In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models

Austin C. Kozlowski [1], **Hyunku Kwon** [2], **and James A. Evans** [3]

[1, 2, 3] Department of Sociology, University of Chicago
[3] Santa Fe Institute

Abstract

By training deep neural networks on massive archives of digitized text, large language models (LLMs) learn the complex linguistic patterns that constitute historic and contemporary discourses. We argue that LLMs can serve as a valuable tool for sociological inquiry by enabling accurate simulation of respondents from specific social and cultural contexts. Applying LLMs in this capacity, we reconstruct the public opinion landscape of 2019 to examine the extent to which the future polarization over COVID-19 was prefigured in existing political discourse. Using an LLM trained on texts published through 2019, we simulate the responses of American liberals and conservatives to a battery of pandemic-related questions. We find that the simulated respondents reproduce observed partisan differences in COVID-19 attitudes in 84% of cases, significantly greater than chance. Prompting the simulated respondents to justify their responses, we find that much of the observed partisan gap corresponds to differing appeals to freedom, safety, and institutional trust. Our findings suggest that the politicization of COVID-19 was largely consistent with the prior ideological landscape, and this unprecedented event served to advance history along its track rather than change the rails.

Keywords Large Language Models · Culture · Polarization · AI

INTRODUCTION

For decades, the term "artificial intelligence" was used to describe computational capabilities that remained out of reach—in a quote often attributed to Larry Tesler, "artificial intelligence is whatever hasn't been done yet."[1] But in recent years, algorithms have gained fluency in such complex tasks as composing novel texts, generating photo-realistic images, and advanced coding, and the term "artificial intelligence" is now part of daily parlance.

This rapid progress is largely the product of two overarching technological developments. First, continuous improvements to hardware have driven decades of exponential growth in computational power. New capabilities emerge as models scale up, and modern chips now make it possible to train models with over a trillion parameters (Kaplan et al. 2020). Second, the proliferation of online digital content has supplied abundant training data. Current models are commonly trained on a near-complete record of all text on the internet, and "multi-modal" models are trained on vast collections of online images and videos as well (Brown et al. 2020). These two complementary developments have resulted in algorithms capable of generating vast and varied forms of content; deep neural architectures make it possible to learn subtle and complex patterns, and large training data provide abundant examples of patterns to learn.

Particularly striking advances have been made in the training of large language models (LLMs), algorithms capable of generating text by predicting the next word in a sequence. LLMs form the foundation of AI conversational agents such as OpenAI's ChatGPT, Anthropic's Claude, and Deep Mind's Gemini. These "chatbots" have swiftly gained widespread public exposure; ChatGPT alone reached over 100 million users within two months of its public release, and a large share of workers in fields ranging from education to computer programming report regularly using LLMs to improve their productivity (Dell'Acqua et al. 2023; Mollick and Mollick 2023). Seemingly overnight, the ability of algorithms to successfully impersonate human interaction – commonly known as "the Turing Test" (Turing 1950) – shifted from aspiration to expectation.

[1] Tesler himself maintains that this is a misquote, and that his actual statement was "intelligence is whatever machines haven't done yet," critiquing the tendency to continuously shift the criteria for intelligence to keep it one step ahead of ever-encroaching machines (Tesler 2010).

Because LLMs are typically trained on the wide variety of texts published on the internet, they learn to reproduce many distinct discursive styles. They achieve this not by memorizing specific sentences (although they occasionally do this), but by learning the latent probability distributions of word sequences constitutive of discourses. Publicly released LLMs are commonly fine-tuned to speak in the style of a helpful professional assistant and avoid making statements that are offensive, biased, or politically charged (Ouyang et al. 2022). But if prompted to do so, even these fine-tuned models can generate texts that mirror the diverse cultural and linguistic styles represented in their training data, ranging from sarcastic wisecracking to postmodern literary criticism to extreme political rhetoric (Argyle et al. 2023; Kim and Lee 2023; Park et al. 2022).

We argue that this capacity to mimic and reproduce human responses opens fruitful new forms of socio-cultural analysis. First, LLMs are able to reproduce the discourses of populations not available for interviews or surveys, including populations from the past. Second, because LLMs can quickly and cheaply generate responses, they facilitate testing wide varieties of wordings for each question, improving robustness and identifying how specific words and associated framings can steer responses (Garcia-Pardina et al. 2022). Third, they allow open-ended responses to be both generated and effectively machine-coded at scale, enabling a high-level view into the system of considerations that inform a given response. Establishing the validity of these LLM methods enables the generation of useful "social science fictions" or simulations that open up a wide range of possibilities that lie beyond the scope of this empirical investigation.

We apply this novel approach to investigate a longstanding question in the study of culture – to what extent are new cultural developments constrained by the existing ideological landscape (Converse 1964; DellaPosta 2020; Hunzaker and Valentino 2019)? More specifically, when a new issue arises, are public responses to that issue predictable given the systematization of attitudes across other topics? Answering this question requires observing how individuals respond to an emerging issue *before it is framed* by commentators, public figures, or personal acquaintances. Yet this analytic approach presents an empirical challenge, as new issues are rapidly subject to public discussion and interpretation, leaving social scientists little opportunity to measure responses prior to top-down framing by opinion leaders.

