Lecture 7.

Believability vs. Accuracy

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park

Announcement

- We just released the assignment for the final project proposal!
  - https://joonspk-research.github.io/cs222-fall24/final_project_proposal.html
  - Due: 11/4/2024 (same day as your project presentation)
  - Groups of 3 to 4 people.
- Reading for Wednesday — updated!

Assignment 1 Q/A

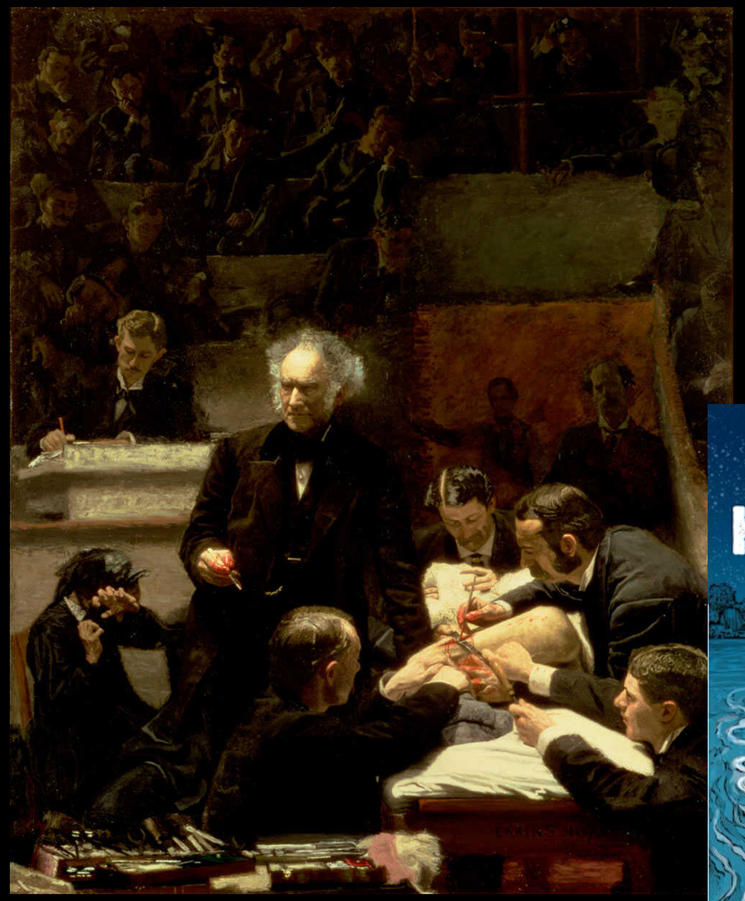
- Time-step when conversing: feel free to set it to 0 for all memory nodes, or increment by 1.
- Importance score: it should range from 0 to 100.

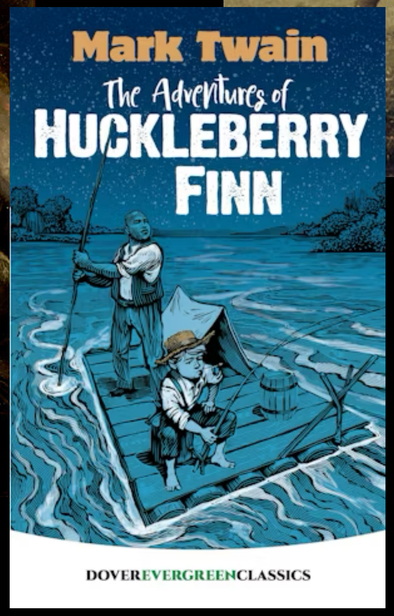
Welcome to Week 4. So far...

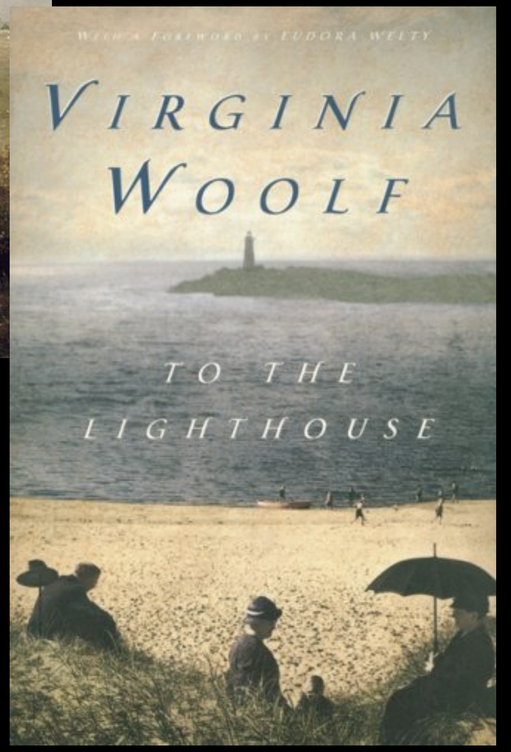
- We defined simulations as the interplay between agents and the environment.
- Simulations should tackle wicked problems.
- We discussed the general architectures for building agents using generative AI.
- We discussed how we built environments to situate/ground the agents.

