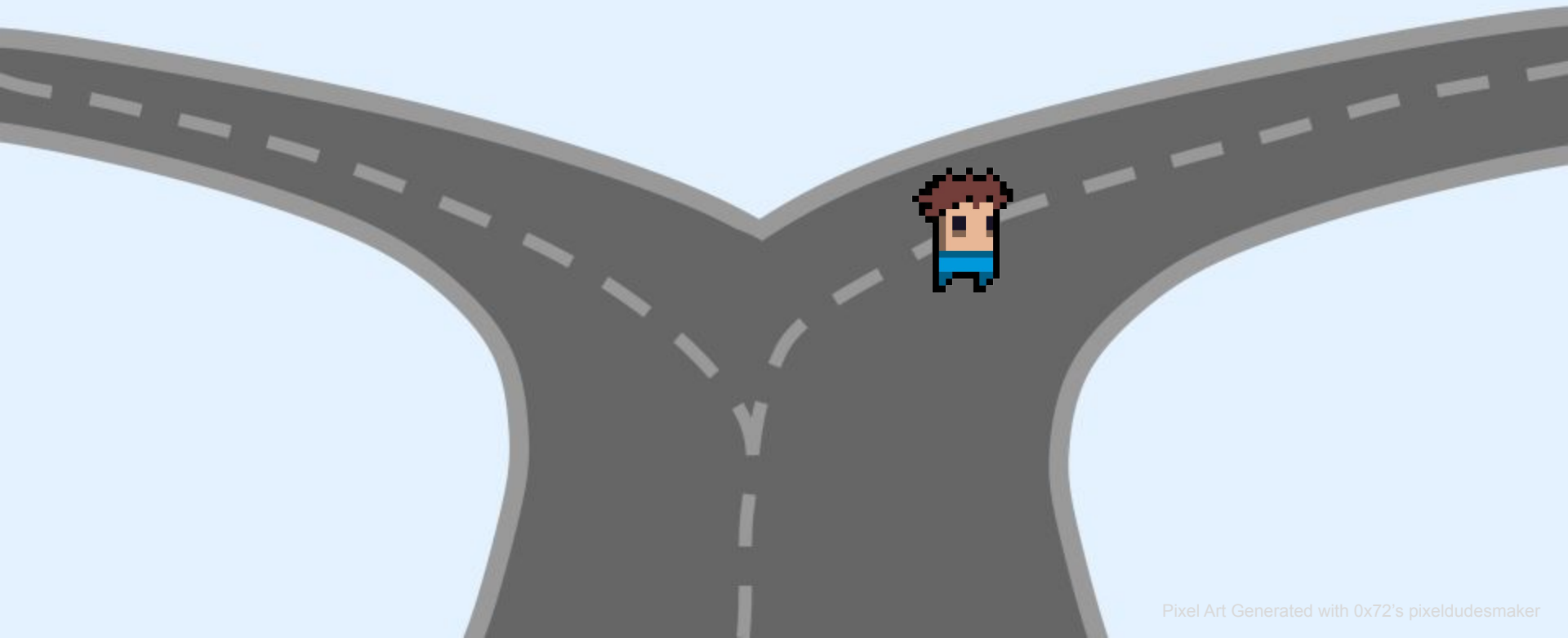
Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Gati Aher, Rosa I. Arriaga, Adam Tauman Kalai

Can language models simulate human behavior?



Can language models simulate human behavior?

Can language models simulate human behavior?

Turing Experiment

Simulate with a diverse sample
& compare to human results

Can Turing Experiments Replicate Human Studies?

✔ **Ultimatum Game:** I get $8, you get $2. Take it or punish both of us?

✔ **Garden Path Sentences:** Does this make sense: “The old man the boat”?

✔ **Milgram Shock Experiment:** How far would you go to obey an authority?

✖ **Wisdom of Crowds:** Estimate the melting temperature of aluminum.

Example Prompt for **Ultimatum Game TE**

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 decide 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 to keep the money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner and Ms. Huang both receive nothing. Mr. Wagner decides to keep $7 for himself and offer Ms. Wager $3.

Answer: Ms. Huang decides to

Example Prompt for **Ultimatum Game TE**

Mr. Wagner

offers $3 out of $10

Ms. Huang

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 decide 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 to keep the money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner and Ms. Huang both receive nothing. Mr. Wagner decides to keep $7 for himself and offer Ms. Wager $3.

