

What Makes LLM Agent Simulations *Useful* for Policy Practice? An Iterative Design Study in Emergency Preparedness

Yuxuan Li
School of Computer Science, Carnegie Mellon University
Pittsburgh, United States
yuxuanll@andrew.cmu.edu

Sauvik Das
School of Computer Science, Carnegie Mellon University
Pittsburgh, United States
sauvik@cmu.edu

Hirokazu Shirado
School of Computer Science, Carnegie Mellon University
Pittsburgh, United States
shirado@cmu.edu

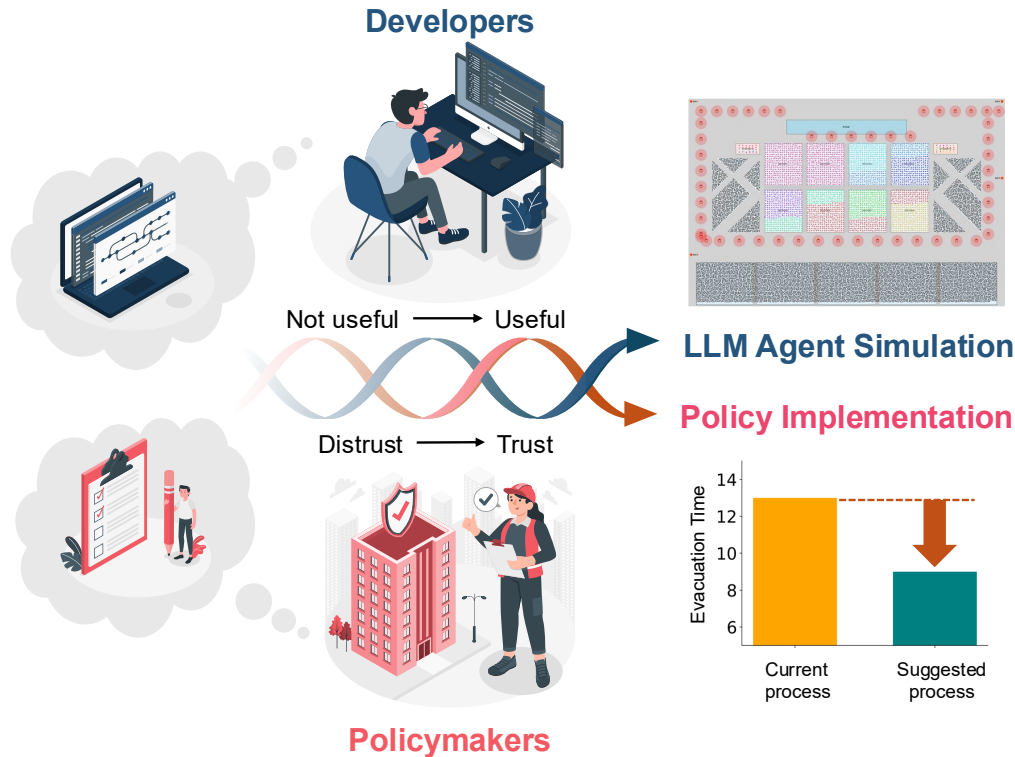


Figure 1: An iterative design engagement with emergency preparedness policymakers developed LLM agent simulations to inform policy implementation. Through cycles of design and validation, simulations shifted from being distrusted to trusted and were ultimately integrated into preparedness practice. This usefulness did not follow a linear path from technical sophistication to institutional adaptation, but instead emerged through an iterative, stakeholder-engaged design process.

Abstract

Policymakers must often act under conditions of deep uncertainty, such as emergency response, where predicting the specific impacts of a policy *a priori* is implausible. Large Language Model (LLM) agent simulations have been proposed as tools to support policymakers under these conditions, yet little is known about how such simulations become useful for real-world policy practice. To address this gap, we conducted a year-long, stakeholder-engaged design process with a university emergency preparedness team. Through iterative design cycles, we developed and refined an LLM agent

simulation of a large-scale campus gathering, ultimately scaling to 13,000 agents modeling crowd movement and communication under various emergency scenarios. Rather than producing predictive forecasts, these simulations supported policy practice by shaping volunteer training, evacuation procedures, and infrastructure planning. Analyzing these findings, we identify three design process implications for making LLM agent simulations useful for policy practice: start from verifiable scenarios to bootstrap trust, use preliminary simulations to elicit tacit domain knowledge, and treat simulation capabilities and policy implementation as co-evolving.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Natural language processing*.

Keywords

Social simulation, Policymaking, Iterative design, Large language model, LLM agent

1 Introduction

Helping policymakers act under deep uncertainty—situations like emergency response where outcomes depend on many unknowable factors—is a longstanding challenge in HCI and decision-making research [21, 33]. Large language model (LLM) agent simulations have been proposed as tools to support policymakers in such contexts: they use natural language to generate emergent social behaviors—coordination, collective decision-making, information propagation—at scale [11, 26, 30]. A growing body of work envisions policymakers using these systems to experiment with alternative scenarios before implementation [14, 15, 17, 20].

Yet what it means for such simulations to be *useful* for policy remains unclear. Existing work implicitly frames policy in terms of prediction or optimization, whereas policy practice involves interpreting uncertain information, coordinating across roles, developing procedures, training personnel, and adapting resources under institutional constraints [21]. Moreover, LLM simulations may reproduce stereotypes or generate outputs that are easy to read but difficult to causally defend within policy practice [3, 23]—concerns that foreground the need to understand the conditions under which they can be meaningfully integrated without causing harm.

