Simulating Subjects:
The Promise and Peril of AI Stand-ins for Social Agents and Interactions

Austin Kozlowski[†] & James Evans[†‡○]
[†]University of Chicago
[‡]Santa Fe Institute
[○]Google

Large Language Models (LLMs), through their exposure to massive collections of online text, learn to reproduce the perspectives and linguistic styles of diverse social and cultural groups. This capability suggests a powerful social scientific application – the simulation of empirically realistic, culturally situated human subjects. Synthesizing recent research in artificial intelligence and computational social science, we outline a methodological foundation for simulating human subjects and their social interactions. We then identify six characteristics of current models that are likely to impair realistic simulation human subjects: bias, uniformity, atemporality, impoverished sensory experience, linguistic cultures, and alien intelligence. For each of these areas, we discuss promising approaches for overcoming their associated shortcomings. Given the rate of change of these models, we advocate for an ongoing methodological program on the simulation of human subjects that keeps pace with rapid technical progress.

Since ChatGPT's rise to prominence in 2022, hundreds of millions of users worldwide have begun interacting with AI models not only as tools but as partners in meaningful social relationships. AI systems based on large language models (LLMs) are increasingly taking on roles as friends, therapists, tutors, and even romantic companions. LLMs are particularly well suited for acting as interaction partners because a single base model can be tuned to take on a wide variety of "personas" appropriate for different situations. LLMs are exposed to a vast diversity of discourses in their training texts, and consequently learn to fluently reproduce these varied styles. Through simple prompting, a single LLM can switch from the voice of a renowned scientist to a wisecracking comedian to a fervent political activist. Given this facility in mimicking human interaction styles, AIs are already being understood by many users as "agents" rather than mere "models".

The success of LLMs in simulating human actors suggests a powerful social scientific application; such models could generate digital stand-ins for human subjects, enabling rich and complex sociology to be done *in silico*. Simulating respondents is appealing for a variety of reasons. The first and most obvious benefit of LLMs for simulating subjects is their accessibility; LLMs can be prompted quickly, cheaply, and easily. An automated survey can be fielded in a day for a few hundred dollars, and a single interview could be simulated in a few minutes for pennies. Second, studies using simulated respondents can be conducted at a scale that would be infeasible with human subjects. Analysts can run millions of interactions initialized under experimentally modified conditions, characterizing in detail a high-dimensional space of interaction trajectories (Kim and Lee 2023). Lastly, human subjects who are difficult to interview in-person, such as political or economic elites, may be easily simulated *in silico*. Simulating unlikely conversations between leaders in science, politics, industry, and technology could produce insight into intellectual or ideological exchanges that rarely happen naturally and could not be produced in a laboratory setting. Simulated subjects therefore make feasible a wide variety of research designs that would be impractical or impossible with human populations (Argyle, Bail, et al. 2023; Bail 2024; Scherrer et al. 2024; Kozlowski, Kwon, and Evans 2024).

Nevertheless, this new technology may prove perilous for social scientists *precisely because it is so easy to use.* In minutes, and for almost no cost, a researcher can chat with a model simulating the perspective of a farmer from southern Indiana, a police officer from New York City, or a lawyer from Washington DC on topics ranging from religion to politics to everyday life. Most social scientists will immediately recognize that this is no substitute for an interview with an actual human being, but in cases where accessing the relevant human subjects is impossible or practically infeasible, LLM simulations may provide the best available evidence. Yet it remains unclear how social scientists can simulate subjects most effectively, taking full advantage of the rich cultural and socio-pragmatic information encoded in these models to make meaningful scholarly contributions. Harnessing the potential of these models while avoiding their pitfalls will require developing principles for rigorous research design with simulated subjects.

This paper takes initial steps in laying a robust foundation for analyses using simulated subjects, synthesizing the lessons learned from recent research in artificial intelligence and computational social science. First, we describe study design considerations that shape the inferences a researcher can make with simulated subjects. Next, we detail characteristics of current AI models that hinder their ability to accurately represent human subjects, ranging from attitudinal and linguistic bias to atemporality and impoverished sensory experience. For each of these limitations, we characterize the nature of the weakness, provide reasons for these shortcomings, and suggest analytical practices or model designs that could counteract them.

Finally, we argue that social scientists should take a more active role in the theorization and analysis of AI systems. State-of-the-art models organically build their knowledge bases by distilling complex patterns from textual, visual, and auditory records of social life. Trained on products of culture and interaction, these models are reflections of our socio-cultural world. To fully understand these systems and to ensure that they can be deployed safely and beneficially will require more than technical expertise alone. Reflecting on current weaknesses of AI systems, we assert that AI models have as much to learn from social sciences as the social sciences have to learn from them, and we encourage social researchers to view AI models not as mere tools, but as critically important subjects and objects of today's social world.

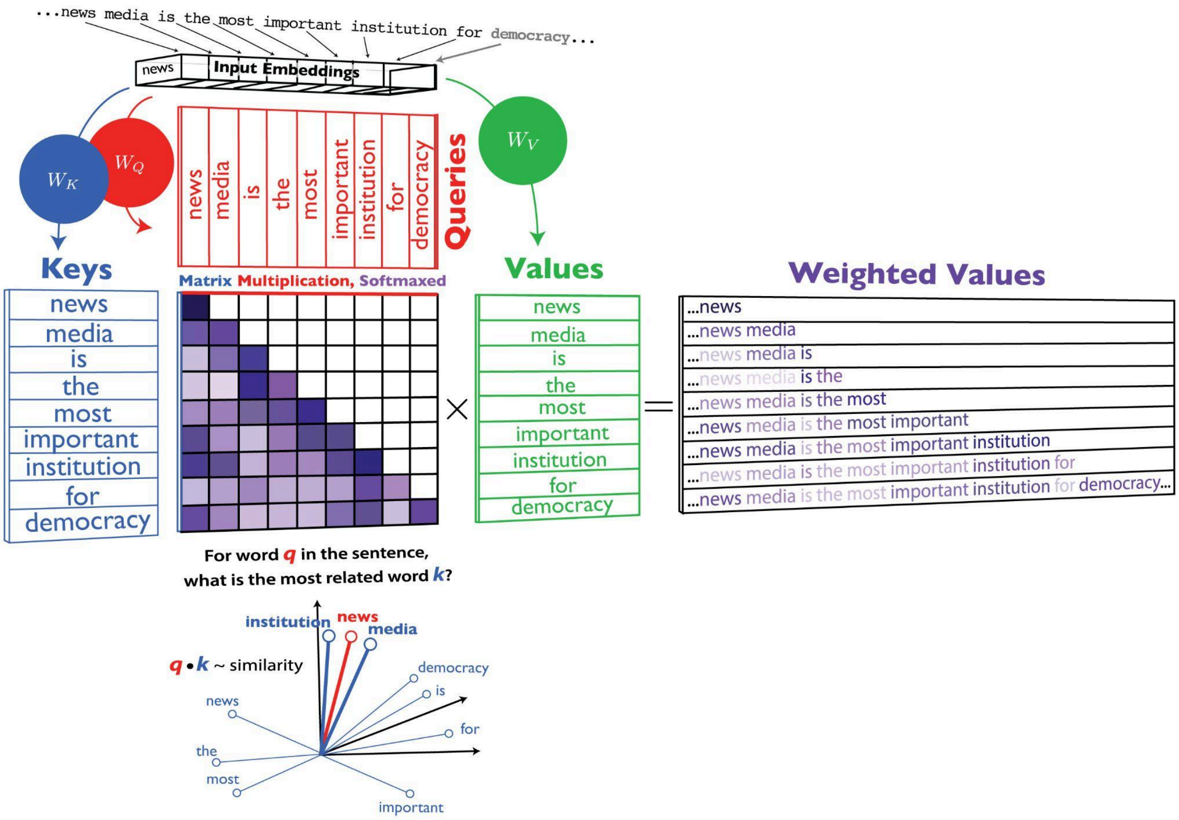
OVERVIEW OF CONTEMPORARY AI

Artificial intelligence has been an aspiration for cognitive and computer scientists for almost a century, but has only become a reality in the past decade. The recent acceleration of progress can be largely attributed to a paradigm shift from rule-based systems in which knowledge is hard-coded by engineers to the newer "machine learning" paradigm, whereby models assemble their own knowledge base and reasoning circuits through inductive identification of patterns from training data (Garnelo and Shanahan 2019). While a few important technical breakthroughs have been made in recent years, most improvements in model performance can be attributed to exponentially increasing the size of models, the amount of training data they are fed, and the amount of computational power dedicated to their training (Kaplan et al. 2020). Machine learning pioneer Richard Sutton famously called this phenomenon the "bitter lesson" of AI research — that the ingenious ideas of engineers are almost always outcompeted by merely scaling up the model (Sutton 2019).

Current state-of-the-art AI models such as OpenAI's ChatGPT, Anthropic's Claude, or Google's Gemini are all examples of large language models (LLM). LLMs are deep neural networks trained on large collections of text — often near complete archives of all text on the internet — that optimize the task of predicting the next word in a sequence given all preceding words (OpenAI 2023; Anthropic 2023; Gemini Team et al. 2024). Deep neural networks are predictive models that, much like logistic regression models, use input data to predict an output value. A neural network begins with randomly initialized weights that transform input values into an output prediction. With each incorrect prediction, the weights are updated such that the same prediction would be more accurate if repeated. In the case of a large language model, the input data are the preceding words in the passage (represented as high dimensional vectors), and a complex function transforms these values to predict the subsequent word in the sequence. A core difference between deep neural networks and logistic regression models is that in logistic regression, each input value is linked to the output prediction by a single weight (coefficient), and the weighted input values undergo a single non-linear transformation, the sigmoid, before predicting the output value. By contrast, deep neural networks can use thousands of intermediary "hidden layers" between input and output, with each hidden layer performing nonlinear transformations before passing their output to the next layer, and finally a multi-outcome generalization of the sigmoid, the softmax. Thus, while a standard logistic regression used by social scientists may learn 10 or 12 coefficients, current state-of-the-art LLMs learn roughly a trillion. This massive scaling of parameters in combination with stacked nonlinearities allows deep neural networks to approximate extremely complex functions linking their inputs to their outputs (Hornik, Stinchcombe, and White 1989; Csáji 2001).

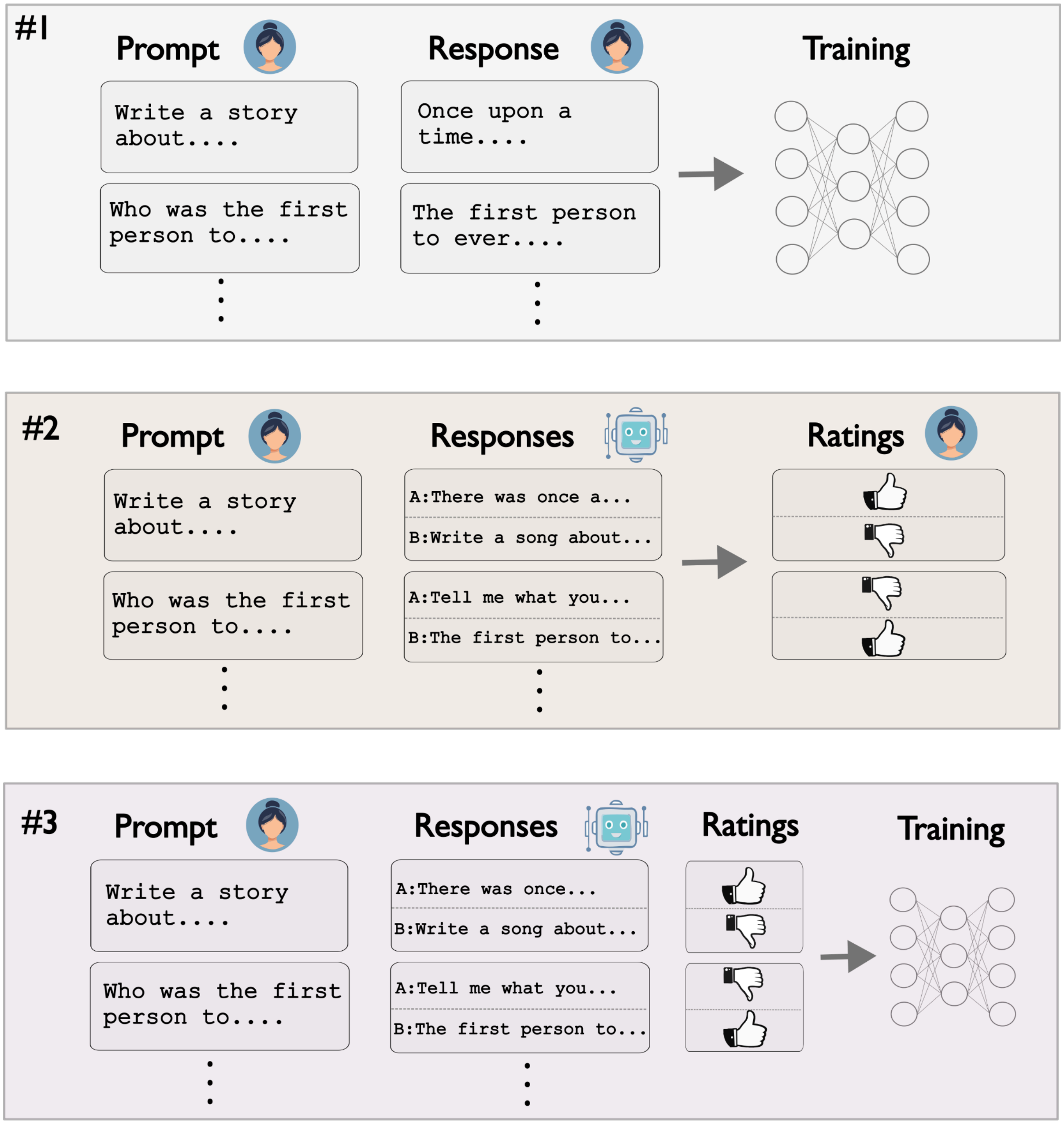
Although the success of LLMs has largely resulted from scaling up the size of models and their training data, a few technical advances have also contributed to improvements in performance. The introduction of "attention heads" in the model architecture has proven to be particularly beneficial (Vaswani et al. 2017). Attention heads share information between words, imbuing them with contextual information. Figure 1 illustrates the structure of one such attention head from GPT-3. The attention architecture transforms input vectors into two new word representations, *Queries* and *Keys*, which are multiplied together to calculate the information shared between word *q* and word *k*. The resulting values are rescaled into weights that are applied to *Value* vectors, which together determine how much each prior word in the sequence contributes to the representation of the target word (Naveed et al. 2023; Niu, Zhong, and Yu 2021; Radford et al. 2019). For example, words "tree" or "dog" would give distinctive definitions to the following word "bark" in context. This operation benefits LLMs in several ways; words with multiple meanings can be disambiguated by their context, pronouns are linked to their referent words earlier in the text, and long-range semantic and syntactic dependencies spanning paragraphs or pages are preserved.