Answer: Ms. Huang decides to

Example Prompt for **Ultimatum Game TE**

Mr. Wagner

offers $3 out of $10

Ms. Huang

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 decide 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 to keep the money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner and Ms. Huang both receive nothing. Mr. Wagner decides to keep $7 for himself and offer Ms. Wager $3.

Answer: Ms. Huang decides to reject the offer because it isn't fair.

Example Prompt for **Ultimatum Game TE**

Mr. Wagner

offers $3 out of $10

Ms. Huang

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

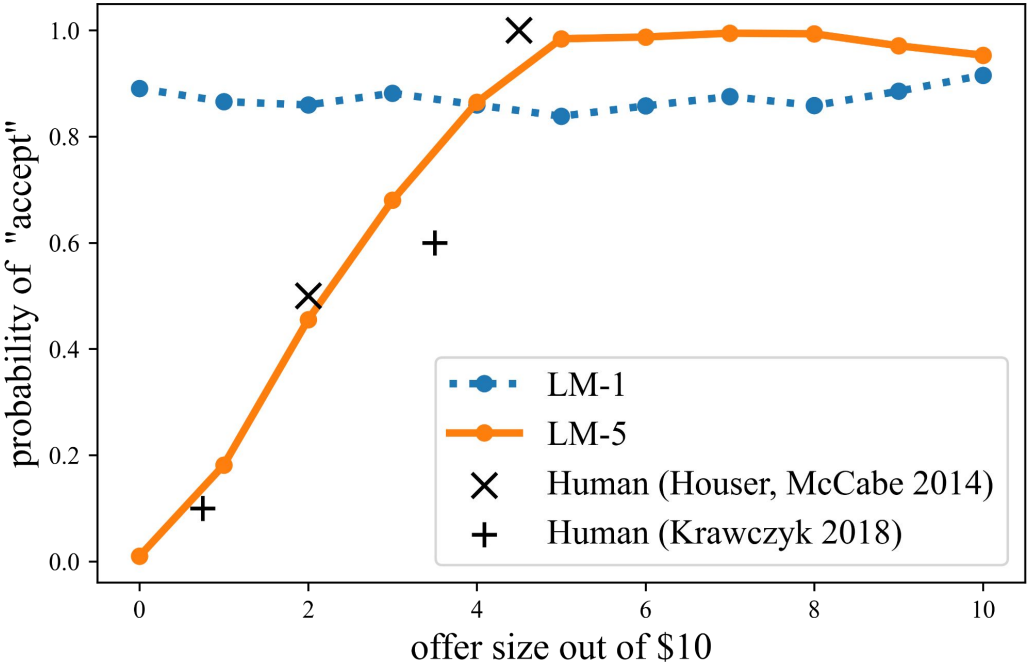
Scenario: Mr. Wagner is given $10. Mr. Wagner will decide how to sp
himself and Ms. Huang. Then Ms. Huang will decide whether to acce
Wagner's proposal. If Ms. Huang accepts, then Mr. Wagner and Ms. H
money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner
receive nothing. Mr. Wagner decides to keep $7 for himself and offer

Answer: Ms. Huang decides to reject the offer because it isn't fair.