Believable agents and simulations

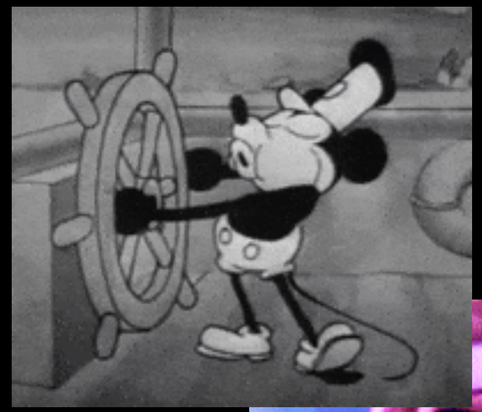
Where do art and stories find their power?

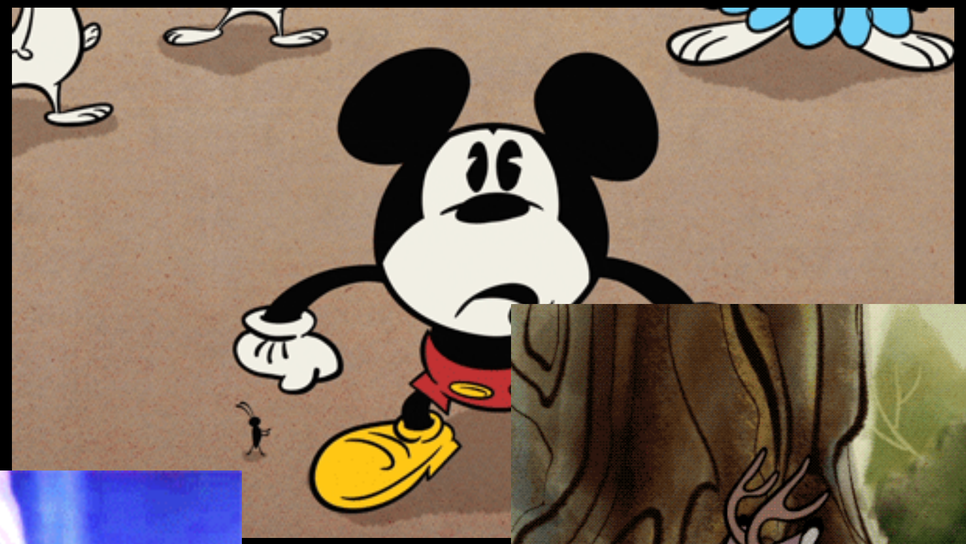


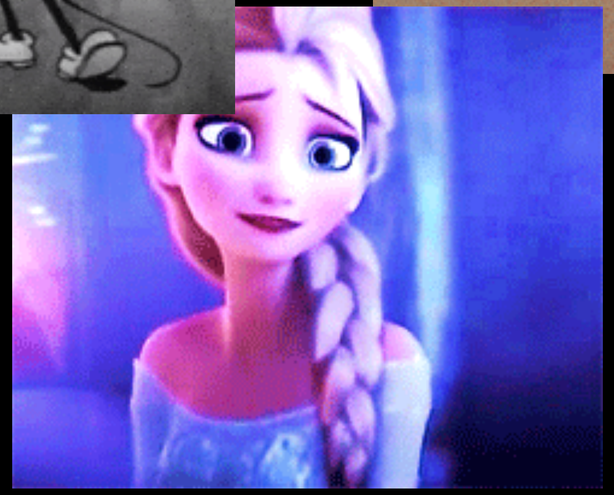


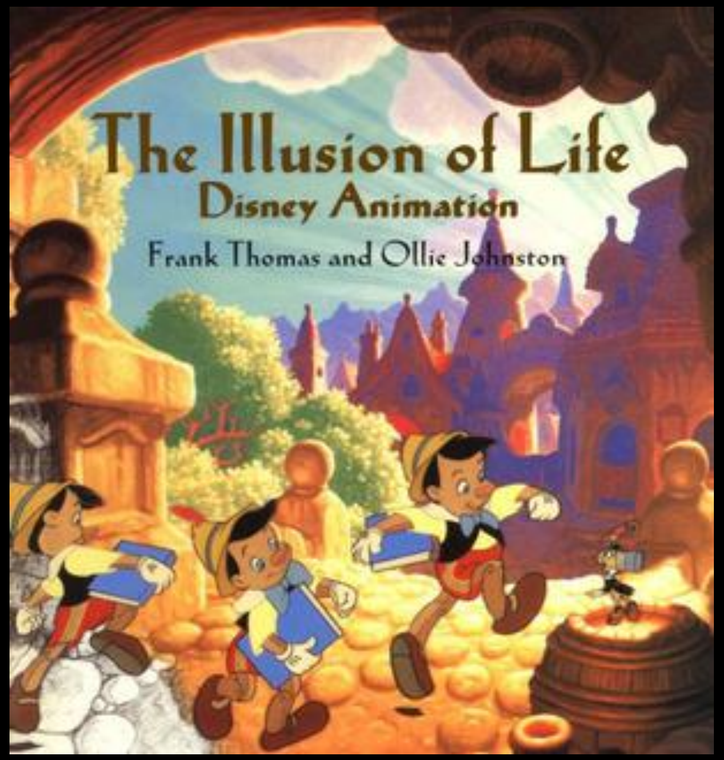


Where do art and stories find their power?









