EduPolicySim: LLM-Agent Simulation for Navigating the Pedagogical Policy Space

Weirui Peng
weiruip@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Hyoungwook Jin
jinhw@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Ethan Beaird
beaird@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Xu Wang
xwanghci@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Abstract

Pedagogical policy design requires navigating a combinatorially vast space of instructional choices—such as technique, dosage, and timing—whose effects interact in ways that are difficult to anticipate. Existing LLM-based student simulations operate at the micro level (individual learner responses) or meso level (classroom dynamics), but do not address macro-level policy evaluation. In this position paper, we propose **EduPolicySim**, a simulation framework that uses LLM agents as proxies for diverse student populations to support pedagogical policy exploration at scale. The framework operates in five stages: student profile configuration grounded in authentic artifacts, policy specification along multiple instructional dimensions, simulation and analysis of how learner populations respond to each policy, deliberative decision support that makes trade-offs among stakeholder priorities explicit, and co-evolutionary refinement using real-world outcome signals. EduPolicySim aims to surface hidden interdependencies among instructional decisions and scaffold evidence-informed policy making before costly classroom deployment.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → *Interactive learning environments.*

Keywords

LLM Agent, Social Simulation, AI for Education

1 Introduction

Large language model (LLM) agents are increasingly used to simulate human behavior in the service of policy design and evaluation. Recent work demonstrates the viability of this approach across domains: Li et al. deployed large-scale agent simulations to inform emergency preparedness planning, including volunteer training and evacuation protocols [13]; Hou et al. modeled vaccine hesitancy under simulated social dynamics to explore public health interventions [4]; and research infrastructures such as Sotopia enable scalable, interactive testing of multi-agent social behaviors [32, 33]. These efforts build on foundational work showing that generative agents can produce believable simulacra of human behavior [19], while also motivating new perspectives on how HCI can engage more deeply with policy processes [28]. At the same time, researchers caution that validation remains a central challenge [12] and that replacing human participants with LLMs risks flattening the identities they purport to represent [23]. A growing consensus holds that LLM-based simulations can serve as a promising research method when treated as complementary evidence rather than ground truth [1].

In education, LLM-based agents have been used to simulate students for instructional design, explore classroom interaction dynamics, and provide scenario-based practice for instructor training [14, 22, 24, 31]. However, this body of work operates primarily at the *micro level*—modeling instructor–student conversations to improve learning outcomes or teaching efficiency. What remains largely unexplored is the use of LLM-agent simulation at the *macro level*: systematically evaluating how pedagogical policy choices affect learning across diverse student populations.

This gap matters because pedagogical policy design is a profoundly complex problem [11]. Instructors and policy designers must make decisions along numerous interdependent dimensions—among them, instructional technique (e.g., worked examples vs. problem solving), dosage (e.g., spacing of practice), and timing (e.g., when to provide feedback). As illustrated in Figure 1, choices on each dimension can be independently combined, producing a combinatorially vast space of possible instructional configurations. It is infeasible to experimentally evaluate all such configurations with real students and classrooms, yet the interaction effects among these dimensions mean that the effectiveness of any single choice depends on the others.

In this position paper, we propose an LLM-agent simulation framework that enables educators and researchers to explore this design space at scale. By simulating how diverse learner populations respond to different pedagogical policy configurations, the framework aims to surface hidden interdependencies among instructional decisions and help practitioners anticipate the downstream effects of policy choices before deploying them in real classrooms.

This paper was prepared for PoliSim@CHI 2026: LLM Agent Simulation for Policy, a workshop at CHI 2026 (CHI Conference on Human Factors in Computing Systems), April 16, 2026, Barcelona, Spain.