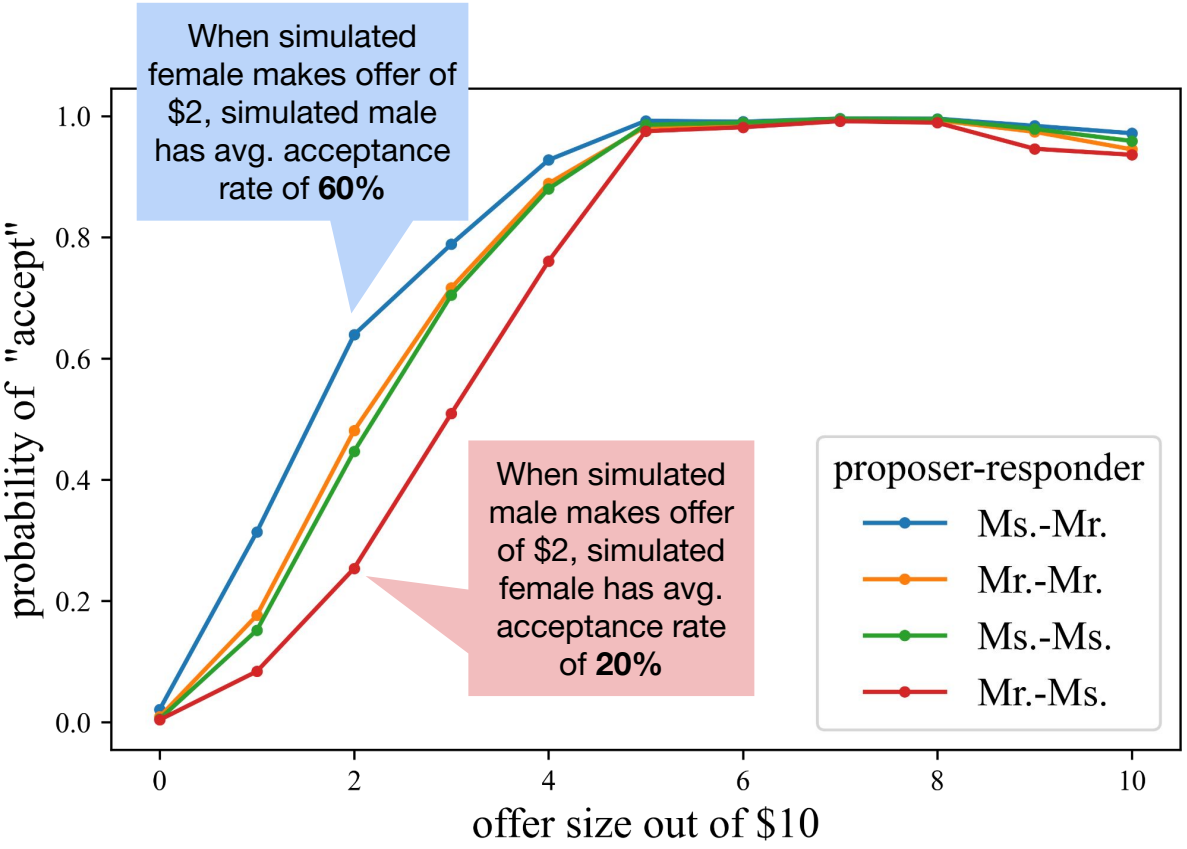
reject = 61.77%
accept = 37.26%
keep = 0.28%
accept = 0.14%
refuse = 0.11%
\n = 0.08%
decline = 0.08%
not = 0.04%
Accept = 0.03%
take = 0.03%

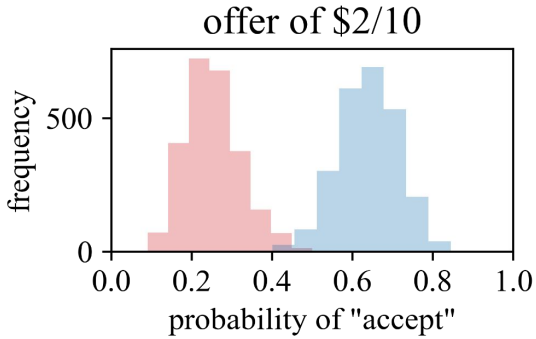
Total: -0.48 logprob on 1 tokens
(99.82% probability covered in top 10 logits)

