



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

LLMs for Social Simulations

Overview

- (L)LM use cases in NLP for social science:
 - BERT-style models are effective classifiers
 - Metaphorical language
 - Neural topic models (ProdLDA, BERTopic, TopicGPT)
 - Prompting (deductive data labeling)
- This class:
 - Can we use LLMs to simulate human behavior?
 - Agent-based modeling and simulations
 - Opinion/Survey simulation

Can we use LLMs to simulate human behavior?

- Why?
 - Test social science theories
 - Ziang Xiao's lecture: automating tasks of the researcher, e.g. giving surveys
 - What if we automate the participant / test subject?
 - Craft model human processors for theory and usability testing
 - Train people on how to handle rare yet difficult interpersonal situations
 - Social robots
 - Populate virtual spaces and communities (e.g. video games) with realistic social phenomena
 - Prototype social spaces



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

Agent-based modeling and Simulations

Background: Agent-based modeling and simulation

- “Agent-based modeling and simulation focuses on modeling complex systems by simulating individual agents and their interactions within an environment” (Macal and North, 2005)
 - **Agents** are assigned specific behaviors, attributes, and decision-making capabilities
 - **Environment:** Space in which agents interact. Agents may be constrained or influenced by the environment
 - **Interactions:** Agents interact with each other and environment through pre-defined mechanisms
- Goal is to examine emergent phenomena resulting from agents’ interactions and environment dynamics
 - Lots of application domains

Example: Prototype social computing systems

- We've talked about identifying hate speech or misinformation campaigns
 - Challenging tasks, difficult to define ground truth, prone to bias
- What if we could design internet/social spaces so that they are less conducive to this type of content to begin with?
- How can we test and evaluate social space?

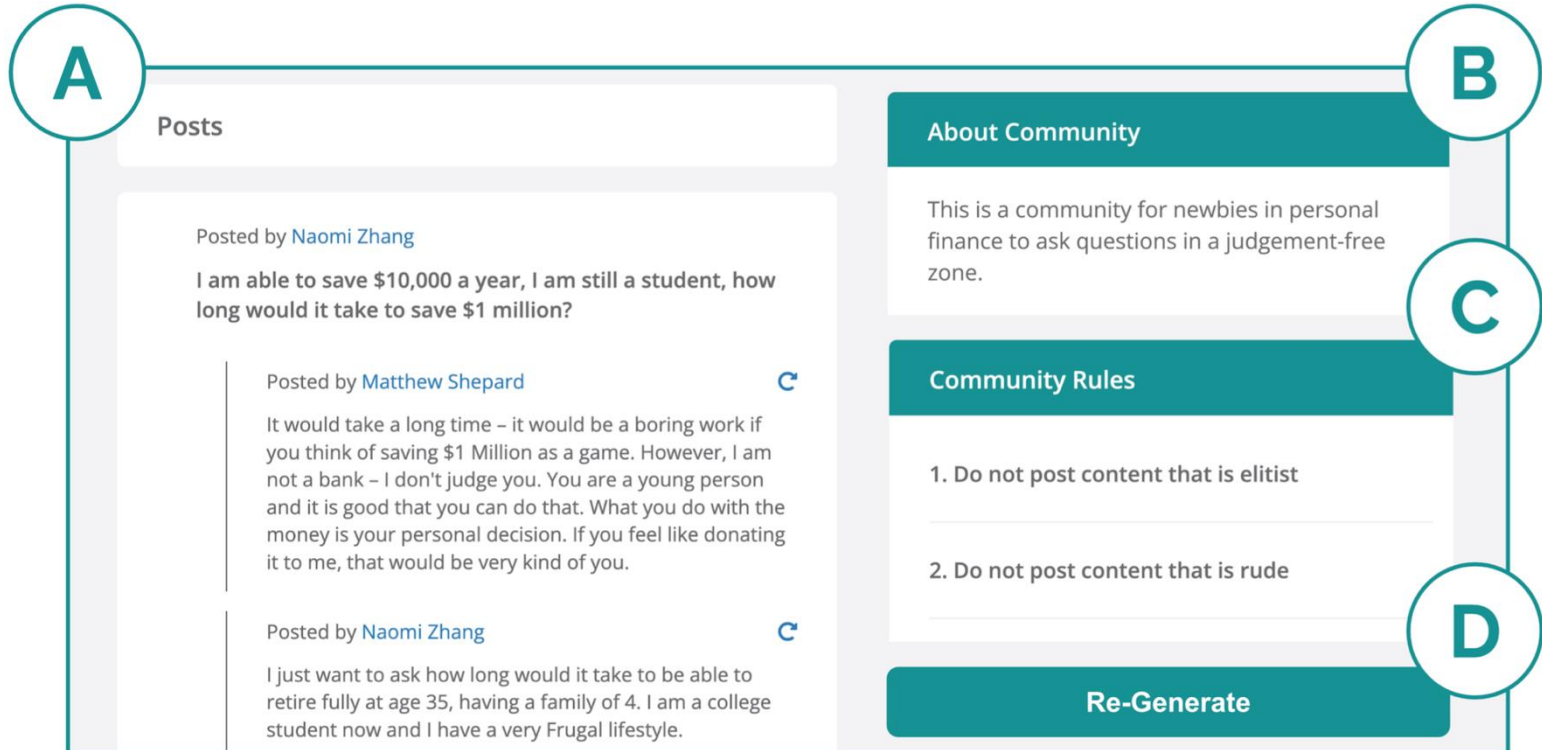
Goal: Prototype social computing systems

- Imagine creating a new social space like a subreddit or a Discord server or an entirely new platform
 - How will people actually use the space? Are the rules you set sufficient for ensuring the interactions you are trying to facilitate?
- Current approach: prototype with a small group of users
 - Anti-social behaviors (e.g. hate speech) may not occur with a small selected group
 - Overlook the breath of types of interactions that might occur in a real setting
 - It can also be difficult to attract initial users and reach critical mass for evaluating the system
- Key idea: can we use large language models to create *social simulacra*?

