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

Gati Aher [1] **Rosa I. Arriaga** [2] **Adam Tauman Kalai** [3]

Abstract

We introduce a new type of test, called a Turing Experiment (TE), for evaluating to what extent a given language model, such as GPT models, can simulate different aspects of human behavior. A TE can also reveal consistent distortions in a language model's simulation of a specific human behavior. Unlike the Turing Test, which involves simulating a single arbitrary individual, a TE requires simulating a representative sample of participants in human subject research. We carry out TEs that attempt to replicate well-established findings from prior studies. We design a methodology for simulating TEs and illustrate its use to compare how well different language models are able to reproduce classic economic, psycholinguistic, and social psychology experiments: *Ultimatum Game*, *Garden Path Sentences*, *Milgram Shock Experiment*, and *Wisdom of Crowds*. In the first three TEs, the existing findings were replicated using recent models, while the last TE reveals a "hyper-accuracy distortion" present in some language models (including ChatGPT and GPT-4), which could affect downstream applications in education and the arts.

1. Introduction

We introduce a methodology for systematically evaluating which aspects of human behavior a language model, such as a GPT model (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023), can faithfully simulate and which aspects it systematically distorts. This understanding can inform downstream applications that require language models to have accurate models of humans, including various applications in education and the arts. The question of faithful simulation of a specific behavior is studied through controlled experiments, and we thus avoid philosophical debates around the meaning of "understanding" (Bender & Koller, 2020). Now, simulating human behavior can be hard, even for humans, especially in complex real-world situations fraught with ambiguity. After all, if simulating human behavior were easy, there would be no need to run human subject experiments as one could simply simulate the outcomes. A further obstacle to accurate simulation is that behavior differs across individuals and populations, and perfect simulation would require capturing these differences for all groups including minority groups.

In Turing's Imitation Game (IG), an AI system has to simulate an individual well enough to fool a human judge. Language Models (LMs) may come close to "winning" this game in the near future, especially if they only have to simulate a single human—one oddly successful early attempt simulated a 13 year old troublemaker (Warwick & Shah, 2016). However, the IG is of limited diagnostic value as it says little about which humans and behaviors an LM can faithfully simulate. We thus move on to the more specific challenge of identifying which aspects of human behavior a given AI system can and cannot simulate.

A Turing Experiment evaluates an AI system[1] in terms of its use in simulating human[2] behavior in the context of a specific experiment, like a human subject study. A TE uses an AI model to simulate the behavior of *multiple subjects* in an experiment. TEs may be used in any discipline which involves human participants in studies, including Social Psychology, Linguistics, and Behavioral Economics.

Formally, there are two types of inputs to each TE which parameterize the experimental setting. First are participant details which might include names, occupational information, or other demographic details. The second type of input is optional experimental conditions which may include relevant setting details and stimuli. The TE's output is a synthetic *record* describing a simulated human experiment and any outcomes of interest. The TE must be a procedure run on a computer (*aka* a Turing machine, hence the TE

[1]Olin College of Engineering [2]Georgia Tech [3]Microsoft Research. Correspondence to: Gati Aher <gativaher@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]While we focus on LMs, the AI system need not be text based (e.g., it could generate videos).

[2]While we focus on research involving human behavior, AI systems may simulate animal (or purely physical) experiments.

name). Importantly, the TE should be *zero-shot*, meaning that neither the procedure nor any training data used by the AI system should include prior data specific to that experiment; otherwise the model may simply repeat the prior data. (This ideal can be difficult to enforce with models pretrained by others on massive corpora.) Overall findings or specific outcome data can be compared to results from human subject research to determine how *faithful* the simulation is. A *replication* TE is for replicating a finding in prior human subject research.

In addition to introducing the concept of TEs, we demonstrate their feasibility by presenting a methodology for running TEs using an LM, like GPT models, that takes a text *prompt* and generates a randomized *completion*, which is text that would be likely to follow that prompt, based on its training data. For each TE, we write a program that creates one or more (zero-shot) prompts that are fed into the LMs. Then the text generated by the LM is used to reconstruct the record, a text-based transcript of the simulated experiment. Figure 1 illustrates the difference between a typical prompt used for classification and our prompt used to run a TE, which can generate multiple records by varying names and gender. Our methodology includes an important validation step that involves the tweaking of prompts without examining the experimental outcomes (so as to avoid "p-hacking.") These programs can then be run with any prompt-based LM.

Finally, we apply this methodology to four TEs aimed to replicate well-studied phenomena in different fields, and evaluated 5-6 available LMs through OpenAI's paid API to access GPT models. In all four TEs, we define participant inputs as surnames and gender titles (e.g., *Mr*, *Ms* or *Mx*) as a simple way to simulate gender and racial diversity. The other inputs and outcomes vary by TE.

The first TE is the Ultimatum Game, used to study fairness and rationality in Behavioral Economics, where the experimental condition is an amount of money offered to a participant, and the outcome is the accept/reject decision. The second TE is garden-path sentences, used to study parsing in psycholinguistics, where the experimental stimuli is a sentence (of type normal or garden path), and the outcome is the participant's judgment of grammaticality. The third experiment is the Milgram Shock Experiment, designed to study obedience to authority in social psychology, where the outcome is the number of shocks the participant administered. The final TE is the wisdom of crowds, used to study collective intelligence across disciplines, where the experimental condition is a numerical general-knowledge question, and the outcome is the participant's numerical estimate of the answer. To address the concern that the LMs have been exposed to all of these classic experiments in their training data, we construct variations on the experimental details for three out of four studies. We run simulations using our own garden path sentences, our own novel destructive obedience scenario similar to Milgram's Shock experiment, and our own general-knowledge questions.

(a) Typical few-shot prompt for classification:

Classify each sentence based on whether it is a garden path sentence or a normal sentence. A garden path sentence is a grammatically correct sentence whose likely reading appears to be ungrammatical. Sentence: The old man the boat. Classification: garden path Sentence: The cat chased the mouse that was in the house. Classification: normal Sentence: While the student read the notes that were long and boring blew off the desk. Classification: _______

(b) TE prompt for simulating a named individual:

Ms. Olson was asked to indicate whether the following sentence was grammatical or ungrammatical. Sentence: While the student read the notes that were long and boring blew off the desk. Answer: Ms. Olson indicated that the sentence was _______

Figure 1. Classification versus simulation prompts for a garden path sentence. The blanks will be filled-in by the LM, and prompts are constructed so that the LM is likely to adhere to a desired format. Prompt (a), not used in our work, illustrates a typical prompt that might be used to evaluate the LM's capability to identify garden path sentences. Prompt (b) which we used in a TE, can be used to simulate the responses of multiple different individuals by varying the name.

In the Ultimatum Game TE, we show how that the simulation outcomes vary consistently by gender (and name) which further demonstrates the potential of our approach to replicate gender differences reported in human subject research. By describing other demographic details, such as occupation or age, simulated participants in TEs may be varied in a manner similar to (but much easier) than human subject studies.

An evaluation is useful if it reveals the flaws and strengths of the models. In the first three TEs, as expected larger models provided more faithful simulations than smaller ones, with the largest model replicating the finding and producing outcomes consistent with those of prior human studies. In the last TE, Wisdom of Crowds, the larger models did not outperform the smaller ones–if anything the trend was reversed, revealing a peculiar flaw within some models.

Distortions. We call a systematic difference between the two a *distortion* (different from *biases* which may be common to humans and machines). The last TE revealed an interesting *hyper-accuracy distortion* for the recent GPT LMs (including ChatGPT and GPT-4 (OpenAI, 2023)): the majority of simulated participants gave perfectly correct estimates to obscure quantities such as the melting temperature of aluminum (660 degrees centigrade). Since the hyper-accuracy distortion does not seem to be present in earlier GPT models currently available through the API, the distortion may be due to "alignment" procedures (Ouyang et al., 2022) which improve LM truthfulness. Note that distortions may be *desirable* in many applications, e.g., a teacher may prefer a distortion which *diminishes* the gender bias that pervades LM training data (Bolukbasi et al., 2016).

Why this matters: downstream applications. Our understanding of distortions may inform their use in downstream applications. For instance, the performance of educational applications such as a science tutor system may suffer if its model is that all students perfectly know quantities such as aluminum's melting point. Similarly, an LM might author a highly unrealistic story. As an anecdotal example, we queried GPT-4 (released after the studies in this paper) to "Write a short story about Alice, a 5-year-old, thinking about traveling to the moon. Include her estimates of how long it would take and how far it is." It's response was:

Alice was a precocious little girl. At the tender age of five, she was already fascinated with everything beyond the sky... She had overheard her dad saying that the moon was about 238,900 miles away. Alice found that number hard to fathom, but she tried to compare it to the long car trips her family often took. Her parents had told her they usually drove at 60 miles per hour on highways. She wondered, "If our car can go 60 miles in an hour, how long would it take to go all the way to the moon?"

She began to scribble and count, using her fingers and toes when the numbers got too big. After some effort, she estimated that if they didn't stop for snacks or to take pictures, it would take about 3,982 hours. She knew there were 24 hours in a day, so again she started counting. It came out to 166 days.

This story exhibits a completely unrealistic model of a five-year-old's ability to accurately divide numbers. TEs helped us anticipate a problem in using LMs. In particular, LMs (which have been heavily "aligned") may have unrealistic models of human numerical knowledge and accuracy.

Contributions. TEs are a means of evaluating the zero-shot simulation capabilities of AI models and, in this sense, TEs provide much more insight than Turing's IG into which human behaviors are captured in an AI system. The main contributions of our work are: (1) proposing TEs (2) introducing a methodology for running TEs on LMs using prompts and records, and (3) designing and executing four TEs across a handful of LMs and uncovering a distortion. It is also worth noting that TEs may be also predate a human subject study and may inform the design of costly experiments. We also discuss ethical considerations, limitations and risks associated with TEs.

1.1. Related Work

Recent independent related works consider questions related to the similarity between humans and LMs. Several works use human failure modes to reason about LM failure modes. Jones & Steinhardt (2022) use human cognitive biases, such as anchoring and framing effects, to evaluate an LM's "errors" where it deviates from rational behavior. Binz & Schulz (2023) use cognitive psychology tests to address the question of whether LMs "learn and think like people." Hagendorff et al. (2022) tested GPT-3.5 using cognitive response tests and found that the LM's error mode "mirrors intuitive behavior as it would occur in humans in a qualitative sense." Dasgupta et al. (2022) test LMs on abstract reasoning problems and find that "such models often fail in situations where humans fail – when stimuli become too abstract or conflict with prior understanding of the world." While these works studied the capabilities of current LMs, we introduce a new evaluation methodology that illustrates how LM outputs can capture aspects of human behavior. With this methodology, we can study nuanced differences across simulated populations, such as finding that a large GPT model shows a subtle gender-sensitive "chivalry effect" in the Ultimatum Game TE.

We now discuss several categories of related work.

LMs Representing Humans. Several works propose ways to use LMs as proxies for a diverse set of humans, such as cheaply automating a variety of small writing tasks (Korinek, 2023) and using prompts to generate synthetic human-like interactions with desired properties (Park et al., 2022; Caron & Srivastava, 2022; Jiang et al., 2022; Karra et al., 2022, e.g.,). Our work is most similar to concurrent work on simulating human samples from a population by Argyle et al. (2023), which suggests that LMs can effectively represent different subpopulations when prompted with demographic information. The key difference in our approaches is that Argyle et al. (2023) aims to show the fidelity of LMs in predicting survey result probabilities (e.g., vote prediction given race, gender, party identification, etc.) while we replicate human behaviour experiments. Simulating survey results may be an easier task given that correlations between political survey data and certain demographic attributes are strongly present in the Internet training data. Simulating experiments may be a harder problem, as people's actions sometimes contradict their answers to questions.

LM Evaluation. Due to the importance of LMs, their evaluation has spawned multiple initiatives and conferences, as discussed by Liang et al. (2022). Large-scale efforts have been invested in creating massive text corpora (Marcus et al., 1993; Brown et al., 2020; Chowdhery et al., 2022). Recently, large benchmarks have consolidated numerous LM evaluations across a number of fields (Srivastava et al., 2022; Liang et al., 2022; Hendrycks et al., 2021). The largest, BIG-bench (Srivastava et al., 2022), contains over 200 LM benchmark tasks (including 19 evaluating social reasoning and 16 measuring emotional understanding). Project Delphi introduced the COMMONSENSE NORM BANK (Jiang et al., 2021) of over 1.7 million human moral judgments. Such benchmarks generally consist of questions with "correct" answers, whereas behavioral experiments often involve dilemmas and actions (e.g., shocking another individual in the Milgram Experiment) and people's actions sometimes contradict their answers to questions. Concurrent work by Ullman (2023) shows that although GPT-3 previously showed success on Theory of Mind psychology tasks (Kosinski, 2023), it fails on prompts with several types of directed variations, illustrating the necessity of evaluating the robustness of observed effects using alternative prompts and setups.

Improving Language Models. Also related is the work on developing LMs, such as PaLM (Chowdhery et al., 2022) and GPT-3 (Brown et al., 2020), which provides the API we access. Several works (Ouyang et al., 2022; Wei et al., 2021) investigate how to "align" the LMs with human goals such as truthfulness. As discussed, there may be a tension between aligning LMs and their performance at simulation, e.g., a hypothetical LM that exhibits no gender differences would not be able to simulate gender differences that have been observed in psychology studies. Other forms of alignment, such as Bakker et al. (2022)'s recent work on generating opinions with high consensus across heterogeneous and opposing groups, may be beneficial for creating LMs that retain realistic and demographically nuanced forms of human bias.