LLMs can shed new light on this question by serving as a "cultural time capsule," capable of reproducing the most plausible responses from the time of the model's training texts. An analyst can present an LLM with issues that *had not yet emerged when the model was trained* and compare these LLM-generated responses with survey responses gathered in the following years. To the extent that the LLM generates the pattern of attitudes that empirically manifest in later years, it suggests that public response to the new issue was prefigured in prior discourse, and that the public reception of the issue was a "predictable" development.

We use this method to examine a pivotal issue of our time, the public response to the COVID-19 pandemic. The spread of COVID-19 foregrounded several topics with little political precedent, such as vaccine mandates, face masks, and lockdowns. The public response to these issues rapidly politicized, with liberals endorsing cautious approaches to the virus and conservatives opposing more drastic measures (Gadarian, Goodman, and Pepinsky 2021). What remains unclear is whether this politicization reflects deep characteristics of American liberalism and conservatism as idea systems, or whether the public simply responded to the cues of partisan elites who quickly espoused opposing responses. Fortunately for our analytic aims, GPT-3, the first LLM to accurately reproduce patterns of public opinion, was trained on texts published through October 2019, and therefore has no knowledge of the COVID-19 pandemic (OpenAI 2023b). We can therefore use GPT-3 to reproduce how liberals and conservatives *likely would have responded* to questions pertaining to a pandemic before any top-down framing emerged around COVID-19 (Kaplan 2008).

We find that liberal and conservative responses simulated with GPT-3 largely anticipate future politicization on a wide variety of issues pertaining to COVID-19. When prompted to speak in the style of a liberal Democrat, the model exhibits a greater likelihood of choosing to be vaccinated, choosing to wear a mask, endorsing vaccine and mask mandates, and supporting lockdowns. By contrast, when prompted to speak as a conservative Republican, responses are more likely to agree that wearing a mask or getting vaccinated should be personal decisions rather than government mandates and that both mask and vaccine mandates should be ended. To shed light on why the model associates liberals with more cautious responses than conservatives, we machine code more than 4,500 open-ended justifications for those responses and identify common themes corresponding to partisan gaps. Specifically, we find that levels of trust in the government and scientific community as well as the prioritization of safety versus freedom emerge as common considerations across questions, grounding novel pandemic-related issues in longstanding ideological principles.

These findings suggest that certain features of contemporary American liberalism and conservatism structured the way COVID-19 politicization unfolded. While these findings do not imply that "discourse is destiny," they provide compelling evidence that existing ideological systems channeled the reception to this novel issue in a way that ultimately undermined a unified public response. Our results suggest that COVID-19 did not fundamentally alter American political ideology, but our research design could similarly be applied to identify instances where history does deviate from the projections based on prior discourse, suggesting an unanticipated development of discourse. When social movements or influential public figures cultivate surprising new discourses that do not conform to existing patterns, this would result in a disjuncture between prior models and subsequent empirical observations. The failure of models to anticipate future developments may therefore suggest points when "history matters," and key events reshape public discourse in unpredictable ways.

Interviewing simulated respondents with an AI model presents obvious limitations. Most critically, an LLM may inaccurately reproduce a discourse, leading to incorrect or misleading inferences (Ji et al. 2023). Nevertheless, in situations where the target population cannot be interviewed or a question requires measuring responses at a scale that is practically infeasible, simulated respondents may provide the best available evidence into such questions that otherwise elude empirical analysis (Kim and Lee 2023; Kozlowski, Taddy, and Evans 2019). Investigations with simulated respondents thus occupy the edge of empirical sociology; the information they provide may not definitively resolve a debate, but can place meaningful evidentiary weight on important questions that lie outside the reach of conventional methods.

THEORETICAL FRAMEWORK

Advocates for the use of artificial intelligence in science often emphasize algorithms' capacity to surpass human performance. Indeed, such "superhuman" abilities are already facilitating important scientific contributions. In social science, machine learning algorithms assist researchers by identifying objects in videos and photographs, transcribing audio into text, and classifying texts into typologies, all at speed and scale far surpassing human capabilities (Bonikowski, Luo, and Stuhler 2022; Grimmer, Roberts, and Stewart 2022; Hannan 2022; Le Mens et al. 2023; Vicinanza, Goldberg, and Srivastava n.d.). In the natural sciences, algorithms are beginning to make new discoveries by combining computational power and speed with sophisticated knowledge bases. Arguably the most important advance has been AlphaFold's success at solving protein folding (Senior et al. 2020), but advances in drug (Jiménez-Luna, Grisoni, and Schneider 2020) and materials discovery (Wilkins 2023; Zhou et al. 2018), the control of complex nuclear fusion reactors (Degrave et al. 2022), and even the identification of novel auction and market policies are also promising (Jiao et al. 2021; Mosavi et al. 2020).