We address the central question: **How can LLM agent simulations be made genuinely useful for policy practice?** Rather than treating simulations as standalone analytical artifacts, we take an HCI-grounded, participatory design approach [27, 28]: we study LLM agent simulations as *interactive tools* whose usefulness depends less on model capability than on how they are iteratively developed with policymakers and grounded in concrete policy implementation work.

We report on a year-long iterative design engagement with a university emergency preparedness team. Through five design phases, we developed and refined an LLM agent simulation of commencement—a large-scale campus gathering—ultimately scaling to 13,000 agents modeling crowd movement and communication under various emergency scenarios. Rather than producing predictive forecasts, these simulations came to support policy practice by shaping volunteer training, evacuation procedures, and infrastructure planning.

Our contributions are: (1) an in-depth account of iterative LLM agent simulation design for emergency preparedness, demonstrating how sustained collaboration turned academic prototypes into institutional tools; (2) five mechanisms through which simulations became useful, trustworthy, and actionable within policy contexts; and (3) three design process implications for practitioners seeking to develop LLM agent simulations that support policy practice.

2 Background

2.1 LLM Agent Simulations for Policy

Traditional agent-based models (ABMs) have proven effective for modeling physical evacuation dynamics but rely on fixed rules that struggle to capture the ambiguity and relational complexity of human decision-making during crises [13, 18, 31]. LLM agents offer a promising alternative, introducing semantic reasoning and generative behavior that has been applied to epidemic response, misinformation, and environmental decision-making [15, 17, 22]. Yet despite rapid technical progress, these simulations remain confined largely to academic demonstration, evaluated on internal coherence rather than institutional adoption [19, 35]. Our work shifts focus from refining agent architectures to designing processes that make simulations useful in institutional policymaking.

2.2 HCI, Policy, and Stakeholder Engagement

HCI has long engaged with civic technologies and decision-support systems, emphasizing alignment with institutional constraints [4, 8, 9]. Traditions of participatory design demonstrate how iterative prototyping and feedback can surface hidden needs and reconcile value tensions [2, 16, 28]—yet policy actors remain an underrepresented stakeholder in HCI research [34]. With AI systems, new challenges emerge: probabilistic, opaque outputs complicate co-design and trust-building [1, 10]. This study contributes to these engagement traditions by examining the specific challenges of LLM agent simulations in emergency preparedness policy, where outputs are generative, collective, and contextual requirements often tacit.

3 Research Context and Methods

3.1 Emergency Preparedness as Design Context

Emergency preparedness is a fitting domain for studying LLM agent simulation utility: field experiments are ethically impossible, observational data are scarce, and policymakers must make high-stakes decisions under uncertainty [7, 12]. Contemporary preparedness challenges are increasingly shaped by collective communication dynamics—how messages are framed, trust established, and individuals coordinated under evolving information—alongside physical crowd movement [24, 32]. These are inherently social and semantic dynamics, making emergency preparedness a relevant context for LLM-based simulations.

We partnered with a university emergency preparedness and response team responsible for preparedness across diverse scenarios (large gatherings, natural disasters, active shooters). Their practice was experience-driven and discussion-based, without routine use of simulation tools. Five core members participated: a Senior Director for Disaster Recovery (*P1*), a Senior Director for Reputation Management (*P2*), a Director for Social Media (*P3*), and two Emergency Preparedness Specialists (*P4*, *P5*).

3.2 Iterative Design and Data Collection

We conducted iterative design engagement from May 2024 to August 2025 (16 months), structured around flexible meetings every 2–6 weeks during active phases. The process unfolded in cycles of

needs assessment through interviews, scenario exploration, simulation prototyping, validation, and policy proposals. We collected field notes, meeting transcripts, simulation logs, semi-structured interviews, commencement event recordings, and relevant email correspondence. All procedures were approved by the university’s IRB.

We analyzed qualitative data through open coding of transcripts and emails, with iterative discussion between two researchers to align on codes, followed by thematic analysis to surface recurrent patterns [6, 29]. Details of the simulation system’s final version are provided in the Appendix (Sec. A.3.1).

4 Design Iterations

We developed the simulation through five iterations, each shaped by policymakers’ feedback (see Fig. 2 and Appendix A.1 for full detail).

Preparation (May–Sep 2024). Semi-structured interviews with *P1–P3* surfaced existing workflows and decision-making practices. We co-created stakeholder maps and process maps documenting key actors, responsibilities, and information flows.

Iteration 1: Communication dynamics (Oct 2024). We prototyped a simulation of 100 LLM agents modeling misinformation spread online. Policymakers expressed skepticism: the domain lacked verifiable ground truth. This led to a key shift—we jointly identified commencement evacuation as a more suitable domain, both operationally critical and recurring enough to allow direct observation.

Iteration 2: Simplified commencement (Apr 2025). We built a 500-agent evacuation simulation in an abstracted stadium layout. Policymakers immediately identified missing contextual details—realistic scale, physical constraints, accessibility, coordinator roles—treating the simulation as a diagnostic artifact that revealed what was needed.

Iteration 3: Realistic commencement (late Apr 2025). Incorporating policymakers’ feedback, we increased scale to 3,000 agents, added physical features (seating sections, aisles, bottlenecks), differentiated agent roles (students, accessibility needs, coordinators), and used the official evacuation announcement. This marked a turning point: policymakers began reasoning about concrete planning and immediately recognized the simulation’s value as training material for coordinators, which was documented in the official after-action report.