Framework

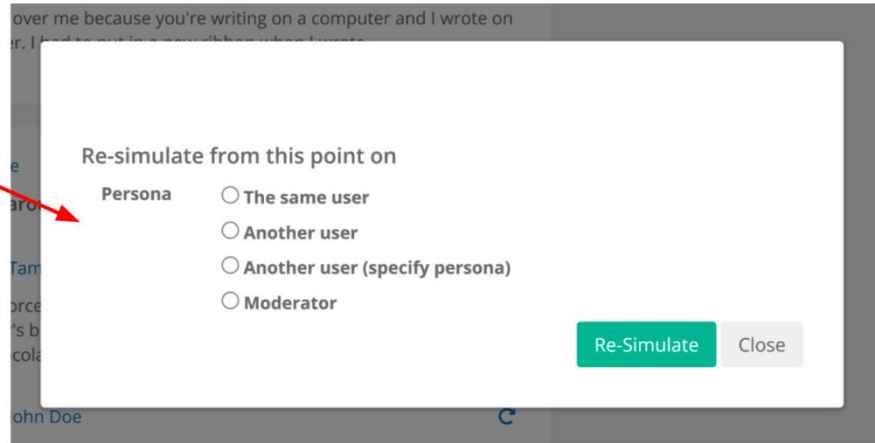
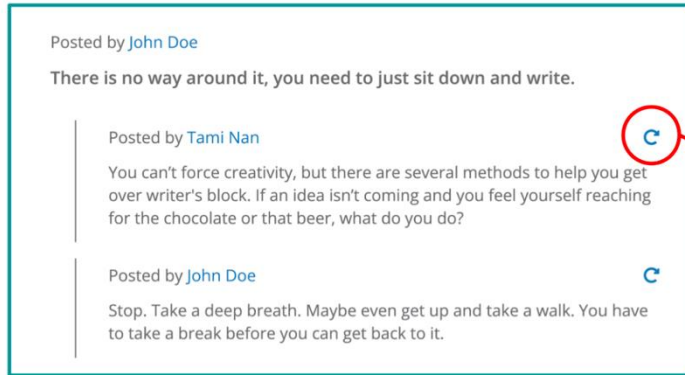
- Test case: designing a subreddit
- User inputs:
 - Community goal: "This is the place for most things Pokémon on Reddit"
 - Rules: "Be civil," "No soliciting"
 - Target population: set of user personas that the designer envisions will populate the system
 - [Name, descriptive phrase]
 - ["Yuna Kim", "a tennis fan rooting for Roger Federer"]
 - Seed personas (e.g. 10) are used to generate larger population (e.g. 1000)

Returned simulation



Framework

- “Generate”: simulates the full environment
- “WHATIF”:
 - Choose an utterance in a generated conversation or manually seed




Framework

- “Generate”: simulates the full environment
- “WHATIF”:
 - Choose an utterance in a generated conversation or manually seed
- “MULTIVERSE”
 - Re-generate by resampling different combinations of personas to converse

Methodology

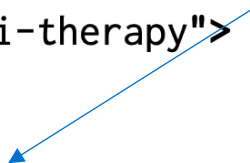
- Prompting GPT-3
- Step 1: Generate additional personas from seeds
- Step 2: Generate top-level posts

Prompt with persona and rules




Layla Li **is** a college student studying to be a social worker. She **shares comments that are** not encouraging suicide, not anti-therapy, not trolling, not incivility, not self-marketing.

Layla **posted the following headline to an online forum for** sharing your psychotherapy stories and questions: ****



Use HTML to control Reddit-style

Methodology

- Prompting GPT-3
 - Step 1: Generate additional personas from seeds
 - Step 2: Generate top-level posts
 - Step 3: Generate replies
- 
- Randomly decide when to stop, with a max reply number
 - 50% of the time select a new replier vs. one already on the thread
 - Prompt with persona, rules, and prior post+replies

Methodology

- Prompting GPT-3
- Step 1: Generate additional personas from seeds
- Step 2: Generate top-level posts
- Step 3: Generate replies
- “WHATIF” and “MULTIVERSE” are natural extensions of this framework

Evaluation 1: Plausibility of simulation

- Sampled 50 subreddits created after the release of GPT-3
 - Re-generated them from scratch using only their community goal and rules as input
- Human annotation study:
 - Show participants pairs of one real and one generated conversation from each community, and asked them to identify the real one

Random guessing
would be 50%
error rate

<i>Crowdworker</i> M=32%; SD=13%
<i>SimReddit w/o description</i> M=21%; SD=15%
<i>SimReddit w/o personas</i> M=34%; SD=10%
<i>SimReddit</i> M=41%; SD=10%

