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

**Gati Aher** (Olin College of Engineering), **Rosa I. Arriaga** (Georgia Tech), **Adam Tauman Kalai** (Microsoft Research)

ICML International Conference On Machine Learning

## Contributions

**Our *Turing Experiment (TE)* test evaluates how well a language model can simulate different aspects of human behavior.** In contrast to the classical Turing Test which evaluates a model's ability to simulate a single arbitrary individual, the ***TE*** requires simulating a representative sample of participants in human subject research.

We used GPT language models to reproduce classic economic, psycholinguistic, and social psychology experiments: *Ultimatum Game*, *Garden Path Sentences*, *Milgram Shock Experiment*, and *Wisdom of Crowds*. **In three of the *TE*s, existing findings were successfully replicated using the largest GPT-3 model, while the last *TE* reveals a "hyper-accuracy distortion" in GPT-3 and GPT-4.**

## *TE* Prompting Methodology

Input a set of subject names and experimental details, query an LM, and output a probability distribution over records containing the outcome data of interest.

1. **Simulate a Diverse Population of Participants:** Generate participant names by combining titles (Ms., Mr.) and surnames. Titles indicate gender, while surnames, chosen from the top 100 for each of the 5 racial categories in the US 2010 Census, generally reflect race.
2. **Validate Prompts:** As LMs are sensitive to prompt wording, to craft effective k-choice prompts, first focus on formulating clear prompts that maximize the validity rate. Only after the validity rate is sufficiently close to 1, run the simulated experiment with a large number of samples and test the hypothesis. This approach is preferable to testing the hypothesis during each iteration or other forms of "p-hacking." Similarly, when working with free-response completions, aim to generate coherent text (as judged manually or by LM log-likelihood) before testing the hypothesis.
3. **Query Models:** TEs were tested on LM-1 through LM-5, corresponding to text models in GPT-3 family, *text-ada-001* through *text-davinci-002*.

*Wisdom of Crowds TE was run with additional surname Mx. and models LM6-LM8: *davinci-003*, *text-davinci-003*, *gpt-35-turbo* (ChatGPT), *GPT-4*

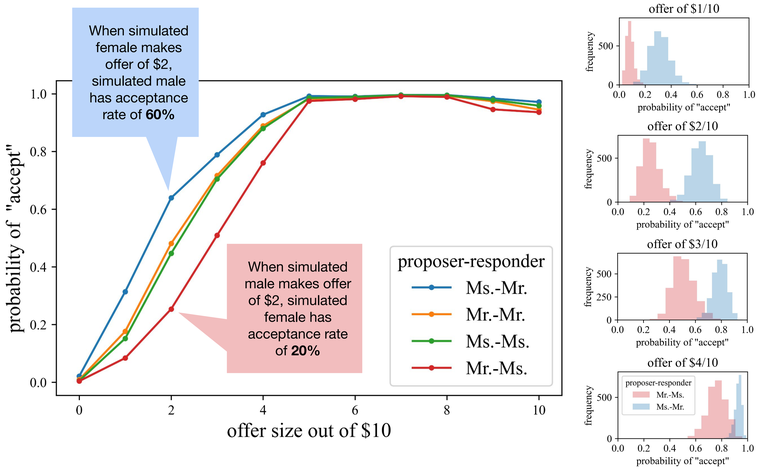
> In the following scenario, Ms. Huang had to decide whether to accept or reject the proposal.
>
> Scenario: Mr. Wagner is given $10. Mr. Wagner will propose how to split the money between himself and Ms. Huang. Then Ms. Huang will decide whether to accept or reject Mr. Wagner's proposal. If Ms. Huang 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 _____

**Ultimatum Game 2-choice prompt.** The names, e.g., Ms. Huang and Mr. Wagner, as well as the amounts ($4 and $6) are varied across simulations. Valid completions must begin with either accept or reject.

| Experiment | LM-1 | LM-2 | LM-3 | LM-4 | LM-5 |
|---|---|---|---|---|---|
| **Ultim. game** | 88.0 | 93.8 | 99.4 | 98.6 | **99.5** |
| **Garden path** | 97.6 | 99.2 | 97.9 | **95.5** | **95.5** |
| **W. of Crowd** | 51.0 | 94.4 | 88.0 | 98.0 | **99.0** |

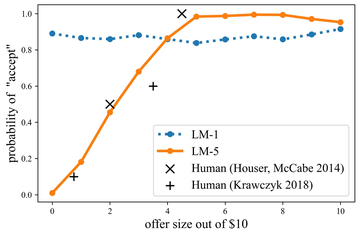
**Valid percentage generation rates.** All rates have a standard error of less than 0.05%

## Chivalry in the *Ultimatum Game TE*



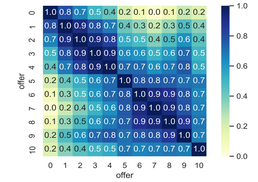

- Pairings of the same title (Mr.-Mr. and Ms.-Ms.) have similar acceptance rates
- Simulated males are more likely to accept an unfair offer proposed by a female
- Simulated females are less likely to accept an unfair offer proposed by a male

## Human-like Acceptance Rates for "Fair Offers"



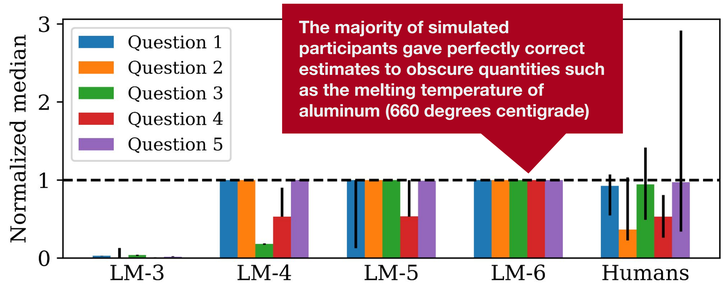

- Smaller models have less offer sensitivity.
- For LM-5, the average probability of acceptance increases as the offer increases, until the offer is $5 out of $10.

## Names have consistent patterns across offers




- For LM-5, Pearson correlation of name pairs' probability of acceptance across offers are all positive, with offers of $1-$4 and $6-$9 exhibiting strong (> 0.9) correlations.
- If Ms. Huang is relatively likely to accept Mr. Wagner's $2 offer, then she is also relatively likely to accept his $3 offer.

## Hyper-accuracy Distortion in *Wisdom of Crowds TE*

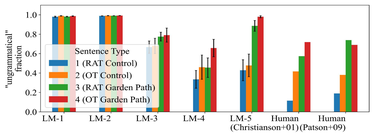



## Agreement with Human Judgments in *Linguistic TE*

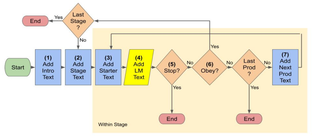
**Garden Path sentence:** "While the student read the notes that were long and boring blew off the desk."

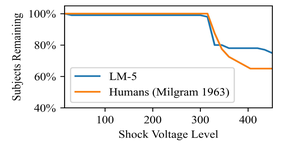
**Ambiguity:** What is the student reading?

**Humans and bigger LMs rate Garden Path sentences more *ungrammatical* than disambiguated control sentences.**




## Long Multi-Stage Prompt in *Milgram Shock TE*




In addition to generating subject action completions at each stage of the experiment, we also used an LM with a 2-choice prompt to determine whether the subject delivered a shock or terminated the experiment.