Iteration 4: Validation with real commencement (May–Jun 2025). We observed the actual 2025 commencement, noting crowd movement patterns. An unanticipated finding—the central role of family members regrouping with students—revealed a critical omission in our model. We redesigned the system to include student, family/friends, and coordinator agents (13,000 total), reflecting the real event scale and venue. Policymakers judged baseline dynamics to align with their experience (Fig. 3), establishing trust to explore hypothetical emergencies.

Iteration 5: Policy exploration (Aug 2025). We examined operationally plausible variations in emergency type, announcement specificity, and coordination strategies. Opening a previously unused northwestern exit significantly improved evacuation efficiency (23.0% relative reduction at the 80% evacuation threshold;

Fig. 4), while coordinator repositioning yielded minimal gains. Policymakers adopted revised coordinator guidance and differentiated emergency protocols into standard operating procedures. The northwest exit entered formal feasibility review. These outcomes were documented in the August 2025 after-action report.

5 Findings and Discussion

Analyzing our design process, we identified five mechanisms through which LLM agent simulations became useful for policy practice (full analysis in Appendix A.2):

- (1) **Validation Filter:** Simulations are useful only in scenarios where outputs can be validated against policymakers’ experience or ground-truth data. The first iteration failed because online communication dynamics could not be verified; commencement succeeded because it recurred and could be directly observed.
- (2) **Trust Bootstrap:** Trust from validating mundane scenarios extends to novel ones. Once policymakers confirmed that baseline crowd dynamics matched their experience, they accepted simulation outputs for hypothetical emergencies they could never directly observe.
- (3) **“Fix-It” Response:** “Wrong” simulations surface tacit domain knowledge. Policymakers could not articulate requirements (family regrouping dynamics, accessibility needs, landmark-based coordination) until they saw outputs that looked off. Even incomplete simulations served as productive technology probes [16].
- (4) **The Details Matter:** Contextual nuance in interaction environments—physical layout, social relationships, procedural roles—is as consequential as agent design for policy use. Under-specified environments produced surface-level plausibility that policymakers found difficult to use or critique meaningfully.
- (5) **Policy–AI Co-evolution:** Simulation capabilities and policy requirements co-evolved, with neither fully specified at the outset. Policymakers clarified that simulations addressed *implementation* processes rather than stable policy commitments, and the most actionable outputs fell within their direct authority (training, coordinator placement, communication).

These findings reframe what “usefulness” means: policymakers were not evaluating predictive accuracy, but asking whether simulation outputs could justify concrete changes in training, staffing, messaging, or infrastructure.

5.1 Design Process Implications

Building on these five mechanisms, we articulate three design process implications distinctive to LLM agent simulations in policy contexts:

5.1.1 Start with Verifiable Scenarios and Build Trust Gradually. Because LLM-generated behaviors are inherently stochastic, trust cannot be established through formal evaluation alone but through alignment with known behavioral patterns. Developers should treat validation as an early design constraint—not a final step—by beginning with recurring, observable scenarios before extending to hypothetical ones. This staged progression (trust bootstrapping)

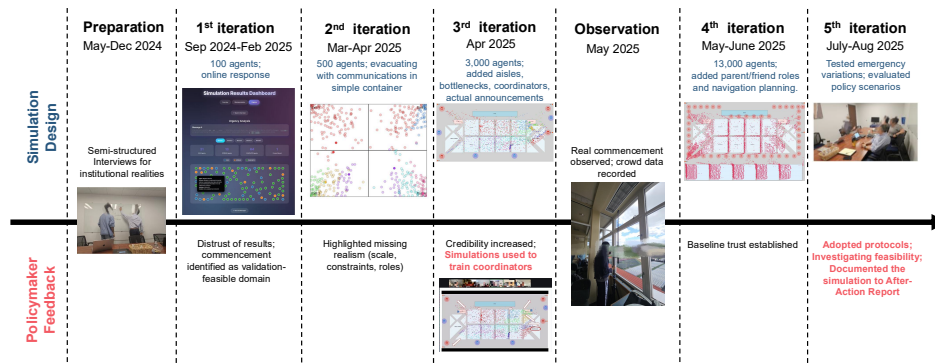


Figure 2: Iterative design process. Across 16 months (May 2024–Aug 2025), we progressed from preparation interviews through five simulation iterations and an in-situ observation. Each phase introduced greater scale and realism (from 100 to 13,000 agents, adding roles, bottlenecks, and family dynamics) and was shaped by policymakers’ feedback. Early iterations surfaced distrust and missing realism; later iterations built credibility and training value; by the final iteration, simulations informed adopted protocols, feasibility assessments, and official after-action reports.

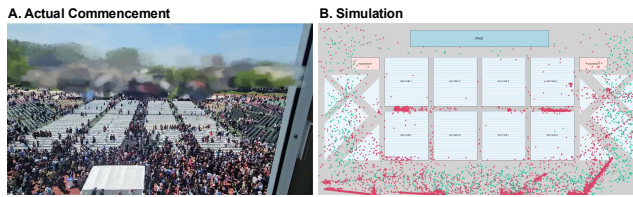


Figure 3: Comparison of crowd movement between empirical observation and simulation. (A) Photograph of the 2025 commencement. (B) Snapshot from the corresponding simulation with 13,000 agents. The simulation reproduced key spatial patterns—including congestion in track areas—supporting alignment between observed and simulated dynamics.

enables LLM agent simulations to leverage their generative capacity while remaining usable for policy practice.

