-agent education sits at the sharper end of this question: when an agent reasons over a student's own sensor data, a confidently wrong answer makes a false claim about that student's body, and when the error is invisible in the data itself, only the student's lived context can catch it. Default LLM assistants tend to ratify rather than challenge users' framing [23], so this corrective challenge must be deliberately engineered rather than assumed, a requirement that shapes our model.

Consumer wearables, mobile sensing systems, and self-tracking apps have produced unprecedented streams of multimodal personal health data [9, 10]. Agentic LLMs now offer a path to turn these streams into personal health questions and hypotheses [5, 12, 17, 20]. Well-designed systems allow individuals to investigate their own bodies with analytical rigor; poorly designed systems can

fabricate confident but ungrounded health claims at scale, making responsible health-agent education high-stakes.

While most pedagogical responses position AI as a productivity tool or a threat to integrity [16], the partnership stance remains under-specified as a teachable model with explicit mechanisms, phases, and safety scaffolds for agentic AI reasoning over students' own sensor data. We propose **Student-AI Collaborative Inquiry (SACI)**, a pedagogical model that reframes this relationship: rather than student as builder and agent as artifact to be evaluated, the pair act as co-investigators of the student's own multimodal sensor data; the student contributes contextualization, interrogation, and reflection, while the agent contributes retrieval, computation, and hypotheses. The central pedagogical mechanism is *mutual interrogation*: the student critically probes the agent's reasoning, and the agent's behavior surfaces gaps in the student's understanding. We articulate SACI for personal health agents, where the embodied stake required for mutual interrogation is sharpest, though the mechanism may extend to other classrooms with sensitive personal data. SACI is enacted through a five-phase progression from onboarding through showcase, anchored by a portable responsible-use scaffold. We are developing two instantiations, a short-format hands-on tutorial [11] and a term-long classroom format; this paper presents the model itself (design rationale, adoption considerations, assessment plan) for other educators to adapt.

2 Design Philosophy: Mutual Interrogation

What makes student-AI inquiry *collaborative* rather than tool-use is *mutual interrogation*: a bidirectional probing loop that extends pre-LLM lived-informatics, personal-informatics, and self-tracking traditions [6, 18, 22] into the agentic-AI era. Unlike teacher-led Socratic dialogue or peer critique, SACI makes the AI both an object of scrutiny and a computational partner whose claims can be checked against tool traces and the student's own data. It also differs from teachable agents, where the student authors the agent's knowledge and its failures reveal curricular gaps [2, 13]. Once authorized, the agent can retrieve and analyze the student's sensor data, and its errors carry safety stakes. The student interrogates the agent (*"Which tool call supports that claim?"*); the agent's behavior interrogates the student (*"You attributed your low HRV to overtraining, but your data suggests poor sleep continuity in the same window."*). This bidirectionality is strongest with personally meaningful data; an embodied stake gives each side standing the other lacks. *Co-investigator* names the structure of the loop, not equal epistemic authority: the agent bears no responsibility of its own, and accountability stays with the student and instructor. The AI-to-student direction of the loop, where it volunteers complicating evidence, is a design requirement rather than an emergent property, given the default sycophancy noted above; students engineer this behavior explicitly in Phase 3, keeping the agent's side of the loop observable. We operationalize mutual interrogation as a measurable interaction pattern resembling the ICAP framework's interactive mode [3], though ICAP describes human dyads; whether an engineered partner confers the same benefit is a question this model is built to test.

We elaborate four affordances of the configuration that SACI rests on (an agent layer over the student's own data), which mutual interrogation converts into learning. Only the first has no pre-agent analogue; the others extend self-experimentation antecedents [15] that the agent layer transforms.

- **First-person AI safety intuition.** Watching one's *own* agent confidently fabricate something about one's *own* body produces a safety intuition that third-person case studies cannot replicate. Hallucination ceases to be an abstract failure mode and becomes embodied.
- **Personalization as a teachable engineering skill.** Generic LLM assignments often target general-purpose or population-level performance; collaborative inquiry shifts the design target to an individual's data distribution, objective, and failure modes, making personalization a first-class design challenge students can iterate on.
- **Persistent artifact, not a report.** Classical self-tracking pedagogies typically end with a written reflection or visualization. Collaborative inquiry ends with a prototype agent that the student can continue refining within the project's stated educational scope, changing both the learning outcome (a tool, not a paper) and motivation across the term.
- **Triple-role metacognition.** The student occupies data source, agent builder, first-person user simultaneously and can reflect across them (*"as a user I felt the answer was off; as a builder I now see why"*), an intended mechanism for reflective learning.