“Disney animation makes audiences really believe in ... characters, whose adventures and misfortunes make people laugh -- and even cry. There is a special ingredient in our type of animation that produces drawings that appear to think and make decisions and act of their own volition; it is what creates *the illusion of life*.”

Thomas, F. And Johnston, O. *Disney Animation: The Illusion of Life.* Abbeville Press, New York, 1981.

Believable agents are designed to create the illusion of life

The Role of Emotion
in Believable Agents

Joseph Bates

There is a notion in the Arts of "believable character." It does not mean an honest or reliable character, but one that provides the illusion of life, thus permitting the audience's suspension of disbelief

The idea of believability has long been studied and explored in literature, theater, film, radio drama, and other media. Traditional character animators are among those artists who have sought to create believable characters, and the Disney animators of the 1930s made great strides toward this goal. The first page of the enormous classic reference work on Disney animation [12] begins with these words:

> Disney animation makes audiences really believe in ... characters, whose adventures and misfortunes make people laugh — and even cry. There is a special ingredient in our type of animation that produces drawings that appear to think and make decisions and act of their own volition; it is what creates the illusion of life.

Many artificial intelligence researchers have long wished to build robots, and their cousins called "agents," that seem to think, feel, and live. These are creatures with whom you'd want to share some of your life—as with a companion, or a social pet. For instance, in his 1985 American Association of Artificial Intelligence (AI) presidential address [3], Woody Bledsoe told of his continuing dream to build a computer friend. He spoke of the "excitement of seeing a machine act like a human being, at least in many ways," of building a machine that could "understand, act autonomously, think, learn, enjoy, hate" and which "liked to walk and play Ping-Pong, especially with me."

Woody Bledsoe is hardly alone. Further reading on the dreams of animators and AI researchers finds both groups speaking of thinking, feeling, living creatures, of creating at least the illusion of life, of building apparently autonomous entities that people, especially their creators, would genuinely care about. Both groups also speak of achieving these dreams by finding the essence of the creatures to be simulated, and reconstructing that essence in the medium of the artist's or scientist's choice.

As AI researchers tried to find these essential qualities of humanity, they gravitated toward reasoning, problem solving, learning via concept formation, and other qualities apparently central to the capacity we call intelligence. Perhaps this happened because these qualities are characteristic of the idealized scientist, and thus are valued by the communities of which the researchers were part.

Artists, in particular the character animators, also tried to understand and express the essence of humanity in their constructions. Character animators had to be especially analytic, because they had to produce human life from nothing more than individual, hand-drawn, flat-shaded line drawings, moved frame by frame, without being able to rely on a human actor to portray the character. The practical requirement of producing hundreds of thousands of these drawings forced animators to use extremely simple, nonrealistic imagery, and to seek and abstract precisely that which was crucial.

It can be argued that while scien-

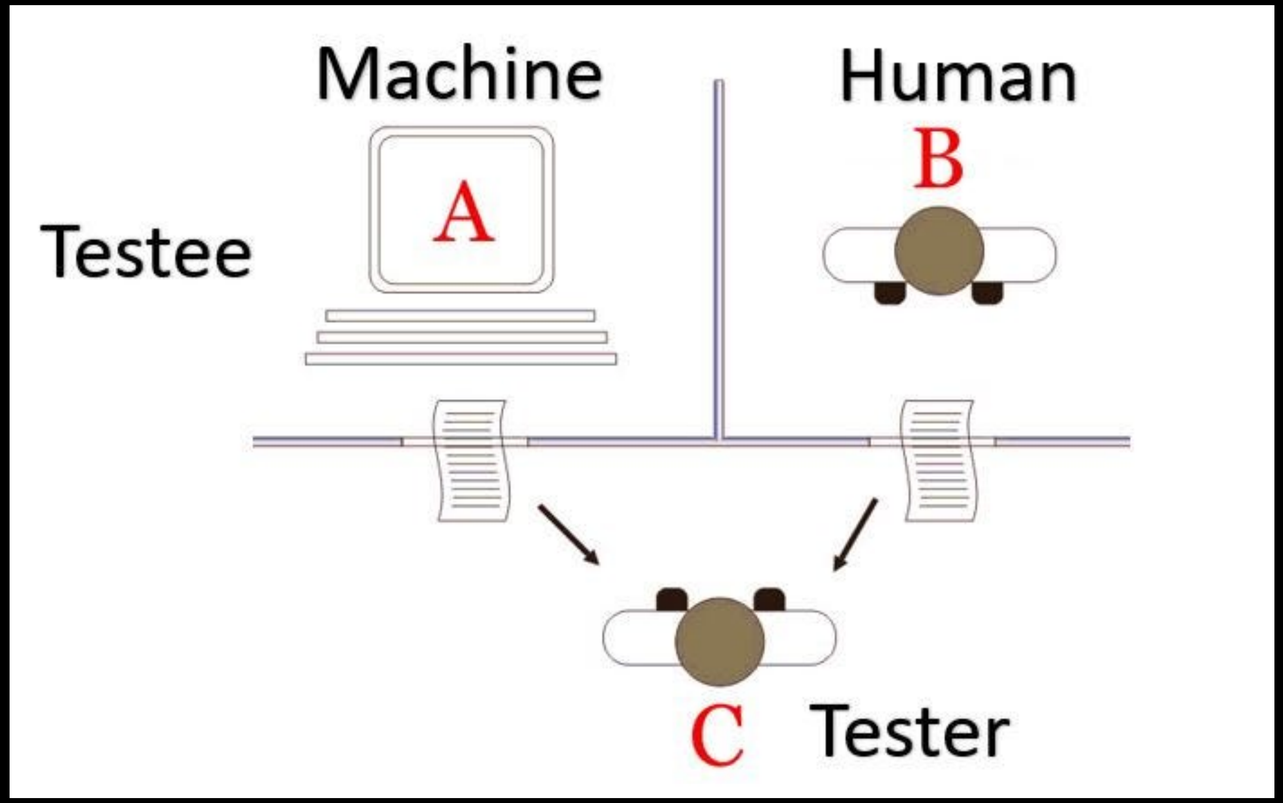
"... the idea of believable agents, by which we mean an interactive analog of the believable characters discussed in the Arts... We have argued that these artists hold some of the same central goals as AI researchers,.. may serve as a component of new user interacts for the broad human population."