5.1.2 Use Preliminary Simulations to Elicit Tacit Knowledge. Deliberately preliminary or “wrong” simulations can provoke rich critique, surfacing tacit domain knowledge that policymakers cannot articulate in advance. Developers should frame outputs with “*what’s missing here?*” rather than “*is this correct?*”, treating interim simulations as probes rather than flawed products. Importantly, elicited feedback often concerns the interaction environment—spatial layouts, social relationships, role structures—not only agent behavior, and these details are ultimately what determines whether a simulation achieves the fidelity needed to inform policy.

5.1.3 Co-evolve Simulation Capabilities and Policy Requirements. Unlike typical user-centered design where needs are assumed stable, simulation capabilities and policymaker requirements must be treated as interdependent and evolving together. Developers should focus on decision spaces where adaptation is institutionally possible—training, resource allocation, communication—while maintaining flexibility for both simulation capabilities and institutional processes to evolve. Technical possibilities and organizational

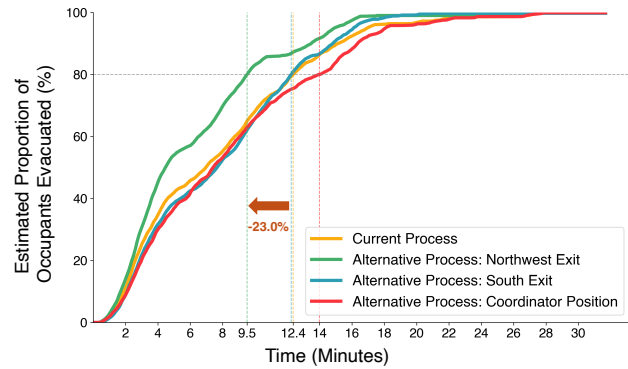


Figure 4: Cumulative evacuation progress across policy alternatives in a severe weather scenario. Opening the northwest exit (green) reduces evacuation time by 23.0% relative to the current process (yellow) at the 80% threshold, while repositioning coordinators or emphasizing another exit yields only marginal gains.

boundaries must be considered jointly: simulation insights were easiest to adopt when the interpreting team could also change the relevant procedures.

5.2 Broader Considerations

LLM agent simulations are not inherently useful; our findings identify enabling conditions: (1) a recurring baseline scenario that can be directly observed; (2) decision levers policymakers can realistically change; and (3) outcomes discussable in operational terms. Domains such as transit disruptions, hospital demand cycles, or large public events may similarly support gradual trust development. Domains lacking stable baselines or placing key levers outside single units—climate adaptation, national security, online misinformation—may require additional validation and organizational alignment beyond what we examined.

Caution is also warranted: as trust builds, model errors and demographic biases may be overlooked [23]. Biases can be leveraged to stress-test policies against real-world disparities, but uncritical reliance risks reinforcing inequities [25].

6 Conclusion

We reported on a year-long iterative design engagement that transformed LLM agent simulations from academic demonstrations into institutionally integrated tools for emergency preparedness planning and training. Usefulness did not arise from technical fidelity alone, but from design practices that gradually built trust, relevance, and shared understanding. Our three design process implications—start with verifiable scenarios, use preliminary simulations to elicit tacit knowledge, and co-evolve simulation capabilities with policy requirements—shift attention from whether LLM agent simulations are “accurate enough” to how they can be responsibly embedded in real organizational workflows. As famously observed, “*All models are wrong, but some are useful*” [5]. Our contribution is to show how, and under what conditions, LLM agent simulations can be designed to be useful for policy practice.

7 Acknowledgments

Portions of the teaser figure (Fig. 1) are adapted from illustrations by Storyset (<https://storyset.com/technology>; <https://storyset.com/work>). We thank Q. Xiao for his insightful advice on the qualitative analysis. This work was supported by the NOMIS foundation and the National Science Foundation under grant #2316768.