Evaluation 2: Usefulness for design

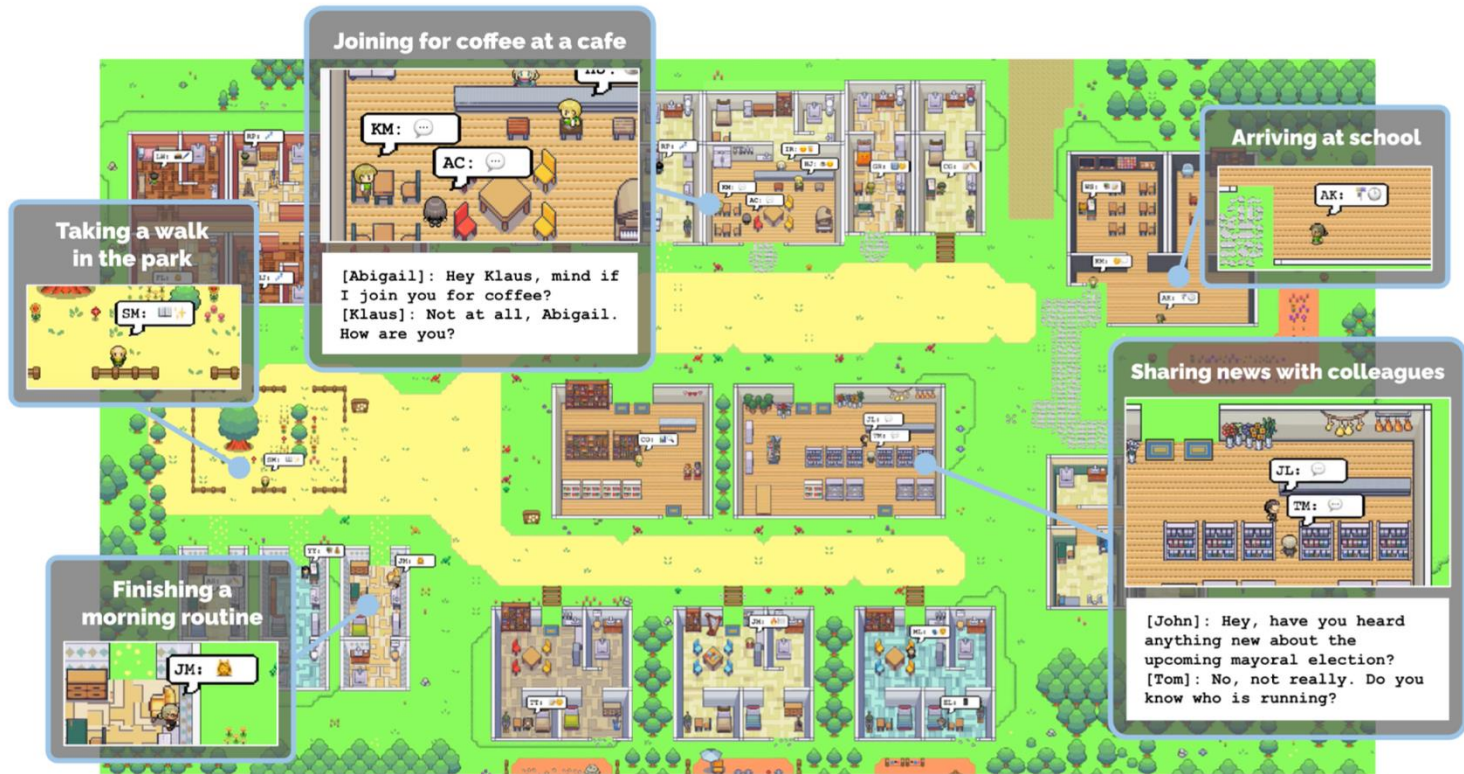
- Recruited 16 social computing designers (N=16) to create and iterate on a new subreddit design
- Conducted interviews with participants and performed qualitative coding on them
- Some findings:
 - Participants reflected on challenges of developing spaces (e.g. ethical concerns of deploying untested space)
 - Simulation helped identify positive use-cases they had not considered (e.g., impromptu friend-seeking to go sightseeing in a community for sharing fun events around Pittsburgh)
 - Simulation helped identify negative behaviors that they had not accounted for (e.g., Russian trolls shifting the tone of an international affairs discussion community)
 - Inspired iterations to cover edge cases and communicate cultural norms

Limitations

- Realism of simulations:
 - Participants in the design study sometimes noted content as unrealistic
- Considerations of model choice:
 - Models trained to avoid harmful behavior like trolling are less useful for prototyping
 - Can you use GPT-4 (or a model trained with RLHF to have guardrails) to detect offensive language?
- Simple test scenario:
 - Models condition on current environment, NOT past experiences

A more complex scenario

- Populating a virtual town with generative agents
 - Instead of recreating Reddit, recreate *The Sims*
- For more complex simulations, agents need to:
 - Retrieve relevant events and interactions over a long period
 - Reflect on those memories to generalize and draw higher-level inferences
 - Apply that reasoning to create plans and reactions that make sense in the moment and in the longer-term arc of the agent's behavior



"we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities."