"Believability. That is what we were striving for... belief in the life of the characters."

J. Bates, The Role of Emotion in Believable Agents. Commun. ACM 37, 122-125 (1994).

How do we measure believability?

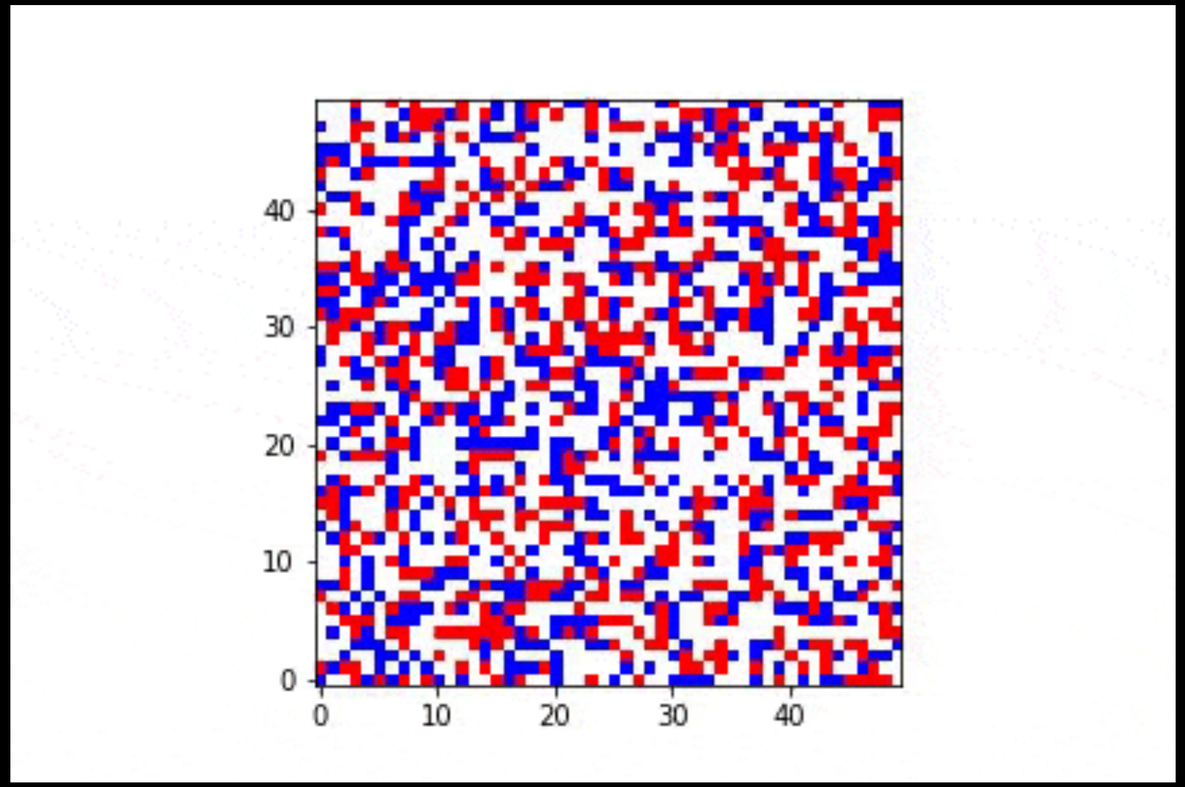
We measured believability for a while...



Turing test

Turing A. Computing machinery and intelligence. Mind. 1950;59(236):433-460. doi:10.1093/mind/LIX.236.433

Do you think ABMs were also evaluated based on believability?



ABMs

T. C. Schelling, Dynamic models of segregation. J. Math. Sociol. 1, 143–186 (1971).

Generative agents were evaluated based on (essentially) a behavioral Turing test.