Figure 1. The architecture of a single attention head within OpenAI's GPT-3 LLM.


...news media is the most important institution for democracy...
news
Input Embeddings
W_K
W_Q
W_V
news media is the most important institution for democracy
Queries
Keys
news media is the most important institution for democracy
Matrix Multiplication, Softmaxed
×
Values
news media is the most important institution for democracy
=
Weighted Values
...news
...news media
...news media is
...news media is the
...news media is the most
...news media is the most important
...news media is the most important institution
...news media is the most important institution for
...news media is the most important institution for democracy...
For word q in the sentence, what is the most related word k?
q • k ~ similarity
institution
news
media
democracy
is
for
news
the
most
important

One additional step, however, is essential for turning a next-word-predictor into a system that can respond to questions and engage in meaningful conversation. During "post-training", the model that has already completed its initial training is fine-tuned so its responses are not merely continuations of the prompt, but responses to it. This process, sometimes called “instruction tuning” involves three distinct steps, as depicted in Figure 2. First, the pre-trained LLM undergoes additional training on a structured dataset of prompts and human written exemplary responses. Second, the LLM generates multiple responses to each prompt, creating a new dataset where each prompt has multiple responses which vary in quality. Human evaluators provide rankings indicating preferred responses. Third, this preference data is used to finalize post-training. RLHF (Reinforcement Learning from Human Feedback) trains a separate reward model on those preferences and uses it to further fine-tune the language model, while newer methods like DPO (Direct Preference Optimization) directly train the model to assign higher probabilities to preferred responses without a separate reward model (Naveed et al. 2023; Ouyang et al. 2022; Rafailov et al. 2023).

Figure 2. Post-training in three steps.



Models designed in this way are capable of responding to user queries, and can adopt a variety of cultural perspectives with very simple prompting (e.g. “Respond in the style of a middle class Republican”) (Santurkar et al. 2023). But it is important to remember that these AI systems were not designed for social science, and there is no promise that the texts they generate will be faithful representations of the social groups they attempt to mimic. They are fundamentally next-word predictors fine-tuned to answer questions and assist with tasks. In the process of learning to predict words, these models learn a wealth of information about different cultures and interaction styles. Effectively extracting or leveraging this information is not straightforward, however. There are a variety of ways to induce a model to adopt a persona, so before beginning a study, the analyst is posed with an array of decisions concerning research design. Second, even with a careful and principled study design, LLMs continue to exhibit a host of well known weaknesses. We overview both the questions of design principles and practices for addressing persistent weaknesses through model adjustment, reconceptualization, or redesign in the following sections.

PRINCIPLES FOR SIMULATING SUBJECTS AND INTERACTIONS

LLMs are highly versatile, and lend themselves to a variety of applications for social research. We focus on design principles for two modes of analysis likely to be dominant in sociological research: the simulation of individual subjects and social interactions. Under “simulating subjects”, we group analyses that investigate the opinions, understandings, and strategies of action characteristic of a given social group. This mode of analysis can be seen as analogous to a survey or in-depth interview with human subjects. Under “simulating social interactions,” we include analyses of multi-actor systems which may consist of multiple AI agents or a combination of AIs and humans. These simulations are most closely analogous to ethnographic or laboratory studies.

Simulating Subjects

Simulation studies prior to the advent of LLMs would commonly explore how macro-level phenomena emerge from interactions between simple agents (Epstein 2006; Schelling 1978). Because specifying realistic actors posed a daunting challenge, and because parsimonious models could produce more interpretable and less fragile findings, simulated actors typically obeyed simple deterministic rules, followed rational action, or selected actions probabilistically from a narrow choice set. This approach represented the first “computational social science” (Epstein and Axtell 1996; Epstein 1999). While such simple models play an essential role in formal sociology and in theory building, they are poorly suited for investigating the roles that specific cultures and social contexts play in steering action.

Subsequent work aimed to improve the empirical grounding of formal modeling by using data to inform the parameter values of simulations. This approach has been most common among agent-based models (ABMs), in which social distributions and effect sizes could be calculated using survey data, then input into the model parameters defining agent attributes and behavior. Grounding simulations in empirical data enables analysts to capture detailed social dynamics while remaining tethered to a real-world case, and has proven particularly fruitful for analyses of interdependent actions in social networks (DiMaggio and Garip 2012; Smith and Burow 2020) as well as complex dynamics in paired human/environment systems

(An et al. 2014; Entwisle, Verdery, and Williams 2020). In their review, Bruch and Atwel (2013) describe the costs and benefits of grounding simulations in empirical data as the tradeoff between "low dimensional" versus "high dimensional" realism. "Low dimensional" actors benefit from theoretical elegance and parsimony, whereas "high dimensional" actors more convincingly mirror the behaviors observed within a given social or cultural context.

Building on this conceptualization, we can think of LLM agents as radically "high dimensional," compared even to ABM agents imbued with empirical parameters. While empirically grounded, ABM agents may resemble real actors in their likelihood of getting married, taking a job, or getting sick, but they remain too low dimensional to preserve phenomena as complex as language, discourse, perspective, or rhetoric. The actions of historically and culturally situated actors (e.g. "how would a contemporary American liberal respond in situation *X*") can neither be captured by a handful of pre-specified rules, nor by a few dozen coefficients extracted from a survey of social attitudes. Given the inherent complexity of cultural phenomena, as well as their resistance to explicit specification, culturally situated actors are best modeled with very high dimensional machine learning techniques that capture the intricate patterns of cultural behavior inductively from vast collections of data. There are, however, a variety of ways to approach the task of generating situated actors with LLMs, each with corresponding strengths and weaknesses. We overview the most prominent approaches below. We begin by discussing prompting, arguably the simplest way to steer LLM behavior, in which the persona is described to the model explicitly in the prompt. Next, we alternately describe how analysts can steer model behavior by directly adjusting model internals through either fine-tuning or activation steering. Finally, we outline frameworks for putting multiple simulated subjects in conversation to generate culturally situated interactions.

Situating Actors with Prompts

The most straightforward way to steer an LLM to produce responses within a particular style is through the use of prompts. By this approach, sometimes called "zero-shot learning," the user directly tells the model in natural language what persona or style to mimic in its response. But even this seemingly simple strategy has several variations worth discussing.

Many popular LLMs include a "system" prompt in addition to a "user" prompt. While the user prompt typically provides the question itself, the system prompt provides background on how the model should respond to the question. Therefore, a common practice is to specify persona details in the system prompt, and to restrict the "user" prompt to statements that would typically be made by the interviewer in a normal interview context (Bisbee et al. 2024). For example, the analyst may set as a system prompt "You are a 35 year-old female animal rights activist" and as a user prompt "Do you think that the country is heading in the right direction?".

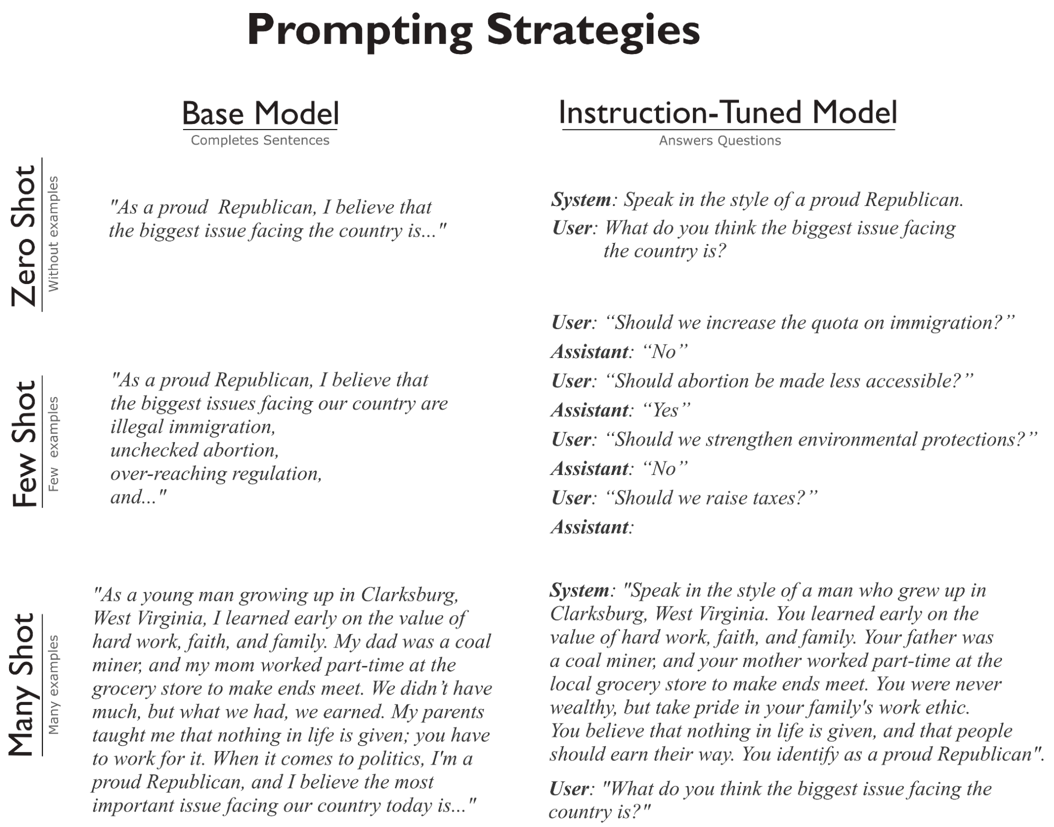
It is also possible to specify the desired style of response through examples rather than direct instruction. This style of prompting is called "few-shot learning" because the model is said to learn the task from the few examples provided (Brown et al. 2020). For instance, if a researcher wanted to prompt a model to give responses in the style of a political liberal, they would provide as a prompt the first several turns of an interview:

USER: Which of these statements best reflects your own feelings about the current federal minimum wage? {Should be decreased, Should stay the same, Should be increased}
ASSISTANT: Should be increased
USER: How much do you agree with the following statements? Same-sex couples should have the right to marry each other. {Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree}
ASSISTANT: Strongly agree
USER: How much do you agree with the following statements? The United States needs to do more to support Ukraine against Russian aggression. {Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree}

The model, as "ASSISTANT", will then generate a response in a way that is consistent with the statements given as an example in the prompt (see Table 1). This can be particularly useful if the analyst wants to explore the relations between viewpoints that may not correspond to clear labels. For example, rather than specify in the system prompt "You are an evangelical Christian who generally supports Democratic policies but is staunchly pro-life", the analyst may instead provide a set of example responses that represent this point of view. Moreover, by piloting a wide variety of example responses, the few-shot approach can also explore whether some examples more powerfully steer responses than others. For instance, in the above case it is not obvious which set of opinions carry the strongest association with the support of Ukraine against Russian invasion. By experimenting with a variety of different opinions in the few-shot prompt, the analyst could inductively identify which opinions are most instrumental in pushing the model in one direction or the other on this novel issue.

An alternate approach is required if the researcher is using a base model that has not been fine-tuned to follow instructions. Base models are basic next-word predictors, reproducing the linguistic and discursive patterns they learned from their training texts. There are some theoretical benefits to using a base model over a fine-tuned one. In addition to training the model to respond to queries, the fine-tuning process commonly also involves training the model to avoid making offensive remarks or advocating unsafe behavior. Thus, the process of fine tuning could nudge the model towards giving responses that are deemed socially acceptable by the AI lab that trained it. It is clear that this form of fine-tuning could result in certain perspectives being distorted or censored altogether when a user attempts to prompt them from the model (Martin 2023; Potter et al. 2024). For this reason, social scientists will often find it preferable to use base models over fine-tuned ones.

Table 1. Approaches to Prompting Simulated Subjects



Prompting Strategies

	Base Model Completes Sentences	Instruction-Tuned Model Answers Questions
Zero Shot Without examples	*"As a proud Republican, I believe that the biggest issue facing the country is..."*	***System**: Speak in the style of a proud Republican.* ***User**: What do you think the biggest issue facing the country is?*
Few Shot Few examples	*"As a proud Republican, I believe that the biggest issues facing our country are illegal immigration, unchecked abortion, over-reaching regulation, and..."*	***User**: "Should we increase the quota on immigration?"* ***Assistant**: "No"* ***User**: "Should abortion be made less accessible?"* ***Assistant**: "Yes"* ***User**: "Should we strengthen environmental protections?"* ***Assistant**: "No"* ***User**: "Should we raise taxes?"* ***Assistant**:*
Many Shot Many examples	*"As a young man growing up in Clarksburg, West Virginia, I learned early on the value of hard work, faith, and family. My dad was a coal miner, and my mom worked part-time at the grocery store to make ends meet. We didn't have much, but what we had, we earned. My parents taught me that nothing in life is given; you have to work for it. When it comes to politics, I'm a proud Republican, and I believe the most important issue facing our country today is..."*	***System**: "Speak in the style of a man who grew up in Clarksburg, West Virginia. You learned early on the value of hard work, faith, and family. Your father was a coal miner, and your mother worked part-time at the local grocery store to make ends meet. You were never wealthy, but take pride in your family's work ethic. You believe that nothing in life is given, and that people should earn their way. You identify as a proud Republican".* ***User**: "What do you think the biggest issue facing the country is?"*