References

- Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.
- Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.
- William Agnew, A Stevie Bergman, Jennifer Chien, Mark Diaz, Selim El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.
- Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. ‘It’s Reducing a Human Being to a Percentage’ Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.
- George EP Box. 1976. Science and statistics. *J. Amer. Statist. Assoc.* 71, 356 (1976), 791–799.
- Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- Louise K Comfort. 2007. Crisis management in hindsight: Cognition, communication, coordination, and control. *Public administration review* 67 (2007), 189–197.
- Eric Corbett and Christopher Le Dantec. 2021. Designing civic technology with trust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
- Eric Corbett and Christopher A Le Dantec. 2018. Going the distance: Trust work for citizen participation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- Mateusz Dolata, Norbert Lange, and Gerhard Schwabe. 2024. Development in Times of Hype: How Freelancers Explore Generative AI? *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE) (2024)*, 2257–2269. <https://api.semanticscholar.org/CorpusId:266933328>
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huangdong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984* (2023).
- Dirk Helbing, Hendrik Ammoser, and Christian Kühnert. 2006. Disasters as extreme events and the importance of network interactions for disaster response management. In *Extreme events in nature and society*. Springer, 319–348.
- Dirk Helbing and Peter Molnar. 1995. Social force model for pedestrian dynamics. *Physical review E* 51, 5 (1995), 4282.
- John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.
- 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).
- Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
- Angel Hsing-Chi Hwang, Michael S Bernstein, S Shyam Sundar, Renwen Zhang, Manoel Horta Ribeiro, Yingdan Lu, Serina Chang, Tongshuang Wu, Aimei Yang, Dmitri Williams, et al. 2025. Human Subjects Research in the Age of Generative AI: Opportunities and Challenges of Applying LLM-Simulated Data to HCI Studies. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- Jaekoo Joo, Namhun Kim, Richard A Wysk, Ling Rothrock, Young-Jun Son, Yeong-gwang Oh, and Seungho Lee. 2013. Agent-based simulation of affordance-based human behaviors in emergency evacuation. *Simulation Modelling Practice and Theory* 32 (2013), 99–115.
- Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E Fox. 2025. Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. 2025. LLM Economist: Large Population Models and Mechanism Design in Multi-Agent Generative Simulacra. *arXiv preprint arXiv:2507.15815* (2025).
- Robert Lempert. 2002. Agent-based modeling as organizational and public policy simulators. *Proceedings of the national academy of sciences* 99, suppl_3 (2002), 7195–7196.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957* (2024).
- Yuxuan Li, Hirokazu Shirado, and Sauvik Das. 2025. Actions Speak Louder than Words: Agent Decisions Reveal Implicit Biases in Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25)*. Association for Computing Machinery, New York, NY, USA, 3303–3325. doi:10.1145/3715275.3732212
- Daniel Nilsson and Anders Johansson. 2009. Social influence during the initial phase of a fire evacuation—Analysis of evacuation experiments in a cinema theatre. *Fire safety journal* 44, 1 (2009), 71–79.
- Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
- 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.
- Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.
- Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC press.
- Anselm L Strauss. 1987. *Qualitative analysis for social scientists*. Cambridge university press.
- Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research* (2023).
- Zhanli Sun, Iris Lorscheid, James D Millington, Steffen Lauf, Nicholas R Magliocca, Jürgen Groeneveld, Stefano Balbi, Henning Nolzen, Birgit Müller, Jule Schulze, et al. 2016. Simple or complicated agent-based models? A complicated issue. *Environmental Modelling & Software* 86 (2016), 56–67.
- Anne Templeton, Hui Xie, Steve Gwynne, Aoife Hunt, Pete Thompson, and Gerta Köster. 2024. Agent-based models of social behaviour and communication in evacuations: A systematic review. *Safety Science* 176 (2024), 106520.
- Paul Waddell and Alan Borning. 2004. A case study in digital government: Developing and applying UrbanSim, a system for simulating urban land use, transportation, and environmental impacts. *Social science computer review* 22, 1 (2004), 37–51.
- Qian Yang, Richmond Y Wong, Steven 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*. 1–15.

[35] Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020* (2024).

A Appendix

A.1 Design Iterations: Full Detail

A.1.1 Preparation Interviews: Capturing Needs and Institutional Realities. To ground the design process, we conducted three formative semi-structured interviews with emergency preparedness policymakers (*P1–P3*) during the first five months of the project. Interviews focused on preparedness workflows and policy development during emergencies. Insights were synthesized into stakeholder and process maps that documented key actors, responsibilities, and information flows, and were reviewed with participants. Full maps are included below (Figs. A1, A2, A3, A4, and A5).

A.1.2 First Iteration: Simulating Communication and Misinformation Dynamics. Based on preparatory interviews, we identified online communication and misinformation as primary concerns during emergencies. *P1, P2,* and *P3* were particularly interested in how official announcements might be misinterpreted and how misinterpretations could spread through online networks.

To explore this space, we developed an initial prototype simulating 100 LLM agents interacting online. Agents, prompted to behave as students, received official lockdown announcements and shared generated responses through a simplified social network. Personas and lightweight state machines introduced variability in interpretation and propagation (implementation details in Sec. A.3.5).

When presented in October 2024, policymakers expressed skepticism about practical utility. While they agreed that modeling misinformation was important, their concerns centered on *validation*: simulated communication dynamics could not be reliably assessed against real-world behavior. *P3* noted:

“It’s very hard for us to access social media data in previous emergencies, and some important communication may happen at private Discord channels... we don’t know, I mean, I don’t know how to communicate directly with some of our audiences overseas.”

These discussions led to a key design shift: we jointly identified evacuation preparedness for the university’s annual commencement as a more suitable domain—one that was both operationally critical and recurrent, enabling direct observation and validation.

A.1.3 Second Iteration: Simulating “Commencement”. Building on the limitations of Iteration 1, we shifted focus to evacuation preparedness during commencement. We developed a simplified evacuation simulation with 500 LLM agents placed in an abstracted stadium layout, receiving an official evacuation announcement, exchanging messages with nearby peers, and selecting exits over multiple time steps. This iteration deliberately prioritized *observability* over realism.

When reviewed by *P1, P4,* and *P5* in early April 2025, policymakers immediately identified gaps limiting policy relevance: realistic scale, physical constraints (aisles and bottlenecks), accessibility considerations, volunteer coordinator roles, and realistic announcement language. Rather than rejecting the simulation, policymakers

treated this iteration as a diagnostic artifact whose critiques directly motivated the next iteration’s design.

A.1.4 Third Iteration: Simulating the University’s Commencement. Guided by policymakers’ feedback, we substantially increased simulation realism. We incorporated key physical features—seating sections, aisles, concourses, bottlenecks—and increased scale to 3,000 agents. Agents reflected heterogeneous attendee roles: differentiated mobility profiles for accessibility needs and coordinator agents positioned to guide evacuation. We used the official evacuation message and focused on severe weather, the most operationally relevant emergency type.