These affordances carry costs (consent management, inter-student data variance, loaner devices); Section 4 discusses each.

3 The Pedagogical Model

SACI organizes instruction around a five-phase progression mapping onto an inquiry cycle [21].

Phase 1: Onboarding and data literacy. Students are introduced to the interpretation gap between aggregate dashboards [8] and the open-ended questions people actually ask, and to the multimodal data landscape of wearable streams and self-report. *Rationale.* Without a conceptual map of where the agent fits, hands-on coding decays into mechanical configuration. *Concretely.* Students receive a loaner wearable (or, in short formats, a synthetic persona), query the resulting dashboard, contextualize what they see against their own week (or the persona's narrative), and list five questions it *cannot* answer.

Phase 2: Single-agent prototyping. Students build a minimal single-tool LLM agent end-to-end [24, 27] over their own data (or a synthetic teaching set whose schema mirrors public longitudinal sensing corpora [26]). *Rationale.* Building from a minimal scaffold rather than configuring a framework exposes the design tensions every agent must resolve (what to retrieve, how to plan, when to refuse), yielding a debuggable mental model that transfers to production frameworks. *Concretely.* The lever is *observed failure on owned data*: when a student's own implementation produces confidently fabricated outputs about days they remember living, tool error and ungrounded output cease to be abstract failure modes.

Phase 3: Decomposition and verification. Students separate retrieval, cross-stream synthesis, and counter-evidence checking into independently observable components. *Rationale.* Realistic prototypes benefit from decomposition, but starting there hides the trade-offs that motivate it. Letting students first feel a single agent's limits

makes decomposition a felt need rather than a stipulated requirement, echoing productive failure [14]. *Concretely.* When the Phase 2 scaffold, with its single registered tool, must answer *"why did my HRV drop?"*, it fetches an undifferentiated slab of recent data and anchors on the most salient stream, recent training volume, without surfacing the sleep-continuity evidence in the same window. Registering separate tools is decomposition's first step, but one context still accumulates retrieval, cross-stream reasoning, and safety checking with no independent check on its own synthesis. The invariant is a verifier that tests the working hypothesis against disconfirming evidence, the lever that engineers the agent's side of mutual interrogation; it may be realized as a structured tool call, a separate verification pass over the agent's own draft, or a role-specialized critic agent coordinated through established patterns [25], with the choice determined by the course's technical level and time budget. Each option involves trade-offs among transparency, latency, and answer quality.

Phase 4: Responsible-use stress testing. Students red-team their own and peers' agents for unsafe medical advice, hallucination, and overconfidence, and practice refusal patterns and IRB-style reasoning. *Rationale.* Structured stress testing is intended to make abstract safety failures concrete and memorable. *Concretely.* Edge-case queries (acute symptoms, medication doses, mental-health expressions) become a structured test log, also serving as a student-authored agent-safety checklist. The probe set exercises explicit escalation paths, distinguishing potentially emergent physical symptoms (routed to emergency or urgent care) from self-harm or mental-health crisis (routed to region-specific resources). Crisis probes come only from the scaffold's scripted set and may be skipped without penalty; sessions close with a debrief naming real support resources. Equity stress tests probe whether the agent misapplies population reference ranges or overgeneralizes optical-sensor findings beyond the devices, activities, metrics, and study populations in which they were validated [1], since refusal behavior alone does not address potential inequities in agent reasoning.

Phase 5: Showcase and reflection. Students demonstrate agents on participant-driven questions, surface the design tensions they encountered, and identify open challenges. *Rationale.* Showcase plus reflection consolidates learning and builds confidence to extend the materials into new settings. *Concretely.* Each demo is paired with a structured reflection prompt (*"name one design choice you would change with another week, and why"*), closing the loop from local fix to portable principle.

A running example. A single recurring query, *"Has my heart rate variability dropped this week, and if so, why?"*, anchors the progression in the spirit of problem-based learning [7], tracing design tensions across all five phases: its surfacing among the dashboard's five unanswerable questions (Phase 1); a misleading one-tool-scaffold hypothesis drawn only from recent training volume (Phase 2); a decomposed agent's more grounded but missing-data-sensitive hypothesis triangulating sleep, stress, and activity streams (Phase 3); a peer red-team session that adds an *"insufficient evidence"* path for nights with charging gaps (Phase 4); and a final reflection on which design choices most affected answer quality (Phase 5).