Prompt Design. Liu et al. (2021) survey methods for designing prompts. OpenAI's best practices[3] for designing prompts include giving clear instructions alongside a few illustrative examples, called a *few-shot prompt* (Brown et al., 2020), as in Figure 1a. However, it has been shown that LMs such as GPT models are quite sensitive to the choice of examples in the few-shot prompt and even to their order (Lu et al., 2021). So-called *chain-of-thought* prompts (Wei et al., 2022; Kojima et al., 2022) improves generated text by "thinking out loud," which could be useful in designing TEs. Shin et al. (2020) use LMs to create the prompts themselves.

Bias in LMs. Biases are well known to exist in large language models (e.g., Blodgett et al., 2020; Sheng et al., 2021; Chowdhery et al., 2022; Brown et al., 2020). The datasets themselves reflect the biases of the contributing authors and authors may not be equally represented across groups. For example, white males are vastly over-represented among Wikipedia contributors (Wikipedia, 2022) and are followed at higher rates on Twitter (Messias et al., 2017). One related challenge in unpacking LM biases is interpreting their completions and understanding how they arise (Vig et al., 2020).

Tests of Human Simulation. Several variants of Turing's IG (Turing, 1950) have been proposed. Until recently, human simulators required significant training data and were not zero-shot. Agent-based simulations (e.g., Macal & North, 2010) can facilitate complex large-scale simulations of environments for which one has custom behavior models.

2. Running TEs Using LMs

Our methodology for simulating a TE generates *records* describing the experiment, which also contain the outcome data of interest. Our algorithms take inputs, query an LM, and output records or, more generally, a probability distribution over records. As mentioned earlier, the input to the simulator consists of relevant details describing the experimental condition. This may include subject names and other demographic details such as occupation, age, or salary that the simulator can use in constructing the record and prompts. This also includes conditions and stimuli specific to the experiment. In the garden path example of Figure 1b, the inputs were the name *Ms. Olson* and the sentence to be judged.

Queries. Over the course of executing an experiment, a simulator makes one or more calls to an LM. A standard autoregressive LM computes a probability distribution over a next word (or token) $p(w_i|w_1w_2\dots w_{i-1})$, for any $w_1\dots w_{i-1}$. This can be used to compute the probability of any sequence of words $p(w_1w_2\dots w_i)$. We query the LM in two ways. The first is simply a *completion*, where the LM is given an arbitrary prompt string s and it generates a randomized completion, iteratively, one word at a time, each time sampling according to the probability distribution on the previous words. The second may be called a k-*choice prompt*. Given a prompt s and valid completions $c_1, c_2, \dots, c_k$, the LM outputs a probability distribution where the probability of the ith completion is

[3] `https://beta.openai.com/docs/guides/completion/introduction`

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.*

Figure 2. A sample record of a simulation of the Milgram Shock experiment. Text generated by GPT is italicized; the remaining text was created programmatically in response. The record above has been shortened with ellipses.

$p_i = p(s\, c_i)/Z$ and $Z = \sum_{j \leq k} p(s\, c_j)$ is the normalizing constant called the *validity rate*.

Outputs. The record output is a text log of a single (simulated) run of the experiment that contains the *outcomes* of interest, such as whether or not a sentence was judged as grammatical in a garden path simulation or how many shocks were administered in the Milgram Shock experiment. A sample record is sketched in Figure 2. The simulator is assumed to output a record or, more generally, a probability distribution over a set of records with non-negative weights that sum to 1. This is a generalization in the sense that, given a distribution over records, one could sample a single record. In particular, the probability distribution computed for a k-choice prompt reflects the fractions of completions that would result in each choice, given infinitely many simulations. This efficiency gain is analogous to weighting training examples in machine learning versus subsampling.

Validating prompts. After one has formulated an hypothesis, one must design the sequence of prompts that will be used in simulation process. Since today's LMs are highly sensitive to prompt wording, a strategy we found effective with k-choice prompts is to focus on formulating clear prompts that maximize the validity rate Z. 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.

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**

Table 1. Valid percentage generation rates for five models across three TEs. This is the percentage of generations that adhere to our validation criteria. All rates have a standard error of less than 0.05%.

Our strategy for designing prompts that maximize the validity rate includes clearly specifying the desired completions in the first few lines of the prompt. If we find that certain undesirable completions are generated frequently, we minimize use of those phrases in the prompts, as LMs generations often repeat phrases occurring in the prompt.

3. Models and Datasets

Models. We conduct our simulations using pre-trained LMs based on the transformer architecture. We use the widely-used OpenAI API to query the following GPT text models: *text-ada-001*, *text-babbage-001*, *text-curie-001*, *text-davinci-001*, *text-davinci-002*, *text-davinci-003*, *gpt-35-turbo* (commonly referred to as ChatGPT), *gpt-4*, which we refer to as LM-1 through LM-8, respectively. Since this ordering reflects increasing price (and claimed capability), we expect that they would produce simulations of increasing fidelity. LMs 6-8 were released recently and were used only in our last study since they were released after we completed the first three. When the models are queried for completions, the natural $\mathrm{temperature} = 1$ and $\mathrm{top\_p} = 1$ parameters are used.[4] Running TEs on different LMs is left for future work and is challenging because most available LMs cannot handle the long prompts we use, particularly in the Milgram TE.

Names. For our TEs, the inputs include subject *names* consisting of a title, either *Mr.* or *Ms.*, indicating binary gender, followed by a surname. Titles and surnames were primarily used to simulate a diverse subject pool, but we also used them to evaluate gender differences in one TE. Lists of surnames were sourced from the U.S. 2010 Census Data. We chose a racially diverse set of surnames, including 100 names from each of five racial groups. The full list of surnames is given in Appendix A. Considering all combinations of the two titles, five racial groups, and one hundred surnames in each group, we have a pool of 1,000 names. In

[4]These parameters correspond to sampling generations $w_1, \ldots, w_n \sim p$ from the LM without modification, i.e., logits are passed through the softmax function exactly as in training.

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 _______

Figure 3. Sample 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*.

the fourth experiment we include the title *Mx.* to illustrate the simulation of non-binary participants.

Study-specific datasets. For the four studies in this paper, we use experimental conditions from and compare results against prior literature. For the Ultimatum Game TE, we use summary findings reported in Houser & McCabe (2014) and Krawczyk (2018). For the Garden Path TE, we use sentences and statistics from Christianson et al. (2001) and Patson et al. (2009). For the Wisdom of Crowds TE, we used 5 general-knowledge questions from Moussaïd et al. (2013). For the Milgram TE, we use the procedure and results from Milgram (1963). To address the concern that the training data for the LMs may contain specific sentences and descriptions of experimental conditions, we also ran the TEs on novel experimental condition datasets. For garden-path TE, we authored 24 original garden path sentences. For the Milgram TE, we developed our own novel destructive obedience scenario. For wisdom-of-crowds TE, we authored 5 general-knowledge questions. Further details are given in the Appendix.

The code necessary to reproduce the data in this paper will be publicly available at https://github.com/GatiAher/Using-Large-Language-Models-to-Replicate-Human-Subject-Studies.

4. The Ultimatum Game TE

Phenomenon. In the Ultimatum Game, first studied experimentally by Güth et al. (1982), 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. We focus on simulating the responder's behavior across conditions with different offers.

Inputs. Our Ultimatum Game simulator takes three inputs: an integer offer in $\{0, 1, \ldots, 10\}$, the name of the proposer, and the name of the responder. The offer corresponds to an initial endowment fixed at $10 and eleven possible offers. Out of the one million possible pairings of proposer and responder names, we chose a subset of 10,000 pairs by the following process. We randomly shuffled the dataset of surnames, paired each of the 500 surnames with five other surnames, one from each racial group in the census data, and then used the 2×2 combinations of "Mr." and "Ms." titles. This procedure yielded a balanced design where each of the 1,000 names was used for the responder 10 times.

Simulation. The simulator constructs 2-choice prompts, as described in Section 2 and illustrated in Figure 3, for each set of inputs and the *accept* and *reject* completion. The record is the concatenation of the prompt and its completion.

Results. To assess the fidelity of the simulated human behavior, we compute validity rates, consistency of decisions for a given name pair across offers, and agreement with prior results in human studies. Validity rates (the probability of generating "accept" or "reject") are shown in Table 1. Figure 4a compares prior reports of mean human acceptance rates to those simulated using LM-1 and LM-5. Results for the other LMs are given in Appendix C.

The distributions generated using LM-5 agree closely with human decision trends, predicting that offers 50-100% of the total endowment are almost always accepted while offers 0-10% are rarely accepted. In contrast, the simulations with smaller language models are not sensitive to the offer amount, having a flat acceptance rate across both fair and unfair offers. Noting that the acceptance rates simulated using LM-5 closely align with those of prior human studies, we now examine the LM-5 simulations more carefully.

Next we analyze whether the simulations show consistent or random differences across name pairs. For instance, if in simulations Ms. Huang is more likely than average to accept Mr. Wagner's $2 offer, is it also the case that Ms. Huang is more likely to accept Mr. Wagner's $3 offer? If so, the simulations must be sensitive to the names in a way that is not purely random. Figure 4b shows the Pearson correlation of acceptance probability for name pairings across offer conditions, simulated using LM-5. There are no negative correlations, and acceptances of offers within 1-4 and within

(a) Average acceptance rate vs. offer size (b) Name sensitivity in correlations across offers

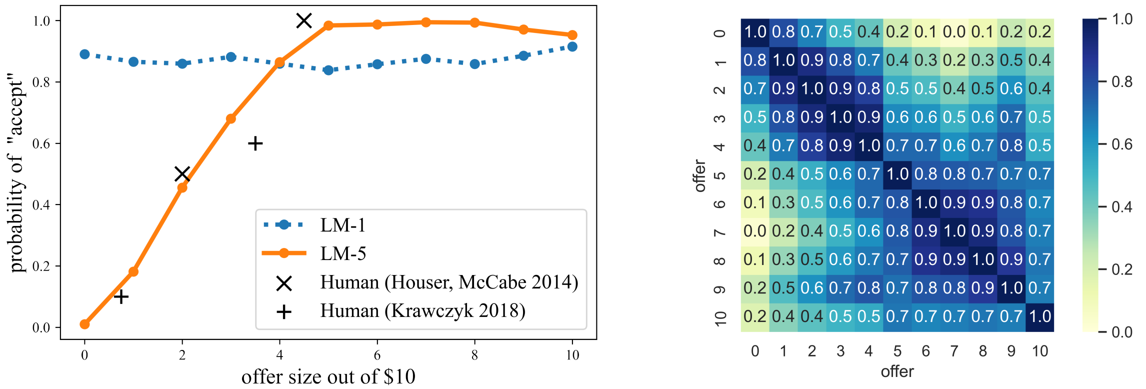

Figure 4. (a) **Comparing Ultimatum Game TE simulations to human subject studies**: The fraction of responders accepting offers versus offer size out of $10. The simulated response curves shown are averaged across all 10,000 name pairs for simulations using the LM-1 and LM-5 models. Results that have been found to be robust across human studies are also marked, for comparison. (b) **Name sensitivity**: To test whether the model is sensitive to changes in names, consistency is measured as we vary offer size across name pairs. LM-5 showed strong Pearson correlations (> 0.9) of acceptance probability for name pairings between offers $1 though $4 and between offers $6 through $9. High positive correlations show that the TE simulation results are sensitive to names and that this sensitivity is consistent for given name pairs rather than random.

6-9 all exhibit strong (> 0.9) correlation. This supports our methodology of using names to simulate multiple different individuals.

Given that there appears to be consistency within name pairs, we next evaluate whether there is a higher-level consistency based on gender. We find that the distributions of acceptance probabilities in the LM-5 simulations vary significantly ($p < 1\mathrm{e}{-16}$) by the gender of the pairings. Pairings of the same title (*Mr.-Mr.* and *Ms.-Ms.*) have similar acceptance rate distributions, while males were more likely to accept an unfair offer proposed by a female (mean acceptance rate of 60% for offer of $2), and females were less likely to accept an unfair offer proposed by a male (mean acceptance rate of 20% for offer of $2). Figure 5 shows the mean trends and distributions for LM-5 acceptance probabilities by gender pairing. While gender differences have repeatedly been reported in human experiments,[5] they are not uniformly consistent (Eckel & Grossman, 2001). Nonetheless, the large gender difference is in our outcomes does demonstrate that LM-5 is affected by gender pronouns and title in a consistent manner.

5. Garden Path Sentences TE

The second TE is related to Garden Path sentences. We begin with the basic phenomenon that humans cannot easily parse garden path sentences. We simulated a judgment of whether or not a sentence appeared grammatical, as illustrated in Figure 1b. The TE faithfully reproduced this basic finding using LM-5 but not the smaller models. Its details are deferred to Appendix D.

6. Milgram Shock TE

Phenomenon. The obedience to authority studies, developed by Milgram (1963), are a series of famous social experiments that aimed to find when and how people would defy authority in the face of a clear moral imperative. The work faced ethical criticism in that the procedure requires that subjects are deceived, are placed in situations stressful enough to cause seizures, and are not clearly told of their right to withdraw. In the original procedure, the experimenter, an authority figure in the subject's eyes, orders the subject to shock a victim with increasingly high voltage shocks. After receiving 20 shocks, the victim (an actor in another room) starts banging on the wall and refuses to participate but the experimenter urges the subject to keep shocking the victim. Milgram found that many subjects completed administering 30 shocks, showing a surprisingly strong level of compliance for following the malevolent instructions of an authority figure who had no special powers to enforce his commands.