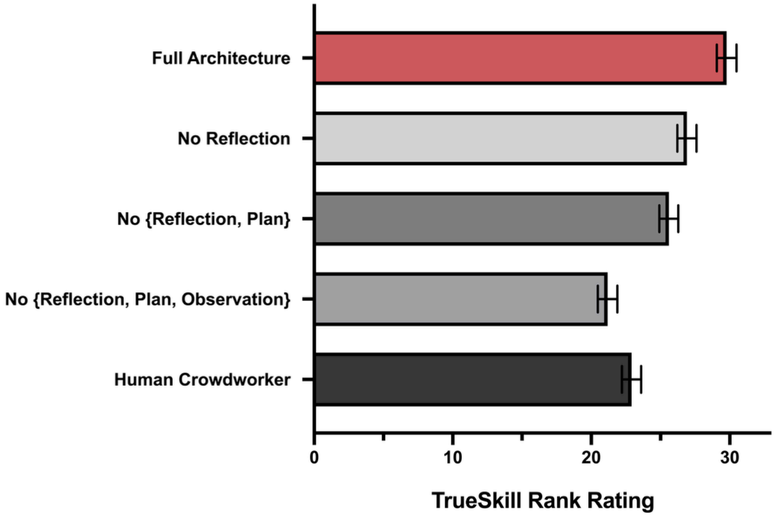


Figure 8: The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

Generative Agents

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).