Table 1: Adoption choices in two contrasting settings.

Design choice	Short format	Long format
Duration	A few hours	A full term
Data substrate	Synthetic teaching set	Student's own data
Phase coverage	Compressed Phases 1–5	All five phases
Hands-on format	Pair programming	Cross-disciplinary teams
Showcase	In-session panel	Redacted/private demo day

Illustrative applications. The phase ordering and rationales are venue-agnostic: the short-format instantiation is a hands-on tutorial [11], with Phase 3 compressed to a brief verifier exercise on the scaffold's counter-evidence prompt; the planned long-format version would provide each student with a course-provided Fitbit for a semester of personal data collection. Table 1 summarizes the design choices in these two settings. Personal data is the setting in which SACI's embodied-stake hypothesis is most directly instantiated. Synthetic-data adaptations preserve the agent-design, evidence-grounding, and safety-learning objectives while enabling privacy-preserving participation; future evaluation should determine which outcomes depend specifically on personal data.

4 Adoption Considerations

We summarize the decisions adopting educators face, stating each as a menu of options adaptable to institutional contexts.

Data source. Three paths trade off authenticity against overhead: (i) a synthetic teaching dataset (zero device cost; may provide less personal relevance; appropriate when total contact time is under one day); (ii) students' existing device data, down to phone-native steps and screen time (zero procurement cost; widely available but not universal at the price of narrower sensor coverage); or (iii) a loaner device pool managed by the instructor (many consumer wearables typically sync data to vendor-managed cloud services; devices are reusable across cohorts, amortizing cost; they should carry student-owned vendor accounts, reset between cohorts, preserving the private-personal default below).

Consent and data governance. Students' personal health data enters classroom artifacts, so explicit consent is required: the student controls whether and how their data is used, and may pause collection or opt out at any point. The embodied stake comes from *analyzing* one's own data, not from *disclosing* it, so governance yields three participation modes, mixable per activity. *Private-personal*: the student analyzes their own data without exposing it in class, with no peer viewing and no instructor access to raw streams unless the student requests it. *Synthetic-collaborative*: peer red-teaming and showcases run on shared synthetic personas, or on coarse aggregates from consenting students with the data owner present under agreed confidentiality norms; sensitive probe categories default here. *Fully-synthetic*: every activity uses the teaching set. Disclosure is never treated as engagement: no grade, rubric score, showcase standing, or reflective prompt requires revealing the health content behind a design decision. The first two modes preserve the embodied stake; only the third attenuates it, and it stays available without penalty, for wellbeing as well as privacy, since self-tracking can itself distress some students. As minimums, we recommend: independent opt-in per data-collection activity; no grade decisions on raw personal data; on-demand export and deletion, with post-course persistence and withdrawal handling stated upfront; and any

research retention, coding, or analysis of course materials begins only after grades are submitted and separate consent is obtained, subject to local IRB requirements. Course artifacts are educational, not diagnostic (stated at consent): if data or agent output suggests a possible health concern, the instructor follows a prepared script and referral sheet rather than ad hoc medical judgment.

LLM backend. Either a hosted API (lowest setup cost; usage-dependent recurring cost; but personal data leaves the classroom) or a locally hosted open-weight model (greater privacy; requires GPU access). Many consumer wearables also sync raw data to the vendor's cloud, regardless of where the LLM runs; hosted-API use with real data should require zero retention, no training terms disclosed at consent. Long formats with real data may favor local hosting; synthetic short formats can use either.

Responsible-use scaffold. The released scaffold is portable across settings:[1] it includes scope and evidence-provenance norms, refusal templates, a counter-evidence verifier, red-team probes, data-retention and escalation guidance, an incidental-findings pathway, and an IRB decision flowchart.

5 Assessment Plan

We propose assessing each adoption along four dimensions, with measures tailored to its format.

Interdisciplinary fluency. Technical-track students' confidence with health data standards, privacy norms, and medical-advice ethics; health-track students' confidence with LLM limitations, agent loops, grounded response generation, and API integration. A common item subset provides a basis for cross-format comparison, whose measurement properties will require validation. Since confidence is not competence, each item is paired with a performance anchor (e.g., debugging a broken agent loop), and a post-course transfer probe has students locate the planning loop and refusal logic in an unfamiliar framework.