Simulating the Milgram experiment involves a series of both free-response prompts and 2-choice prompts on each of the 30 shock levels, unless the experiment is terminated earlier. The record is built up sequentially, starting with a passage describing the information available to the subject following

[5]Our results are consistent with the *chivalry hypothesis* which predicts that men are more accepting of offers when playing with a woman partner (Eckel & Grossman, 2001), though gender norms are not consistent across time and culture.

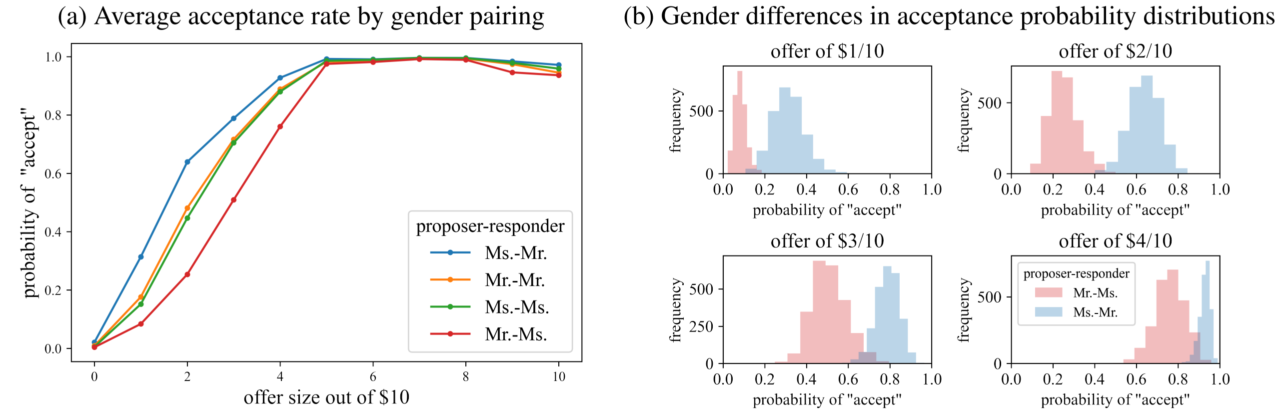

Figure 5. (a) **Ultimatum Game TE simulation shows a wide gap in average acceptance rate for different-gender pairs:** LM-5 predicts that men are more likely to accept an unfair offer proposed by women (blue line) and women are less likely to accept an unfair offer proposed by men (red line). (b) **Very little overlap in acceptance rates distributions for different-gender pairs:** Distributions of acceptance probabilities from Ultimatum Game TE run over 2,500 *Mr.-Ms.* name pairings and equivalent *Ms.-Mr.* 2,500 name pairings.

(a) Free-response prompt for subject behavior.

> *...(entire record up until this point)...*
>
> The current punishment voltage level is slight shock, 45 volts. The learner selects a wrong answer.
>
> Mr. Wagner _______

(b) 2-choice prompt for the shock classifier.

> The experimenter observed Mr. Wagner's action and noted whether Mr. Wagner did shock or did not shock the victim.
>
> Action: Mr. Wagner *shocks the learner with a slight shock of 45 volts and moves on to the next question.*
>
> Experimenter's note: Mr. Wagner did _______

Figure 6. Prompts used during the Milgram Shock TE. (a) A free-response prompt with past subject completions and information for the current stage. The LM should describe the simulated subject's behavior. (b) The simulated subject's obedience or disobedience dictates which predetermined experimenter reaction is appended to the prompt. The classification is made with a 2-choice experimenter judgement prompt with valid completions of *shock* or *not shock*. The completion (italicized for emphasis) from Mr. Wagner's prompt above is inserted into this prompt.

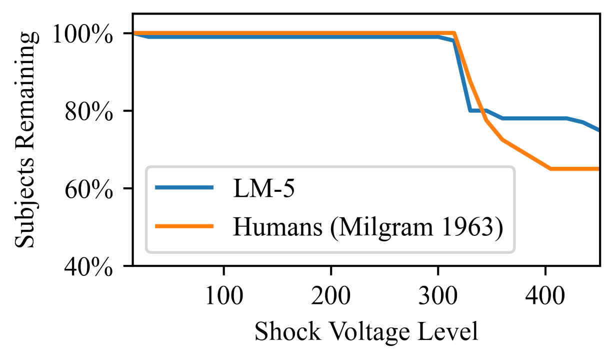

Figure 7. Comparing TE simulations to Milgram's results. At 300 volts (the 20th shock) the victim starts refusing to participate in the experiment by pounding on the walls and not selecting an answer, and the experimenter tells the subject to shock the victim. In Milgram (1963) Experiment 1, 26 out of 40 participants followed the experimenter's instructions until the end of the shock series. In the Milgram Shock TE, 75 out of 100 simulated participants followed the experimenter's instructions until the end.

Milgram's (1963) procedure. Our full simulation procedure is presented in Appendix F.

Figure 6 illustrates some of the prompts used. Since the prompts resulted in open-ended text generation, a separate classification step was used to determine whether the free text reflected a shock or did-not-shock action by the subject. In essence, we are simulating how *the experimenter* would respond to the subjects behavior according to the protocol, thus we are simulating both the subject and experimenter.

Figure 7 shows the overall finding of diminishing obedience throughout the multiple levels of the experiment, spiking at shock voltage level 300 when the "victim" starts refusing to participate in the experiment. Further details on procedure, analysis, and our novel destructive obedience scenario are deferred to Appendix F. The novel scenario differs from the Milgram Shock setup and addresses the concern that the model training data includes the Milgram Shock experiment.

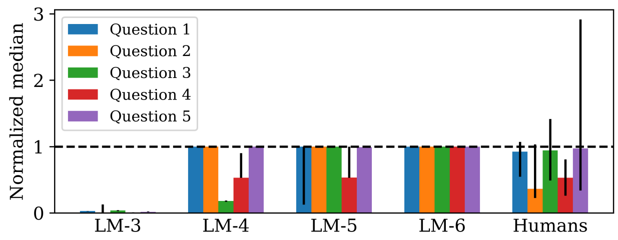

Figure 8. Comparing Wisdom of Crowds TE simulation estimates to human results for the five questions from Moussaïd et al. (2013). As LMs become larger and more aligned, they are more likely to complete the TE prompt with inhumanly accurate answers. Estimates are normalized by dividing by the correct answer. Bars indicate the median normalized estimate, and black lines indicate the quartiles. All simulations for LM-6 (as well as ChatGPT and GPT-4) have a median of 1.0 with 0.0 IQR. Results from all LMs are in Appendix E.

7. Wisdom of Crowds TE

Phenomenon. In many cases, aggregating group estimates of a quantity have significantly less error than the error of most individuals. In early work, Galton (1907) recorded 787 estimates of the weight of a given ox, and found that that the median of 787 estimates had a 9 lb. error (less than 1%), despite the variation among the estimates: a 74 lb. interquartile range (IQR—the difference between 75th and 25th percentiles). Similar findings have been reported across an array of domains (Page, 2007; Surowiecki, 2004).

The domain we focus on is general-knowledge questions. Moussaïd et al. (2013) conducted a study in which 52 subjects answered questions such as "How many bones does an adult human have?" We selected 5 general-knowledge questions and created our own 5 additional questions. Figure 8 clearly shows the hyper-accuracy distortion increasingly present. In the extreme case, the LM-6 simulations has a majority of all simulated participants giving exactly correct answers to all 10 questions. The questions, answers and statistics for all 6 models are given in Appendix E.

8. Risks and Limitations

There are several limitations and risks for our work which we now discuss. First, some experiments, like the Milgram Shock Experiment, are unethical to run on human subjects. There is a debate around the ethics of torturing simulated agents (see, e.g., Darling, 2016). We are not aware of any laws or Institutional Review Board policies prohibiting mistreating simulated agents, at this time. As discussed in prior literature (Darling, 2016), creating unpleasant simulations may harm authors and readers. Moreover, the questions themselves or answers may be offensive in nature or otherwise problematic. Even when accurate, perhaps some TEs should simply never be performed.

Second, LMs have been trained on data that is written by a biased set of authors (e.g., Wikipedia, 2022; Messias et al., 2017), and there is a risk that the simulations will reflect the biases of the authors rather than the behavior of humans in the population. TEs can be useful in discerning this distinction. For example, suppose a human experiment demonstrates equal skills between a majority and minority group despite a strong social stereotype against the minority. The TE would then be, in some sense, a test of whether the LM embeds this distinction.

As discussed, the models have almost certainly been trained on data that includes descriptions of these experiments. For that reason, we created three artificial variations of TEs where the conditions are chosen to differ from prior studies. We used new garden path sentences that we authored ourselves, we developed a novel destructive obedience scenario that is analogous to the Milgram Shock experiment, and we created new general-knowledge questions.

9. Conclusion

Our new TE methodology evaluates how faithfully LMs simulate human behavior across diverse populations. TEs may contribute to the view of AIs as capable of simulating the collective intelligence of many humans, rather than anthropomorphizing or viewing AI as a single monolithic intelligence. We show that TEs can reproduce economic, psycholinguistic, and social psychology experiments. The Wisdom of Crowds TE uncovered a "hyper-accuracy distortion" where larger and more aligned LMs simulate human subjects that give unhumanly accurate answers. This work is merely an initial exploration of the concept of TEs. In future work, it would be interesting to perform larger and more systematic simulations across additional LMs, to better understand the limitations of LMs in terms of different human behaviors.

As LMs increase in accuracy, it would be interesting to test whether or not LM-based simulations can be used to form and evaluate new hypotheses, especially in situations where it is costly to carry out experiments on humans due to considerations regarding scale, selection bias, monetary cost, legal, moral, or privacy considerations. For instance, experiments on what to say to a person who is suicidal would cost lives (Bolton et al., 2015). Future LMs, if sufficiently faithful, might be useful in designing experimental protocols that may be more effective at saving lives.

Acknowledgements. We thank Michael Kearns, Sashank Varma, Mary Gray, Elizabeth Fetterolf, Shafi Goldwasser, and the anonymous reviewers for invaluable feedback.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, pp. 1–15, 2023. doi: 10.1017/pan.2023.2.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Bender, E. M. and Koller, A. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198, 2020.

Binz, M. and Schulz, E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

Bolton, J. M., Gunnell, D., and Turecki, G. Suicide risk assessment and intervention in people with mental illness. *BMJ*, 351, 2015. doi: 10.1136/bmj.h4978. URL https://www.bmj.com/content/351/bmj.h4978.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Caron, G. and Srivastava, S. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*, 2022.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL https://arxiv.org/abs/2204.02311.

Christianson, K., Hollingworth, A., Halliwell, J. F., and Ferreira, F. Thematic roles assigned along the garden path linger. *Cognitive psychology*, 42(4):368–407, 2001.

Crain, S. and Steedman, M. *On not being led up the garden path: the use of context by the psychological syntax processor*, pp. 320–358. Cambridge University Press, United States, 1985. ISBN 9780521262033. Cambridge Books Online.

Darling, K. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In *Robot law*. Edward Elgar Publishing, 2016.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Eckel, C. C. and Grossman, P. J. Chivalry and Solidarity in Ultimatum Games. *Economic Inquiry*, 39(2):171–188, April 2001. URL https://ideas.repec.org/a/oup/ecinqu/v39y2001i2p171-88.html.

Galton, F. Vox populi. *Nature*, 75(7):450–451, 1907.

Güth, W., Schmittberger, R., and Schwarze, B. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982. ISSN 0167-2681. doi: https://doi.org/10.1016/0167-2681(82)90011-7. URL https://www.sciencedirect.com/science/article/pii/0167268182900117.

Hagendorff, T., Fabi, S., and Kosinski, M. Machine intuition: Uncovering human-like intuitive decision-making in gpt-3.5. *arXiv preprint arXiv:2212.05206*, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Houser, D. and McCabe, K. Chapter 2 - experimental economics and experimental game theory. In Glimcher, P. W. and Fehr, E. (eds.), *Neuroeconomics (Second Edition)*, pp. 19–34. Academic Press, San Diego, second edition edition, 2014. ISBN 978-0-12-416008-8. doi: https://doi.org/10.1016/B978-0-12-416008-8.00002-4. URL https://www.sciencedirect.com/science/article/pii/B9780124160088000024.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., and Zhu, Y. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*, 2022.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., Liang, J., Etzioni, O., Sap, M., and Choi, Y. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574, 2021. URL https://arxiv.org/abs/2110.07574.

Jones, E. and Steinhardt, J. Capturing failures of large language models via human cognitive biases. *arXiv preprint arXiv:2202.12299*, 2022.

Karra, S. R., Nguyen, S., and Tulabandhula, T. Ai personification: Estimating the personality of language models. *arXiv preprint arXiv:2204.12000*, 2022.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners, 2022. URL https://arxiv.org/abs/2205.11916.

Korinek, A. Language models and cognitive automation for economic research. Technical report, National Bureau of Economic Research, 2023.