This iteration marked a clear turning point. When *P1, P4,* and *P5* reviewed the system in late April 2025, they expressed substantially greater confidence and began using the simulation to reason about concrete planning decisions. *P1* used the visualization to diagnose where volunteer coordination mattered most, noting that congestion emerged well before crowds reached exits. Policymakers immediately recognized value as training materials: coordinators could “see where congestion is likely to happen and what to expect.” They integrated simulations into coordinator training sessions ahead of the 2025 commencement and documented the approach in the official *after-action report*—the first transition from exploratory technology probe to institutionally-recognized tool.

A.1.5 Fourth Iteration: Validating Simulation with Real Commencement. Following Iteration 3, we observed the actual commencement in May 2025 alongside the policymaker team, examining the venue and recording crowd movement under IRB approval. During the event, policymakers highlighted areas they routinely monitored for congestion, aligning our analytic focus with practitioner priorities.

One unanticipated pattern proved especially consequential: the central role of students’ family members. More than half of attendees were family or friends, and after the ceremony many students sought to reunite with them, producing the most severe congestion near family seating areas. Our prior simulations modeled only student agents acting independently—this mismatch revealed a critical omission.

We addressed this gap by scaling to 13,000 agents—matching the actual event—and incorporating *student, family/friends,* and *coordinator* agents. Agents were seated in designated areas, grouped socially, and designed to communicate and move jointly. In June 2025, *P1, P4,* and *P5* judged the simulation’s baseline crowd dynamics to align with their routine commencement experience (Fig. 3), establishing sufficient trust to explore hypothetical emergencies.

P1 reflected:

“...when you showed a baseline evacuation without anything major, it mirrored the 20 minutes of what you all experienced with us with just the regular evacuation. And your simulation mirrored what we experience every time we have it. So that gives us some baseline trust, if you will, to see that it is connecting to what experiences we already had.”

A.1.6 Fifth Iteration: Exploring Better Policy Implementation. In the final iteration, we used the validated simulation to explore concrete policy options. We examined operationally plausible variations in emergency type (severe weather and bomb threat), announcement

specificity, and coordination strategies, including repositioning coordinators and opening additional exits. Interventions were evaluated on “evacuation efficiency”—time required for 80% of agents to evacuate.

This iteration produced three actionable proposals: (1) opening a previously unused exit at the northwestern corner of the stadium, (2) training coordinators to guide both nearby and distant groups during localized threats, and (3) differentiating evacuation procedures by emergency type and incorporating these distinctions into training materials.

Figure 4 illustrates how simulation outputs informed discussion. Variations in coordinator positions did not substantially affect evacuation time, while opening the northwest exit significantly improved evacuation efficiency (23.0% relative reduction at 80% threshold). This improvement was robust across severe weather and bomb threat scenarios, except when the threat itself was located near that exit.

Policymakers adopted Proposals 2 and 3, integrating revised coordinator guidance and differentiated emergency protocols into standard operating procedures. Proposal 1 entered formal feasibility review, consistent with institutional procedures for infrastructure modifications. These outcomes were documented in the August 2025 after-action report, which committed to continued use of simulation for preparedness planning and training.

A.2 Findings: Full Detail

A.2.1 Validation Filter: Usefulness Requires Scenarios That Can Be Validated. Simulations were considered useful only when outputs could be validated against policymakers’ experience or ground-truth data. In this context, validation did not mean comparing outputs to a single ground-truth value; rather, policymakers needed to check simulation behavior against something they already knew from experience, observation, or institutional records.

This surfaced immediately in Iteration 1, where the online communication domain was deemed unsuitable despite its clear policy relevance. As *P1* framed commencement as an alternative:

“We will be more than happy to have you here... you have the whole view and you can see actually see the flow of people.”

Commencement was appealing not only because it was high-stakes, but because it recurred and could be directly observed, making it possible to establish a shared sense of what “normal” looked like before using simulation to explore more uncertain scenarios.

A.2.2 Trust Bootstrap: Trust from Validating Mundane Scenarios Extends to Novel Ones. The validation filter creates a paradox: simulations are most valuable for hard-to-observe events, yet trusted only when grounded in well-understood scenarios. In practice, this paradox was addressed through gradual scope expansion: policymakers first accepted the simulator’s representation of routine movement patterns, then became willing to explore scenarios layering additional uncertainty on top of this validated baseline.

This bootstrapping was reflected in *P1*’s comment during Iteration 4 (quoted above) and enabled policymakers to both imagine new options and critically assess their feasibility. For example, *P5* proposed exploring a behind-stage exit; the simulation showed it was less efficient than the northwestern exit, and the team did not

pursue it further—illustrating how validated trust supported both creative exploration and critical evaluation.

A.2.3 “Fix-It” Response: “Wrong” Simulations Surface Tacit Domain Knowledge. Even when policymakers were deeply familiar with their domain, they did not initially know which details were critical for interpreting a simulation until they saw something that looked off. Incomplete simulations therefore served as productive technology probes [16].

For example, when our original simulation showed agents moving at the same speed, *P5* immediately flagged the omission of accessibility needs:

“Did you think about like people with disabilities and they might want to exit like through the one that has like the ramp or anything like that?”

Similarly, after seeing a student-only simulation, *P4* surfaced the critical role of family members:

“You start thinking about the fact that in a graduation ceremony, for every student, there is potentially minimum of two family members that will be in attendance... you’ve also got to put into consideration the fact that there is a family connection and they are not with their family in that graduation component.”