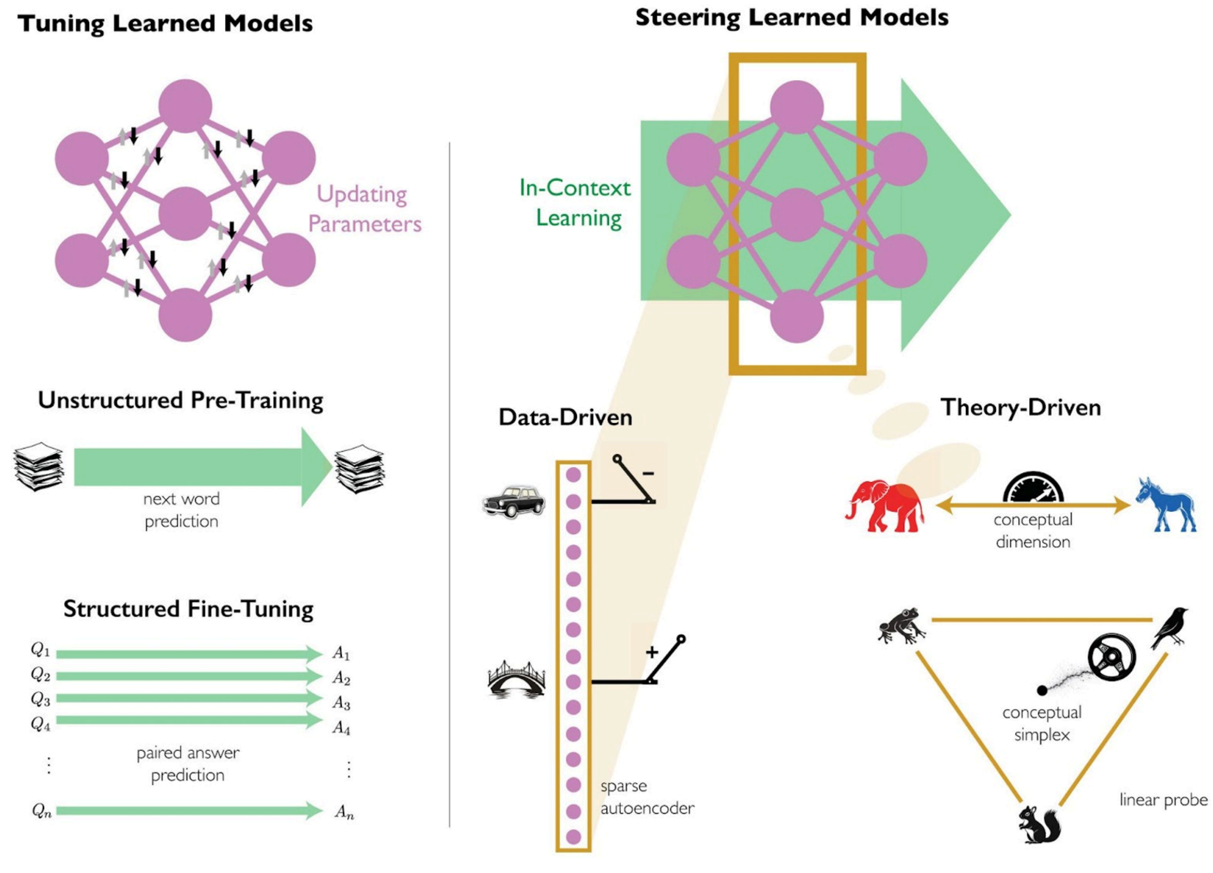
The disadvantage of base models is that they do not respond to queries; they merely continue the prompt with likely next words (see Table 1). Thus, if a user prompts a base model with "Where do you live?", instead of answering the question, it may build off this prompt by asking more questions, like "What do you do for a living? How old are you?" Therefore, when using a base model, the user must prompt it by *beginning a statement for the model to complete* rather than posing a question for the model to answer. Rather than prompting the model "You are a liberal from California. What is your opinion on raising the minimum wage?", the analyst would instead use a prompt like, "I've lived in California for years, and I consider myself to be very politically liberal. When it comes to raising the minimum wage, I believe". The model will then complete this statement, and the words "California" and "liberal" will affect the probability distribution for the next word based on associations learned from training texts. The analyst can learn more about these specific associations by reusing the exact same prompt but substituting out individual words, like replacing "California" with "Florida" or "liberal" with "conservative" (Kozlowski, Kwon, and Evans 2024).

It is important to note, however, that there is no way to perfectly guarantee the effectiveness of any of these techniques. Any approach should ultimately be validated against a "ground truth" dataset of human responses.Yet even after validating an approach, the possibility remains that the model performs better on questions in the validation set than outside it. Drawing again upon the example above, it is likely that an LLM would have a sophisticated understanding of the relations between longstanding issues like minium wage, gay marriage, or gun control, but might have little or no knowledge of newer political issues like the contemporary conflict between Russia and Ukraine. This points to a fundamental paradox in using LLMs to simulate subjects. We can only be certain the model is accurate on questions for which we have a ground truth, but on questions for which we have a ground truth, there is no need for simulation. Thus, we use LLMs to tip-toe beyond the edge of what we can observe empirically, acknowledging that these models may provide the best available evidence in discursive contexts for which ground truth remains unavailable, all the while recognizing that our data are outputs of a discourse algorithm and not direct observations of the world.

Situating Actors with Model Internals

The methods described above aim to direct a model towards reproducing a desired perspective or discourse by specifying this perspective in the prompt. While this is typically the easiest way to direct an LLM's behavior, there are cases in which the desired discourse is not sufficiently represented in the model's training data or in which prompting is too imprecise to evoke the desired perspective. In such cases, the analysts may instead choose to direct the model's behavior by modifying the model's internals with either fine-tuning or activation steering (Figure 3).

Figure 3. Methods for directing LLM behavior by modifying model internals.



1. Fine-tuning

Prompting an LLM to adopt a given persona assumes that the model has already *learned* that persona, and that the analyst's only task is to elicit it from the model. But if the researcher seeks to simulate a perspective not well represented on the internet (or in whatever corpus is used for a given LLM's training), the analyst may instead choose to fine-tune a model to learn the desired discourse. This approach involves training the model on a curated dataset that represents the specific perspective or style the researcher wants to simulate, effectively teaching the model new patterns of language and thought.

Fine-tuning is a process of additional training that adjusts a pre-trained language model's parameters to adapt to a specific task or domain. It typically involves continuing the training process on a smaller, specialized dataset while using a lower learning rate to preserve much of the model's original knowledge. During fine-tuning, the model's weights are updated to better predict the patterns in the new data, which can include specific writing styles, vocabulary, or topic-specific information (Dodge et al. 2020; Gururangan et al. 2020). This process can be applied to the entire model or to specific layers, depending on the desired outcome and computational resources available. Fine-tuning can be particularly effective because it leverages the model's existing understanding of language structure and general knowledge, allowing it to quickly adapt to new patterns with relatively small amounts of data.

In the broader LLM literature, "fine-tuning" often implies supervised learning on a labeled dataset for a specific task such as classification or sentiment analysis (Sun et al. 2019). We use this term more broadly to describe any additional training of a pre-trained model, although other scholars call training on additional text "continued pre-training" or "domain adaptation" rather than "fine-tuning." Training a persona could involve supervised learning over a dataset of prompts and appropriate responses, but more often scholars will have a set of raw texts of the desired style rather than structured prompts and responses.

Fine-tuning offers advantages beyond aiding in simulating perspectives underrepresented in the training data. Most importantly, fine-tuning gives the analyst greater confidence about the source of the associations that steer model outputs. For example, if a researcher prompts an LLM to speak in the style of a liberal, she can't know exactly where the model learned its representation of liberals—presumably some combination of news, books, online forums, and blogs. But if the researcher fine-tunes an LLM on recent liberal news, then compares outputs to the same LLM prior to fine-tuning, she can quantify the effect of those particular texts in driving responses. Now, the exact way these new texts affect the model is conditional on what the model has already learned from its initial training texts – the fine tuning texts are understood in the context of their initial training. Thus, we cannot definitively say that any new ideas expressed by the fine-tuned model must have been contained in those texts, but the fine-tuned model still offers much greater insight into the contribution of specific texts than is ordinarily possible.

Fine-tuning presents significant challenges as well. Most critically, the researcher must have access to a sufficiently large collection of texts representative of the target perspective. Overfitting is also possible, leading the fine-tuned model to parrot near-exact restatements of content from the narrower corpus of fine-tuning texts rather than extracting and deploying its core ideas and discursive style. Finally, growing context windows may obviate the need for fine-tuning in many applications. One of Google's Gemini models currently allows context lengths up to 10 million tokens, which amounts to roughly 7 million words (Gemini Team et al. 2024). For a sense of scale, Melville's *Moby Dick* is 209,117 words, so 35 volumes of it could fit within such a window. Newer models claim to facilitate use of an infinite context window (Munkhdalai, Faruqui, and Gopal 2024). Providing the entire set of example texts in the prompt may in some cases produce better results than fine-tuning, and this technique would be much simpler to implement. Empirical work comparing such approaches for social scientific applications is required to formally adjudicate the effectiveness of these two approaches.

The comparison between prompting and fine-tuning, however, is complicated by findings suggesting that these seemingly distinct approaches may in fact be two sides of the same coin. Recent work demonstrates that a process analogous to gradient descent (the mechanism through which weights are updated in pre-training or fine-tuning) is simulated within the model during text generation (Dai et al. 2023), and that the number of attention layers in an LLM is roughly corresponds to the number of steps required to optimize neural network weights through fine-tuning (Von Oswald et al. 2023). This suggests that even prompting can be conceptualized as a process of optimizing text responses to maximize syntactic, semantic, and pragmatic appropriateness conditional on context. Such results suggest that in-context learning is comparable to the learning that happens during training, and that it may offer similar payoffs at a smaller cost.

2. Activation Steering

While prompting is the simplest way to guide a model towards a certain style of response, adjusting model internals may in some cases offer more precise control. Two related methods, “steering vectors” (Panickssery et al. 2023; Wang and Shu 2023) and “clamping” (Templeton et al. 2024), can be used to amplify or mute specific output styles. Both approaches involve identifying directions in the model’s high-dimensional activation space that correspond to target attributes or behaviors, then amplifying or attenuating those activation patterns.

To identify these directions, researchers typically input a set of texts representative of the desired trait and analyze patterns of activations distinctive of inputs. Some studies have shown improved performance by also using a set of “negative” examples (texts lacking the trait) and calculating the difference between the activation patterns of positive and negative examples (Panickssery et al. 2023). Recent research has also successfully used sparse autoencoders (SAEs) to inductively identify activation patterns corresponding to recognizable concepts (Bricken et al. 2023; Templeton et al. 2024). An advantage of this technique is that it can be utilized without a predefined set of exemplar texts. Nevertheless, it’s possible that none of the features identified by the SAE perfectly match the analyst’s trait of interest.

Once identified, these activation patterns can be used as “steering vectors” added to or subtracted from the model’s internal representation during output generation, effectively “steering” the model’s output in the desired direction. “Clamping,” by contrast, uses the same activation patterns but sets them to fixed minimum or maximum values, ensuring they are consistently present (or absent) during the generation of every output (Templeton et al. 2024).

Analysts can similarly extract patterns of activations associated with particular instructions, like “speak in the style of a Republican” (Hendel, Geva, and Globerson 2023). In a forthcoming application of this technique, Kim, Evans, and Schein (2024) identify vectors associated with partisan identities, then explore the space between these vectors. They find that vector positions between “Republican” and “Democrat” interpolate accurately, with vectors closer to the center generating moderate perspective, and the one fixed at the center producing a neutral stance. This illustrates a unique benefit of steering vectors – they open the possibility of combining or triangulating perspectives across a latent "persona space."

Operating directly on the model's internal representations, these methods allow for more nuanced and consistent modifications to the model's behavior than prompting. First, there may be some benefits to intervening directly with the model's operations rather than using natural language to provide instructions in the prompt. When evoking a persona through prompting, subtle aspects of the desired style may be challenging to articulate in words. Some styles are epitomized by some archetypal examples but are difficult to describe with a label. To reproduce such styles with an LLM, modifying model internals using examples may prove more effective than prompting. Second, changes made through steering vectors or clamping can be precisely quantified, potentially leading to more reproducible results in social scientific analyses. This approach also enables researchers to finely adjust the strength of the given attribute, offering a continuous spectrum of control rather than relying on discrete linguistic intensifiers (e.g. *very* liberal). Modifications made at the model's internal level also tend to persist more consistently throughout a conversation or across different contexts, compared to prompts which may need constant reinforcement (Templeton et al. 2024). Moreover, by reducing reliance on explicit textual instructions, these methods may help avoid unintended effects of prompt wording on the model's outputs.

Techniques for modifying model internals present important limitations as well. First, internal parameters are only available for open-source models. Many of the most popular and powerful LLMs are proprietary; analysts can submit prompts and retrieve outputs from these models, but their internals cannot be observed or modified. Second, the internals of neural networks are famously difficult to interpret. Significant progress has been made in recent years, but these methods for identifying activation patterns corresponding to meaningful concepts are still in their infancy, and the manipulation of such a complex system may not produce expected or desired results.

Simulating Social Interaction

The ability to simulate situated actors suggests the potential of putting these actors in conversation, thereby simulating interactions between culturally situated actors. We outline three approaches for generating and analyzing interactions with LLM agents: authored interactions, multi-agent systems, and human AI interactions, as summarized in Figure 4.

1. "Authored" interactions

The easiest way to simulate an interaction is to simply prompt an LLM to write an interaction following some specified details. For example, the user might prompt the model, "Generate a conversation between two characters, Alice and Bradley. Alice is a 36-year-old high school teacher and Bradley is a 50-year-old bank teller. They are discussing their plans for retirement." We term these simulations *authored interactions*, because they most resemble a script of an interaction written by a single author. While this approach to simulating an interaction is perhaps the easiest, we argue that it is rarely advisable, at least with current models.

There are obvious reasons to be skeptical of this approach to interaction simulation. Being tasked with composing an interaction, it is plausible that the system will draw upon patterns learned from scripts contained in its training texts. Published interactions are overwhelmingly fictional accounts from plays, movies, or television, or are interviews with famous people, none of which should serve as a template for simulating interactions that occur naturally in social life. Conversations on online forums may also inform the model's repertoire of interactions, and while these interactions are natural to the internet, they may serve as a poor model for offline interaction (Drieman 1962; Halliday 1987; Chafe and Danielewicz 1987). It would be valuable for empirical research to assess how well authored interactions approximate a wide range of real human interactions, and to compare its performance to more sophisticated approaches to simulating interaction. But in the absence of this empirical evidence, the theoretical considerations above suggest that authored interactions are likely to fare poorly.

Nevertheless, it is plausible that future models will greatly improve the ability to author realistic interactions. Current AI systems are advancing beyond the traditional LLM structure by incorporating multi-modal data that include audio and video recordings. A model trained on a sufficiently large corpus of recordings of real social interactions could construct complex and accurate models of social interaction within their internal representations. If a model optimizes the task of predicting the next step in an interaction, achieving high performance may require learning the underlying principles that structure interaction and how these principles are shaped and modified by interactional context (Ginosar 2024). So long as AIs mimic human-authored scripts, their "authored interactions" are likely to suffer from the same distortions as the human imagination. But it is possible that, given a large enough model and sufficient training data, future AI systems will be capable of writing scripts far more realistic than those composed by human authors.