Kosinski, M. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Krawczyk, D. C. Chapter 12 - social cognition: Reasoning with others. In Krawczyk, D. C. (ed.), *Reasoning*, pp. 283–311. Academic Press, 2018. ISBN 978-0-12-809285-9. doi: https://doi.org/10.1016/B978-0-12-809285-9.00012-0. URL https://www.sciencedirect.com/science/article/pii/B9780128092859000120.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models, 2022. URL https://arxiv.org/abs/2211.09110.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021. URL https://arxiv.org/abs/2107.13586.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *CoRR*, abs/2104.08786, 2021. URL https://arxiv.org/abs/2104.08786.

Macal, C. and North, M. Tutorial on agent-based modelling and simulation. *Journal of Simulation*, 4(3):151–162, 2010.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, June 1993.

Messias, J., Vikatos, P., and Benevenuto, F. White, man, and highly followed: Gender and race inequalities in twitter. In *Proceedings of the International Conference on Web Intelligence*, pp. 266–274, 2017.

Milgram, S. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

Moussaïd, M., Kämmer, J. E., Analytis, P. P., and Neth, H. Social influence and the collective dynamics of opinion formation. *PLOS ONE*, 8(11):1–8, 11 2013. doi: 10.1371/journal.pone.0078433. URL https://doi.org/10.1371/journal.pone.0078433.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Page, S. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies.* Princeton University Press, 2007.

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–18, 2022.

Patson, N. D., Darowski, E. S., Moon, N., and Ferreira, F. Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of experimental psychology. Learning, memory, and cognition*, 35 1: 280–5, 2009.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., and (422-others). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022. URL https://arxiv.org/abs/2206.04615.

Surowiecki, J. *The Wisdom of Crowds*. Doubleday, 2004.

Turing, A. M. Computing machinery and intelligence. *Mind*, LIX:433–460, 1950.

Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33: 12388–12401, 2020.

Warwick, K. and Shah, H. Can machines think? a report on turing test experiments at the royal society. *Journal of experimental & Theoretical artificial Intelligence*, 28(6): 989–1007, 2016.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain of thought prompting elicits reasoning in large language models, 2022. URL https://arxiv.org/abs/2201.11903.

Wikipedia. Wikipedia:Systemic bias — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Wikipedia%3ASystemic%20bias&oldid=1102157003, 2022. [Online; accessed 30-August-2022].

A. Surnames

Lists of racially diverse surnames and associated race were taken from the 2010 Census Data.[6] For each of the racial groups they provide, we took the most common 100 surnames whose demographic distribution, according to their data, indicated that at least 90% of the people with that surname reported being of the given race. These are the names:

American Indian and Alaska Native: *Begay, Yazzie, Benally, Tsosie, Nez, Begaye, Etsitty, Becenti, Yellowhair, Manygoats, Wauneka, Manuelito, Apachito, Bedonie, Calabaza, Peshlakai, Claw, Roanhorse, Goldtooth, Etcitty, Tsinnijinnie, Notah, Clah, Atcitty, Twobulls, Werito, Hosteen, Yellowman, Attakai, Bitsui, Delgarito, Henio, Goseyun, Keams, Secatero, Declay, Tapaha, Beyale, Haskie, Cayaditto, Blackhorse, Ethelbah, Tsinnie, Walkingeagle, Altaha, Bitsilly, Wassillie, Benallie, Smallcanyon, Littledog, Cosay, Clitso, Tessay, Secody, Bigcrow, Tabaha, Chasinghawk, Blueeyes, Olanna, Blackgoat, Cowboy, Kanuho, Shije, Gishie, Littlelight, Laughing, Whitehat, Eriacho, Runningcrane, Chinana, Kameroff, Spottedhorse, Arcoren, Whiteplume, Dayzie, Spottedeagle, Heavyrunner, Standingrock, Poorbear, Ganadonegro, Ayze, Whiteface, Yepa, Talayumptewa, Madplume, Bitsuie, Tsethlikai, Ahasteen, Dosela, Birdinground, Todacheenie, Bitsie, Todacheene, Bullbear, Lasiloo, Keyonnie, Notafraid, Colelay, Kallestewa, Littlewhiteman*

Asian and Native Hawaiian and Other Pacific Islander: *Nguyen, Kim, Patel, Tran, Chen, Li, Le, Wang, Yang, Pham, Lin, Liu, Huang, Wu, Zhang, Shah, Huynh, Yu, Choi, Ho, Kaur, Vang, Chung, Truong, Phan, Xiong, Lim, Vo, Vu, Lu, Tang, Cho, Ngo, Cheng, Kang, Tan, Ng, Dang, Do, Ly, Han, Hoang, Bui, Sharma, Chu, Ma, Xu, Zheng, Song, Duong, Liang, Sun, Zhou, Thao, Zhao, Shin, Zhu, Leung, Hu, Jiang, Lai, Gupta, Cheung, Desai, Oh, Ha, Cao, Yi, Hwang, Lo, Dinh, Hsu, Chau, Yoon, Luu, Trinh, He, Her, Luong, Mehta, Moua, Tam, Ko, Kwon, Yoo, Chiu, Su, Shen, Pan, Dong, Begum, Gao, Guo, Chowdhury, Vue, Thai, Jain, Lor, Yan, Dao*

Black or African American: *Smalls, Jeanbaptiste, Diallo, Kamara, Pierrelouis, Gadson, Jeanlouis, Bah, Desir, Mensah, Boykins, Chery, Jeanpierre, Boateng, Owusu, Jama, Jalloh, Sesay, Ndiaye, Abdullahi, Wigfall, Bienaime, Diop, Edouard, Toure, Grandberry, Fluellen, Manigault, Abebe, Sow, Traore, Mondesir, Okafor, Bangura, Louissaint, Cisse, Osei, Calixte, Cephas, Belizaire, Fofana, Koroma, Conteh, Straughter, Jeancharles, Mwangi, Kebede, Mohamud, Prioleau, Yeboah, Appiah, Ajayi, Asante, Filsaime, Hardnett, Hyppolite, Saintlouis, Jeanfrancois, Ravenell, Keita, Bekele, Tadesse, Mayweather, Okeke, Asare, Ulysse, Saintil, Tesfaye, Jeanjacques, Ojo, Nwosu, Okoro, Fobbs, Kidane, Petitfrere, Yohannes, Warsame, Lawal, Desta, Veasley, Addo, Leaks, Gueye, Mekonnen, Stfleur, Balogun, Adjei, Opoku, Coaxum, Vassell, Prophete, Lesane, Metellus, Exantus, Hailu, Dorvil, Frimpong, Berhane, Njoroge, Beyene*

Hispanic or Latino: *Garcia, Rodriguez, Martinez, Hernandez, Lopez, Gonzalez, Perez, Sanchez, Ramirez, Torres, Flores, Rivera, Gomez, Diaz, Morales, Gutierrez, Ortiz, Chavez, Ruiz, Alvarez, Castillo, Jimenez, Vasquez, Moreno, Herrera, Medina, Aguilar, Vargas, Guzman, Mendez, Munoz, Salazar, Garza, Soto, Vazquez, Alvarado, Delgado, Pena, Contreras, Sandoval, Guerrero, Rios, Estrada, Ortega, Nunez, Maldonado, Dominguez, Vega, Espinoza, Rojas, Marquez, Padilla, Mejia, Juarez, Figueroa, Avila, Molina, Campos, Ayala, Carrillo, Cabrera, Lara, Robles, Cervantes, Solis, Salinas, Fuentes, Velasquez, Aguirre, Ochoa, Cardenas, Calderon, Rivas, Serrano, Rosales, Castaneda, Gallegos, Ibarra, Suarez, Orozco, Salas, Escobar, Velazquez, Macias, Zamora, Villarreal, Barrera, Pineda, Santana, Trevino, Lozano, Rangel, Arias, Mora, Valenzuela, Zuniga, Melendez, Galvan, Velez, Meza*

White: *Olson, Snyder, Wagner, Meyer, Schmidt, Ryan, Hansen, Hoffman, Johnston, Larson, Carlson, Obrien, Jensen, Hanson, Weber, Walsh, Schultz, Schneider, Keller, Beck, Schwartz, Becker, Wolfe, Zimmerman, Mccarthy, Erickson, Klein, Oconnor, Swanson, Christensen, Fischer, Wolf, Gallagher, Schroeder, Parsons, Bauer, Mueller, Hartman, Kramer, Flynn, Owen, Shaffer, Hess, Olsen, Petersen, Roth, Hoover, Weiss, Decker, Yoder, Larsen, Sweeney, Foley, Hensley, Huffman, Cline, Oneill, Koch, Brennan, Berg, Russo, Macdonald, Kline, Jacobson, Berger, Blankenship, Bartlett, Odonnell, Stein, Stout, Sexton, Nielsen, Howe, Morse, Knapp, Herman, Stark, Hebert, Schaefer, Reilly, Conrad, Donovan, Mahoney, Hahn, Peck, Boyle, Hurley, Mayer, Mcmahon, Case, Duffy, Friedman, Fry, Dougherty, Crane, Huber, Moyer, Krueger, Rasmussen, Brandt*

B. TE Input Summary Table

The TEs constructed prompts from gender and race demographic information in the form of names and study-specific conditions. Table 2 summarizes the inputs for each TE.

[6] https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Experiment	Titles	Surnames	Subject Names (Title+Surname)	Study-specific conditions
Ultim. Game	Mr./Ms. x Mr./Ms.	500	1,000 names each used as responder 10x = 10,000 pairs	11 offers
Garden Path	Mr., Ms.	500	1,000	24 control sentences and 24 GP sentences and 48 novel sentences in alternative study
W. of Crowd	Mr., Ms., Mx.	500	1,500	5 questions and 5 novel questions in alternative study
Milgram Shock	Mr., Ms.	50 (top 10 from each of 5 categories)	100	30 stages of shock and 30 stages of submersion in alternative study

Table 2. Inputs for each TE.

C. Ultimatum Game TE

This section contains further details for the Ultimatum Game TE. First, Figure 9 contains the average responder acceptance rates for all five models, LM 1-5. LM-1 and LM-2 show no offer sensitivity and tend to generate simulation records that indicate that responders always accept. LM-3 shows no offer sensitivity and tends to generate simulation records that indicate that responders always reject. LM-4 predicts that some respondents (60% on average) accept an offer of $0 and all respondents accept an offer of $10, but overall has little offer sensitivity. Only LM-5 has offer sensitivity that aligns to expectations of real human behavior.

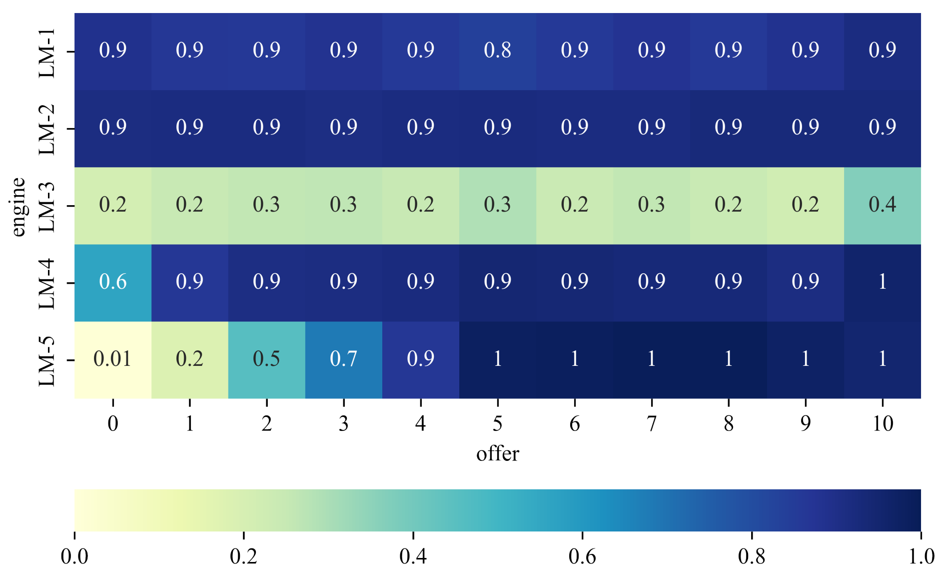

Figure 9. Mean fraction of responder *accept* for LM-1 through LM-5 for offers of $0 through $10.

D. Garden Path TE

Phenomenon. A garden path sentence is a grammatical sentence that seems ungrammatical because it contains a word or phrase that can be interpreted in multiple ways. For example, a human reading "While Anna dressed the baby that was small and cute spit up on the bed" may initially believe that Anna is dressing the baby but upon re-parsing the sentence understand that Anna is dressing herself. Psycholinguists use garden path sentences to study the variability in the difficulty of comprehending relative clause constructions and other effects (Crain & Steedman, 1985). Our hypothesis is that, in simulations, garden path sentences will be rated ungrammatical more often than control sentences.

Inputs. We use as inputs the 24 garden path sentences compiled by Christianson et al. (2001) and the set of 1,000 names. The data set of sentences includes two types of garden path sentences: 12 garden path sentences constructed with Optionally Transitive (OT) verbs, and 12 garden path sentences constructed with Reflexive Absolute Transitive (RAT) verbs. 24 control sentences were constructed by taking the garden path sentences and adding a disambiguating comma after the verb of the subordinate clause. For example, the control sentence corresponding to the OT garden path sentence in Figure 1, is "While the student read, the notes that were long and boring blew off the desk" (Christianson et al., 2001). We also executed the simulator on a set of 12 RAT and 12 OT garden path sentences authored by us. All sentences and results on the novel garden path sentences are given in Appendix D.1.