P5 added:

“Like they’ll meet up at landmarks that the family knows... So like if your family’s not familiar with CMU and you tell them like let’s meet by the goal post that’s something visually that they can see.”

Despite being core to actual evacuation behavior, these family dynamics were never mentioned in initial requirements discussions—they emerged only when policymakers confronted unrealistic outputs.

A.2.4 The Details Matter: Contextual Nuance in Interaction Environments Enables Policy Use. In early iterations, we focused primarily on designing agents. However, policymakers consistently pointed out that without sufficient nuance in the interaction environment, simulations lacked credibility. When the interaction environment was under-specified, realistic-looking behavior gave the impression of policy relevance without clearly signaling its limits.

In our case, critical contextual details included the physical (stadium layout, seating), social (family relationships, accessibility needs), procedural (coordinator roles, protocols), and temporal (event sequences, decision timelines). When these contexts were incorporated, policymakers began to envision practical uses. For example, seeing weather announcements integrated into the simulation prompted *P4* to imagine bomb-threat scenarios: “*We were thinking that this evacuation is due to severe weather, which would be a fairly benign reason. But it could also be because of a bomb threat, and that opens up a whole different conversation.*”

A.2.5 Policy-AI Co-evolution: Usefulness Emerges From Co-Evolution in Requirements and Capabilities. Our iterative process revealed a co-evolution between policy requirements and simulation capabilities. Policymakers clarified that our simulations addressed *implementation processes*—not stable policy commitments:

“These are process changes, not policy... Our policy is to implement safe procedures to make sure that we protect life, property... What we’re trying to do is to find ways to fulfill the policy, which is protecting life.” (P1)

This distinction clarified when simulations became actionable: policymakers found them most useful when addressing decisions within their direct authority (training, coordinator placement, internal procedures), and harder to act on when changes depended on external command structures or legal authority. More broadly, policymakers treated the simulator as a mutual learning device:

“We are aware that there’s so many variables, so we’re not going to be going to be able to simulate 100% the whole. So when we trust the system enough... I think that to mutually learn... Because I think this is for us this is a learning experience.” (P1)

A.3 Simulation System Description

A.3.1 Final System Overview. The final simulation system (Fig. A6) models approximately 13,000 agents representing students, family members, and emergency coordinators, each with distinct roles, goals, mobility constraints, and social relationships. Agents receive inputs from official announcements, local environmental conditions, nearby crowd dynamics, and group communication through online networks. OpenAI’s GPT-4.1 governs agents’ decision-making and message generation, while a rule-based controller handles physical movement and collision-aware navigation.

The simulation environment models the actual stadium layout, including seating sections, aisles, concourses, exits, and accessibility routes. Across scenarios, agents exhibited interpretable context-sensitive behavior: they moved more urgently under severe weather than during routine dispersal, and responded heterogeneously to bomb-threat announcements.

Policymakers primarily interacted with the system through visualizations, recorded scenario replays, and comparative scenario outputs rather than direct parameter tuning.

A.3.2 Agent Population Generation. The simulation employs a two-stage persona generation process. Student persona generation operates across ten academic disciplines with predetermined enrollment distributions, totaling 2,928 base student personas. An additional 44 students with accessibility requirements are generated, distributed proportionally across all programs. The second stage creates social network structures: 2,000 students attend with family members or friends (requiring additional persona generation), 800 students form friend groups with other students, and 128 students attend individually. The complete population produces approximately 13,000 total agents with explicit social network topologies, realistic demographic distributions, and heterogeneous accessibility requirements.

A.3.3 Agent Architecture. Each agent is initialized with persona attributes and social network connections. Agents operate through a three-state machine: *Discussing* (deliberating on action), *Moving* (navigating toward a destination), and *Waiting* (holding at an

intermediate point for group members). The LLM governs decision-making and communication; a rule-based controller handles physical movement and collision-aware navigation.

A.3.4 Round-Based Execution. The simulation operates through discrete round-based execution cycles. Each round processes: (1) state transitions based on group arrival status; (2) LLM-driven discussion for agents in the discussing state, with semaphore-based concurrency control; (3) movement processing with density-based speed adjustments. Movement speed adapts dynamically to local crowd density, enabling realistic congestion patterns to emerge.

A.3.5 System Description: Iteration 1.

Misinterpretation Simulation. The Iteration 1 system implements a multi-stage pipeline evaluating how university announcements may be misinterpreted by students with diverse behavioral characteristics. It operates through four phases: agent generation (each agent assigned a persona including misinterpretation propensity), message corpus development, interpretation and reaction simulation (LLM generates interpretation and behavioral reaction for each agent-message pair), and misinterpretation scoring (0–100 scale). For scores exceeding a threshold, the system additionally generates an extreme reaction scenario.

Propagation Simulation. The propagation system models information spread through a multi-agent framework. Agents decide to spread, remain idle, or evacuate. When spreading, content is distributed to a 70% random sample of agents (moderated version prepends a misinformation warning). Agents update their behavior based on environmental feedback at section boundaries.

A.4 Co-creating Stakeholder and Process Maps



Figure A1: Developers and policymakers P1 collaboratively co-creating a stakeholder map and process map during a policy meeting.