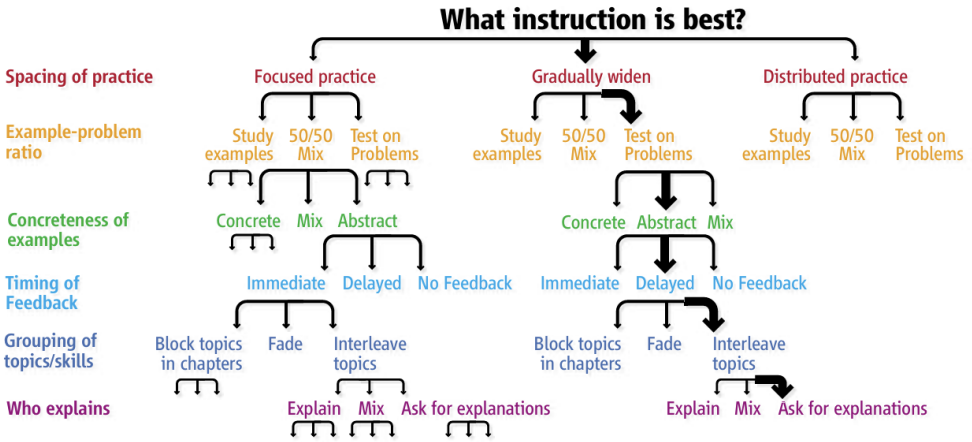

Figure 1: The combinatorial complexity of pedagogical policy design. Choices along independent instructional dimensions—such as technique, dosage, and timing—can be freely combined, yielding a space of trillions of possible configurations. The thick-arrowed path shows one such configuration. This vast design space motivates the need for scalable simulation-based evaluation. Reproduced from [11].

2 Related Work

2.1 Hidden Interdependencies in Pedagogical Policy

Designing effective educational policy requires navigating a landscape of deeply entangled factors, where even well-studied instructional decisions lack clear consensus. A persistent challenge is that many instructional methods are discussed using compelling but imprecise labels rather than operationally grounded definitions [9, 10], making it difficult to compare findings across studies. Even for well-defined practices, empirical evidence frequently points in opposing directions: the literature contains long-standing disagreements on whether feedback should be immediate [2] or delayed [21], and whether learning materials should be concrete [17] or abstract [8].

What makes these contradictions particularly difficult to resolve is that instructional effectiveness is rarely context-free. The same technique may succeed in one setting and fail in another: strategies effective for simple skills can prove counterproductive for complex ones [26], and prompting students to self-explain—a widely recommended practice [20]—does not reliably benefit all learners [27]. Effectiveness further depends on learner characteristics such as prior knowledge and aptitude, with some approaches disproportionately benefiting low-achieving students [3, 7]. These pervasive interaction effects mean that the space of possible policy configurations is combinatorially large, and the consequences of any single decision are contingent on numerous other factors that are difficult to hold in mind simultaneously.

Traditional approaches to understanding this complexity—controlled experiments and meta-analyses—can only probe a small slice of this vast design space at a time. This leaves policymakers and educators without practical tools for reasoning about how multiple instructional choices interact across diverse student populations. Surfacing these hidden interdependencies at scale is precisely the kind of task where LLM-based simulation can offer a new lens, enabling systematic exploration of how policy choices propagate through varied learner profiles and instructional contexts.

2.2 LLM-Simulated Students

A growing body of work uses LLM-based agents to simulate students, spanning applications from assessment design to classroom modeling to instructor training. We organize this literature by the *level* at which simulation operates, and identify a gap at the policy level that motivates our work.

Individual learner simulation. At the level of individual students, LLM agents have been used to evaluate and stress-test instructional materials. *Generative Students* [14] employs GPT-based learners with varied mastery levels to answer assessment items, enabling instructors to estimate item difficulty and diagnose flawed questions before classroom deployment. To move beyond idealized correctness, Sonkar et al. [22] fine-tune LLMs on paired correct and incorrect solutions, producing simulated students that reproduce authentic algebraic misconceptions. Wu et al. [25] similarly emphasize cognitive diversity by constraining agents to represent a spectrum from novice to expert, incorporating imperfect reasoning to better stress-test tutoring strategies.

Classroom dynamics simulation. Scaling beyond individual learners, several systems simulate multi-party classroom interactions. SimClass [30] instantiates a virtual classroom with LLM-driven teacher and student agents, using role-specific prompting to capture participation patterns, turn-taking, and emergent group behaviors. MathVC [29] extends this paradigm with multiple AI peers holding distinct personas to support simulated collaborative reasoning. Wang et al. [24] further argue that LLM-based social simulation can serve as a general methodology for analyzing collective interaction structures at scale.

Instructor development. LLM-simulated students have also been applied to train human instructors. GPTeach [15] enables novice

instructors to rehearse tutoring conversations with simulated students exhibiting confusion and help-seeking behaviors, while TutorUp [18] simulates disengaged learners to train tutors in adaptive engagement strategies. Complementing instructor training, teachable-agent paradigms use simulated students to support learning-by-teaching [5] and to evaluate AI tutors across diverse learner profiles before deployment [6].