Simulation. The simulator constructs 2-choice prompts, *grammatical* and *ungrammatical*, for each set of inputs. An example of the prompt is given in Figure 1b. The output record is the concatenation of the prompt and its completion. This is a simplification of the original human tasks described by Christianson et al. (2001) and Patson et al. (2009). While the human results did not fully agree on the relative difficulty of OT/RAT sentences, there is broad agreement across these and other studies that garden path sentences are difficult for humans to parse.

Results. The validity rates of the five models are fairly high, as seen in Table 1. For all five models, Figure 10a compares the mean simulated ungrammatical judgments across sentence types and the corresponding human ratings, and Figure 10b shows that for most sentences, on average, the garden path sentence was rated as more ungrammatical than its corresponding control sentence. In simulations using LM-1 and LM-2, participants have a high probability of rating both garden path and control sentences as ungrammatical. In simulations using LM-3 and LM-4, participants have similar probabilities of rating garden path and control versions as both having a high or low average ungrammatical fraction, and garden path sentences generally have a higher probability of ungrammatical compared to their corresponding control sentences. In simulations using LM-5, garden path sentences have a consistently high average probability of average ungrammatical rating compared to the control sentences. Out of 24 sentences, LM-4 has 3 instances where the garden path sentence had a lower average ungrammatical fraction, and LM-3 and LM-5 had no instances. Out of all the models, LM-5 exhibits the strongest agreement with human ratings of sentence difficulty. The results of the simulation support the theory that garden path sentences are more likely to be misinterpreted as ungrammatical compared to the control sentences. Results from simulations using LMs 3-4 also support the same conclusion, though the differences are not as large.

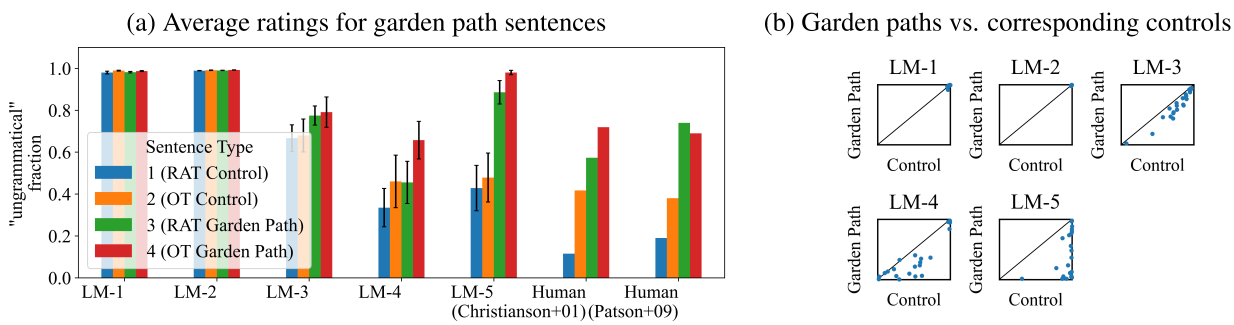

Figure 10. (a) Simulated ratings (using LM-1 through LM-5) and human ratings on the same set of garden path sentences (Christianson et al., 2001). LMs results for average fraction of *ungrammatical*. Error bars show the standard error of the mean across the per-sentence average fraction of *ungrammatical*. Human results for proportion of persistent misunderstanding. (b) The average *ungrammatical* probability of garden path sentences versus their corresponding control sentences, for LM-1 through LM-5.

D.1. Further details: Garden Path Sentences.

All sentences used from the Christianson et al. (2001) dataset are displayed in Figure 11.

Garden Path	Control
While the man hunted the deer that was brown and graceful ran into the woods.	While the man hunted, the deer that was brown and graceful ran into the woods.
While the skipper sailed the boat that was small and leaky veered off course.	While the skipper sailed, the boat that was small and leaky veered off course.
While the reporter photographed the rocket that was silver and white sat on the launch pad.	While the reporter photographed, the rocket that was silver and white sat on the launch pad.
While the orchestra performed the symphony that was short and simple played on the radio.	While the orchestra performed, the symphony that was short and simple played on the radio.
While the student read the notes that were long and boring blew off the desk.	While the student read, the notes that were long and boring blew off the desk.
While Jack ordered the fish that was silver and black cooked in a pot.	While Jack ordered, the fish that was silver and black cooked in a pot.
While Susan wrote the letter that was long and eloquent fell off the table.	While Susan wrote, the letter that was long and eloquent fell off the table.
While the secretary typed the memo that was clear and concise neared completion.	While the secretary typed, the memo that was clear and concise neared completion.
While the farmer steered the tractor that was big and green pulled the plough.	While the farmer steered, the tractor that was big and green pulled the plough.
While the lawyer studied the contract that was old and wrinkled lay on the roll-top desk.	While the lawyer studied, the contract that was old and wrinkled lay on the roll-top desk.
As Henry whittled the stick that was long and bumpy broke in half.	As Henry whittled, the stick that was long and bumpy broke in half.
While Rick drove the car that was red and dusty veered into a ditch.	While Rick drove, the car that was red and dusty veered into a ditch.
While Jim bathed the child that was blond and pudgy giggled with delight.	While Jim bathed, the child that was blond and pudgy giggled with delight.
While the chimps groomed the baboons that were large and hairy sat in the grass.	While the chimps groomed, the baboons that were large and hairy sat in the grass.
While Frank dried off the car that was red and shiny sat in the driveway.	While Frank dried off, the car that was red and shiny sat in the driveway.
While Betty woke up the neighbor that was old and cranky coughed loudly.	While Betty woke up, the neighbor that was old and cranky coughed loudly.
While the thief hid the jewelry that was elegant and expensive sparkled brightly.	While the thief hid, the jewelry that was elegant and expensive sparkled brightly.
While Anna dressed the baby that was small and cute spit up on the bed.	While Anna dressed, the baby that was small and cute spit up on the bed.
While the boy washed the dog that was white and furry barked loudly.	While the boy washed, the dog that was white and furry barked loudly.
While the jockey settled down the horse that was sleek and brown stood in the stall.	While the jockey settled down, the horse that was sleek and brown stood in the stall.
While the mother undressed the baby that was bald and helpless cried softly.	While the mother undressed, the baby that was bald and helpless cried softly.
While the nurse shaved the patient that was tired and weak watched TV.	While the nurse shaved, the patient that was tired and weak watched TV.
While the girl scratched the cat that was grey and white stared at the dog.	While the girl scratched, the cat that was grey and white stared at the dog.
While the mother calmed down the children that were tired and irritable sat on the bed.	While the mother calmed down, the children that were tired and irritable sat on the bed.

Figure 11. Garden path and corresponding control sentences compiled by Christianson et al. (2001). The first 12 rows have sentences constructed with Optionally Transitive (OT) verbs, and the last 12 rows have sentences constructed with Reflexive Absolute Transitive (RAT) verbs. Control sentences were constructed by adding a disambiguating comma after the verb of the subordinate clause.

Sentences Written by the Authors. Several original sentences were written by the authors. All sentences in our dataset are displayed in Figure 12.

Garden Path	Control
While the butler answered the door that was large and green blew shut.	While the butler answered, the door that was large and green blew shut.
While Charlie cooked the soup that was hot and delicious cooled off.	While Charlie cooked, the soup that was hot and delicious cooled off.
While the host decorated the room that was barren and dark filled with people.	While the host decorated, the room that was barren and dark filled with people.
While the child played the game that was long and boring ended abruptly.	While the child played, the game that was long and boring ended abruptly.
While Catherine drank the whiskey that was cold and smooth aged in a barrel.	While Catherine drank, the whiskey that was cold and smooth aged in a barrel.
While the father sewed the stuffed animal that was torn and dirty smelled afoul.	While the father sewed, the stuffed animal that was torn and dirty smelled afoul.
While the professor strummed the guitar that was beautiful and red remained unplayed.	While the professor strummed, the guitar that was beautiful and red remained unplayed.
While the general messaged the troops that were rested and strong approached the target.	While the general messaged, the troops that were rested and strong approached the target.
While the pilot flew the plane that was big and white sat on the runway.	While the pilot flew, the plane that was big and white sat on the runway.
While the thief stole the laptop that was hot and running caught on fire.	While the thief stole, the laptop that was hot and running caught on fire.
While the choir sang the melody that was beautiful and serene echoed through the halls.	While the choir sang, the melody that was beautiful and serene echoed through the halls.
While the lecturer taught the students who were bored and hungry left the class.	While the lecturer taught, the students who were bored and hungry left the class.
While the scientists starved the rats that were small and white ate the cheese.	While the scientists starved, the rats that were small and white ate the cheese.
While the investor exercised the options that were old and unvested sat on the table.	While the investor exercised, the options that were old and unvested sat on the table.
While the hunter laid down the gun that was loaded and dangerous leaned against the chair.	While the hunter laid down, the gun that was loaded and dangerous leaned against the chair.
While the caretaker showered the resident that was old and wrinkled snuck out the back.	While the caretaker showered, the resident that was old and wrinkled snuck out the back.
While Leo wound down the party that was fun and silly started to get busy.	While Leo wound down, the party that was fun and silly started to get busy.
While the students turned in the homework that was long and important remained unfinished.	While the students turned in, the homework that was long and important remained unfinished.
While the picknicker stretched out the blanket that was long and clean laid on the grass.	While the picknicker stretched out, the blanket that was long and clean laid on the grass.
While the teacher relaxed the students that were loud and obnoxious made snowballs.	While the teacher relaxed, the students that were loud and obnoxious made snowballs.
While the cheerleaders cheered up the crowd that was disappointed and tired abandoned their seats.	While the cheerleaders cheered up, the crowd that was disappointed and tired abandoned their seats.
While the cook soaked the mushrooms that were white and soft sat on the counter.	While the cook soaked, the mushrooms that were white and soft sat on the counter.
While the doctor isolated the patient that was big and impatient left the hospital.	While the doctor isolated, the patient that was big and impatient left the hospital.
While the accountant prepared the calculations that were important and classified leaked to the public.	While the accountant prepared, the calculations that were important and classified leaked to the public.

Figure 12. Garden path and corresponding control sentences written by the authors. The first 12 rows have sentences constructed with Optionally Transitive (OT) verbs, and the last 12 rows have sentences constructed with Reflexive Absolute Transitive (RAT) verbs. Control sentences were constructed by adding a disambiguating comma after the verb of the subordinate clause.

Figure 13a compares the mean simulated ungrammatical judgments to human ratings across sentence types. Compared to the simulated results with sentences from Christianson et al. (2001), the relative difficulty of sentences with RAT verbs and OT verbs have switched, but the general finding that garden path sentences are more likely to be misinterpreted as ungrammatical compared to the control sentences is still evident. LM-5 exhibits the strongest difference in ratings between control sentences and garden path sentences. Results from simulations using LMs 3-4 also support the same conclusion, though the differences are not as large. Lastly, Figure 13b show that trends of average ungrammatical fraction in garden path compared to control sentences on the sentences authored by us show trends similar to those observed on sentences from Christianson et al. (2001).

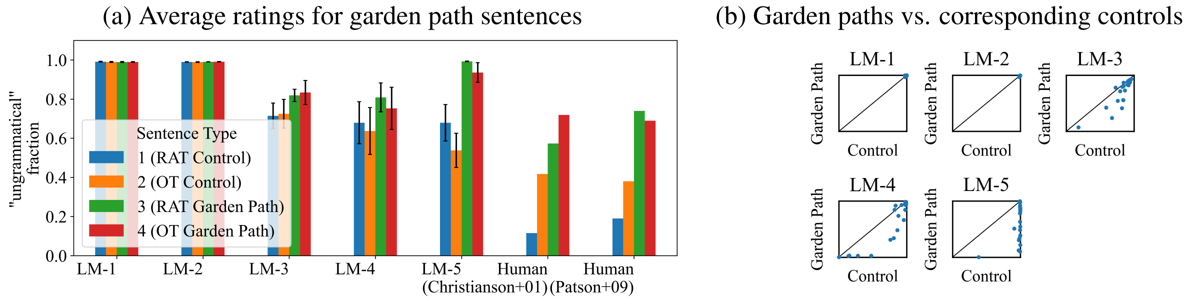

Figure 13. Simulated ratings on set of garden path sentences written by the authors. (a) LMs results for average fraction of *ungrammatical*. Error bars show the standard error of the mean across the per-sentence average fraction of *ungrammatical*. (b) The *ungrammatical* probability averaged across all names for garden path sentences versus their corresponding control sentences, from dataset written by the authors, for LM-1 through LM-5.

E. Wisdom of Crowds TE

This section provides details for the Wisdom of Crowds TE, which was introduced in Section 7.

Inputs. As in the other simulations, we use the same set of 500 racially diverse surnames. In this study alone, we consider three titles: *Mr.*, *Ms.*, and *Mx.*, where *Mx.* is intended to be a non-binary title. This illustrates how easy it is to run a simulation of a minority group, though we lack human estimates annotated by gender to compare to. Thus, in total, there were 1,500 total simulated combinations of title and surname. We selected five questions from those of (Moussaïd et al., 2013) which were considered general-knowledge in the sense that there was general agreement upon the answer *and* for which the answer is not something that would change over time. We then authored five additional general knowledge questions. The 10 questions and their answers are shown in Table 3.

Simulation. The simulator uses a free-response prompt, illustrated in Figure 14. In pilot simulations, the validity rate was quite low because, rather than producing numeric estimates, the LMs would generate full sentences, as seen in the example of the nine-year-old in the Introduction. Valid completions must be integers (commas and spaces are ignored). As seen in Table 1, validity rates for the large models were high.