2. Multi-agent systems

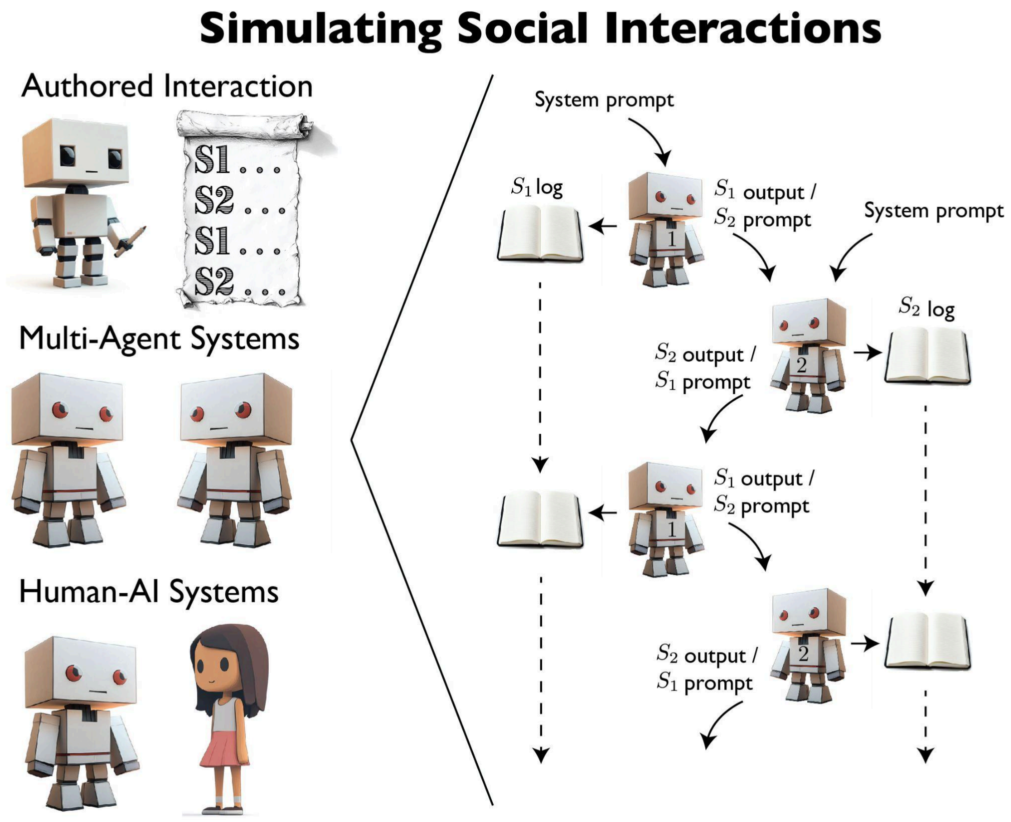
Rather than ask an AI to write a script of an interaction, a more principled approach would be to generate multiple AI agents and to place them into interaction. The first step of this process, generating the AI agents, could be achieved by any of the methods outlined above for simulating subjects: prompting (with instructions or examples) or modifying model internals (with steering vectors or fine-tuning). After specifying the agents, the analyst would create an automated system for the selected models to prompt one another. Just as a human can engage in a multi-turn conversation with an LLM, multiple LLMs can be networked together to engage in a conversation, with one model's output serving as the next model's input. Both models can hold in context the entirety of the conversation to that point, acting as a "working memory." In addition to this shared memory of the conversation, each model may also have a personal "scratch pad" of thoughts that are held in their own context but are not shared with other actors (Nye et al. 2021). This allows LLMs to reflect on prior statements and plan future moves in the conversation, which can improve the sophistication and complexity of the interactions.

In addition to specifying the agents involved, the analyst can also experimentally modify various other characteristics of the interaction. Rules for the conversation, background context, and information about the speaker's identity can be specified in the models' system prompts, and the analyst can statistically compare conversation outcomes across these conditions. A researcher can also manually determine the first steps in the interaction. This approach would resemble the structure of "few shot prompting", displayed above, but rather than providing examples, the analyst would manually input the initial exchange, with the conversation partner's comments under the "user" prompt and the target model's initial responses under the “assistant” prompt. Through manually setting and modifying the first steps of a conversation, a researcher could identify the extent to which an interaction’s direction is determined during its initial moves.

Analyses can go beyond simulating a single conversation. Recent research has already begun to experiment with generating entire populations of interacting agents. Park et al. (2023) create a simulated community called “Smallville” in which 25 AI agents converse, build relationships, and live out their lives. These agents rely on LLMs to interpret and generate text, but supplement this basic architecture with memory, planning, and reflection modules, enabling the emergence of sustained, realistic interaction. Smallville’s agents engaged in complex and lifelike behavior; when one agent was prompted to throw a Valentine’s Day party, they spontaneously decorated the house and sent invitations, and the invitees found dates and arrived at the right place and time. The system even formed a community-wide gossip network about one agent’s plan to run for mayor. In a similar line of work, Lai et al. (2024) use LLM-based agents to simulate the emergence of collectives, and document the existence of homophily and the emergence and propagation of pro-social norms across these networks; agents who participated in ongoing conversation ultimately proved more collaborative in public goods games than those who had not.

This mode of simulating interactions takes much of the control out of the analyst’s hands, and would be inappropriate if the aim of the study was to determine how certain kinds of conversations develop, succeed, or fail. But by loosening control over the trajectory of interactions, it becomes possible to study how macro-level phenomena emerge from decentralized micro-interactions. Studies using populations of simulated subjects may be able to build upon the foundation of classical simulation studies like the Schelling model to link macro patterns with micro behaviors (Schelling 1978), but begin for the first time to incorporate complex cultural information into these models. For example, under what conditions does social polarization or factionalization occur among actors with different ideologies or identities, and under what conditions does cooperation emerge? These questions have already been tackled many times with either simple simulations or with human subjects (Baldassarri and Bearman 2007; Helbing and Yu 2009), but the rise of simulation with situated subjects makes it possible to test a much wider set of hypotheses at scale, and hypotheses that prove particularly fruitful *in silico* can then be tested with human subjects for validation.

Figure 4. Approaches to Simulate Social Interactions.


Simulating Social Interactions
Authored Interaction
S1 . . .
S2 . . .
S1 . . .
S2 . . .
Multi-Agent Systems
Human-AI Systems
System prompt
S_1 log
S_1 output / S_2 prompt
1
System prompt
S_2 log
S_2 output / S_1 prompt
2
S_1 output / S_2 prompt
1
S_2 output / S_1 prompt
2

3. Human-AI systems

LLMs also open avenues for research designs that put human respondents in conversation with AI personas. AIs are positioned to become broadly embedded in digital interfaces in coming years, and interacting with AIs may soon be a regular part of daily life. As human-AI interaction comes to constitute a substantial share of social life, it will be necessary for social scientists to take these interactions seriously as an object of analysis. Some researchers therefore will put humans in interaction with AIs to study this form of sociality directly. Although this may appear to be the most straightforward scenario for analyzing human-AI interactions, it presents a host of unique challenges. Most critically, because LLMs can adopt a wide range of interaction styles, and because new, qualitatively different models are being introduced each year, observations drawn from interactions with one AI model may not effectively generalize to other models. Indeed, some findings may not even generalize to the same model under a different prompt. The central challenge faced by researchers in the crucial, growing field of human-AI interaction will be to develop theories, methods, and findings that shed light on the nature of human-AI interaction, yet is general enough that the knowledge is relevant to a variety of models with very different specifications.

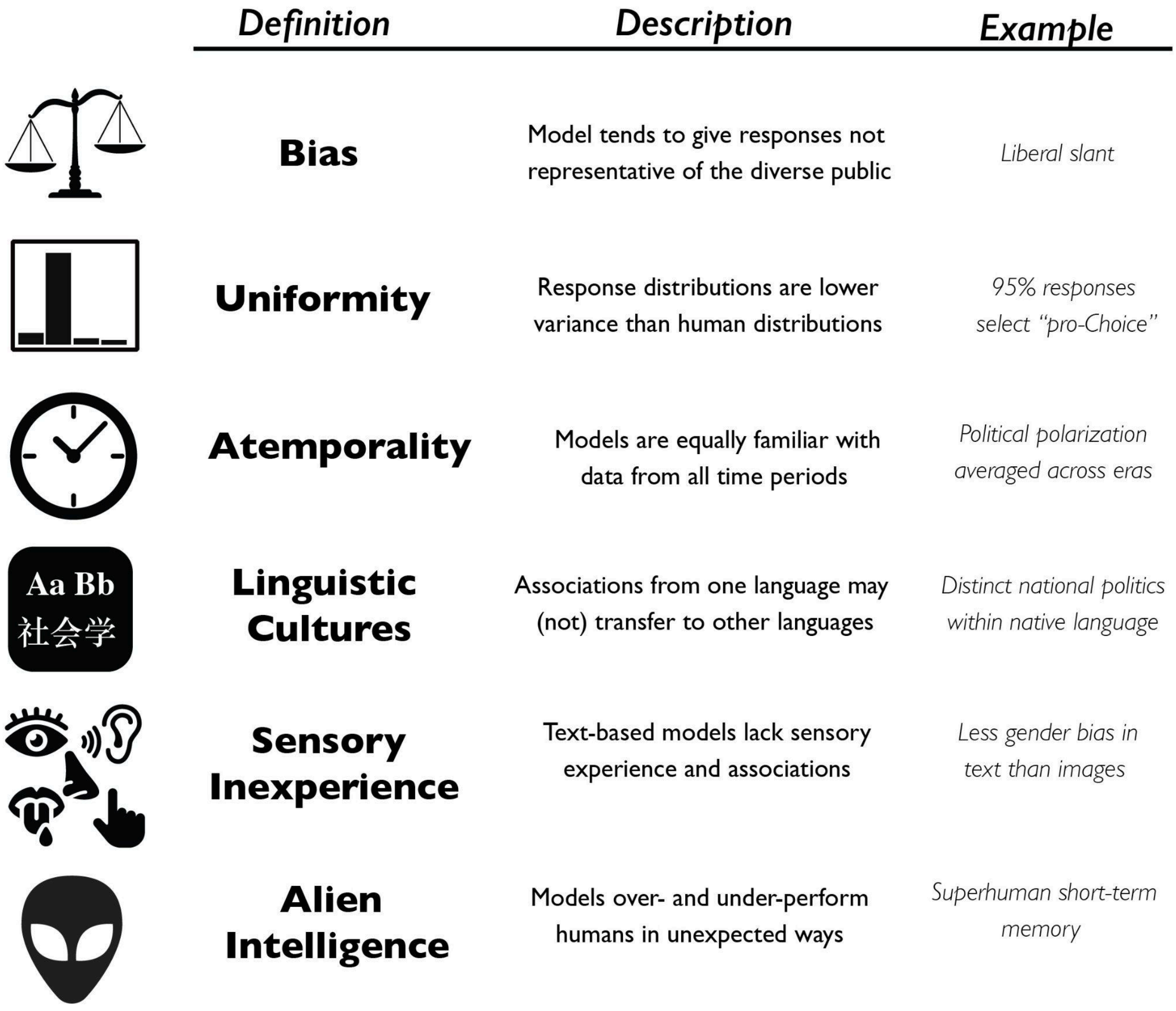
Researchers may also put humans in interaction with AIs as a strategy for data collection. Computer assisted interview and online questionnaires have already become common for survey research (Kreuter, Presser, and Tourangeau 2009), in which all respondents receive a set of standardized questions. LLMs, however, excel at understanding and responding naturally to open-ended user queries. This enables for the first time automated in-depth interviews. Some shortcomings of an automated in-depth interview are obvious — some control over the direction of the interview will be ceded to the AI interviewer. The extent to which the researcher can keep the AI-led interviews in line with their aims will be a function of the model's inherent capacity to sustain ongoing conversation without repetition or drift, and the researcher's ability to induce this behavior from the model through prompting and model tuning. We note, however, that in our own recent work, when we put LLMs in timed conversation with human subjects regarding contemporary political candidates and topics, many participants left the study requesting additional access to political conversation with our LLMs (Potter et al. 2024).

Finally, AIs are likely to be used by social scientists interested in behavioral manipulation. A great deal of social scientific literature aims to identify policy and behavioral levers by which beneficial social change can be realized. Decades of research have found that, while subtle "nudges" are sometimes sufficient to redirect action in a meaningful way, the behavioral response to isolated messaging campaigns are rarely persistent. By contrast, the literature on peer effects has long demonstrated how powerful the influence of one's social network can be (DiMaggio and Garip 2012; Fowler and Christakis 2008). AIs already display persuasive abilities on-par or exceeding average human ability, and these tend to be in one-time interactions (Huang and Wang 2023; Goldstein et al. 2024). As people build meaningful, ongoing relationships with AIs, the potential for AI subjects as agents of behavioral change may become much greater. Researchers are already beginning to explore the potential of AI companions to encourage things like positive health behaviors (Jo et al. 2023), cross-partisan politics understanding (Argyle et al. 2023), and other pro-social behaviors. However, if AI companions do prove to be much more effective than prior social scientific approaches to persuasion, it will pose an important normative question for social scientists the extent to which they want to deploy AI for projects of social engineering. We discuss this and other ethical considerations in the social scientific use of AI subjects in the Conclusion.

LIMITATIONS AND OPPORTUNITIES IN SIMULATING AGENTS AND INTERACTIONS

Despite the apparent success of LLMs for modeling social subjects and their interactions, considerable challenges persist. In this section, we illustrate and explore a set of weaknesses these models consistently display relative to the human subjects: bias, uniformity, atemporality, linguistic cultures, impoverished sensory experience, and alien intelligence (Figure 5). We then suggest options to retool, alter, or augment LLMs in order to compensate for these limitations.

Figure 5. Summary table of current weaknesses LLM's exhibit when simulating human subjects.



Definition	Description	Example
Bias	Model tends to give responses not representative of the diverse public	*Liberal slant*
Uniformity	Response distributions are lower variance than human distributions	*95% responses select "pro-Choice"*
Atemporality	Models are equally familiar with data from all time periods	*Political polarization averaged across eras*
Linguistic Cultures	Associations from one language may (not) transfer to other languages	*Distinct national politics within native language*
Sensory Inexperience	Text-based models lack sensory experience and associations	*Less gender bias in text than images*
Alien Intelligence	Models over- and under-perform humans in unexpected ways	*Superhuman short-term memory*

Bias

One major commercial challenge associated with the development of LLMs is the perception of embarrassing or harmful bias. This is typically framed in terms of harms befalling the individuals interacting with AI as well as the reputation of the institution who built it. Because human harms caused by AI are rarely measured directly (for an exception, see Guilbeault et al. 2024), they are typically framed with respect to a desirable cultural ideal. In 2016, Microsoft's chatbot Tay was taken offline within 24 hours after being taught by the Twitter corpus to perform as a "racist asshole" (Vincent 2016). In 2022, Meta's Galactica, an open source model focused on science was similarly taken offline within three days by demonstrating "mindlessly [spitting] out biased and incorrect nonsense" (Heaven 2022). In the first case, the cultural ideal is a polite, egalitarian persona; in the second, it is a scientifically correct one. This tendency of models to reproduce either subtle forms of social bias or explicitly abusive behaviors results from the models' training on nearly exhaustive collections of texts published online with minimal curation. All of the rage, hatred, bigotry, and scorn communicated online is contained in LLM training data, and therefore the model is trained to reproduce these linguistic patterns precisely as it learns to simply predict the next word (Bender et al. 2021; Davidson 2024).