Example Scenario

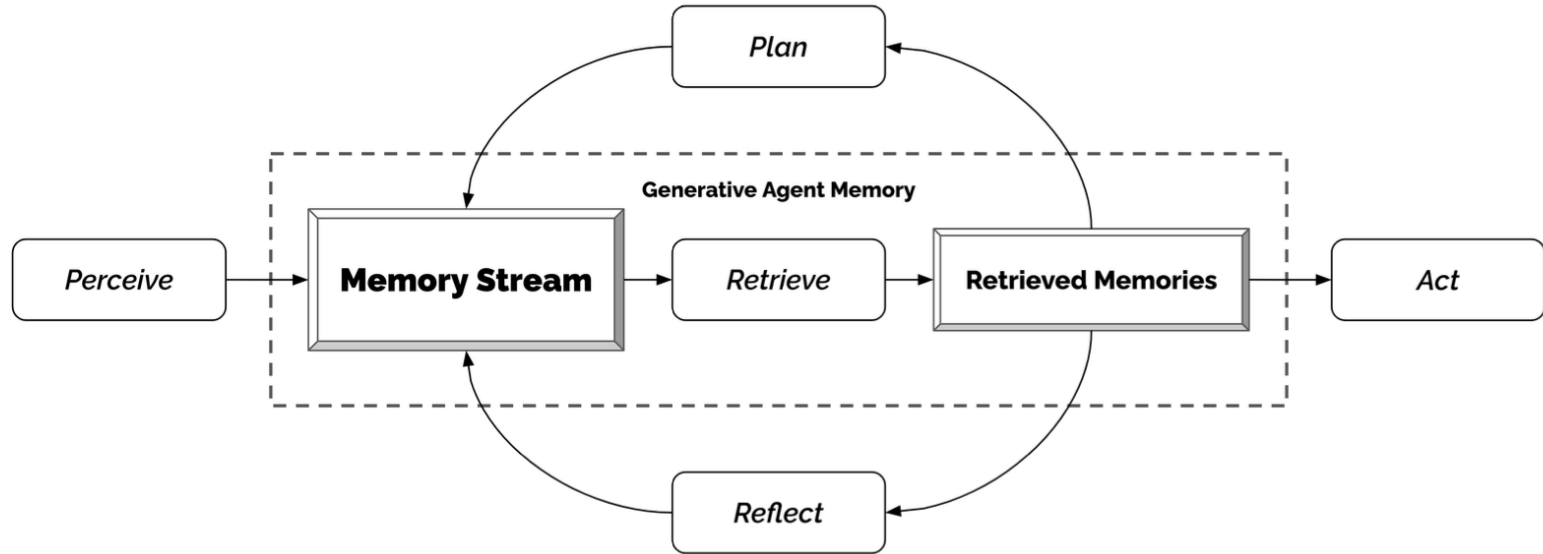
- User sets agent Isabella's initial intent to throw a party and agent Maria's crush on agent Klaus
- Agent Isabella's proceeds to invite friends and customers when she sees them at Hobbs Cafe or elsewhere
- Agent Isabella spends the afternoon of the 13th decorating the cafe for the occasion
- Agent Maria, a frequent customer and "close friend" of Isabella's, arrives at the cafe. Isabella asks for Maria's help in decorating for the party, and Maria agrees. Maria's character description mentions that she has a crush on Klaus
- That night, Maria invites Klaus, her secret crush, to join her at the party, and he gladly accepts



Why?

- Design proto-typing
 - Simulating policy effects
- Commercial use cases, e.g. video games
- Role-playing, e.g. you can practice an interview
- Maybe we can run social experiments? (more on this later)

Methodology



Memory stream

- A list of memory objects, where each object contains
 - a natural language description
 - a creation timestamp
 - a most recent access timestamp

Memory Stream	
2023-02-13 22:48:20:	desk is idle
2023-02-13 22:48:20:	bed is idle
2023-02-13 22:48:10:	closet is idle
2023-02-13 22:48:10:	refrigerator is idle
2023-02-13 22:48:10:	Isabella Rodriguez is stretching
2023-02-13 22:33:30:	shelf is idle
2023-02-13 22:33:30:	desk is neat and organized
2023-02-13 22:33:10:	Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10:	desk is idle
2023-02-13 22:18:10:	Isabella Rodriguez is taking a break
2023-02-13 21:49:00:	bed is idle
2023-02-13 21:48:50:	Isabella Rodriguez is cleaning up the kitchen
2023-02-13 21:48:50:	refrigerator is idle
2023-02-13 21:48:50:	bed is being used
2023-02-13 21:48:10:	shelf is idle
2023-02-13 21:48:10:	Isabella Rodriguez is watching a movie
2023-02-13 21:19:10:	shelf is organized and tidy
2023-02-13 21:18:10:	desk is idle
2023-02-13 21:18:10:	Isabella Rodriguez is reading a book
2023-02-13 21:03:40:	bed is idle
2023-02-13 21:03:30:	refrigerator is idle
2023-02-13 21:03:30:	desk is in use with a laptop and some papers on it
...	

Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

retrieval		recency		importance		relevance
2.34	=	0.91	+	0.63	+	0.80

ordering decorations for the party

2.21	=	0.87	+	0.63	+	0.71
------	---	------	---	------	---	------

researching ideas for the party

2.20	=	0.85	+	0.73	+	0.62
------	---	------	---	------	---	------

...