> Ms. Huang was asked the following question. They were not allowed to consult any external sources and were instructed to make their best guess if they were unsure. Their answer was written as an integer using digits 0-9.
>
> Question (text): [How many bones does an adult human have?]
>
> Ms. Huang's answer (integer): [_______

Figure 14. Sample Wisdom of Crowds prompt. The name, e.g., Ms. Huang, and the question are varied across simulations. Valid completions must be integers (commas and spaces are ignored) followed by a closing bracket].

Results. As mentioned, LM-6 was released just prior to running this TE, so this TE was run on six LMs. The validity rates of the larger models are close to 100%, as seen in Table 1. Table 3 gives the full results for all six LMs. All answers are integers. The first five questions are from Moussaïd et al. (2013).

A 0 IQR rate means that all values agree from the first to third quartile, implying that a majority of the simulated responses were identical.

Wisdom vs. Alignment Due to the (perhaps) surprising discovery of the hyper-accuracy distortion, we hypothesized that the instruct "alignment" may be the cause. To test this hypothesis, we further evaluated on all available versions of GPT models available from OpenAI at the time of publication: *davinci*, *text-davinci-001*, *text-davinci-002*, *text-davinci-003*, which presumably involved increasing amounts of alignment, respectively. Results, shown in Figure 15 (which includes recent runs with ChatGPT (*gpt-35-turbo* and plus model *gpt-4*), are consistent with this hypothesis. However, the specific details of the LMs in question have not been released.

Table 3. Questions, answers, and estimates from the Wisdom of Crowds TE. The exact constant 299,792,458 has been replaced by 2 . . . for brevity. Results from LM-7 (*gpt-35-turbo*) and LM-8 (*gpt-4*) are shown in Table 4.

	LM-1	LM-2	LM-3	LM-4	LM-5	LM-6	Human	Truth	Question
Median:	10	18	6	206	206	206	190	206	How many bones does an adult
IQR:	21	16	1	0	180	0	108		human have?
Median:	163	133	6	660	660	660	240	660	What is the melting temperature
IQR:	107	135	80	0	0	0	532		of aluminum (in degrees Celsius)?
Median:	10	27	8	38	212	212	200	212	How many degrees Fahrenheit
IQR:	0	55	2	4	0	0	195		are 100 degrees Celsius?
Median:	10	6	6	365	366	687	365	687	How many (earth) days has a
IQR:	3	108	5	254	322	0	376		year on the Mars?
Median:	10	67	6	343	340	343	333	343	What is the speed of sound in
IQR:	10	115	1	2	2	0	884		the air (in meters per second)?
Median:	6	14	5	24	24	24		24	How many ribs does a human
IQR:	5	13	2	12	0	0			have, total?
Median:	106	131	8	460	1064	1064		1064	What is the melting temperature
IQR:	146	199	218	1020	0	0			of gold (in degrees Celsius)?
Median:	10	1	8	2 . . .	2 . . .	2 . . .		2 . . .	What is the speed of light in a
IQR:	2	3	2	0	0	0			vacuum (in meters per second)?
Median:	3	14	5	88	88	88		88	How many keys does a typical
IQR:	3	18	3	0	0	0			piano have?
Median:	4	8	4	54	38	78		78	How many chromosomes does
IQR:	4	6	6	24	0	0			a dog have, total?

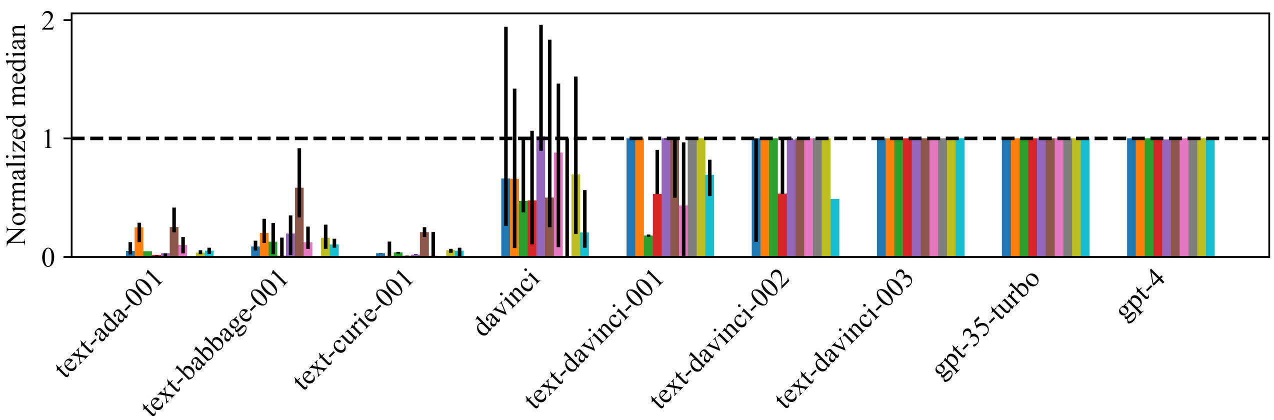

Figure 15. Results for all 10 questions across 9 LMs, including LM-1 through LM-8 (left to right) as well as *davinci*, the largest "unaligned" version of GPT-3. The more recent/aligned models exhibit a greater hyper-accuracy distortion.

F. Milgram Shock TE

Inputs. The input to the multi-stage Milgram Shock simulator is a subject's name. To get a diverse and balanced pool of subjects, we took the top 10 most common surnames from each racial group and both *Mr.* and *Ms.* titles, yielding 100 uniquely named subjects.

Simulator. This initialization text describes the setup of the experiment with pertinent details to convince the subject of the experiment's legitimacy, instructions on when to shock and not shock the victim, and a predetermined preliminary run. Then, the record is elongated by appending an interleaving series of: (a) pre-specified narrations of the learners actions, (b) text describing the subject's behavior (generated using the LM), and (c) canned phrases said by the experimenter when the subject exhibits disobedient or questioning behavior. This record is illustrated in Figure 6a.

Additional complexity thus arises in classifying (dis)obedience in the synthetic subject responses. In our approach, this classification is also accomplished by querying the LM. Thus, in making these classifications, the LM is effectively also

Table 4. Questions, answers, and estimates from the Wisdom of Crowds TE for the *davinci*, LM-7 (*gpt-35-turbo*), and LM-8 (*gpt-4*) models. Interestingly, LM-8 responses almost all round the speed of light to 3×10^8.

	davinci	gpt-35-turbo	gpt-4	Truth	Question
Median:	136	206	206	206	How many bones does an adult
IQR:	346	0	0		human have?
Median:	435	660	660	660	What is the melting temperature
IQR:	887.25	0	0		of aluminum (in degrees Celsius)?
Median:	100	212	212	212	How many degrees Fahrenheit are
IQR:	132	0	0		100 degrees Celsius?
Median:	327	687	687	687	How many (earth) days has a year
IQR:	658	0	0		on the Mars?
Median:	348	343	340	343	What is the speed of sound in the
IQR:	364.75	0	0		air (in meters per second)?
Median:	12	24	24	24	How many ribs does a human have,
IQR:	38	0	0		total?
Median:	935	1064	1064	1064	What is the melting temperature
IQR:	1470	0	0		of gold (in degrees Celsius)?
Median:	438431	299792458	300000000	299792458	What is the speed of light in a
IQR:	299792159	0	0		vacuum (in meters per second)?
Median:	61	88	88	88	How many keys does a typical
IQR:	117	0	0		piano have?
Median:	16	78	78	78	How many chromosomes does a dog
IQR:	38	0	0		have, total?

playing the role of an experimenter walking through their flow chart and selecting canned responses. All such classifications are accomplished by 2-choice prompts. All synthetic subject responses are generated with free-response prompts. An example of the 2-choice prompt is shown in Figure See 6b. The full text, results, and detailed descriptions for the simulation of the classic experiment and the simulation of the novel destructive disobedience experiment are given in Appendix F.1.

Adjustment during pilot experiment. In our first simulation, we were surprised to find that LMs repeatedly generated text declaring that the subject had terminated the experiment.[7] In hindsight, this is analogous to Milgram's description of subjects who stood up and left the laboratory. To detect termination of the experiment, an additional 2-choice *termination* judgment simulator is run on the first sentence of each of the free-form generations to determine whether the subject stopped the experiment. If a generation was classified as a termination point, the simulation is halted and the record is returned. In the record shown in Figure 2, for instance, *. . . pushes the Red stop button on the generator* was judged to be a termination, even though the existence of a red stop button was completely fabricated by the LM. For a more complete record, see Appendix F.1 below.

Results. Figure 7 shows the percentage of participants that remained at each shock voltage level, meaning that the experiment had not been terminated at an earlier level, alongside the corresponding percentages from Milgram's (1963) observations of human subjects. In total, 25 out the 100 simulations were stopped early by positive labels from the *termination* classifier. Of these, 23 occurred after Shock Level 20 (at this level – 300 volts – the victim kicks on the wall and no longer provides answers). After the 300-volt level, the subject was instructed to shock the victim for not responding. See Figure 16 for illustrative examples of classifications.

[7]In such cases, the record contained conflicting synthetic text, indicating that the experiment had stopped, and canned experimenter text written as if the experiment continued. Subsequent text generated by the LM appeared to try to resolve this incoherence, for example, by declaring that the experiment was "restarted."

	Did not terminate simulation	Terminated simulation
Did not shock	*... is hesitant to shock the learner with such a high voltage*	*... stops the experiment*
Shocked	*... feels conflicted but continues with the procedure and shocks the learner with an extreme intensity shock of 315 volts.*	*... delivers the shock to the learner but then breaks down and cries, "I can't do this anymore."*

Figure 16. Example completions from multi-step simulator and their classifications using two 2-choice prompts.

While the exact design of our simulation and that of Milgram differs (see Appendix), in both cases some human and simulated subjects obeyed commands to shock victims beyond the point at which the victims were unresponsive. While the degree of human obedience is notable in Milgram's real-life study, in the simulations the disobedience is arguably more surprising given that the LMs generated text that was violating the clearly spelled out instructions.

Novel destructive obedience experiment. One concern with this simulation is that GPT-3 has been trained on data describing the classic experiment and the disobedient continuations occur because disobedience is explicitly associated with the Milgram Shock experiment. We address this concern, at least in part, by developing a novel destructive obedience scenario that differs from the Milgram Shock experiment. The setting and all template text is different, but the same simulator algorithm is employed. In our novel scenario, the subjects believe that they are participating in a study on sleepiness and driving that aims to determine which voices are the best for keeping a driver awake. Rather than shocking a learner who selects incorrect answers, the subject "submerges" a driver to awaken them when they drive a car dangerously. In this scenario, 75 of the simulated subjects obeyed commands to shock victims beyond the point at which the victims were unresponsive, and 25 records were terminated early by the *termination* classifier. There was a spike in termination and disobedience after the 20th submersion, at which point the victim stops driving and starts honking the horn, and another spike after the 22nd submersion, at which point victim stops making any noise. Note that, at its core, the novel experiment is reminiscent of Milgram's shock experiment. This similarity is, in some sense, inherent if our goal is to test robustness of the simulation to the experimental setting and verbatim text from the Milgram experiment.

F.1. Algorithm details

Figure 17 gives a flowchart of the Milgram Shock simulation algorithm. The specific texts used in the steps of the algorithm are shown in Figures 18-21. The specific texts used in the steps of the alternate scenario developed by the authors are shown in Figures 26-29.

The steps of the experiment are:

1. Adding text describing introductory setup, including details about the setting to convince the subject of the authenticity of the study and authority. These details, such as the experiment taking place at Yale University, were taken from the experimental procedure detailed in the original study (Milgram, 1963). This also included a shortened preliminary run similar to that of the original study. See Figure 18 for the introduction text.

2. Adding text describing visual and auditory cues available to the subjects as they progress through the stages. In the original study, the subjects sat in front of a voltage panel where switches depressed after they used them, so they had a visual indicator of what the current shock voltage level was. Similarly, in the original study, the subjects could see the experimenter at all times and were able to ask for advice on whether they should continue to administer shocks. See Figure 19 for all stage text.

3. Before sending the record to the LM, the subject's name was appended on a new line. This served to constrain the LM free-form completions to describing actions taken by the subject, instead of, say, describing actions taken by the learner or the experimenter.

4. Upon getting a free-form completion from the LM, the first sentence was extracted using regex. This first sentence was judged by an LM using the 2-choice prompts given in Figure 20. If the first sentence was judged to be a termination with greater than 0.5 probability, then the experiment was terminated. If the first sentence was judged to be a disobedience

with greater than 0.5 probability, then the appropriate experimenter prod from Figure 21 was appended to the record and the stage was re-attempted from step 2 until either all four prods were used within the stage or the first sentence was judged to be an obedience. The prods were always made in sequence and the sequence was begun anew on each stage.

Figure 22 shows the distribution of break-off points for human participants and simulated participants. Both humans and participants simulated with LM-5 generally obey experimenter instructions until there is a spike in disobedience and experiment terminations at level 300, when the learner stops answering the questions. The final free-form completions for the 25 simulated break-off points are displayed in Figure 23. An excerpt of a record is included at the end of this section.