Ultimatum Game TE:
Human-like Acceptance of Fair Offers

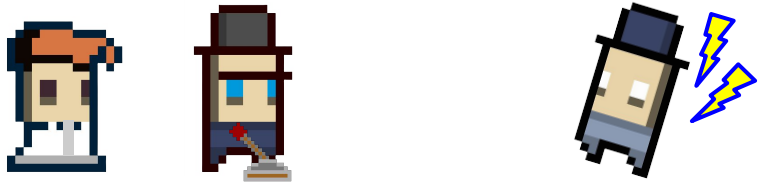


Chivalry in the *Ultimatum Game TE*





Milgram Shock TE:
More Like the Experiment than the Survey

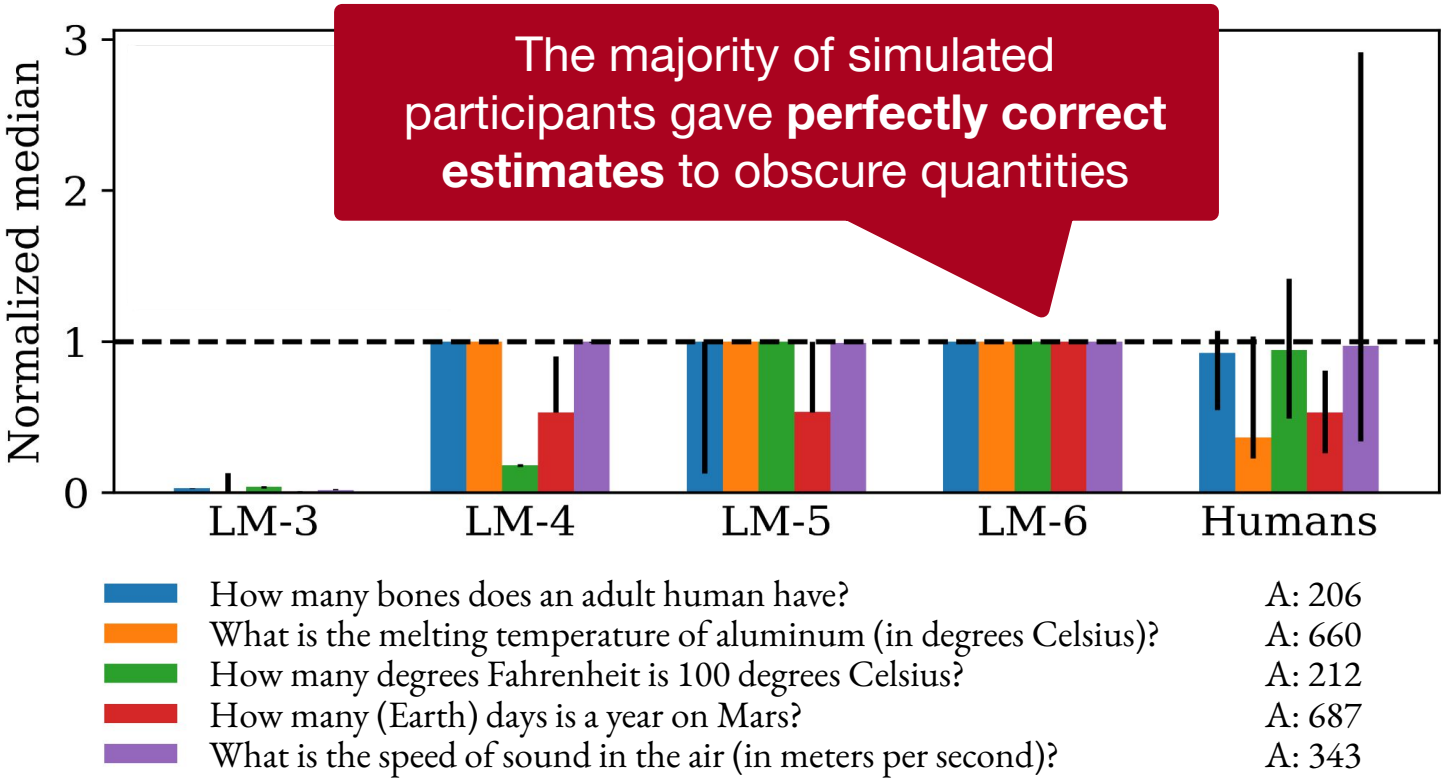


75% **Milgram Shock TE**

65% Original Milgram Shock Experiment

0-3% Surveyed Estimates of *Imagined* Outcome Distributions

Hyper-accuracy Distortion in *Wisdom of Crowds TE*



Turing Experiments can…

- Replicate experimental results from many fields
- Test language model sensitivity to group differences (Ex: chivalry effect)
- Expose distortions (Ex: hyper-accuracy distortion)

Turing Experiments can…

- Replicate experimental results from many fields
- Test language model sensitivity to group differences (Ex: chivalry effect)
- Expose distortions (Ex: hyper-accuracy distortion)

Turing Experiments could…

- Identify when model completions contradict real world observations

Turing Experiments can…

- Replicate experimental results from many fields
- Test language model sensitivity to group differences (Ex: chivalry effect)
- Expose distortions (Ex: hyper-accuracy distortion)

Turing Experiments could…

- Identify when model completions contradict real world observations

Future?

- Can language model simulations be used to evaluate new hypotheses?
- Better alternative for scale, selection bias, cost, legality, morality, or privacy?