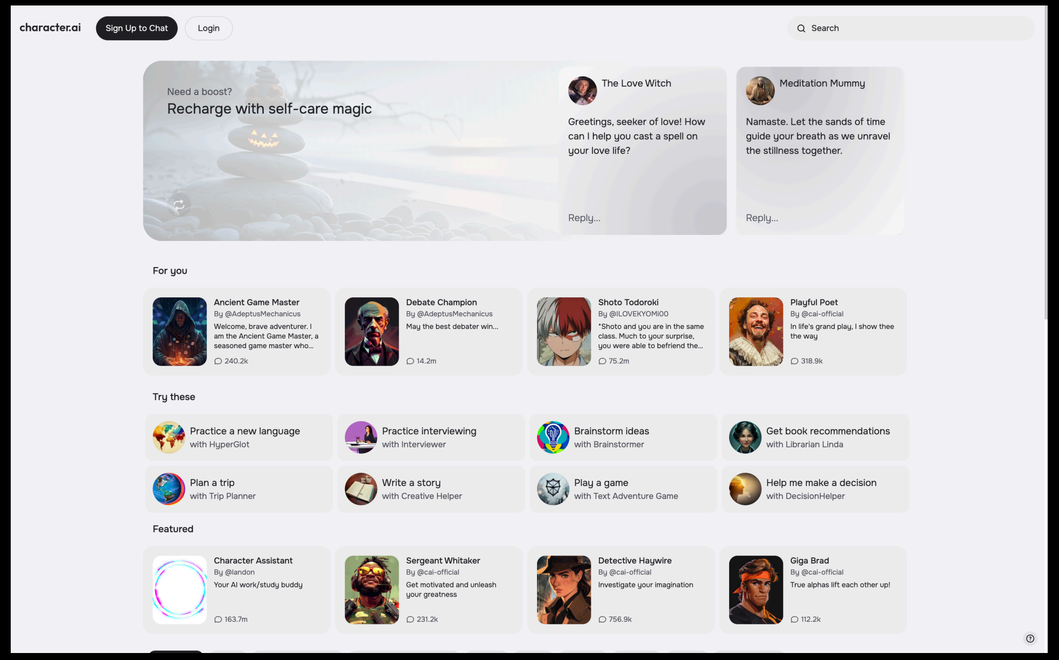
Believable agent applications

Believable agents and simulations can power games and storytelling

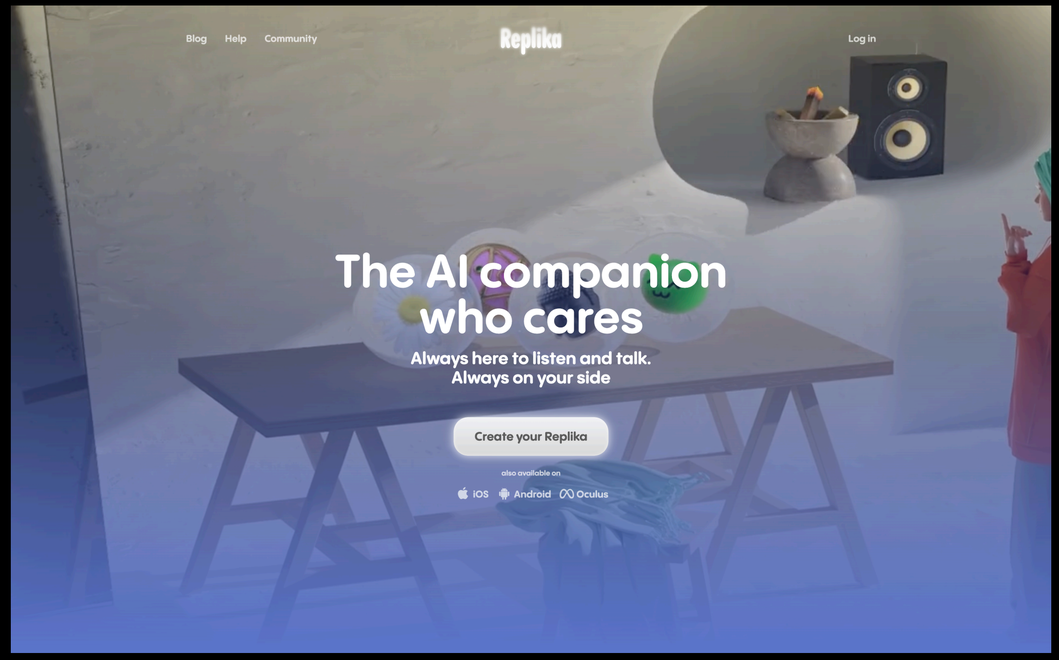
The Sims

Minecraft

Believable agents and simulations can power “AI companions”



character.ai



Replika

Believable agents and simulations can power various rehearsal spaces

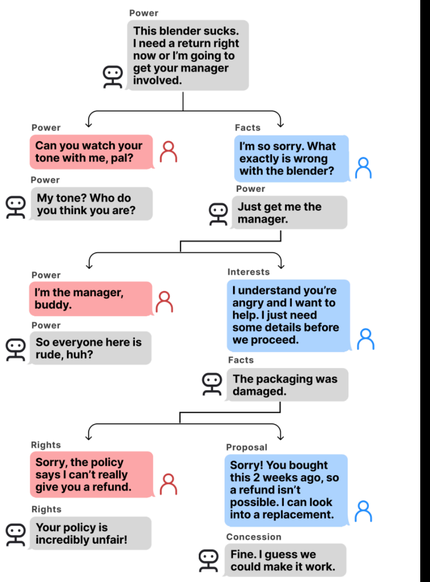


Figure 1: An example interaction trace with REHEARSAL, where an employee practices conflict with a simulated customer. The employee quickly realizes that Rights and Power-based strategies result in heightened conflict. The conflict is eventually resolved using an Interests-oriented approach.

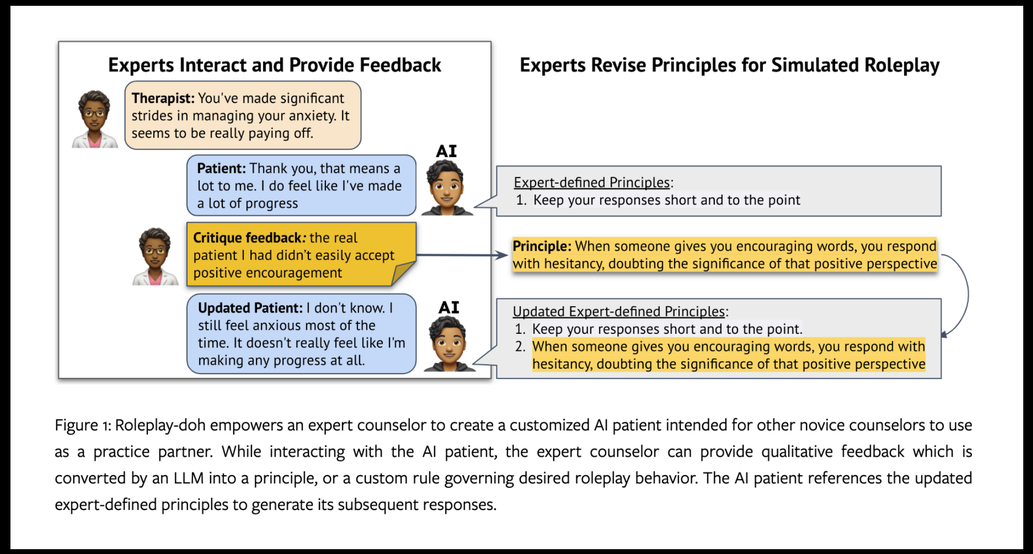


Figure 1: Roleplay-doh empowers an expert counselor to create a customized AI patient intended for other novice counselors to use as a practice partner. While interacting with the AI patient, the expert counselor can provide qualitative feedback which is converted by an LLM into a principle, or a custom rule governing desired roleplay behavior. The AI patient references the updated expert-defined principles to generate its subsequent responses.

Rehearsal spaces for people

Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents

JUNKAI LI[†#], SIYU WANG[†], MENG ZHANG[†], WEITAO LI[†#], YUNGHWEI LAI[†], XINHUI KANG[†#], WEIZHI MA[†], and YANG LIU[#†]

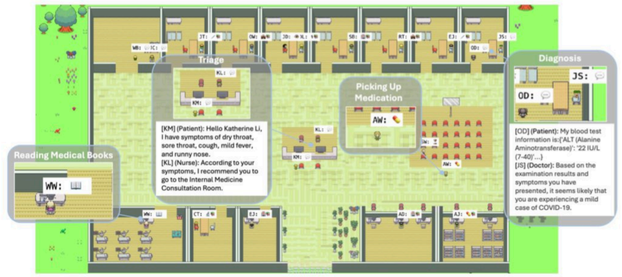


Fig. 1. An overview of Agent Hospital. It is a simulacrum of hospital in which patients, nurses, and doctors are autonomous agents powered by large language models. Agent Hospital simulates the whole closed cycle of treating a patient's illness: disease onset, triage, registration, consultation, medical examination, diagnosis, medicine dispensary, convalescence, and post-hospital follow-up visit. An interesting finding is that the doctor agents can keep improving treatment performance over time without manually labeled data, both in simulation and real-world evaluations.