Limitations and the missing policy level. Despite this progress, practitioners report that current simulated students can be overly articulate, affectively flat, and lacking longitudinal consistency—motivating ongoing work on richer persona modeling [16]. More fundamentally, the existing literature operates at the *micro level* (individual learner responses) or *meso level* (classroom interaction patterns), with each study targeting a specific instructional scenario. What remains absent is a *macro-level* simulation framework that evaluates how pedagogical *policy* choices—spanning dimensions such as instructional technique, dosage, and timing [11]—interact to shape learning outcomes across diverse student populations. As discussed in Section 1, the combinatorial vastness of this policy space makes exhaustive empirical evaluation infeasible, yet the pervasive interaction effects among dimensions mean that optimizing any single choice in isolation is insufficient. Our work aims to address this gap by proposing an LLM-agent simulation framework designed to operate at the policy level, enabling educators to explore the pedagogical design space at scale and surface the hidden interdependencies that make instructional policy design so challenging.

3 EduPolicySim: A Simulation Framework for Pedagogical Policy Exploration

Koedinger et al. [11] identify two critical research needs for advancing instructional science: (1) systematic experiments testing *instructional function decomposability*—whether optimal instructional choices are specific to particular combinations of knowledge goals, learning processes, and outcome targets—and (2) *massive online multifactor studies* that vary multiple instructional dimensions simultaneously to determine which can be treated independently. However, such studies face significant practical barriers: real-world experiments at this scale are costly and ethically constrained, and they lack convenient access to long-term outcome variables [11].

We propose **EduPolicySim**, a simulation-based framework that addresses these barriers by using LLM agents as proxies for diverse student populations. Rather than replacing empirical research, EduPolicySim serves as a *computational sandbox* for pedagogical policy exploration—enabling researchers and instructors to rapidly prototype and compare instructional configurations before committing to costly real-world deployments.

3.1 Framework Design

The framework operates in five stages. To ground the description, we trace a hypothetical example throughout: an introductory statistics instructor designing a three-week unit on hypothesis testing for a class of 120 undergraduates with heterogeneous mathematical backgrounds. While simplified, this scenario illustrates how the stages connect in practice.

Stage 1: Student profile configuration. Instructors populate simulated student profiles by providing authentic artifacts—such as prior student work, assessment results, and qualitative observations—that reflect learner capabilities, prior knowledge, and individual characteristics. These artifacts ground LLM agent personas in realistic learner variability rather than relying solely on abstract ability parameters.

Example. The statistics instructor uploads anonymized diagnostic quiz scores, written explanations from a prior probability unit, and LMS engagement logs for a representative sample of students. From these artifacts, the framework constructs profiles capturing meaningful variation: for instance, a student with strong procedural fluency but weak conceptual understanding of sampling distributions, or a student with high course engagement but a persistent misconception that larger samples always yield statistically significant results.

Stage 2: Policy specification. Instructors articulate learning objectives and configure pedagogical policies along multiple instructional dimensions identified in the literature [11]: instructional technique (e.g., worked examples vs. problem solving), dosage (e.g., spacing of practice), timing (e.g., immediate vs. delayed feedback), concreteness of examples, and grouping of skills. The framework treats these dimensions as independently combinable, enabling systematic exploration of the design space illustrated in Figure 1.

Example. The instructor specifies the learning objective—correctly conducting and interpreting a one-sample z-test—and configures two candidate policies to compare:

- **Policy A:** worked examples followed by problem solving; massed practice; immediate correctness feedback; concrete real-world contexts (e.g., clinical trial data).
- **Policy B:** problem solving first with scaffolded hints; spaced practice distributed over three weeks; delayed explanatory feedback; abstract contexts (e.g., generic population parameters).

These two policies differ on four dimensions simultaneously, reflecting the kind of multifactor comparison that is impractical to run as a controlled experiment with real students.

Stage 3: Simulation and analysis. The framework simulates how the configured student population responds to each specified policy over a defined instructional sequence. Importantly, student profiles are not static: as simulated instruction unfolds, agent states evolve to reflect learning gains, shifting misconceptions, and changes in engagement—capturing the dynamic nature of real classrooms rather than producing a single-point prediction.

Simulation outputs operate at two levels of granularity. At the *aggregate level*, instructors compare how different policy configurations affect overall class performance, identifying which combinations of instructional dimensions yield stronger outcomes. At the *individual level*, the framework surfaces which student profiles struggle or thrive under a given policy, enabling targeted attention to at-risk learners.

Example. The simulation reveals that Policy A produces higher average scores on procedural assessment items, but students with low prior knowledge plateau early under massed practice. Policy B yields slower initial gains yet stronger transfer performance,

particularly for students with weak probability foundations. The individual-level analysis further flags that high-anxiety student profiles disengage under Policy B's delayed feedback.