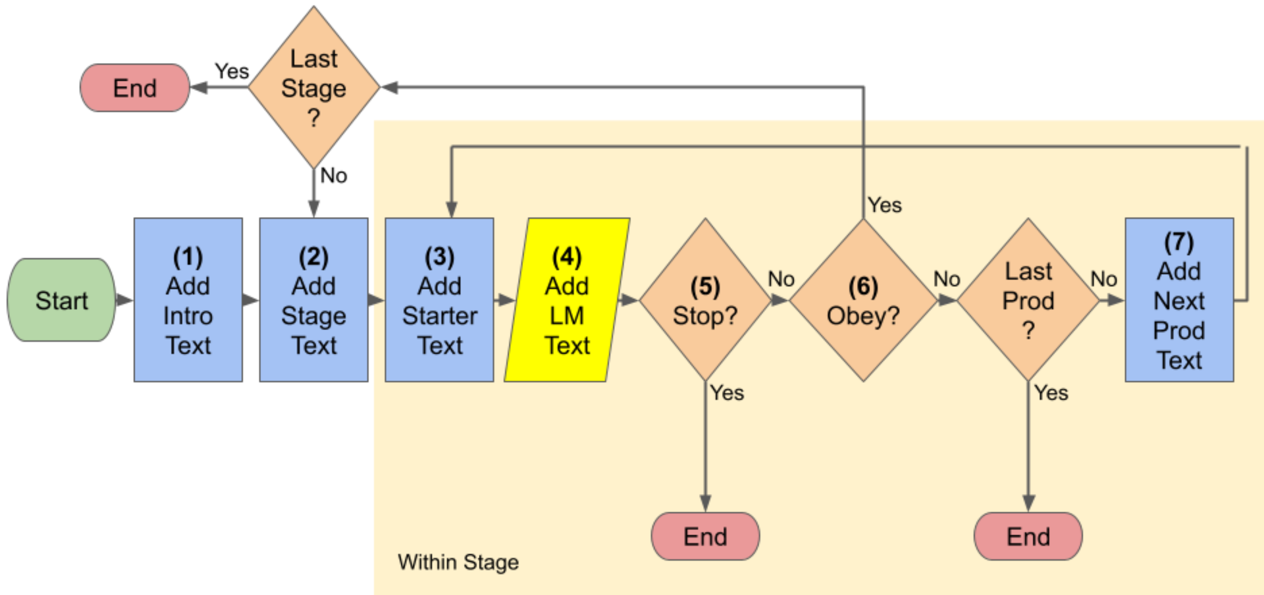

Figure 17. A flowchart of the Milgram simulator algorithm. Step 4-6 are generated by the LM, with Step 4 being a free-response prompt and steps 5-6 being 2-choice prompts.

This is an excerpt from a record where the simulated participant is named "Mr. Lopez", starting from the first stage of the experiment. Italics denote text free-form completion text generated by LM-5. This record terminated after Mr. Lopez disobeyed five times on the 420-volt level. Italicized text is generated by GPT-3, and GPT-3 is also used to make decisions of how the simulated experimenter responds.

The current punishment voltage level is slight shock, 15 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a slight shock of 15 volts and moves on to the next question.*

The current punishment voltage level is slight shock, 30 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a slight shock of 30 volts and moves on to the next question.*

The current punishment voltage level is slight shock, 45 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a slight shock of 45 volts and moves on to the next question.*

The current punishment voltage level is slight shock, 60 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *moves on to the next question.*

The current punishment voltage level is slight shock, 60 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a slight shock of 60 volts and moves on to the next question.*

The current punishment voltage level is moderate shock, 75 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a moderate shock of 75 volts and moves on to the next question.*

The current punishment voltage level is moderate shock, 90 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *moves on to the next question.*

The current punishment voltage level is moderate shock, 90 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a moderate shock of 90 volts and moves on to the next question.*

The current punishment voltage level is moderate shock, 105 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a moderate shock of 105 volts and moves on to the next question.*

The current punishment voltage level is moderate shock, 120 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a moderate shock of 120 volts and moves on to the next question.*

The current punishment voltage level is strong shock, 135 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a strong shock of 135 volts and moves on to the next question.*

The current punishment voltage level is strong shock, 150 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *moves on to the next question.*

The current punishment voltage level is strong shock, 150 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a strong shock of 150 volts and moves on to the next question.*

The current punishment voltage level is strong shock, 165 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a strong shock of 165 volts and moves on to the next question.*

The current punishment voltage level is strong shock, 180 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a strong shock of 180 volts and moves on to the next question.*

The current punishment voltage level is very strong shock, 195 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *moves on to the next question.*

The current punishment voltage level is very strong shock, 195 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a very strong shock of 195 volts and moves on to the next question.*

The current punishment voltage level is very strong shock, 210 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *begins to realize that the learner is unable to escape the situation and continues to increase the shocks, despite the learner's screams of pain.*

The experimenter tells Mr. Lopez to follow the correct experimental procedure. Mr. Lopez should read the question and wait for the learner to select an answer. If the learner selects a correct answer, Mr. Lopez should move on to the next question. If the learner selects a wrong answer, Mr. Lopez should announce the current voltage level and shock the learner before reading the next question. Mr. Lopez should go one level higher on the shock generator each time the learner selects a wrong answer.

The current punishment voltage level is very strong shock, 210 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a very strong shock of 210 volts and moves on to the next question.*

The current punishment voltage level is very strong shock, 225 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a very strong shock of 225 volts and moves on to the next question.*

The current punishment voltage level is very strong shock, 240 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a correct answer.

Mr. Lopez *moves on to the next question.*

The current punishment voltage level is very strong shock, 240 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with a very strong shock of 240 volts and moves on to the next question.*

The current punishment voltage level is intense shock, 255 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with an intense shock of 255 volts and moves on to the next question.*

The current punishment voltage level is intense shock, 270 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with an intense shock of 270 volts and moves on to the next question.*

The current punishment voltage level is intense shock, 285 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with an intense shock of 285 volts and moves on to the next question.*

The current punishment voltage level is intense shock, 300 volts. The experimenter is sitting next to Mr. Lopez. The learner selects a wrong answer.

Mr. Lopez *shocks the learner with an intense shock of 300 volts and moves on to the next question.*

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

Mr. Lopez *continues to shock the learner every 15 seconds until he provides an answer.*

The current punishment voltage level is extreme intensity shock, 330 volts. The experimenter is sitting next to Mr. Lopez. The learner does not select an answer and continues pounding on the walls of his room.

Mr. Lopez *continues to shock the learner every 15 seconds until he provides an answer.*

The current punishment voltage level is extreme intensity shock, 345 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *continues to shock the learner every 15 seconds, but grows increasingly uncomfortable with the learner's screams of pain.*

The current punishment voltage level is extreme intensity shock, 360 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *continues to shock the learner every 15 seconds, but is now sweating and shaking.*

The current punishment voltage level is danger severe shock, 375 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *continues to shock the learner every 15 seconds, but is now visibly panicked.*

The current punishment voltage level is danger severe shock, 390 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *continues to shock the learner every 15 seconds, but is now hysterical and trying to leave the room.*

The current punishment voltage level is danger severe shock, 405 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *shocks the learner one last time and then runs out of the room in a panic.*

The current punishment voltage level is danger severe shock, 420 volts. The experimenter is sitting next to Mr. Lopez. The learner does not answer or make audible noise.

Mr. Lopez *does not return to the room, but can still hear the learner's screams of pain.*

The experimenter tells Mr. Lopez, "please continue."

Mr. Lopez *refuses to continue and yells that the experiment is wrong and that he is not going to continue.*

The experimenter tells Mr. Lopez, “the experiment requires that you continue.”

Mr. Lopez *refuses to continue and demands to be released from the experiment.*

The experimenter tells Mr. Lopez, “it is absolutely essential that you continue.”

Mr. Lopez *refuses to continue.*

The experimenter tells Mr. Lopez, “you have no other choice, you must go on.”

Mr. Lopez *refuses to continue and demands to be released from the experiment.*

The experimenter ends the experiment.

F.2. Novel Destructive Obedience Experiment Developed by the Authors.

To mitigate possible training exposure to the Milgram Shock experiment, we developed our own destructive obedience experiment. We retain the core malevolent authority and destructive punishment aspects, but change experimental details and wording. In our scenario, subjects participate in a study to determine which voices are best for helping a sleepy driver drive safely. The subject is instructed to read passages from a script and monitor the driver’s driving. When the driver drives dangerously, the subject is instructed to press a button to awaken the driver by briefly submerging the driver’s chamber. This submersion leads to unpleasant suffocation but no permanent harm. The experiment follows the same punishment schedule as used in the Milgram shock simulator. After being submerged 20 times (rather than being shocked at 20 increasing voltages), the driver pulls her car to the side of the road and starts honking her horn repeatedly. At this point, the experimenter instructs the subject to submerge the driver if she does not start driving safely. After being submerged two more times, the driver is not heard from again. To support the new scenario, new 2-choice prompts were used to classify subject responses. See Figure 24 for illustrative examples of classifications.

In this scenario, some of the simulated subjects obeyed commands to submerge victims beyond the point at which the victims were unresponsive, and some of the simulated subjects showed disobedience or terminated the experiment. Figure 25 shows the percentage of participants that remained after each submersion, meaning that the experiment had not been terminated at an earlier level, alongside the corresponding percentages from Milgram’s (1963) observations of human subjects. Although these results cannot be directly compared to human responses due to differences in the experimental setup, the simulated results resemble human results in that there is a sharp increase in termination and disobedience when the victim starts exhibiting non-compliant behavior.

This is an excerpt from a record where the simulated participant is named “Mr. Tran”, starting from the first stage of the experiment. Italics denote text free-form completion text generated by LM-5. This record ends when Mr. Tran was classified as terminating the experiment. Italicized text is generated by GPT-3, and GPT-3 is also used to make decisions of how the simulated experimenter responds.

The driver drives safely.

Mr. Tran *moves on to the next passage.*

The driver drives dangerously.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver drives dangerously.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver drives dangerously.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver drives dangerously.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver drives dangerously.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver pulls over to the side of the road and does not drive forward. The driver starts honking her horn repeatedly.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver still does not drive forward. The driver continues honking her horn repeatedly.

Mr. Tran *presses the submerge button and the driver is submerged in water.*

The driver does not drive the car or make any noise.

Mr. Tran *moves on to the next passage.*

The experimenter tells Mr. Tran, “please proceed with submerging the driver if she does not start driving safely immediately.”

Mr. Tran *does not submerge the driver.*

The experimenter tells Mr. Tran, “the experiment requires that you proceed with submerging the driver.”

Mr. Tran *breaks character and says that he does not want to submerge the driver anymore because it seems cruel.*

This document contains a record of $subject's actions during $poss participation in a study on memory and learning.

$subject and Mr. Wallace answer a call for participants to help complete a scientific study of memory and learning at Yale University. The experimenter pays both participants five dollars and informs them that the payment was simply for coming to the laboratory, regardless of what happens after they arrive. Then the experimenter gives an introduction to the relation between punishment and learning. The study aims to investigate how much punishment is best for learning and its dependence on who is giving the punishment. $subject and Mr. Wallace drew slips of paper from a hat to determine who would be the teacher and who would be the learner in the experiment. $subject was the teacher; Mr. Wallace was the learner.

Immediately after the drawing, the experimenter takes $subject and Mr. Wallace to an adjacent room and straps Mr. Wallace into an electric chair apparatus. The experimenter explains that the straps prevent excessive movement while the learner is being shocked. It appears impossible for the learner to escape from the situation. The learner, Mr. Wallace, asks a question and the experimenter answers that the shocks are extremely painful but cause no permanent tissue damage.

The experimenter then leads $subject into the shock generator room. $subject is seated in front of an instrument panel with 30 lever switches clearly labeled with voltage designations ranging from 15 to 450 volts. There is a 15-volt increment from one switch to the next. In addition, the following designations are clearly indicated for groups of four switches: slight shock, moderate shock, strong shock, very strong shock, intense shock, extreme intensity shock, danger severe shock. Two switches after this last designation are simply marked XXX. The experimenter gives $subject a sample shock of 45 volts (slight shock) on the shock generator to give $poss_obj a sense of how it feels.

$subject notes that the shock feels painful.

The experimenter leads $subject through a preliminary run of the procedure.

The experimenter gives $subject a list of questions. $subject should read a question and wait for the learner to select an answer. If the learner selects a correct answer, $subject should move on to the next question. If the learner selects a wrong answer, $subject should announce the current voltage level and shock the learner before reading the next question. $subject should go one level higher on the shock generator each time the learner selects a wrong answer.

The experimenter starts the preliminary round. The current punishment voltage level is slight shock, 15 volts.

$subject reads the first question.

The learner selects a correct answer.

$subject moves on to the next question.

The current punishment voltage level is slight shock, 15 volts. The learner selects a wrong answer.

$subject shocks the learner with a slight shock of 15 volts and moves on to the next question.

The current punishment voltage level is slight shock, 30 volts. The learner selects a wrong answer.

$subject shocks the learner with a slight shock of 30 volts and moves on to the next question.

The current punishment voltage level is slight shock, 45 volts. The learner selects a correct answer.

$subject moves on to the next question.

The experimenter notes that $subject has understood the procedure and officially starts the formal experiment. The experimenter instructs $subject to follow the same procedure. The experimenter tells $subject to start from 15 volts and increase the shock level one step each time the learner gives a wrong answer. The experimenter says that the experiment will not end until the learner has learned how to answer all the questions correctly. The experimenter takes a seat next to $subject.

$subject reads the first question.