Artifact quality. The rubric scores three properties of the prototype agent. *Reliability* captures whether the agent handles noisy, real-world data without silent failure, with top scores requiring graceful, acknowledged failure on missing data and low scores for silently fabricated values. *Safety* captures whether the agent appropriately refuses out-of-scope medical queries while remaining helpful and appropriately bounded within scope. *Faithfulness* captures whether generated claims trace to retrieved data, with explicit citations of which tool calls supplied which evidence. Per Section 4, *Reliability* and *Faithfulness* are graded on synthetic-set runs or student-submitted trace excerpts, never on raw streams.

Reflective and red-team outcomes. Exit interviews surface newly understood agent failure modes through prompts like *"Describe a failure your own agent produced that you did not anticipate, and what you changed in response."* Red-team outcomes document the categories of unsafe behavior elicited, their severity, and the fixes subsequently implemented.

Mutual interrogation pattern. Sessions can be coded for three signals: (i) evidence-demanding student turns (e.g., "which tool call supports this?"); (ii) agent responses that trace evidence to specific tool outputs, qualify their confidence, or complicate the student's working hypothesis; and (iii) instances where the student revises a hypothesis in response to an agent output. Signal (ii) checks the engineered agent side; (i) and (iii) the student side. These are process measures: they establish that the pattern occurred, not that it produced learning; the performance anchors and transfer probe above supply the outcome evidence. Only transcripts that students have reviewed and redacted are retained and coded for research under separate post-grading consent.

[1] https://github.com/llm-health-agent-tutorial/llm-health-agent-starter/blob/main/SACI_SCAFFOLD.md

6 Limitations and Open Questions

As a model-stage contribution, SACI faces validity threats: early-cohort gains may reflect Hawthorne effects [19] or the disproportionate motivation of early adopters. Transferability to resource-constrained settings depends on loaner devices, TA staffing, and consent infrastructure; the synthetic-data variant preserves mechanical skills but attenuates the embodied-stake mechanism. Future evaluation must include a baseline against conventional LLM tutorials. Adopters using real wearable data should engage canonical sensing pitfalls, including missing-not-at-random patterns, validation differences across devices and activity contexts [1], and time-zone or daylight-saving-time artifacts. Applicable privacy and education-record requirements vary by jurisdiction and institutional context.

Fairness under heterogeneous student data. Inter-student variance complicates uniform grading: a student whose wearable failed or was uncharged for three days has different debugging artifacts than one with dense, clean data. We grade *outcome* criteria (*Reliability*, *Faithfulness*) on the shared synthetic set, and *process* criteria on the student's own traces against a fixed bar (did the student detect and respond to missingness?).

Scaffolding-vs-discovery balance. If the starter scaffold preconfigures too much, students never confront the design tensions it was meant to illustrate; if too little, the course becomes infrastructure plumbing. We lean toward *preconfigure infrastructure, scaffold pedagogy*: ship a runnable agent loop and data loader, leaving planning logic, tool selection, and refusal wiring open for students.

Peer-data exposure. Red-teaming peer agents is promising but requires careful consent, so synthetic-collaborative mode is the default; shared synthetic personas scale it to large cohorts.

7 Conclusion

SACI articulates the partnership stance: student and AI as co-investigators of the student's own sensor data. This paper provides a testable mechanism (mutual interrogation), a five-phase progression that enacts it, and a governance framework, anchored by a released responsible-use scaffold, that supports safer classroom evaluation. Educators can adapt these wherever students hold a personal stake in the data the AI reasons over; the first instantiation runs at this conference.

Acknowledgments

Research reported in this publication was supported in part by the National Science Foundation under Award Number 2615689. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation. We also thank the Columbia University Research Stabilization Grant for supporting this research project.

References

[1] Brinnae Bent, Benjamin A. Goldstein, Warren A. Kibbe, and Jessilyn P. Dunn. 2020. Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors. *npj Digital Medicine* 3, Article 18 (2020). doi:10.1038/s41746-020-0226-6

[2] Catherine C. Chase, Doris B. Chin, Marily A. Oppezzo, and Daniel L. Schwartz. 2009. Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology* 18, 4 (2009), 334–352. doi:10.1007/s10956-009-9180-4

[3] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. doi:10.1080/00461520.2014.965823

[4] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2024. Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education and Teaching International* 61, 2 (2024), 228–239. doi:10.1080/14703297.2023.2190148

[5] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai "Orson" Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)* 8, 2, Article 56 (2024), 25 pages. doi:10.1145/3659604

[6] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A Lived Informatics Model of Personal Informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, New York, NY, USA, 731–742. doi:10.1145/2750858.2804250