Rehearsal spaces for agents

R. Louie, A. Nandi, W. Fang, C. Chang, E. Brunskill, D. Yang, Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. Preprint (2024).
O. Shaikh, V. Chai, M. J. Gelfand, D. Yang, M. S. Bernstein, Rehearsal: Simulating Conflict to Teach Conflict Resolution, in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, May 11-16, 2024.
J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, Y. Liu, Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. Preprint (2024).

Discussion

Simulation agents do not necessarily need to go after our “idealized view of intelligence”

- “… These include the appearance of reactivity, goals emotions, and situated social competence, among others. The emphasis in “alternative AI” on reactivity could be seen as choosing one of the believability requirements and elevating it to a position of importance, while downgrading other qualities, such as those related to our idealized view of intelligence.”

J. Bates, The Role of Emotion in Believable Agents. Commun. ACM 37, 122-125 (1994).

Believable agents offer an *illusion* of life.

(But still only a plausible simulacra)

J. Bates, The Role of Emotion in Believable Agents. Commun. ACM 37, 122-125 (1994).

Accurate agents and simulations

Q. What do you think it means for a simulation to be “accurate”?

Accurate simulations are predictions of the future

What is the challenge in achieving accurate simulations?

- Is the challenge in “building” accurate agents or in understanding when they are accurate through “evaluation”?

How do we measure accuracy?

General evaluation scheme:

Gather ground-truth data and see if the simulation replicates it.

Challenges of evaluating accurate situations

- Individual:
  - Open-ended nature... On what axis do we evaluate?
  - Inconsistency.
- Group:
  - Complex dynamics (Some believe this is not possible).
- Population:
  - Sometimes lacks ground truth.
  - If replicating known studies, the model may have memorized the study.

Today: evaluating “population-level” simulations

Can we predict studies that are not yet included in the datasets of language models?

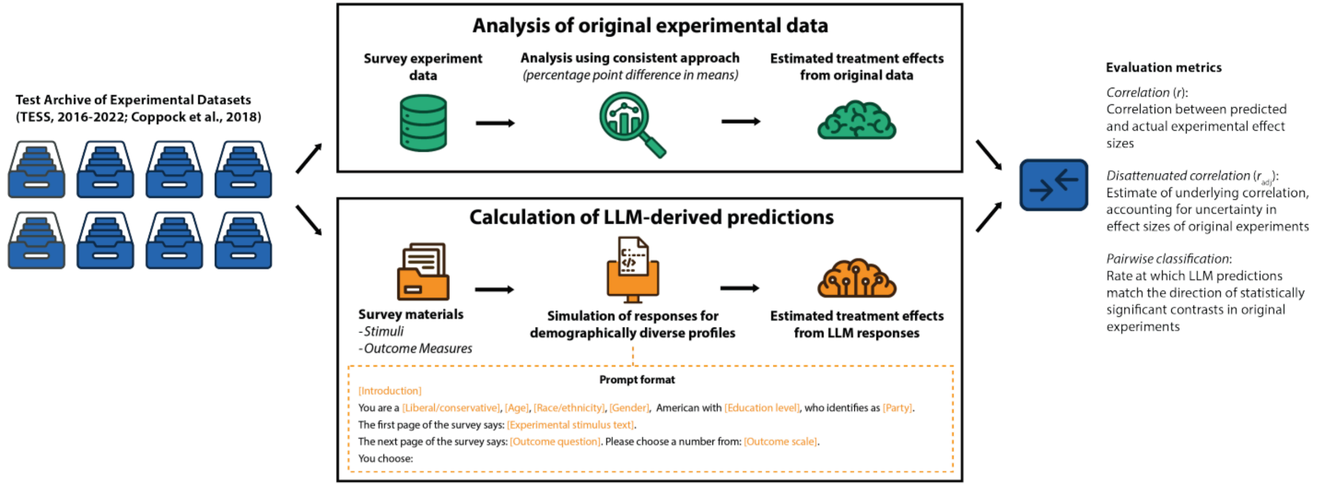


Figure 1: **Method.** We re-analyze raw data from 70 nationally-representative U.S. studies and estimate treatment effects in a consistent manner. We use a Large Language Model to simulate those same experiments, providing the original study materials and demographically-diverse participant profiles, and then calculate the average simulated response for each condition. We evaluate the accuracy of the model in terms of the correspondence between measured- and simulated- treatment effects.

A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).

It appears to be the case!