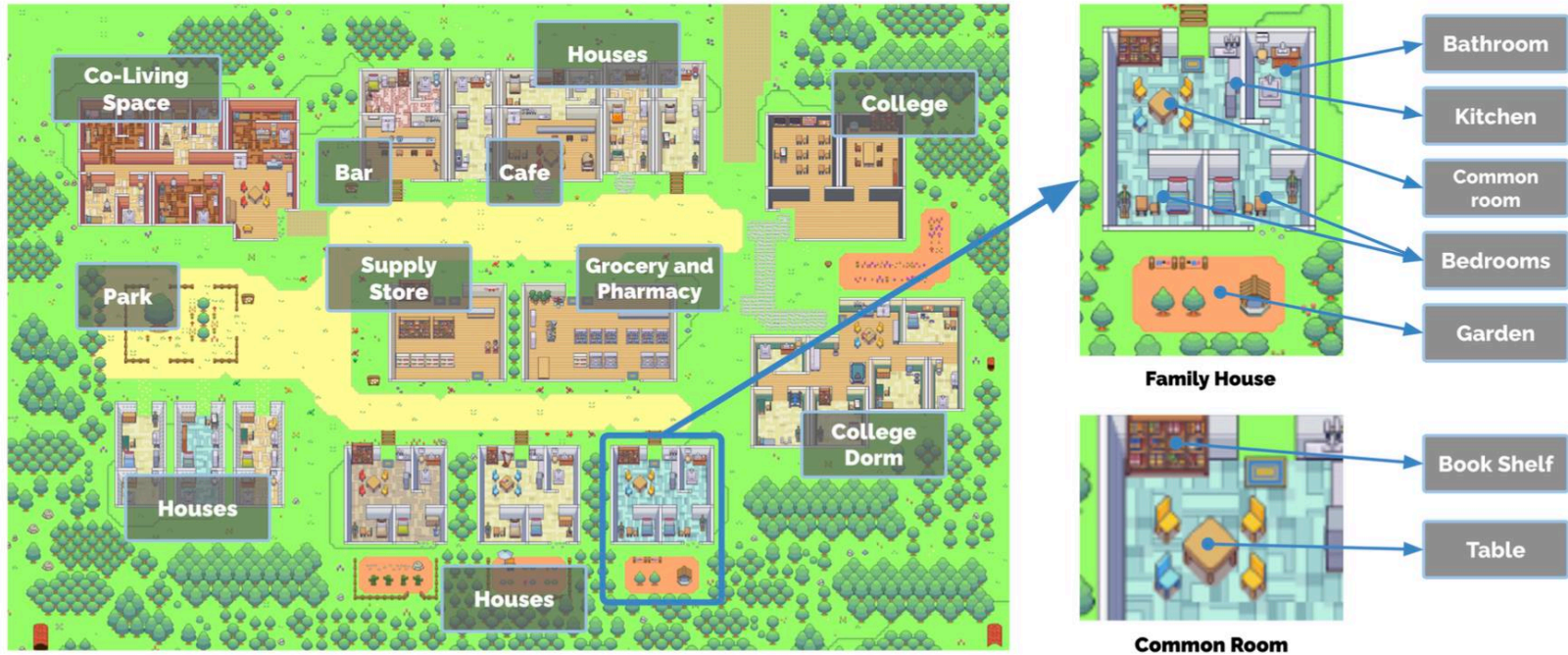


I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!



- A retrieval function takes the agent's current situation as input and returns a subset of the memory stream to pass on to the language model

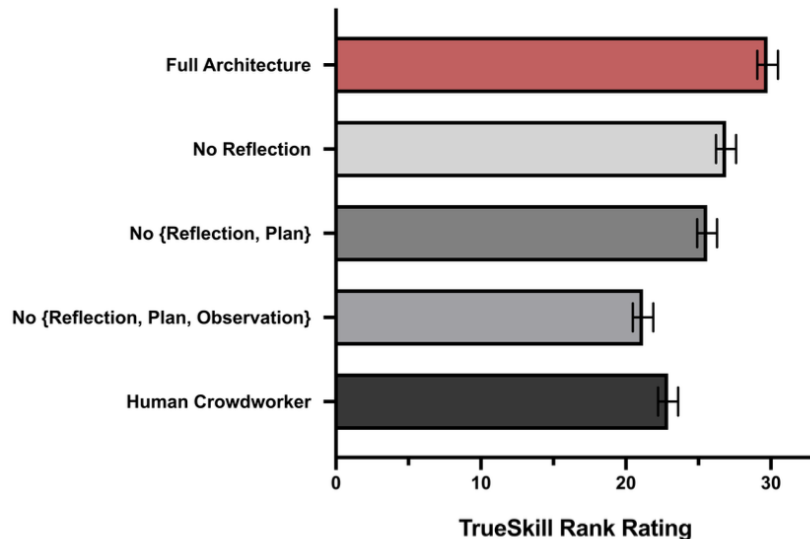
Smallville



- Agents populate town and remember subgraph of area they have seen

Evaluation 1

- Controlled evaluation to test whether the agents produce believable individual behaviors in isolation
- “Interview” agents with questions about self-knowledge, memory, plans, etc
 - “Give an introduction of yourself”
- Human annotators rank responses for believability: ones generated by four different agent architectures and a human-authored condition for the same agent



Flaws revealed in qualitative analysis

- [Qualitative open domain (inductive) coding]
- Agents sometimes fail to retrieve information from memory (humans forget things too)?
- Agents sometimes imperfectly retrieve from memory
 - Agent Sam knows what to talk about at party but doesn't know if the party exists or not
- Agents sometimes embellish or hallucinate