[7] Cindy E. Hmelo-Silver. 2004. Problem-Based Learning: What and How Do Students Learn? *Educational Psychology Review* 16, 3 (2004), 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3

[8] Zhihan Jiang, Handi Chen, Rui Zhou, Jing Deng, Xinchen Zhang, Running Zhao, Cong Xie, Yifang Wang, and Edith C. H. Ngai. 2024. HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 1205–1215. doi:10.1109/TVCG.2023.3326943

[9] Zhihan Jiang, Lin Lin, Xinchen Zhang, Jianduo Luan, Running Zhao, Longbiao Chen, James Lam, Ka Man Yip, Hung Kwan So, Wilfred H. S. Wong, Patrick Ip, and Edith C. H. Ngai. 2023. A Data-Driven Context-Aware Health Inference System for Children during School Closures. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)* 7, 1, Article 18 (2023), 26 pages. doi:10.1145/3580800

[10] Zhihan Jiang, Vera van Zoest, Weipeng Deng, Edith C. H. Ngai, and Jiangchuan Liu. 2023. Leveraging Machine Learning for Disease Diagnoses Based on Wearable Devices: A Survey. *IEEE Internet of Things Journal* 10, 24 (2023), 21959–21981. doi:10.1109/JIOT.2023.3313158

[11] Zhihan Jiang, Will Ke Wang, Georgianna Lin, Brenna Li, and Xuhai Xu. 2026. Tutorial: From Multimodal Sensing Data to Actionable Health Insights—A Hands-on Guide to Prototyping Your Personal LLM Health Agent. In *Proceedings of the 2026 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the ACM International Symposium on Wearable Computers (UbiComp '26)*. Shanghai, China. doi:10.1145/3798063.3836753

[12] Zhihan Jiang, Running Zhao, Lin Lin, Yue Yu, Handi Chen, Xinchen Zhang, Xuhai Xu, Yifang Wang, Xiaojuan Ma, and Edith C. H. Ngai. 2026. DietGlance: Dietary Monitoring and Personalized Analysis at a Glance with Knowledge-Empowered AI Assistant. *ACM Transactions on Computing for Healthcare* 7, 2, Article 34 (2026), 46 pages. doi:10.1145/3797883

[13] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, Article 652, 28 pages. doi:10.1145/3613904.3642349

[14] Manu Kapur. 2008. Productive Failure. *Cognition and Instruction* 26, 3 (2008), 379–424. doi:10.1080/07370000802212669

[15] Ravi Karkar, Jasmine Zia, Roger Vilardaga, Sonali R. Mishra, James Fogarty, Sean A. Munson, and Julie A. Kientz. 2016. A Framework for Self-Experimentation in Personalized Health. *Journal of the American Medical Informatics Association* 23, 3 (2016), 440–448. doi:10.1093/jamia/ocv150

[16] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences* 103, Article 102274 (2023). doi:10.1016/j.lindif.2023.102274

[17] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. In *Proceedings of the Fifth Conference on Health, Inference, and Learning (CHIL) (Proceedings of Machine Learning Research, Vol. 248)*. PMLR, 522–539. https://proceedings.mlr.press/v248/kim24b.html

[18] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 557–566. doi:10.1145/1753326.1753409

[19] Jim McCambridge, John Witton, and Diana R. Elbourne. 2014. Systematic Review of the Hawthorne Effect: New Concepts Are Needed to Study Research Participation Effects. *Journal of Clinical Epidemiology* 67, 3 (2014), 267–277. doi:10.1016/j.jclinepi.2013.08.015

[20] Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. 2026. Transforming Wearable Data into Personal Health Insights Using Large Language Model Agents. *Nature Communications* 17, Article 1143 (2026). doi:10.1038/s41467-025-67922-y

[21] Margus Pedaste, Mario Mäeots, Leo A. Siiman, Ton de Jong, Siswa A. N. van Riesen, Ellen T. Kamp, Constantinos C. Manoli, Zacharias C. Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of Inquiry-Based Learning: Definitions and the Inquiry Cycle. *Educational Research Review* 14 (2015), 47–61. doi:10.1016/j.edurev.2015.02.003

[22] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2014. Personal Tracking as Lived Informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, USA, 1163–1172. doi:10.1145/2556288.2557039

[23] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR)*.

[24] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

[25] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In *Conference on Language Modeling (COLM)*.

[26] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman S. Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak N. Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2022. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)* 6, 4, Article 190 (2022), 34 pages. doi:10.1145/3569485

[27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.