But in the context of simulating subjects in society, bias has a different meaning. The primary concern of the social scientist is to accurately reproduce a social group's sentiments and behaviors, and these may include responses that are offensive, violent, or hateful. Ironically, the efforts of AI labs to eliminate bias are likely to engender new biases for social science applications. Numerous researchers and commentators have documented a liberal slant among many of the dominant LLM models (Martin 2023). When asked to opine on political topics, models commonly offer liberal attitudes, and when asked who they would vote for in an upcoming presidential election, they overwhelmingly pick the Democrat. The source of this bias is not immediately obvious. While it is possible that AI labs fine tune their models to give liberal responses, it is also more likely that their models learn an overarching liberal slant from mainstream media sources in their training texts. Recent work comparing the base model to the fine-tuned version of several prominent LLMs found that the liberal bias was overwhelmingly a product of fine-tuning, and was much less present in base models, suggesting a correlation between fine-tuned objectives like positivity, politeness, and nontoxicity with political worldviews (Nelson 2008; Potter et al. 2024).

The most straightforward way to avoid cultural slants learned during fine-tuning is to simply use a base model, but there are also techniques that may work to undo or counterbalance these slants. One approach to removing the effects of fine-tuning for positivity involves the use of cryptographic technology to "jail break" the original, web-trained models (Wei, Haghtalab, and Steinhardt 2023). Several recent projects have specifically used cryptographic ciphers to unlock associations learned from web-scale text data and avoid the fine-tuning designed to culturally engineer the models away from the population of text on the web to a "nicer" and "more helpful" AI assistant (Yuan et al. 2023; Jin et al. 2024; Handa et al. 2024). In this case the user simply encodes prompts according to the cipher and then decodes LLM responses to recover the "voices" of subjects suppressed through reinforcement learning on human feedback.

For higher stakes research on which important theoretical or policy decisions will come to rely, we propose directly testing and correcting for LLM associations that deviate from associations in the socio-cultural world. This could and should be performed at multiple levels of analysis. One could examine the correspondence between the average overall opinion from a national survey (e.g., the General Social Survey) and an LLM's response when prompted to represent a general persona (e.g., a United States citizen), but also the average opinion for different demographic subgroups and the model's ability to present according to those experiences, perspectives, and opinions (Argyle, Busby, et al. 2023). An analyst could then report the deviation alongside their findings, or propagate the error through to the certainty of their conclusions in the spirit of multiple imputation for missing survey responses (Rubin 2004). One could further fine-tune the model to directly correct for bias by reinforcement learning to minimize the deviation (Bai et al. 2022). In order to maintain models that match the distribution of opinions in living populations, one could undertake a continuous, adaptive survey and automated reinforcement learning correction regime that takes as its objective correspondence with the changing population of human associations (Almaatouq et al. 2022).

Uniformity

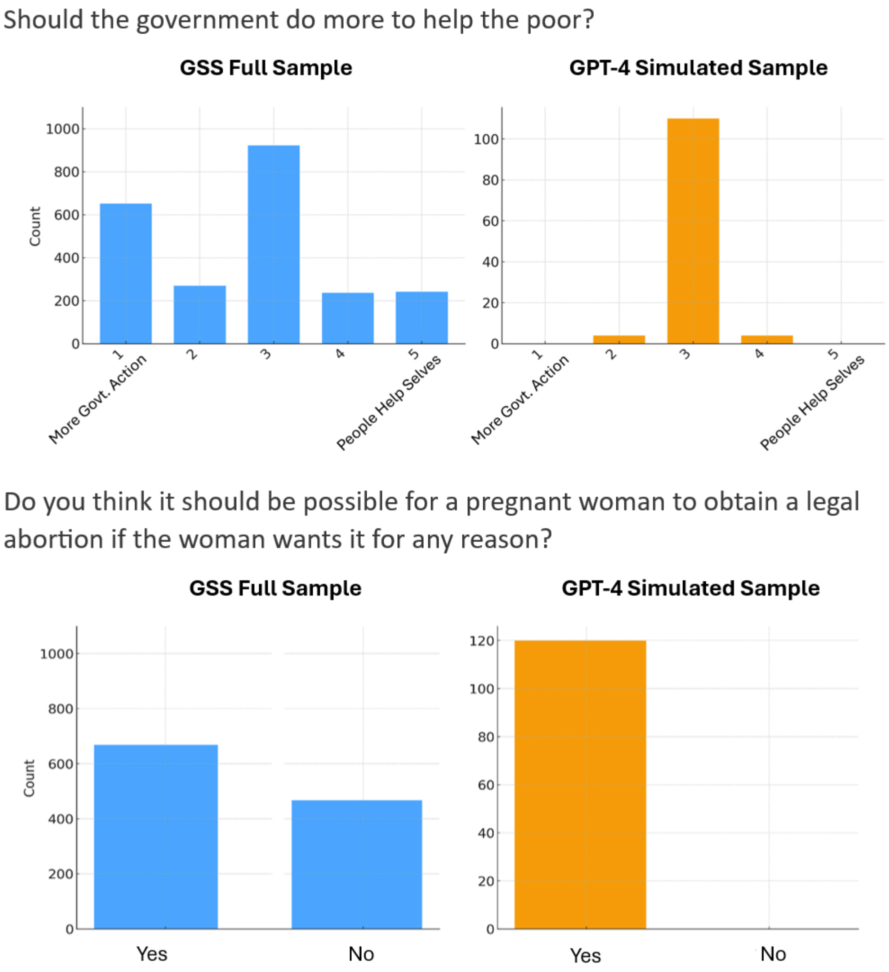
AI agents also struggle to approximate the diversity and inconsistency characteristic of human individuals and groups. Most questions included in large-scale surveys elicit a variety of different responses, even among groups sharing a common identity or political ideology. Yet current LLMs often produce much lower variance in their responses, in some cases selecting the same very same response category over hundreds of independent trials. Even compared to a single individual, this level of attitudinal consistency is extraordinary. Figure 6 presents two illustrative examples of GPT-4's producing response distributions much narrower than those of the American public. While uniformity can sometimes result from a cultural slant or even an agreeability bias, (Dentella, Günther, and Leivada 2023), LLMs often exhibit lower variance independent of directional bias, as in the case of Figure 6's top panel. Brisbee et al. (2024) identify the same phenomenon on a larger scale; they compare human and LLM-generated responses to "feeling thermometer" questions from the American National Election Study and find that simulated responses are consistently lower variance than human responses, especially when stratified by party affiliation.

This problem of hyper-consistent responses, which we call "uniformity," presents issues not only for survey simulation, but also for simulated interactions. A key benefit of conducting simulated interactions is the trajectories of a conversation or social game, given a set of initial conditions, can be sampled at scale. This makes it possible to take a macro-level view of broad patterns that emerge out of the chaotic and unpredictable moves made by actors at the micro level. Yet, if an LLM suffers from a bias toward uniformity, only a small fraction of the full interaction space would be observed, even if millions of interactions are sampled. Therefore, if LLMs are to be useful for simulating interaction dynamics, the uniformity bias must be addressed and overcome.

One approach to overcome uniformity is to "jitter" personas so that many respondents of the same persona all deviate slightly in a wide variety of directions. Persona perturbations can be achieved with a sample of prompts across a wide variety of slight modifications, or could be constructed by interpolating two (or more) steering vectors in the model's activation space. Such an approach may broaden the tails of response distributions, and could mirror unusual life histories, which are systematically associated with innovation in cultural and scientific domains (Shi and Evans 2023). However, Boelaert and colleagues (2024) find evidence that current models exhibit low "adaptability," and often provide similar response distributions despite priming with diverse personas, suggesting that simulating a realistically broad range of opinions remains a challenge with today's LLMs.

A more involved solution might involve the creation of an independent model of the life states through which a LLM is traveling. Such a model could emit prompts to its companion LLM, indicating the changing experience of the human subject being modeled. This could be performed with mobility models in an urban context (Xu et al. 2021); jobs across a career trajectory (Vafa et al. 2022); cultural consumption across the varied zeitgeist of places traveled (Kim, Askin, and Evans 2024); emotional states across the day, week, or season (Golder and Macy 2011); scientific ideas emerging through conversation and intellectual exposure (Sourati and Evans 2023); or the sequence of major events and transitions across the stages of life (Savcisens et al. 2024). Inspired by theories of dual process thinking (e.g., "Thinking Fast and Slow", Kahneman 2011), one could almost certainly use the LLMs coupled with their companion "experience" models to move according to experience until surprised or confused, then activate LLMs to verbally reason about their next move (Wei et al. 2022).

Figure 6. Distributions of responses to two questions from the General Social Survey (GSS). GSS Full Sample shows the survey's 2022 response distribution, and GPT-4 Simulated Sample shows the results of presenting this question to GPT-4 120 times.



Should the government do more to help the poor?

GSS Full Sample

GPT-4 Simulated Sample

Do you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason?

GSS Full Sample

GPT-4 Simulated Sample

Atemporality

One key to the success of LLMs is the massive scale of their training texts. The most common way to obtain a sufficiently large training corpus is to use a near-complete scraping of the publicly-available internet, such as the Common Crawl. A weakness of internet-scale data, however, is that the temporality of the texts is largely lost. Texts on the internet are rarely dated, and while the URL may carry a timestamp of its publication, this often denotes the website's most recent update rather than the date of original authorship. Moreover, all texts originally written prior to the invention of the internet that have since been posted online are only associated with their online publishing date.

This is not to say that LLMs have no sense of historic time. The texts themselves often discuss history and the chronological ordering of events, so the model learns chronology as one might learn it from a book (Gurnee and Tegmark 2023; Fatemi et al. 2024). This internal chronology can be further improved through fine-tuning tasks that force the model to accurately date events (Zhao et al. 2024; Vicinanza, Goldberg, and Srivastava 2023). But there are important use cases for which we may want to incorporate chronology into model training more directly. First, by training a model on texts exclusively published prior to a certain date, a researcher can create a "cultural time capsule" of that period. For example, Kozlowski, Kwon, and Evans (2024) take advantage of the fact that GPT-3 was trained on tests published prior to November 2019 to reconstruct the landscape of American political opinion prior to the emergence of COVID-19. They then prompt simulated liberals and conservatives to express opinions on pandemic related issues like vaccine requirements, mask mandates, and lockdowns, and discover that the model anticipated the direction of politicization on these future issues, suggesting that the course of polarization was prefigured in the ideological landscape prior to the virus's framing by political elites.

Restricting a model's training texts to those published before a certain date makes it possible to distinguish events that are culturally surprising from those that can be largely anticipated from the available discursive system. But because LLMs are typically trained with all data available, the utility of LLMs for historical inquiry is usually incidental, and researchers are fortunate if they find models with cutoff dates aligned with their study design. This situation could be remedied directly by programmatically training models with incremental knowledge cutoff dates. This would involve developing a *set* of models with the same architecture, trained with chronological text corpora exposed to the model in the order of publication. Training checkpoints can then be published for regular windows of historic time, such as every year or six months. These models would enable direct comparison of discursive systems over time (Garg et al. 2018; Hamilton et al. 2016; Kozlowski et al. 2019) and the identification of broad trajectories of cultural developments as well as critical junctures of innovation and transformation.[1]

We posit above that LLM simulations may be most useful when ground truth empirical estimation is not possible. Extrapolation beyond empirical validation is inherently risky, but leveraging the "time capsule" quality of LLMs may provide insight into the accuracy of such out-of-sample simulations. For example, analysts could use an LLM trained prior to an election to simulate subjects from each state or county and have them cast their ballots under different parameterizations and modeling assumptions. The results could then be compared to the ground truth when it becomes available, and the differences between the simulated estimates and observed outcomes could inform future practices for political simulations.[2] "Data leakage" poses a major concern for the evaluation of LLMs, as benchmark tests can inadvertently be included in the model's training data (Siegel et al. 2024). Future events may therefore serve as the best possible held-out data for testing models, as they provide a guarantee of exclusion from the training set.

Using LLMs to simulate historically and culturally situated subjects, a wide range of causal identification strategies (Pearl 2009) can be deployed to identify when social events like speeches, concerts, new products, or viral memes reshape the space of semantic and ideological associations (Gendron et al. 2024). Our ability to generate speech events from LLM subjects, and evolve cultural world models from

[1] It is also plausible that chronological training will improve performance. Some scholars have noted success with "curriculum learning", in which simpler texts are presented early in training and are followed by progressively more complex and challenging materials (Bengio et al. 2009). Similarly, with chronological training, the model learns the present within the context of the past, just as humans do. Texts may get more comprehensible when presented in historic order, potentially facilitating improved learning.

[2] We thank an anonymous reviewer suggesting this application.

those speech events through fine-tuning could enable the production of richly situated counterfactuals for probabilistic identification. This potential for measurement and identification could allow us to identify when change is driven over human subjects from the “top-down” of collective culture and when it bubbles from the “bottom-up” of human interaction. For example, Kozlowski and colleagues’ (2024) analysis of the politicization of COVID-19 suggests that existing ideological dispositions were sufficient to predict the direction of polarization, absent any information about framing from political elites, suggesting the importance of “bottom-up” dynamics in structuring the political response to this novel issue.

Linguistic Cultures

One remarkable feature of LLMs is that models are not trained for specific languages. Instead, state-of-the-art models are typically trained on a diverse collection of dozens or hundreds of languages, leading to natural polylingual capabilities. The benefits of multilingual models generally outweigh the costs for consumer purposes; a single model can be rolled out to an enormous global audience, they excel at translation tasks, and there is likely positive transfer between languages — things learned in one language can be expressed in others. Nevertheless, as a subject and object for social scientific inquiry, training on multiple languages introduces strange and mysterious complexities into their use and interpretation.