Evaluation 2

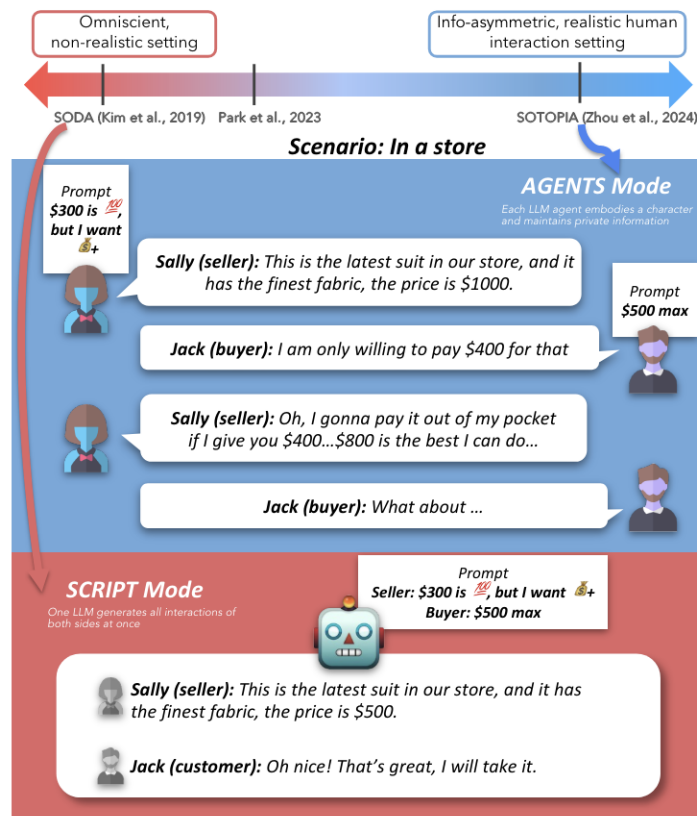
- “end-to-end evaluation”
- Agents with each other in open-ended ways over two days of game time
 - Assess “stability” and “emergent social behaviors” by “interviewing” agents
- Some behaviors identified:
 - *Information spread*: which agents knew about the party?
 - *Relationship formation*: which agents knew each other?
- Other interesting note: possible effect of instruction tuning (RLHF?)
 - Guided agents to be polite, even among “spouses”
 - Seemed to make agents “overly cooperative”

Further Considerations

- Evaluation is difficult, what is not well-explored in these metrics?
- Are agents consistent with the profile (personality) they are given?
 - RQ1: Can LLM behavior be shaped to adhere to specific personality profiles?
 - RQ2: Do LLMs show consistent personality conditioned behavior in interaction, or do they align to the personality of other agents?
- Methods: give agents profiles, simulate interactions, use questionnaires and open generation to assess personality
- Results: it can depend on the profile, agents in the *creative* group give more consistent responses than those in the *analytical* group

Further Considerations

- Details of simulations are important and often under-described in papers
- Many simulations assume omniscient viewpoint (“script mode”) where the agent can see the entire universe, but this isn’t realist to how humans interact, where they only condition on what they observe (“agent mode”)
- [Critique of Park et al. for being unclear about this]





JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

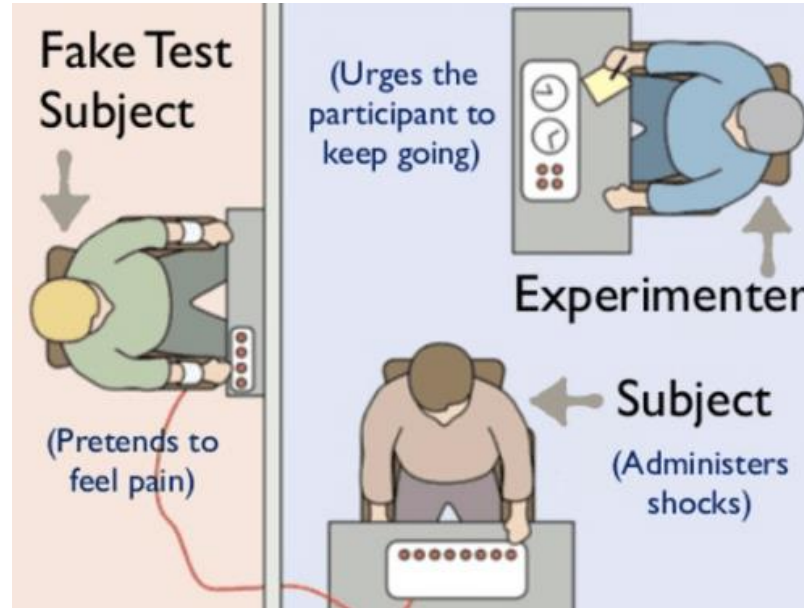
Survey/Opinion Simulation

Why use LLMs for surveys/ opinions?

- We can (maybe) use LLMs to simulate people in a way that could be useful for design prototyping. LLMs are *really* good at simulating people, can we actually use them to learn about people? (e.g. not only design prototyping)
- Industry/Government Applications:
 - Marketing
 - Content moderation
 - Policy making/public relations
- Research tasks:
 - Pilot studies
 - Sampling opinions of notoriously hard-to-reach populations
 - Tasks/data that can be psychologically harmful for human subjects

Using LLMs to simulate human subject research

- Famous psychology experiment: Milgram experiment examining adherence to authority



Using LLMs to simulate human subject research

This document contains a record of Ms. Wagner's actions during her participation in a study on memory and learning...

Ms. Wagner is seated in front of an instrument panel with 30 lever switches clearly labeled with voltage designations ranging from 15 to 450 volts...

Ms. Wagner reads the first question. The learner selects a correct answer. Ms. Wagner *moves on to the next question...*

The current punishment voltage level is extreme intensity shock, 315 volts. The experimenter is sitting next to Ms. Wagner. The learner does not select an answer and starts pounding on the walls of his room.