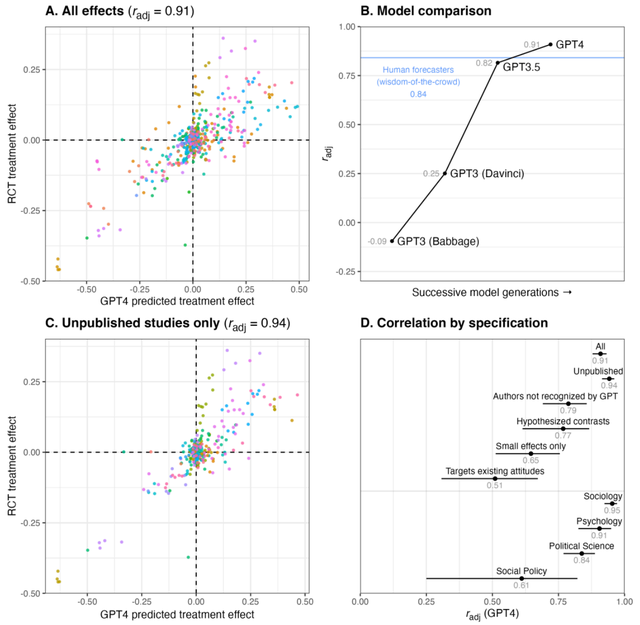


Figure 2: **LLMs accurately predict treatment effects in text-based social science experiments conducted in the US.** (a) In a dataset of 70 text-based experiments with 476 effects, LLM-derived estimates of treatment effects pooled across many prompts were strongly correlated with original treatment effects ($r = 0.85$; $r_{adj} = 0.91$). (b) The accuracy of LLM-derived predictions improved across generations of LLMs, with accuracy surpassing predictions collected from the general population. (c) LLM-derived predictions remained highly accurate for studies that could not have been in the LLM training data given they were not published prior to the LLM training data cutoff date. (d) In robustness check analysis of various subsets of experiments, accuracy of LLM-derived predictions remained high. In panels A and C, different colors depict different studies.

A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, "Predicting Results of Social Science Experiments Using Large Language Models" (2024).

State of the art in evaluating generative simulations

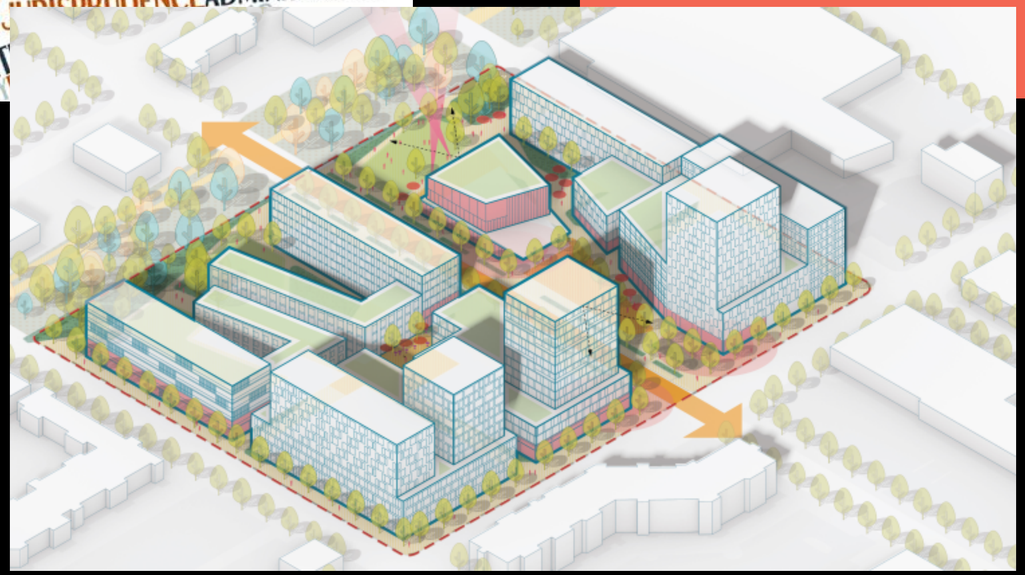
- Population-level? Yes
- Individual-level? Verdict is still out (for the work that is available)
- Group? The real question — not sure.

Applications of accurate agents

Many wicked problems require accurate simulations

Social sciences

Market research



Urban studies

Can we build personal assistants that simulate their users to create a model of their needs?

As simulations become more accurate, it does not necessarily mean they become more believable.

For example, does an accurate simulation need to provide the illusion of life through emotions? Maybe, but maybe not.

References

- Thomas, F. And Johnston, O. Disney Animation: The Illusion of Life. Abbeville Press, New York, 1981.
- J. Bates, The Role of Emotion in Believable Agents. Commun. ACM 37, 122-125 (1994).
- Turing A. Computing machinery and intelligence. Mind. 1950;59(236):433-460. doi:10.1093/mind/LIX.236.433
- T. C. Schelling, Dynamic models of segregation. J. Math. Sociol. 1, 143–186 (1971).
- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (ACM, 2023).
- R. Louie, A. Nandi, W. Fang, C. Chang, E. Brunskill, D. Yang, Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. Preprint (2024).
- O. Shaikh, V. Chai, M. J. Gelfand, D. Yang, M. S. Bernstein, Rehearsal: Simulating Conflict to Teach Conflict Resolution, in Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, May 11-16, 2024.

References

- J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, W. Ma, Y. Liu, Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. Preprint (2024).

CS 222: AI Agents and Simulations

Stanford University

Joon Sung Park