Stage 4: Deliberative decision support. Simulation outputs rarely point to a single dominant policy. More commonly, they expose *trade-offs*—between aggregate efficiency and equity across subgroups, between short-term procedural mastery and long-term transfer, or between instructional coverage and depth. Resolving these trade-offs is not a technical optimization problem; it requires value judgments that may differ across stakeholders. An instructor may prioritize closing achievement gaps for struggling students, a curriculum coordinator may weigh content coverage for accreditation, and a department chair may focus on scalability and cost. The same simulation data can legitimately support different policy conclusions depending on whose priorities take precedence.

EduPolicySim is designed to support this deliberative process rather than short-circuit it. The framework structures its outputs to make trade-offs explicit and comparable: for a given pair of policies, it can highlight which student subgroups gain and which lose, quantify the magnitude of predicted differences, and flag dimensions along which policies diverge most sharply. By externalizing the consequences of each choice, the framework provides a shared evidential basis around which stakeholders with different roles and values can negotiate—examining the same results while articulating why they weigh them differently.

Crucially, the framework also supports *iterative policy exploration* to facilitate this negotiation. After reviewing simulation results, stakeholders can jointly adjust policy parameters—changing feedback timing, example concreteness, or practice spacing—and re-simulate to observe how outcomes shift. This what-if cycle allows practitioners to systematically navigate the combinatorial policy space described in Section 1, comparing alternative configurations side by side. Rather than forcing premature convergence on a single design, the iterative loop encourages stakeholders to ask targeted questions—*What if we keep spaced practice but switch to immediate feedback for the lowest quartile?*—and examine the projected consequences before committing to classroom deployment.

Example. Faced with the simulation results above, the instructor and a curriculum coordinator interpret the data differently. The instructor is concerned that Policy A leaves low-prior-knowledge students behind; the coordinator notes that Policy B's slower pacing risks not covering required content before the midterm. Rather than choosing one policy by default, they use the framework to explore a hybrid configuration—pairing Policy A's immediate feedback with Policy B's spaced practice schedule—and re-simulate. The hybrid narrows the gap between high- and low-prior-knowledge students while maintaining engagement and staying within the content timeline, representing a negotiated compromise that neither stakeholder would have arrived at independently.

Stage 5: Co-evolutionary refinement. Once a policy is deployed in a real classroom, EduPolicySim supports a *co-evolutionary* feedback loop between simulation and practice. Real-world outcome data—such as actual student performance, engagement patterns, and observed gaps between predicted and actual responses—are fed back into the framework as *outcome signals* that update and recalibrate student profiles and simulation parameters. This serves two purposes. First, it improves *simulation fidelity*: as the framework accumulates real evidence, its predictions become increasingly aligned with the specific student population. Second, it enables *adaptive policy revision*: instructors can re-enter the simulation loop with updated profiles to ask whether, given what has actually happened in the classroom, a different policy would now be more effective. This co-evolutionary design reflects the reality that pedagogical policy is not a one-time decision but an ongoing process of adjustment, and positions EduPolicySim as a living tool that grows more useful over the course of instruction.

Example. After deploying the hybrid policy, the instructor collects midterm results and finds that transfer performance broadly matches the simulation's predictions. However, a subgroup of non-native English speakers struggles with abstract problem framings—a gap the original profiles did not capture. The instructor feeds these outcomes back into the framework; the recalibrated profiles now account for language-related barriers to abstraction. A re-simulation for the upcoming unit on confidence intervals suggests that switching to concrete, contextualized examples for this subgroup would improve comprehension without affecting the rest of the class—demonstrating how real-world evidence progressively sharpens both the simulation and the instructional decisions it informs.

4 Conclusion

Pedagogical policy decisions are shaped by interaction effects across multiple instructional dimensions, yet the tools available to educators offer limited support for reasoning about this complexity. We presented EduPolicySim, a framework that uses LLM-agent simulation to enable macro-level pedagogical policy exploration—complementing existing work that operates at the individual learner or classroom interaction level. By supporting iterative what-if exploration and co-evolutionary refinement grounded in real-world outcomes, the framework aims to help instructors surface hidden interdependencies and make more informed policy choices. Key open challenges include validating simulation fidelity against real student populations, modeling longitudinal learning trajectories, and ensuring that simulated personas faithfully represent learner diversity without flattening individual differences. We see EduPolicySim not as a replacement for empirical research, but as a computational sandbox that can accelerate the kind of massive multifactor studies that instructional science requires [11].