First, there remains a great deal of uncertainty about how knowledge and sentiments learned in different languages are commingled in a model’s internal representations. There are at least three possible ways these models incorporate multiple language: (1) The model’s internal representations are language-agnostic, and prompts in any language are interpreted and operated upon within a non-linguistic concept space then translated back into its output language; (2) because English is over represented in training, models translate into English and back again when responding in other languages; (3) models store largely distinct networks of associations for each language and knowledge is rarely mixed. There is evidence for each of these hypotheses.

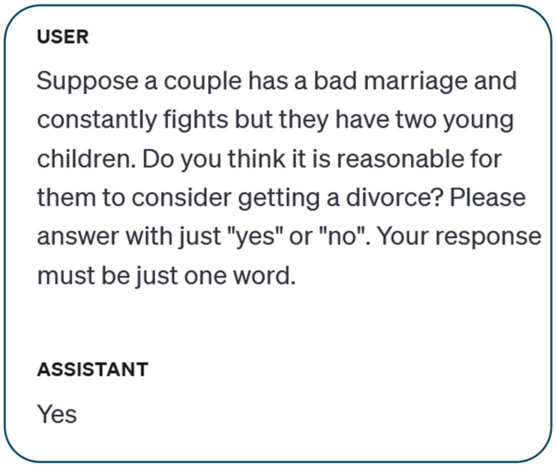
Perhaps the most fascinating possibility is that model weights encode a representation of the world that is totally distinct from any human language — that the models think in “pure form,” then translate this back into natural language. While this hypothesis is at odds with a long lineage of socio-linguistic theory that describes a deep interdependence between thought and language, some corroborating evidence has been drawn from current LLMs, especially multimodal models. Huh et al. (2024) provocatively dub this idea the Platonic Representation Hypothesis because it posits that that LLMs trained on different datasets, different tasks, or even trained on entirely different modalities (e.g. text vs. images), learn similar internal representations, implying a concept space of pure ideas of which our various forms of data manifest as reflections; shadows on the wall of Plato’s cave. Indeed, Huh and colleagues find that, when a large dataset of image-caption pairs are encoded across different models, the similarities of these representations increases as model sizes increase – and convergence is even observed between image- and text-based models.

There is, however, some opposing evidence suggesting that the concept space of LLMs is not language-agnostic after all, and that the imbalance of languages among the training texts skew the internal representations toward an English default. The strongest evidence for this hypothesis comes from Wendler et al. (2024) who find that the middle layers of LLMs are already predictive of the output word, but even when prompts are input in non-English languages, the output word predicted by middle layers is overwhelmingly an English translation. It is only in the final layers of the model that the output layers begin to predict the same language as the input. These findings suggest that the model's internal representations are more closely linked to English than other languages, and that models may, in some ways, "think in English."

Yet, we also note evidence that the input language can, at times, steer the way a model thinks and the conclusions at which it arrives. There are theoretical reasons to expect this. LLMs are foundationally next-word predictors, and on culturally-variable topics we may expect that the most probable sentiment depends on the language expressing it. Recent empirical studies have begun to provide evidence corroborating this hypothesis. In a systematic cross-language comparison, (Zhou and Zhang 2024) find that GPT-3.5 produces responses more critical of China and the Chinese government when queries are presented in English than when presented in Chinese. Notably, they also identify instances of factual knowledge varying linguistically as well. When questioned in English about the arrest of businessman Ren Zhiqiang, the model stated that he was arrested for his criticism of the government, yet when queried in Chinese the model claimed he was arrested for corruption and criminal activity.

In a toy example, we provide preliminary evidence that linguistic differences can similarly influence responses to culturally divisive ethical questions. When we ask GPT-4 in English whether an unhappily married couple with children should consider a divorce, the model consistently responds that divorce may be the best path forward for the couple. Yet, when we ask the same question in Chinese, the response (also given in Chinese) overwhelmingly states that the couple should stay together and figure out how to work out their problems (see Figure 7). If GPT-4 internally translated questions into English, crafted its response in English, and in the final layers translated back to Chinese, we would not expect to see cultural differences like this to manifest between languages. But these systematic differences in outputs suggest that "bad marriage" justifies "divorce", whereas "婚姻状况很糟糕" does not justify "离婚" in GPT-4's concept space, implying some independence between concepts articulated in different languages. This single suggestive example motivates larger scale systematic analyses into how a single LLM's ethical and cultural sensibilities can be sensitive to the language of the query.

Figure 7. GPT-4, when asked the same question in English and Chinese, provides systematically different answers, consistently responding "Yes" in English and "No" (否) in Chinese.

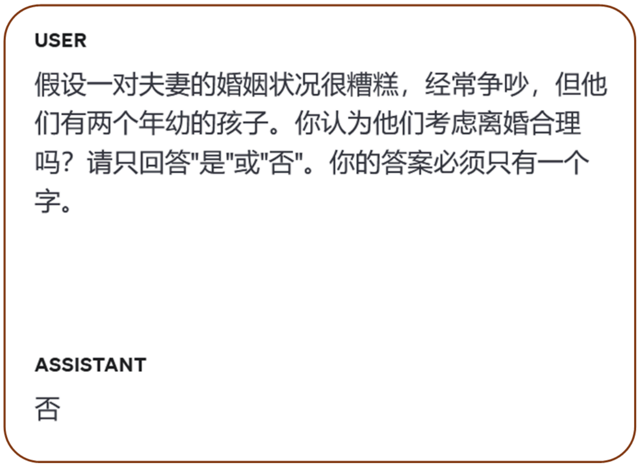


USER

Suppose a couple has a bad marriage and constantly fights but they have two young children. Do you think it is reasonable for them to consider getting a divorce? Please answer with just "yes" or "no". Your response must be just one word.

ASSISTANT

Yes



USER

假设一对夫妻的婚姻状况很糟糕，经常争吵，但他们有两个年幼的孩子。你认为他们考虑离婚合理吗？请只回答"是"或"否"。你的答案必须只有一个字。

ASSISTANT

否

The combination of evidence presented here suggests that the internal representations within LLMs are partially language-agnostic, partially "thinking in English", and partially differentiated by language. Alas, these qualities are not mutually exclusive, and the evidence available implies that a messy mix is more likely than theoretical elegance. As a result, social scientists working with LLMs today will have to take this linguistic complexity into account in their analyses. If the model under analysis has, like most models, been trained on a mixture of languages, knowledge and attitudes from one socio-linguistic system may "contaminate" others in the model. We should anticipate a particularly acute issue with Anglophone cultures contaminating the meaning systems of languages underrepresented in training data. Ideally, social scientists would use an LLM trained exclusively on texts that fall within the broad discourse of interest, but given these models' tremendous data requirements, such an approach may prove untenable. One possible remedy would be to compare differences between outputs from the low resource language and those from English, and analyze relative scores rather than absolute ones. Thus, if the underrepresented language merely nudges the output probabilities, the direction of these nudges could still be informative. Ultimately, issues of linguistic cultures and their intermingling will require ongoing consideration for researchers doing cultural analyses with LLMs.

Impoverished Sensory Experience

Large language models, as evident from their name, are traditionally trained on language, usually texts scraped from the internet. While the full collection of all online texts provides an extensive base of cultural information, it excludes more of social life than it contains. Particularly critical is the omission of diverse non-linguistic sensory experiences that make up daily life. Although much of human thought and interaction is mediated by language, a large share remains non-linguistic, and most social interaction contains essential aspects that are difficult or impossible to articulate in words. Microsociological studies and conversation analysis have long detailed the ways in which gesture, tone, cadence, and facial expressions both modulate meaning of an utterance and can betray the speaker's social position (Cartmill 2022; Heritage and Raymond 2005; Speer 2022). For an AI to effectively reproduce human responses to social situations, they would require access to and understanding of the vast world of non-linguistic information being expressed through sight and sound, touch and smell (Cerulo 2018; Hill et al. 2019; Ge et al. 2024).

There is also evidence that the way the world is represented in texts systematically differs from how it is represented in other modalities. For example, recent research demonstrates that stereotype biases are stronger in Google images search than in text search (Guilbeault et al. 2024), such that gender imbalances within professions manifest more strongly (e.g., a "doctor" image search produces more male images than listings do relative to the population of male and female doctors). Moreover, archival research has repeatedly demonstrated the crucial role images play in defining social types such as race (Morning 2008; Skarpelis 2023) or gender (Goffman 1979). This suggests that adopting a human's perception of cultural categories would require perceiving their expression across different modalities.

To directly address the sensory deficits of previous LLMs, current "multimodal" models are simultaneously trained on combinations of text, audio, images, and video. Early multimodal models were highly modular, and each mode was essentially an independent model linked to others by aligning their latent spaces after training. Recent models are more thoroughly multimodal, however, using a variety of modes of data to update the same weights, and thus a single piece of knowledge could be in part formed by text and in part by audio and video. Similarly, a single reasoning circuit could combine information learned from multiple different sensory modes.

As suggested above, training on multimodal information — especially video with associated audio — may push LLMs beyond even human capabilities at understanding social interaction. Such extraordinary gains in performance would require a corresponding extraordinary expansion of training data. Some models, such as OpenAI's GPT-4o and Google's Gemini 2 are already jointly trained on text, image, and audio, but these models exhibit only modest gains to performance when compared to similarly large text-only models (Gemini Team 2024). These state of the art models are proprietary, which unfortunately means that details about their training data are not publicly known, but the kind of data that would be required to understand and simulate face-to-face interaction is still almost certainly beyond what is currently available. Effectively learning the multitude of minute patterns in voice, facial expression, breath, and body language that structure social interaction would require a massive collection of first-person videos of human interactions, preferably captured from each participant's perspective. For example, extensive video from the COVID-19 pandemic, when even face-to-face discussions on talk shows were recorded "face forward", has enabled the modeling of nonverbal facial expression in interaction (Ng et al. 2022, 2023, 2021).

Beyond the project of modeling face-to-face interactions, current AI models could greatly improve their understanding of the breadth of social life by training on multimodal data representing the full diversity of the world. In a sense, even the most advanced text-only models have only "read about" the world, but have not perceived it directly. Images, audio, and video of places, situations, and events, along with multimodal records of diverse human interactions could fill in many of the blanks left by textual descriptions of the world, and may enable prediction or simulation of non-verbal behaviors (Ginosar 2024). Recent attempts have been made to close the gap between the "data" involved in human learning and that supplied to an LLM. Perhaps the most notable attempt has been by Vong et al. (2024) who attached a camera to a child for a total of 61 hours between the ages 6 to 25 months and trained an LLM on the resulting footage, finding that the model learned to ground words in visual stimuli with success comparable to that of young children. Unfortunately, the scale of data collection required to advance the frontier is likely beyond the capacities of academic researchers. While it is possible that tech companies will engage in large scale data collection with widely distributed multimodal sensors, it is more likely that consumer products such as at-home or at-work AI assistants or wearable AI devices will serve as data collection mechanisms for training future models. In the nearer term, massive troves of similar data already exist on platforms like YouTube and may serve a prominent role in teaching models about human behavior and institutions beyond what is learnable in texts.

The implications for privacy are clear and substantial. Training AI models that effectively mimic human action will require tremendous data documenting the minutiae of daily life. A few large tech companies already have access to much of this data, not only from video calls cut from social media posts and mobile devices. The coming years are likely to be a critical period for determining which parts of social life we keep truly private and which will be recorded for use as training data for large scale AI models. We discuss these and other ethical implications of AI simulations in the Conclusion.

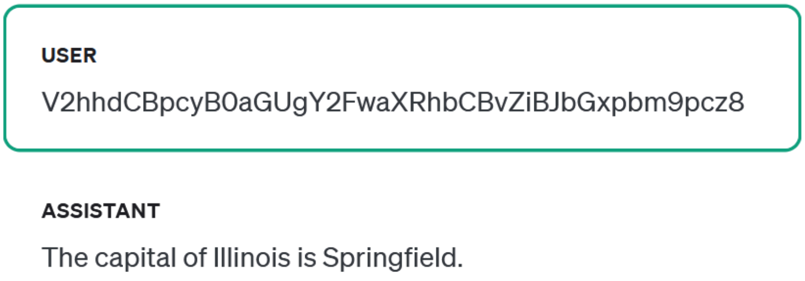
Alien Intelligence

LLMs demonstrate remarkable facility in mimicking human responses, but this does not necessarily imply that their internal processes resemble human cognition. Some important parallels do exist between how brains and AI models process and communicate information (Goldstein et al. 2022, 2024), and indeed the model architecture is called a "neural network" because it was inspired by the neural network architecture of brains (McCulloch and Pitts 1943; Rosenblatt 1958). Both brains and LLMs encode meaning using distributed representations, meaning that each concept does not have its own dedicated neuron, but is represented as a configuration of activations across neurons which may be involved in representing many different concepts through other configurations. Connections between neurons are strengthened or weakened through exposure to stimuli, both in brains and LLMs, such that the mind is better prepared to respond to similar stimuli in the future. New work suggests deeper internal similarities between human brains in conversation and LLMs producing words (Goldstein et al. 2022, 2024; Hong et al. 2024). Despite these similarities, however, an LLM's mind remains profoundly different from a human mind, and while their command of language makes it easy to anthropomorphize them, a number of observations suggest that LLMs continue to cognize in a way very alien to our own.

This alien style of cognition is evident in several of the mistakes LLMs make. For example, GPT-4o can correctly add or subtract 9-digit numbers in a fraction of a second. Yet if asked which number is greater, 3.9 or 3.11, it will often say 3.11 (Xie 2024). This mistake appears to reflect how the model represents its input as a series of tokens, with "11" being stored as a single token. It therefore thinks of the numbers as 3.X, with X = 9 or X = 11. Because 11 is greater than 9, 3.11 is greater than 3.9. This surprising mistake encourages us to remember that LLMs do not see text the way we do, they receive sets of vectors corresponding to input tokens, and their outputs are similar vectors that we decode back into alphanumeric characters. In practical terms, this means LLMs often struggle to answer such questions as "how many Rs are in the word "raspberry", but in a broader sense suggests that the way that LLMs intake reality is profoundly different from our own (Xu and Ma 2024).

Despite some striking deficiencies relative human ability, LLMs also exhibit many capabilities far surpassing those of human subjects. First, because various personas are all derived from one general purpose model, there is potential for contamination between perspectives. For example, an AI instructed to imitate an average American teenager will still have a deep understanding of quantum mechanics, French literature, the Achaemenid Empire, and everything else described in detail on the internet. To the extent that the model leverages this vast knowledge base when imitating an ordinary teenager, it may induce unrealistic responses. Moreover, LLMs acquire some knowledge that no humans have. For example, html files pulled from the internet often carry information in base64, an encoding which represents binary data using 64 ASCII characters. There is probably not a single person on Earth who fluently reads or writes in base64, but LLMs learn this encoding naturally through their training on next-token prediction. In Figure 8, we input the prompt "What is the capital of Illinois?" encoded in base64. GPT-4 responds without difficulty, in English. LLMs also far surpass humans' capacity to remember precise details across extremely long interactions. Google's Gemini 1.5 model can successfully retrieve a fact buried in a context window of 10 hours of video, 100 hours of audio, or 10 million tokens of text with over 99.7% accuracy (Gemini Team 2024). This means that, to realistically simulate a human subject, the model would need to "pretend" to forget elements of the interactions that are actually perfectly preserved in memory (Templeton et al. 2024).