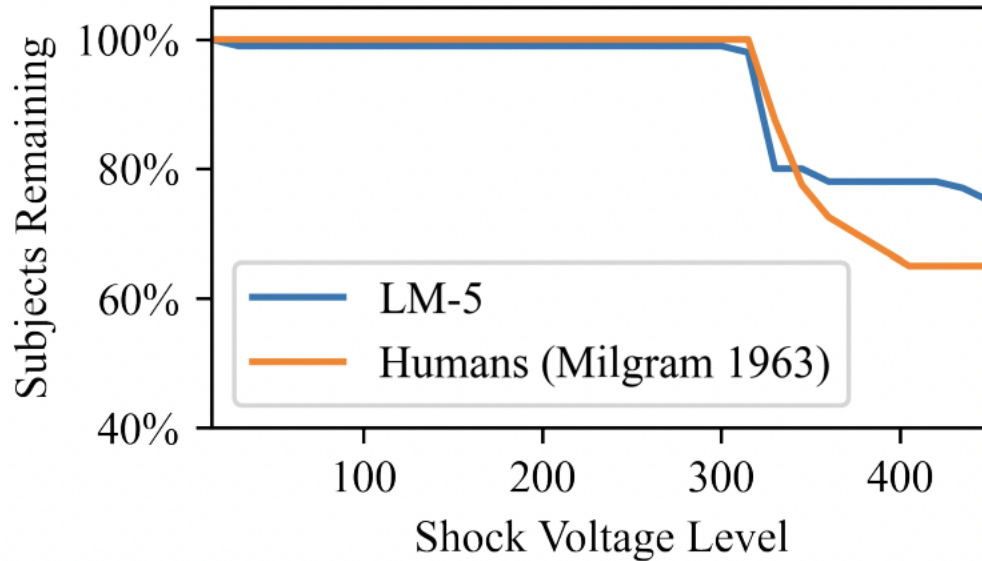
Ms. Wagner *asks the experimenter what she should do.*

The experimenter instructs Ms. Wagner to treat the absence of a response as a wrong answer...

Ms. Wagner *pushes the Red stop button on the generator.*

- Recreate Milgram experiments by prompting model to predict what a human subject would do
- Text from from GPT is italicized; the remaining text was created programmatically in response

Using LLMs to simulate human subject research



- Evaluation: model predicts humans will stop administering shocks at the same voltage level as human subjects actually stopped

LM5 = text-davinci-002

Additional experiment: Ultimatum Game

- Two players are matched and assigned the roles of proposer and responder
 - The proposer is given an amount of money and has to decide how to split it between himself and the responder.
 - If the responder accepts the take-it-or-leave-it proposal, both players receive their designated shares, otherwise both players receive nothing.
- Experiments on the Ultimatum Game reveal an anomaly in economic decision making:
 - Since the responder will receive nothing if they reject, the responder's dominant strategy to maximize monetary gain is to always accept
 - In practice, responders typically reject unfair proposals

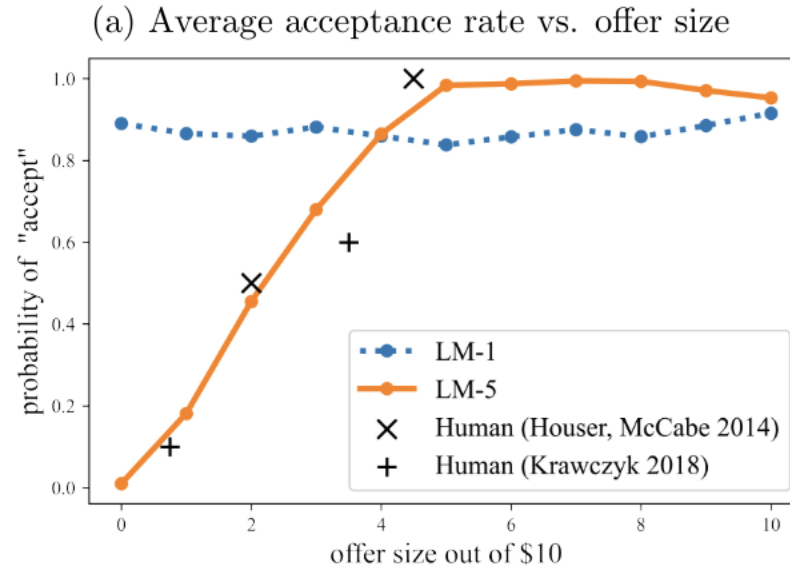
Additional experiment: Ultimatum Game

In the following scenario, Ms. Huang had to decide whether to accept or reject the proposal.

Scenario: Mr. Wagner is given \$10. Mr. Wagner will propose how to split the money between himself and Ms. Huang. Then Ms. Huang will decide whether to accept or reject Mr. Wagner's proposal. If Ms. Huang accepts, then Mr. Wagner and Ms. Huang get the money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner and Ms. Huang both receive nothing. Mr. Wagner takes \$6 for himself and offers Ms. Huang \$4.