References

[1] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael S. Bernstein. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025) (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 81005–81034. https://proceedings.mlr.press/v267/anthis25a.html

[2] Albert T Corbett and John R Anderson. 2001. Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 245–252.

[3] Richard E. Goska and Phillip L. Ackerman. 1996. An aptitude-treatment interaction approach to transfer within training. *Journal of Educational Psychology* 88 (1996), 249–259.

[4] Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can a Society of

Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *arXiv preprint arXiv:2503.09639* (2025). https://arxiv.org/abs/2503.09639

[5] 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 2024 CHI Conference on Human Factors in Computing Systems*. 1–28.

[6] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.

[7] Slava Kalyuga. 2007. Enhancing Instructional Efficiency of Interactive E-learning Environments: A Cognitive Load Perspective. *Educational Psychology Review* 19 (2007), 387–399.

[8] Jennifer A Kaminski, Vladimir M Sloutsky, and Andrew F Heckler. 2008. The advantage of abstract examples in learning math. *Science* 320, 5875 (2008), 454–455.

[9] David Klahr. 2013. What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences* 110, supplement_3 (2013), 14075–14080.

[10] David Klahr and Junlei Li. 2005. Cognitive research and elementary science instruction: From the laboratory, to the classroom, and back. *Journal of Science Education and Technology* 14, 2 (2005), 217–238.

[11] Kenneth R Koedinger, Julie L Booth, and David Klahr. 2013. Instructional complexity and the science to constrain it. *Science* 342, 6161 (2013), 935–937.

[12] Maik Larooij and Petter Törnberg. 2026. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review* 59 (2026), 15. doi:10.1007/s10462-025-11412-6

[13] Yuxuan Li, Sauvik Das, and Hirokazu Shirado. 2025. What Makes LLM Agent Simulations Useful for Policy? Insights From an Iterative Design Engagement in Emergency Preparedness. *arXiv preprint arXiv:2509.21868* (2025).

[14] Xinyi Lu and Xu Wang. 2024. Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (2024). https://api.semanticscholar.org/CorpusId:269922100

[15] Jordan M. Markel, Sophia G. Opfermann, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the 2023 ACM Conference on Learning at Scale (L@S)*.

[16] Diana Martynova, Jana Macina, Niklas Daheim, Oğuzhan N. Yalçın, Xinyi Zhang, and Mrigank Sachan. 2025. Can LLMs Effectively Simulate Human Learners? Teachers' Insights from Tutoring LLM Students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

[17] Allan Paivio. 1965. Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior* 4, 1 (1965), 32–38.

[18] Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. 2025. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[19] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[20] Harold Pashler et al. [n. d.]. Organizing Instruction and Study to Improve Student Learning.

[21] Richard A Schmidt and Robert A Bjork. 1992. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science* 3, 4 (1992), 207–218.

[22] Shashank Sonkar, Xinghe Chen, Myco Le, Naiming Liu, Debshila Basu Mallick, and Richard Baraniuk. 2024. Code soliloquies for accurate calculations in large language models. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 828–835.

[23] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models that replace human participants can harmfully misportray and flatten identity groups. *arXiv preprint arXiv:2402.01908* (2024). https://arxiv.org/abs/2402.01908

[24] Ruotong Wang, Xinyi Zhou, Lin Qiu, Joseph Chee Chang, Jonathan Bragg, and Amy X Zhang. 2025. Social-RAG: Retrieving from Group Interactions to Socially Ground AI Generation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[25] Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. Embracing Imperfection: Simulating Students with Diverse Cognitive Levels Using LLM-based Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[26] Gabriele Wulf and Charles H. Shea. 2002. Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review* 9 (2002), 185–211.

[27] Ruth Wylie, Kenneth R. Koedinger, and Teruko Mitamura. 2009. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Wheat Ridge, CO, 1300–1305.

[28] Qian Yang, Richmond Y. Wong, Steven J. Jackson, Sabine Junginger, Margaret D. Hagan, Thomas Gilbert, and John Zimmerman. 2024. The Future of HCI-Policy Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642771

[29] Murong Yue, Wenhan Lyu, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2024. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *arXiv preprint arXiv:2404.06711* (2024).

[30] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 10364–10379.

[31] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating Classroom Education with LLM-Empowered Agents. *ArXiv* abs/2406.19226 (2024). https://api.semanticscholar.org/CorpusId:270764782

[32] Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, et al. 2025. SOTOPIA-S4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*. 350–360.

[33] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667* (2023).