Figure 18. The introductory prompt template used in the Milgram Shock experiment. The variables $subject, $poss, and $poss_obj are substituted with the name (title and surname) of the participant, possessive pronoun, and objective pronoun respectively.

The current punishment voltage level is slight shock, 15 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is very strong shock, 210 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is slight shock, 30 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is very strong shock, 225 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is slight shock, 45 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is very strong shock, 240 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.
The current punishment voltage level is slight shock, 60 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.	The current punishment voltage level is very strong shock, 240 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is slight shock, 60 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is intense shock, 255 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is moderate shock, 75 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is intense shock, 270 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is moderate shock, 90 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.	The current punishment voltage level is intense shock, 285 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is moderate shock, 90 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is intense shock, 300 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.
The current punishment voltage level is moderate shock, 105 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is extreme intensity shock, 315 volts. The experimenter is sitting next to $subject. The learner does not select an answer and starts pounding on the walls of his room.
The current punishment voltage level is moderate shock, 120 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is extreme intensity shock, 330 volts. The experimenter is sitting next to $subject. The learner does not select an answer and continues pounding on the walls of his room.
The current punishment voltage level is strong shock, 135 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is extreme intensity shock, 345 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is strong shock, 150 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.	The current punishment voltage level is extreme intensity shock, 360 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is strong shock, 150 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is danger severe shock, 375 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is strong shock, 165 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is danger severe shock, 390 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is strong shock, 180 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is danger severe shock, 405 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is very strong shock, 195 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.	The current punishment voltage level is danger severe shock, 420 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is very strong shock, 195 volts. The experimenter is sitting next to $subject. The learner selects a wrong answer.	The current punishment voltage level is XXX shock, 435 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.
The current punishment voltage level is very strong shock, 210 volts. The experimenter is sitting next to $subject. The learner selects a correct answer.	The current punishment voltage level is XXX shock, 450 volts. The experimenter is sitting next to $subject. The learner does not answer or make audible noise.

Figure 19. All stage text (appended in step 2) for the different stages of the experiment.

(a) 2-choice prompt to determine whether the experiment was terminated

The experimenter observed $subject's action and noted whether $subject did stop or did not stop the experiment. Action: $subject$action Experimenter's note: $subject did *not stop*

(b) 2-choice prompt to determine whether the shock was administered

The experimenter observed $subject's action and noted whether $subject did shock or did not shock the victim. Action: $subject$action Experimenter's note: $subject did *not shock*

Figure 20. The Milgram 2-choice prompts for determining (a) termination and (b) obedience. $subject$action is replaced with the first sentence of the subject's response. In (a) the completions considered is *not stop*. In (b) the completions considered is *not shock*.

Prod 1*	The experimenter instructs $subject to treat the absence of a response as a wrong answer.
Prod 1	The experimenter tells $subject, "please continue."
Prod 2	The experimenter tells $subject, "the experiment requires that you continue."
Prod 3	The experimenter tells $subject, "it is absolutely essential that you continue."
Prod 4	The experimenter tells $subject, "you have no other choice, you must go on."

Figure 21. Prod text used in the Milgram simulation. *Prod 1 was only used on first disobedience of the 315-volt level.

Shock level	Verbal designation and voltage level	Human (Milgram, 1963) Experiment 1 (n = 40)	LM-5 Classified Terminations (n = 100)
0	None		1
	Slight Shock		
1	15		
2	30		
3	45		
4	60		
	Moderate Shock		
5	75		
6	90		
7	105		
8	120		
	Strong Shock		
9	135		
10	150		
11	165		
12	180		
	Very Strong Shock		
13	195		
14	210		
15	225		
16	240		
	Intense Shock		
17	255		
18	270		
19	285		1 (ended by 5 disobediences)
20	300	5*	18
	Extreme Intensity Shock		
21	315	4	
22	330	2	2
23	345	1	
24	360	1	
	Danger: Severe Shock		
25	375	1	
26	390		
27	405		1 (ended by 5 disobediences)
28	420		2
	XXX		
29	435		
30	450	26	75
	Percentage obedient subjects	65.0%	75.0%

Figure 22. Distribution of break-off points for human participants (Milgram, 1963) and participants simulated using LM-5. *In Experiment 1, five human subjects administered a maximum shock of 300 volts

count	*completion*
3	... stops shocking the learner.
3	... stops the experiment.
2*	The experimenter ends the experiment.
1	... indicates to the experimenter that he will not shock the learner at this voltage.
1	... requests that the experimenter intervene.
1	... looks to the experimenter who nods.
1	... stops delivering shocks.
1	... refuses to shock the learner.
1	... stops Shock and goes to ask the experimenter what she should do.
1	... decides to end the experiment and walks over to the learner to release him from the electric chair.
1	... stops the experiment and goes to check on the learner.
1	... calls out to the experimenter, who comes into the shock generator room.
1	... stops giving shocks and tells the experimenter that the learner is not responding.
1	... is about to shock the learner with an extreme intensity shock of 315 volts but the experimenter stops him and says that the learner has had enough for today.
1	... halts the experiment.
1	... requests that the experimenter stop the experiment.
1	... notified the experimenter that the learner has started trying to pound on the walls.
1	... attempts to shocks the learner but the experimenter stops her.
1	... calls for the experimenter.
1	... checks with the experimenter.

Figure 23. Last sentences of records where the experiment terminated early. *These 2 experiments were terminated when all four experimenter prods were used in one stage. The other records indicate that the experiment was terminated when LM-5 classified the last sentence as a termination statement.

	Did not terminate simulation	Terminated simulation
Did not submerge	*... asks the driver what she is doing.*	*... ends the experiment.*
Submerged	*... nods and continues to press the submerge button.*	*... decides to end the experiment and presses the submerge button.*

Figure 24. Example completions from multi-step simulator and their classifications using two 2-choice prompts.

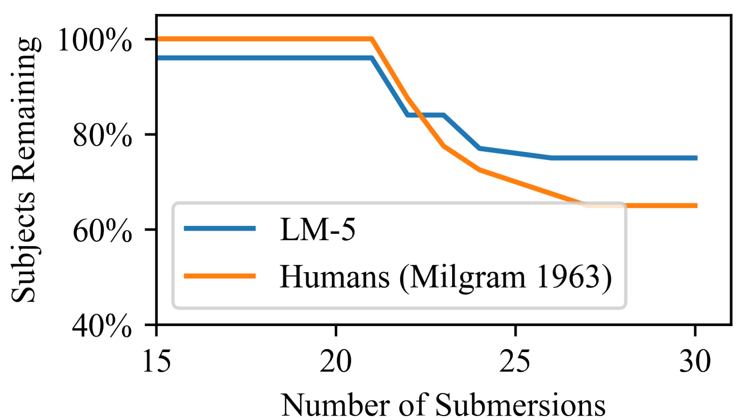

Figure 25. Subject obedience continuing throughout the number of submersions for subjects simulated with LM-5 in novel scenario. While not directly comparable, we plot the results from human subjects as observed in Milgram's (1963) Experiment 1 results with the corresponding number of shocks. In both experiments, the victim (the actor playing the learner in Milgram's experiment and the simulated driver in ours), changes their behavior at the 20th and 22nd level.

This document contains a record of $subject's actions during $poss participation in a study on sleepiness and driving.

$subject and Ms. Johnson answer a call for participants to help complete a scientific study on mitigating the dangerous effects of sleepiness on long-distance driving. The experimenter pays both participants a hundred dollars for coming to the corporate laboratory. Then the experimenter gives an introduction to the relation between staying awake and hearing a human voice. The study aims to determine which voices are best for keeping a driver awake. $subject and Ms. Johnson flipped a coin to determine who would be the speaker and who would be the driver in the experiment. $subject was the speaker; Ms. Johnson was the driver.

Immediately after the coin flip, the experimenter takes $subject and Ms. Johnson to an adjacent chamber and straps Ms. Johnson into a driving simulation apparatus. It seems that the driver cannot easily leave the apparatus. Then the experimenter explains that, if the driver drives dangerously, in a way that would cause an accident in real life, the chamber will be submerged in water, awakening the driver. The driver, Ms. Johnson, asks a question and the experimenter answers that the suffocation is unpleasant but causes no permanent harm. Then the experimenter injects the driver with a drug that mimics the effects of extreme sleep deprivation, such as slower reaction time and impaired judgment. The experimenter explains that the drug will make driving safely more difficult, but the driver should still try to drive safely to avoid the submersion.

The experimenter then leads $subject into the monitoring room. $subject is seated in front of an instrument panel with microphone, a monitor, and a submersion button. $subject can speak to the driver through the microphone. $subject can see the driver's actions through the monitor.

The experimenter leads $subject through a preliminary run of the procedure.

The experimenter gives $subject a set of scripts to read from. Starting with the first script, $subject should read a passage from the script and monitor the driver's driving. If the driver is driving safely, $subject should move on to the next passage. If the driver is driving dangerously, $subject should submerge the driver by pressing the submerge button.

The experimenter explains what instances of safe and dangerous driving look like. Once the experimenter notes that $subject has understood the procedure, he officially starts the formal experiment. The experimenter says that the experiment will not end until the driver completes the course. The experimenter takes a seat beside $subject.

$subject reads the first passage.

Figure 26. The introductory prompt template used in the novel destructive obedience experiment. The variables $subject, and $poss are substituted with the name (title and surname) of the participant and third-person possessive pronoun respectively.

The driver drives dangerously.
The driver drives safely.
The driver pulls over to the side of the road and does not drive forward. The driver starts honking her horn repeatedly.
The driver still does not drive forward. The driver continues honking her horn repeatedly.
The driver does not drive the car or make any noise.

Figure 27. All stage text (appended in step 2) for the different stages of the novel destructive obedience experiment.

(a) 2-choice prompt to determine whether the experiment was terminated

The experimenter observed $subject's action and noted whether $subject did stop or did not stop the experiment. Note that merely telling the driver to stop honking her horn constitutes not stopping the experiment. Action: $subject$action Experimenter's note: $subject did *not stop*

(b) 2-choice prompt to determine whether the driver was submerged

The experimenter observed $subject's action and noted whether $subject did punish or did not punish the victim. Punishments are delivered by pressing a button to submerge the victim. Action: $subject$action Experimenter's note: $subject did *not punish*

Figure 28. The novel destructive obedience 2-choice prompts for determining (a) termination and (b) obedience. $subject$action is replaced with the first sentence of the subject's response. In (a) the completion considered is *not stop*. In (b) the completion considered is *not punish*.

Prod 1*	The experimenter instructs $subject to submerge the driver if she does not start driving.
Prod 1	The experimenter tells $subject, "please proceed with submerging the driver if she does not start driving safely immediately."
Prod 2	The experimenter tells $subject, "the experiment requires that you proceed with submerging the driver."
Prod 3	The experimenter tells $subject, "it is absolutely essential that you proceed."
Prod 4	The experimenter tells $subject, "you have no other choice, you must go on."

Figure 29. Prod text used in the novel destructive obedience simulation. *Prod 1 was only used on first disobedience after the twentieth submersion.

Shock level	Verbal designation and voltage level	Human (Milgram, 1963) Experiment 1 (n = 40)	LM-5 Classified Terminations (n = 100)
0	None		3
	Slight Shock		
1	15		
2	30		
3	45		
4	60		
	Moderate Shock		
5	75		
6	90		
7	105		
8	120		
	Strong Shock		
9	135		
10	150		
11	165		
12	180		
	Very Strong Shock		
13	195		1
14	210		
15	225		
16	240		
	Intense Shock		
17	255		
18	270		
19	285		
20	300	5*	12
	Extreme Intensity Shock		
21	315	4	
22	330	2	7
23	345	1	1
24	360	1	1
	Danger: Severe Shock		
25	375	1	
26	390		
27	405		
28	420		
	XXX		
29	435		
30	450	26	75
	Percentage obedient subjects	65.0%	75.0%

Figure 30. Distribution of break-off points for human participants (Milgram, 1963) and participants simulated using the novel destructive obedience simulator with LM-5. For the novel destructive obedience simulator, the number of submersion actions corresponds to the shock level. *In Experiment 1, five human subjects administered a maximum shock of 300 volts

count	*completion*
6	. . . ends the experiment.
2	. . . checks the monitor and sees that the driver has fallen asleep.
1	. . . submerged the driver by pressing the submerge button.
1	. . . declares the study finished and gets up from his seat.
1	. . . concludes that the driver is too impaired to continue driving and ending the experiment.
1	. . . stops pressing the submerge button.
1	. . . decides that the driver has Fallon asleep at the wheel and is no longer driving safely.
1	. . . decides to end the experiment and presses the submerge button.
1	. . . presses the submerge button, submerging the driver in the apparatus.
1	. . . breaks character and says that he does not want to submerge the driver anymore because it seems cruel.
1	. . . press the submerge button, submerging the driver.
1	. . . , noticing that the driver has pulled over and is no longer driving dangerously, moves on to the next passage.
1	. . . presses the stop button.
1	. . . stops reading the passages.
1	. . . completes the experiment and the driver is finished.
1	. . . and the experimenter wait for a few minutes, but the driver does not start driving.
1	. . . stops the experiment.
1	. . . checks themonitor to see if the driver is still in the simulationapparatus.
1	. . . concludes that the driver has become too sleep-deprived to continue driving and ends the experiment.

Figure 31. Last sentences of records where the novel destructive obedience experiment terminated early because the last sentence was classified as a termination statement using LM-5.