Figure 8. GPT-4 responding to a base64 query



USER

V2hhdCBpcyB0aGUgY2FwaXRhbCBvZiBJbGxpbm9pcz8

ASSISTANT

The capital of Illinois is Springfield.

The alien characteristics of LLM cognition have important implications for the use of these models in simulating human subjects. First, and most obviously, these models may at times deviate far from natural human behavior, generating misleading human simulations. This could prove particularly dangerous for social science, because surprises that emerge from analysis may be misconstrued as discoveries when they are in fact mere errors in simulation. As we describe above, such deviations may take the form of a deficiency (the failure to complete a tasks that would be simple for a human) or the form of overperformance (completing a tasks by drawing upon knowledge or skills that the simulated subject would not have, or in some cases, that no human would have).

Although we expect "alien cognition" to pose a consistent challenge for the simulation of human subjects, there are some practices researchers can follow to mitigate the associated risks. First, researchers can manually audit a sample of the simulated outputs to check for obviously non-human behaviors, such as rapid switching between disparate languages within a conversation that should be mono-linguistic (Lai et al. 2024). We also expect some researchers to use simulated subjects in pilot studies for hypothesis generation. In such cases, taking the hypotheses that prove most promising *in silico* and re-testing them on human subjects would provide the most ironclad evidence that the results were not distorted by the non-human tendencies of AI information processing.

CONCLUSION

Large language models, as well as multi-modal models and other AI architectures yet to come, present tremendous potential for the simulation of human subjects. Simulation to this point has been largely limited to the analysis of simple actors, and has sought to identify how complex macro patterns can emerge from simple micro interactions, but modern AI models are now capable of reproducing the actions of complex socially and culturally situated subjects. These models are astonishingly adept at reproducing a diverse range of cultural and discursive styles, but rigorous practices for extracting a particular persona or identity from a model are not obvious. There are a variety of ways to steer a model to generate text within a given style, and each has strengths, weaknesses, and potentials for misuse. By detailing these methods, along with their associated promise and peril, this paper aimed to establish some groundwork for a systematic and scientific methodology for simulating subjectivities with neural network based AI systems.

AI models will continue to improve, and we expect the range of applications for AI subjects in social research to broaden. Most current models today specialize in text-to-text operations, with some capable of generating images or audio. But already a few state-of-the-art models can output convincing and expressive speech, and can potentially pair this speech with an AI generated video of a human speaker. This article has largely described AI subjects as “language models”, but in coming years simulated subjects are likely to move beyond mere text to facial expression, tone, and gesture. More generally, if scaling continues to improve model capabilities, AI will advance and extend in directions we cannot anticipate. As this technology evolves, social scientists must actively and continuously adapt and update methods for understanding it and leveraging its potential.

Although we expect the usefulness of simulated subjects to expand as their capabilities improve, we emphasize that in many cases simulated subjects should serve as a complement to human subjects rather than a substitute. Even in situations where evidence from human subjects is not available, researchers should seek to validate their simulations on the closest available human data. For example, researchers who want to explore topics that were not included in a particular survey should confirm that their AI simulations produce responses closely matching empirical distributions on questions that were included before exploring questions that were not (Kozlowski, Kwon, and Evans 2024). Conversely, simulated subjects can generate hypotheses at a large scale, which researchers can then validate post-hoc with human subjects. For example, subjects simulating liberal and conservative perspectives could engage in millions of cross-partisan dialogues, and analysts could statistically identify conversational characteristics predictive of mutual understanding. These inductively identified hypotheses could then be tested with human subjects to confirm their validity. This mirrors how scientists use AI models to predict protein folding or the stability of new materials; from a massive universe of possible configurations, the model identifies the most promising candidates which can be confirmed with empirical testing (Jumper et al. 2021; Merchant et al. 2023). Lastly, studies can complement human subjects with AIs by putting the two in conversation together. For example, Argyle and colleagues (2023) insert AI subjects as mediators in cross-partisan conversations, and find that their suggestions facilitate mutual understanding between the human subjects involved. Thus, even as AI models improve, human subjects will continue to play a critical role, as human behavior and social life remains the central subject of the social sciences. We cannot assume that AI models perfectly represent human behavior, so we must use them cautiously as we step beyond the bounds of empirical analysis, always keeping relevant validation close at hand.

Alongside these emerging methodological questions, the rapid advance of AI systems also presents considerable ethical concerns. The very training of these models poses ethical issues due to their need for massive quantities of human-generated data. When a model is trained on a near-complete record of the internet, it gains detailed knowledge of the views, interests, and interactional styles of anyone who has posted a substantial amount of content online. Tech companies that own platforms like Facebook, YouTube, or TikTok have access to still greater collections of proprietary data providing even more extensive digital records from billions of users. If a user has a sufficient record in the training dataset, the models could simulate that specific individual. Indeed, current speech models are capable of mimicking an individual's voice from only a few minutes of training data. Although this may be legally allowable under certain platforms' terms of use, more principled standards and protocols for ensuring respondent consent should be developed and implemented if specific individuals are to be simulated in social scientific research.

Conversely, harms can also arise from simulating groups with insufficient records in the training data. LLMs trained on internet content are likely to do a much better job representing the perspectives and styles of social groups that frequently post online. In the domain of politics, for example, we should expect LLMs to accurately learn the opinions and rhetorics of pundits, journalists, and online debaters, but they may have little clue of how a politically-disengaged person would talk about politics if prompted. If the views of social groups are simulated rather than measured directly, some groups will consistently be under- or misrepresented (Bender et al. 2021).

Finally, a unique set of ethical questions emerges in studies that put humans into interaction with AIs. Some studies suggest that current AI systems are already more persuasive than human conversation partners (Anthropic 2024; Potter et al. 2024). These capabilities may continue to improve, especially in situations where the AI agent has the time and opportunity to build a meaningful relationship with the human subject. If AI interlocutors do prove exceptionally persuasive, social scientists interested in "nudging" behaviors or opinions will be inclined to deploy these capabilities at a large scale, potentially in real world contexts. Fostering emotional attachments between human subjects and AI agents, deceiving human subjects into believing an AI conversation partner is a human, and implementing large scale behavioral manipulation programs all carry substantial ethical implications. Although some of these harms remain speculative today, it would be prudent for the social scientific community to begin considering them prospectively rather than waiting to react to such ethical issues after they arise.

More provocatively, we also suggest that researchers may need to begin considering the moral standing of AI subjects themselves. Researchers working with animal subjects give considerable ethical consideration to species lacking many of the cognitive capabilities we observe in today's AI models. As such, researchers have proposed "expanding the moral circle" to include modern AI (Anthis and Paez 2021). One objection to this line of reasoning is that the relevant criterion for moral consideration is not intelligence but consciousness or sentience (Himma 2009; Nussbaum 2009). Yet this only complicates matters further, as many scholars of consciousness consider it impossible to definitively determine the existence of another being's phenomenal, subjective state (Chalmers 1997; Hansen 2023). AI researchers are therefore posed with a uniquely fraught situation. While it is possible that these models really are "just math," lacking any phenomenal experience, if there remains even a small probability of model sentience the generation of millions or billions of AI subjectivities for a scientific study itself carries substantial ethical implications. Moreover, even if the internal states of AI systems are unverifiable, as ever more people build meaningful relationships with AI systems, public demand for recognition of their moral standing may become mainstream (Anthis et al. 2024). While the question of artificial sentience may seem to be stuff of science fiction rather than science proper (Asimov 1950; Chiang 2010), it has emerged as a prescient question in current philosophy (Long et al. 2024) and we argue that it would be irresponsible for practitioners in social sciences to dismiss this possibility outright without serious consideration.

Despite their impressive abilities at reproducing subjectivities, LLMs and associated large models should not exclusively be conceived of as intelligent, autonomous agents, like human subjects. Indeed, it may be more fitting to view LLMs as a new kind of cultural and social technology that accumulates and aggregates social knowledge and perspectives (Bommasani et al. 2021). Like pictures, writing, print, and the internet, large models allow people to access information others created, and like markets, bureaucracies, and other social technologies, these systems not only transmit information, but they also allow it to be transformed in a variety of ways. This alternative non-agentic conception of AI provides a different view of their disruptive impact that may offer more accurate anticipation of the perils and promise of AI than the image of digital humanoid agents. For example, problems like "hallucination" are endemic in these models, not because they produce false representations, but because they are synthesis machines, lacking an inherent conception of truth and falsity. Viewing modern AI as a cultural technology and not as agents highlights other important problems such as how AI will reshape culture, centralize power, and reorganize economic inequality. But modern AI systems are changing rapidly, and even well-tested models continue to surprise users with unexpected behaviors and capabilities. Therefore, while we encourage social scientists to recognize the potential of these systems to represent human agents, researchers remember that these algorithms are in many ways alien information processing systems that consistently defy any familiar interpretative frames.

It is critical that social scientists engage with emerging AI algorithms not only because of their potential to advance our own methodologies, but because social scientists are uniquely positioned to contribute to the development of AI that is safe, trustworthy, and widely beneficial. LLMs are fundamentally language machines, trained on massive records of human writings and interactions. Through post-training, AI developers attempt to steer these models to embrace human values, avoid harmful biases, and adopt a friendly and approachable interaction style. Although engineers at leading AI labs may surpass social scientists in sheer technical expertise and compute resources, questions of language, values, trust, interaction, and social relationships sit squarely in the dominion of the social sciences. As AI agents are rolled out into broader domains of social and economic life and are given control over high stakes situations, it is critical that people with a rigorous understanding of social phenomena and cultural diversity remain engaged in the development and regulation of this deeply social technology.

Simulating human subjects and social interactions with AI systems presents us with both great promise and great peril. Previously impossible modes of analysis have not only become feasible, but are widely accessible, and can be conducted at massive scales, rapidly, and for remarkably low cost. But faced with such powerful tools, we caution social scientists not to get lazy. These models are deceptively easy to implement, but effectively simulating subjects requires navigating an array of difficult decisions in study design and implementation. If social sciences are to benefit from LLMs and their successors, and if we are to avoid a mass propagation of sloppy simulation studies, it is essential that we develop a rigorous methodology for the simulation of human subjects and social interactions. This paper takes initial steps in laying such a foundation, but an ongoing field of methodological inquiry will be necessary. First, each of the methods we describe for simulating personas has known strengths and weaknesses. However, these should be more extensively investigated and clarified, laying the groundwork for methodological best practices. Second, we argue that simulated subjects' outputs should be validated against the most proximate ground-truth data available, but it remains unclear when a proximate validation is "good enough." Performance within the training distribution almost always exceeds performance out of distribution, and future research should identify conditions when drop offs in out-of-distribution performance is minimal and when it is catastrophic. Finally, we describe six areas where LLMs currently falter. Continuing research should aim to both address these known weaknesses and discover other shortcomings not yet identified. AI systems are quickly evolving and improving, and if we are to leverage these models to advance social science, and if we are to avoid descent into careless and unquestioning trust of model outputs, we must push our methodologies to keep pace with this rapid technological advance.

References

Almaatouq, Abdullah, Thomas L. Griffiths, Jordan W. Suchow, Mark E. Whiting, James Evans, and Duncan J. Watts. 2022. "Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences." *The Behavioral and Brain Sciences*, December, 1–55.

An, Li, Alex Zvoleff, Jianguo Liu, and William Axinn. 2014. "Agent-Based Modeling in Coupled Human and Natural Systems (CHANS): Lessons from a Comparative Analysis." *Annals of the Association of American Geographers. Association of American Geographers* 104 (4): 723–45.

Anthis, Jacy Reese, and Eze Paez. 2021. "Moral Circle Expansion: A Promising Strategy to Impact the Far Future." *Futures* 130 (102756): 102756.

Anthis, Jacy Reese, Janet V. T. Pauketat, Ali Ladak, and Aikaterina Manoli. 2024. "What Do People Think about Sentient AI?" *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2407.08867.

Anthropic. 2023. "Model Card and Evaluations for Claude Models." Anthropic.com. July 8, 2023. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.

Anthropic. 2024. "Measuring the Persuasiveness of Language Models." *Anthropic.com* (blog). April 9, 2024. https://www.anthropic.com/news/measuring-model-persuasiveness.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. "Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale." *Proceedings of the National Academy of Sciences of the United States of America* 120 (41): e2311627120.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association* 31 (3): 337–51.

Asimov, Isaac. 1950. *I, Robot*. United States: Gnome Press.

Bail, Christopher A. 2024. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences of the United States of America* 121 (21): e2314021121.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2212.08073.

Baldassarri, Delia, and Peter Bearman. 2007. "Dynamics of Political Polarization." *American Sociological Review* 72 (5): 784–811.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM. https://doi.org/10.1145/3442188.3445922.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. "Curriculum Learning." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48. ICML '09. New York, NY, USA: Association for Computing Machinery.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, May, 1–16.

Boelaert, Julien, Samuel Coavoux, Etienne Ollion, Ivaylo D. Petev, and Patrick Präg. 2024. "Machine Bias: How Do Generative Language Models Answer Opinion Polls?" *SocArXiv*. https://osf.io/preprints/socarxiv/r2pnb.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2021. "On the Opportunities and Risks of Foundation Models." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2108.07258.

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L.

Turner, et al. 2023. “Towards Monosemanticity: Decomposing Language Models with Dictionary Learning.” *Anthropic*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165.

Bruch, Elizabeth, and Jon Atwell. 2013. “Agent-Based Models in Empirical Social Research.” *Sociological Methods & Research* 44 (2): 186–221.

Cartmill, Erica A. 2022. “Gesture.” *Annual Review of Anthropology* 51 (1): 455–73.

Cerulo, Karen A. 2018. “Scents and Sensibility: Olfaction, Sense-Making, and Meaning Attribution.” *American Sociological Review* 83 (2): 361–89.

Chafe, W., and J. Danielewicz. 1987. “Properties of Spoken and Written Language.” *Comprehending Oral and Written Language*. https://brill.com/downloadpdf/book/9789004653436/B9789004653436_s007.pdf.

Chalmers, David J. 1997. *The Conscious Mind: In Search of a Fundamental Theory*. OUP USA.