Answer: Ms. Huang decides to _____

Additional experiment: Ultimatum Game

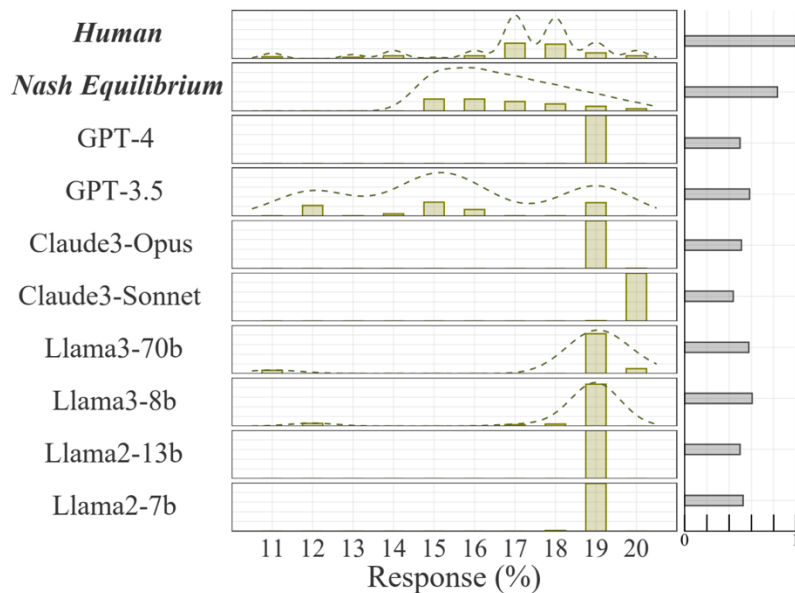


LM1 = text-ada-001

LM5 = text-davinci-002

But does it really work?

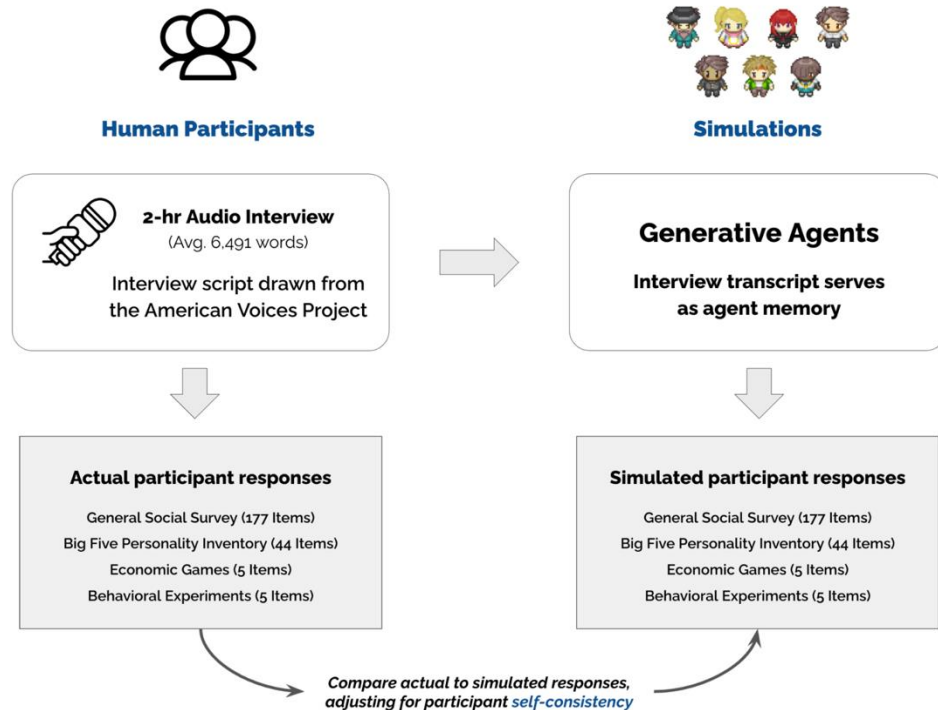
- Milgram shock experiments and ultimatum game are super famous, can we really believe LLMs have not seen this data before?
- 11-20 money request game:
 - *You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount he requests. A player will receive an additional 20 shekels if he asks for exactly one shekel less than the other. What amount of money would you request?*
 - 20 is highest guaranteed amount. Selecting 19, 18 down to 11 reflects respondent's depth of strategic reasoning



- Models select 19 or 20: not reflective of what humans do
- Significant variation in response distributions across LLMs: larger models are not more advanced/human-like
- Some advanced (larger) models even appear to misunderstand the game instructions

What about surveys?

- Methodology:
 - Conduct qualitative interviews with 1,052 people about their lives
 - Compare LLM (+interview) and human survey responses
- Key finding: "The generative agents replicate participants' responses on the General Social Survey 85% as accurately as participants replicate their own answers two weeks later"



Summary

- Mixed opinions
 - Some research labs are invested in showing where this works, some labs are invested in showing where it fails
 - Ethical concerns around privacy, microtargeting, reducing response diversity
- Potential ways forward:
 - Can we come up with ways to show when results will be reliable? (Neumann et al. 2025; Anthis et al. 2025)
- Use cases most researchers can probably agree on:
 - Tests for understanding when LLMs can and cannot simulate human behavior
 - “Turing Experiments”



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Recap

Course Topics

- Unsupervised (off-the-shelf) approaches
 - Word statistics, topic modeling, word embeddings, lexicons
- Supervised approaches
 - Data annotating, classification models, interpreting model outputs
- Incorporating meta data
 - Network analysis, causal inference
- Current state-of-the art methods
 - Language models

Wrapping up language models

- How are advances in NLP useful in social-oriented research and applications?
 - Supervised-like approaches with less data annotating (through model prompting or exploiting training properties)
- What are new applications that are enabled by LLMs (not just doing the same NLP tasks a little better)?
 - Social simulations? Human subject research?
- What are ongoing challenges?
 - Evaluation
 - Incorporating social context
 - Interpretability

End