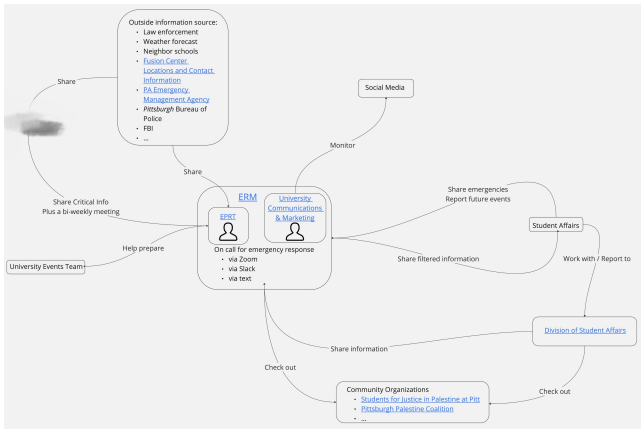


Figure A2: Stakeholder map of emergency preparedness and response at the university, co-created with policymakers.

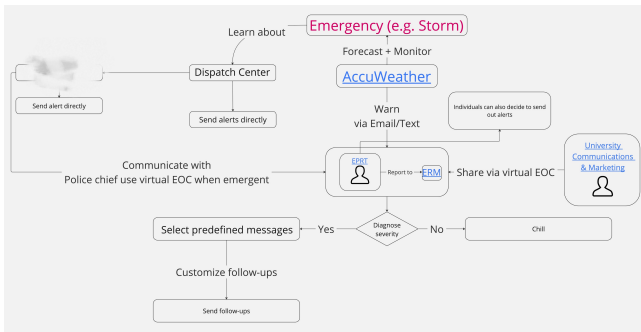


Figure A3: Process diagram of emergency preparedness and response at the university, documenting sequential phases from preparedness planning to active response and recovery.

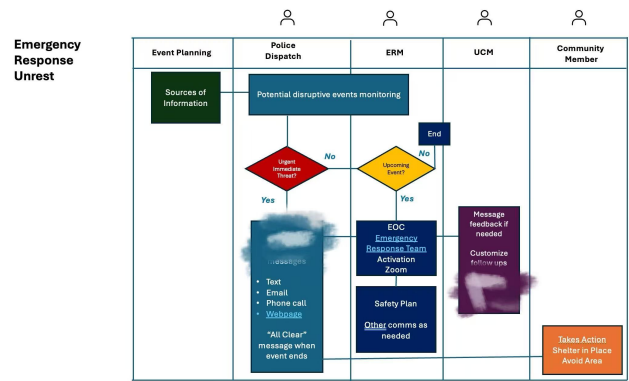


Figure A5: Process diagram of emergency preparedness tailored for unrest or bomb threat scenarios.

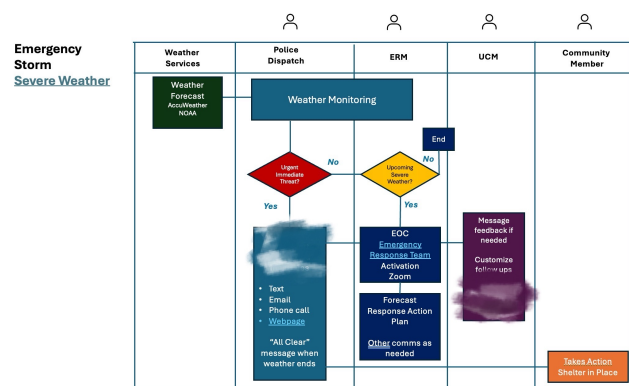
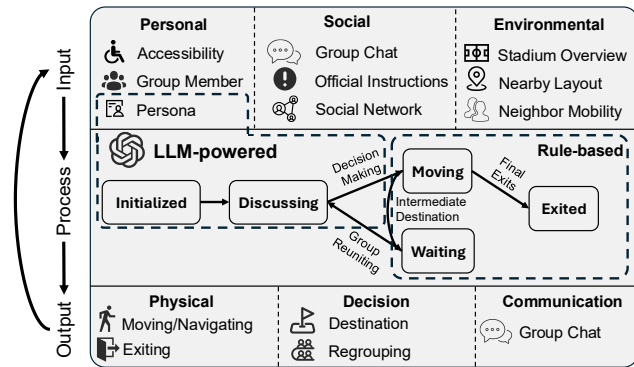


Figure A4: Process diagram of emergency preparedness tailored for severe weather scenarios.

A. LLM Agents Conceptual Diagram



B. System Display

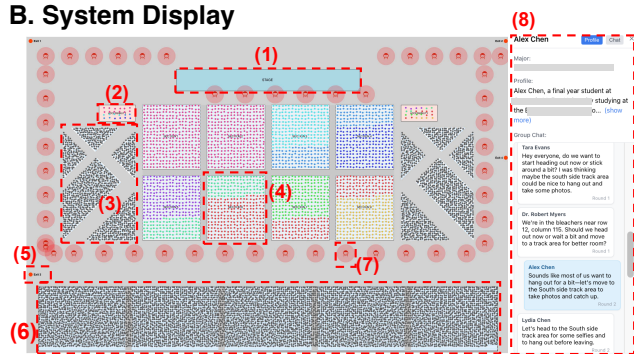


Figure A6: Simulation overview. (A) LLM agents conceptual diagram: Agents receive inputs from Personal (accessibility needs, group membership, persona), Social (group chat, official instructions, social network), and Environmental (stadium overview, local layout, neighbor mobility) sources. An LLM governs decision-making and communication; a rule-based controller handles embodied action. (B) System display: Stadium-scale interface showing physical layouts, agents as colored dots, coordinators, and a per-agent side panel. Together this system supports simulations with up to 13k agents.