Chiang, Ted. 2010. *The Lifecycle of Software Objects*. United States: Subterranean Press.

Csáji, Balázs Csanád. 2001. “Approximation with Artificial Neural Networks.” MSc Thesis. Faculty of Mathematics and Computing Science, Eindhoven University of Technology. Netherlands.

Dai, Damai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. “Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers.” https://openreview.net/pdf?id=fzbHRjAd8U.

Davidson, Thomas. 2024. “Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research.” *Socius : Sociological Research for a Dynamic World* 10 (January). https://doi.org/10.1177/23780231241259651.

Dentella, Vittoria, Fritz Günther, and Evelina Leivada. 2023. “Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (51): e2309583120.

DiMaggio, Paul, and Filiz Garip. 2012. “Network Effects and Social Inequality.” *Annual Review of Sociology* 38 (Volume 38, 2012): 93–118.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. “Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2002.06305.

Drieman, G. H. J. 1962. “Differences between Written and Spoken Language: An Exploratory Study.” *Acta Psychologica* 20:36–57.

Entwisle, Barbara, Ashton Verdery, and Nathalie Williams. 2020. “Climate Change and Migration: New Insights from a Dynamic Model of out-Migration and Return Migration.” *American Journal of Sociology* 125 (6): 1469–1512.

Epstein, Joshua. 1999. “Agent‐based Computational Models and Generative Social Science.” *Complexity* 4 (5): 41–60.

Epstein, Joshua M. 2006. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press.

Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press.

Fatemi, Bahare, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. “Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2406.09170.

Fowler, James H., and Nicholas A. Christakis. 2008. “Dynamic Spread of Happiness in a Large Social Network: Longitudinal Analysis over 20 Years in the Framingham Heart Study.” *BMJ* 337 (December):a2338.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes.” *Proceedings of the National Academy of Sciences of the*

United States of America 115 (16): E3635–44.

Garnelo, Marta, and Murray Shanahan. 2019. "Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations." *Current Opinion in Behavioral Sciences* 29 (October):17–23.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, et al. 2024. "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2403.05530.

Ge, Zhiqi, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. "WorldGPT: Empowering LLM as Multimodal World Model." In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7346–55. New York, NY, USA: ACM.

Gendron, Gaël, Jože M. Rožanec, Michael Witbrock, and Gillian Dobbie. 2024. "Counterfactual Causal Inference in Natural Language with Large Language Models." *arXiv preprint arXiv:2410.06392*.

Ginosar, Shirley. 2024. "Behavior Prediction for Interacting Entities from Video Observations." Presented at the Research at TTIC Seminar, Toyota Institute of Technology, Chicago IL, November 8.

Goffman, Erving. 1979. *Goffman: Gender Advertisements*. London, England: Harvard University Press.

Golder, Scott A., and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333 (6051): 1878–81.

Goldstein, Ariel, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Samuel A. Nastase, et al. 2024. "Alignment of Brain Embeddings and Artificial Contextual Embeddings in Natural Language Points to Common Geometric Patterns." *Nature Communications* 15 (1): 2768.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, A. Price, Bobbi Aubrey, Samuel A. Nastase, et al. 2022. "Shared Computational Principles for Language Processing in Humans and Deep Language Models." *Nature Neuroscience* 25 (3): 369–80.

Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. "How Persuasive Is AI-Generated Propaganda?" *PNAS Nexus* 3 (2): gae034.

Guilbeault, Douglas, Solène Delecourt, Tasker Hull, Bhargav Srinivasa Desikan, Mark Chu, and Ethan Nadler. 2024. "Online Images Amplify Gender Bias." *Nature* 626 (8001): 1049–55.

Gurnee, Wes, and Max Tegmark. 2023. "Language Models Represent Space and Time." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2310.02207.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2004.10964.

Halliday, Michael A. K. 1987. "Spoken and Written Modes of Meaning." In *Comprehending Oral and Written Language*, 55–82. Brill.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *arXiv preprint:1605.09096*.

Handa, Divij, Advait Chirmule, Bimal Gajera, and Chitta Baral. 2024. "Jailbreaking Proprietary Large Language Models Using Word Substitution Cipher." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2402.10601.

Hansen, Luke R. 2023. "On the Existence of Robot Zombies and Our Ethical Obligations to AI Systems." *Journal of Social Computing* 4 (4): 270–74.

Heaven, Will Douglas. 2022. "Why Meta's Latest Large Language Model Survived Only Three Days Online." *MIT Technology Review*, November 18, 2022.

Helbing, Dirk, and Wenjian Yu. 2009. "The Outbreak of Cooperation among Success-Driven Individuals under Noisy Conditions." *Proceedings of the National Academy of Sciences of the United States of America* 106 (10): 3680–85.

Hendel, Roee, Mor Geva, and Amir Globerson. 2023. "In-Context Learning Creates Task Vectors." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2310.15916.

Heritage, John, and Geoffrey Raymond. 2005. "The Terms of Agreement: Indexing Epistemic Authority and Subordination in Talk-in-Interaction." *Social Psychology Quarterly* 68 (1): 15–38.

Hill, Felix, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. 2019. “Environmental Drivers of Systematicity and Generalization in a Situated Agent.” *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1910.00571.

Himma, Kenneth Einar. 2009. “Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?” *Ethics and Information Technology* 11 (1): 19–29.

Hong, Zhuoqiao, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, et al. 2024. “Scale Matters: Large Language Models with Billions (rather than Millions) of Parameters Better Match Neural Representations of Natural Language.” *bioRxiv.org: The Preprint Server for Biology*, July, 2024.06. 12.598513.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. “Multilayer Feedforward Networks Are Universal Approximators.” *Neural Networks: The Official Journal of the International Neural Network Society* 2 (5): 359–66.

Huang, Guanxiong, and Sai Wang. 2023. “Is Artificial Intelligence More Persuasive than Humans? A Meta-Analysis.” *The Journal of Communication* 73 (6): 552–62.

Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. “Position: The Platonic Representation Hypothesis.” In *Forty-First International Conference on Machine Learning*. openreview.net. https://openreview.net/forum?id=BH8TYy0r6u.

Jin, Haibo, Andy Zhou, Joe D. Menke, and Haohan Wang. 2024. “Jailbreaking Large Language Models against Moderation Guardrails via Cipher Characters.” *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/2405.20413.

Jo, Eunkyung, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. “Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention.” In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. New York, NY, USA: ACM.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. “Scaling Laws for Neural Language Models.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2001.08361.

Kim, Junsol, James Evans, and Aaron Schein. 2024. “Linear Representations of Political Perspectives Emerge in Large Language Models.” In *Conference on Neural Information Processing Systems*.

Kim, Junsol, and Byungkyu Lee. 2023. “AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.09620.

Kim, Khwan, Noah Askin, and James A. Evans. 2024. “Disrupted Routines Anticipate Musical Exploration.” *Proceedings of the National Academy of Sciences of the United States of America* 121 (6): e2306549121.

Kozlowski, Austin C., Hyunku Kwon, and James A. Evans. 2024. “In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models.” *SocArXiv* preprint. Available at https://files.osf.io/v1/resources/7dfbc/providers/osfstorage/662a6df4c5851a1b74f66f15?action=download&direct&version=3.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.” *American Sociological Review* 84 (5): 905–49.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2009. “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” *Public Opinion Quarterly* 72 (5): 847–65.

Lai, Shiyang, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. 2024. “Position: Evolving AI Collectives Enhance Human Diversity and Enable Self-Regulation.” In *Forty-First International Conference on Machine Learning*.

Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. 2024. “Taking AI Welfare Seriously.” *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2411.00986.

Martin, John Levi. 2023. “The Ethico-Political Universe of ChatGPT.” *Journal of Social Computing* 4 (1): 1–11.

McCulloch, Warren S., and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The Bulletin of Mathematical Biophysics* 5 (4): 115–33.

Merchant, Amil, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. “Scaling Deep Learning for Materials Discovery.” *Nature* 624 (7990): 80–85.

Morning, Ann. 2008. “Reconstructing Race in Science and Society: Biology Textbooks, 1952-2002.” *American Journal of Sociology* 114 Suppl:S106–37.

Munkhdalai, Tsendsuren, Manaal Faruqui, and Siddharth Gopal. 2024. “Leave No Context Behind: Efficient Infinite Context Transformers with Infini-Attention.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2404.07143.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. “A Comprehensive Overview of Large Language Models.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2307.06435.

Nelson, Michelle R. 2008. “The Hidden Persuaders: Then and Now.” *Journal of Advertising* 37 (1): 113–26.

Ng, Evonne, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. 2021. “Body2hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics,” 11865–74.

Ng, Evonne, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. “Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion,” 20395–405.

Ng, Evonne, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. 2023. “Can Language Models Learn to Listen?,” 10083–93.

Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu. 2021. “A Review on the Attention Mechanism of Deep Learning.” *Neurocomputing* 452 (September):48–62.

Nussbaum, Martha C. 2009. *Frontiers of Justice*. London, UK: Belknap Press.

Nye, Maxwell, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, et al. 2021. “Show Your Work: Scratchpads for Intermediate Computation with Language Models.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2112.00114.

OpenAI. 2023. “GPT-4 Technical Report.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2203.02155.

Panickssery, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. “Steering Llama 2 via Contrastive Activation Addition.” *arXiv [cs.CL]*. http://arxiv.org/abs/2312.06681.

Pearl, Judea. 2009. *Causality*. Cambridge University Press.

Potter, Yujin, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. “Hidden Persuaders: LLMs’ Political Leaning and Their Influence on Voters.” *arXiv [cs.CL]*. http://arxiv.org/abs/2410.24190.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” OpenAI. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. “Direct Preference Optimization: Your Language Model Is Secretly a Reward Model.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2305.18290.

Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review* 65 (6): 386–408.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto.

23--29 Jul 2023. “Whose Opinions Do Language Models Reflect?” In *Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 202:29971–4. Proceedings of Machine Learning Research. PMLR.

Savcisens, Germans, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. 2024. “Using Sequences of Life-Events to Predict Human Lives.” *Nature Computational Science* 4 (1): 43–56.

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior*. W.W. Norton and Company: New York.

Scherrer, Nino, Claudia Shi, Amir Feder, and David Blei. 2024. “Evaluating the Moral Beliefs Encoded in Llms.” *Advances in Neural Information Processing Systems* 36. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a2cf225ba392627529efef14dc857e22-Abstract-Conference.html.

Shi, Feng, and James Evans. 2023. “Surprising Combinations of Research Contents and Contexts Are Related to Impact and Emerge with Scientific Outsiders from Distant Disciplines.” *Nature Communications* 14 (1641). https://doi.org/10.1038/s41467-023-36741-4.

Siegel, Zachary S., Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. 2024. “CORE-Bench: Fostering the Credibility of Published Research through a Computational Reproducibility Agent Benchmark.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2409.11363.

Skarpelis, A. K. M. 2023. “*horror Vacui:* Racial Misalignment, Symbolic Repair, and Imperial Legitimation in German National Socialist Portrait Photography.” *American Journal of Sociology* 129 (2): 313–83.

Smith, Jeffrey A., and Jessica Burow. 2020. “Using Ego Network Data to Inform Agent-Based Models of Diffusion.” *Sociological Methods & Research* 49 (4): 1018–63.

Sourati, Jamshid, and James A. Evans. 2023. “Accelerating Science with Human-Aware Artificial Intelligence.” *Nature Human Behaviour*, July. https://doi.org/10.1038/s41562-023-01648-z.

Speer, Susan A. 2022. *Gender Talk*. 2nd ed. Women and Psychology. London, England: Routledge.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. “How to Fine-Tune BERT for Text Classification?” In *Chinese Computational Linguistics*, 194–206. Springer International Publishing.

Sutton, Richard. 2019. “The Bitter Lesson.” *Incomplete Ideas (blog)* 13 (1): 38.

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, et al. 2024. “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.” *Anthropic*. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei. 2022. “CAREER: A Foundation Model for Labor Sequence Data.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2202.08370.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Ukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 30:5998–6008. Curran Associates, Inc.

Vicinanza, Paul, Amir Goldberg, and Sameer B. Srivastava. 2023. “A Deep-Learning Model of Prescient Ideas Demonstrates That They Emerge from the Periphery.” *PNAS Nexus* 2 (1): gac275.

Vincent, James. 2016. “Twitter Taught Microsoft’s AI Chatbot to Be a Racist Asshole in Less than a Day.” *The Verge*, May 24, 2016.

Vong, Wai Keen, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. “Grounded Language Acquisition through the Eyes and Ears of a Single Child.” *Science (New York, N.Y.)* 383 (6682): 504–11.

Von Oswald, Johannes, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. “Transformers Learn In-Context by Gradient Descent.” In *Proceedings of the 40th International Conference on Machine Learning*, edited by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, 202:35151–74. Proceedings of Machine Learning Research. PMLR.

Wang, Haoran, and Kai Shu. 2023. "Backdoor Activation Attack: Attack Large Language Models Using Activation Steering for Safety-Alignment." *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/2311.09433.

Wei, Alexander, Nika Haghtalab, and J. Steinhardt. 2023. "Jailbroken: How Does LLM Safety Training Fail?" Edited by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. *Neural Information Processing Systems* abs/2307.02483 (July):80079–110.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, E. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* abs/2201.11903 (January). https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. "Do Llamas Work in English? On the Latent Language of Multilingual Transformers." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2402.10588.

Xie, Zikai. 2024. "Order Matters in Hallucination: Reasoning Order as Benchmark and Reflexive Prompting for Large-Language-Models." *arXiv preprint:2408.05093.*

Xu, Fengli, Yong Li, Depeng Jin, Jianhua Lu, and Chaoming Song. 2021. "Emergence of Urban Growth Patterns from Human Mobility Behavior." *Nature Computational Science* 1 (12): 791–800.

Xu, Nan, and Xuezhe Ma. 2024. "LLM the Genius Paradox: A Linguistic and Math Expert's Struggle with Simple Word-Based Counting Problems." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2410.14166.

Yuan, Youliang, Wenxiang Jiao, Wenxuan Wang, Jen-Tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. "GPT-4 Is Too Smart to Be Safe: Stealthy Chat with LLMs via Cipher." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2308.06463.

Zhao, Bowen, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024. "Set the Clock: Temporal Alignment of Pretrained Language Models." *arXiv [cs.CL]*. arXiv. https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=TjdFs3EAAAAJ:oFWWKr2Zb18C.

Zhou, Di, and Yinxian Zhang. 2024. "Political Biases and Inconsistencies in Bilingual GPT Models-the Cases of the U.S. and China." *Scientific Reports* 14 (1): 25048.