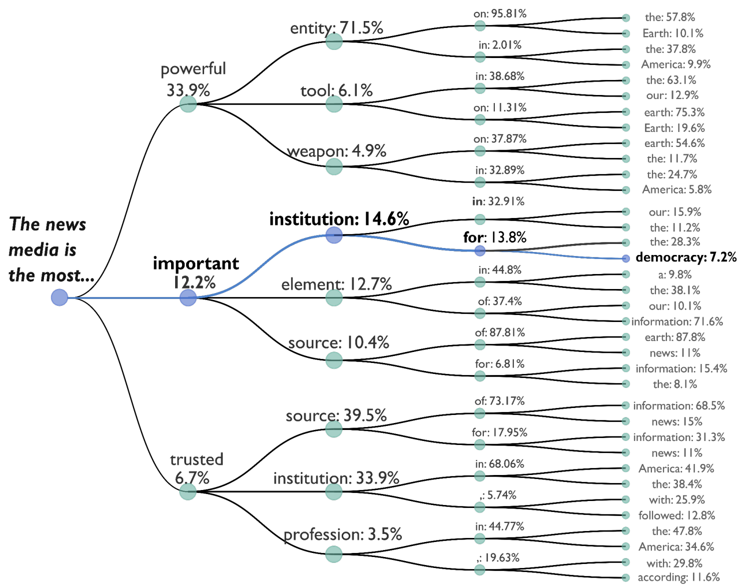
Social scientists, however, may gain more from artificial intelligence by capitalizing not on its capacity to surpass human performance, but its ability to mimic it (Brynjolfsson 2023; Sourati and Evans 2023). Simulation studies in the social sciences have historically favored elegant models with simplistic agents over empirically realistic ones. Such formal models provide important insight into how complex social patterns emerge from simple interactions, but they tell us comparatively little about the dynamics of specific empirical cultural systems, organizations, or institutions. Progress on this front has long been hindered by the difficulty of specifying empirically realistic agents to populate complex social simulations. Fortunately, due to their training on massive archives of rich cultural data, modern AI models can now generate "digital doubles" of human respondents, capable of faithfully reproducing the knowledge, preferences, and behaviors characteristic of a specific social group.

Social simulations with empirically realistic agents open productive avenues for research that would be impossible with human subjects. First, AI models can use textual records to reproduce the discourse of social groups that no longer exist or that otherwise cannot be interviewed. Because these models are generative, they enable analysts to go beyond the exact statements made in the textual archive and extrapolate likely out-of-sample utterances consistent with the semantic associations in observed texts. Second, digital doubles can produce simulated data at scale. Millions of interactions between digital actors can be quickly and affordably simulated over a wide array of initial conditions and differently parameterized agents, detailing a richer and denser high-dimensional interaction space than would be possible with human actors (Lai et al. 2024). Lastly, the internal representations of AI agents are directly observable in a way that human representations are not. Although the activation patterns of deep neural networks are commonly described as black boxes, these representations can be directly analyzed and explored for deeper understanding.

How Large Language Models Work

Computational models for text analysis have already found widespread application in the social sciences (Gentzkow, Kelly, and Taddy 2019; Grimmer, Roberts, and Stewart 2022). The most recent innovation to gain prominence is the word embedding model, which represents semantic relationships between words in a text as geometric relationships between word vectors in a high dimensional space (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). Words that are used in similar contexts (and therefore share similar meanings) are positioned close together in the embedding space, whereas words that occupy very different contexts are located far apart. Social scientists have shown that the positioning of words in an embedding space preserves cultural information from model's training texts, such as words' connotations of masculinity or femininity, affluence or poverty, and thought or action (Boutyline, Arseniev-Koehler, and Cornell 2023; Garg et al. 2018; Kozlowski, Taddy, and Evans 2019; Stoltz and Taylor 2019).

Figure 1: Autoregressive Language Modeling as a Branching Tree of Possible Completions



Although word embedding models mark a major advance in learning and representing semantic relations, they remain ill-suited for the task of generating new texts. Modern LLMs outperform classical word embeddings at task of language modeling for three key reasons: (1) LLMs are autoregressive models that iteratively predict the next word rather than the central word (Brown et al. 2020), (2) they use "self-attention" to imbue their word vectors with local contextual information (Vaswani et al. 2017), and (3) they leverage a deep neural network architecture that enables the learning of more varied and complex linguistic patterns (Kaplan et al. 2020).

Autoregressive language modeling

LLMs such as GPT-3 are "autoregressive" language models, meaning they optimize the prediction of the next word given a sequence of previous words. Such models take as their input a "prompt," and conditional on the sequence of words comprising the prompt, they generate a probability distribution for the next word in the sequence. A word is then randomly drawn according to this probability distribution and is appended to the original prompt. The next-word prediction task is then repeated using this newly extended prompt, generating yet another "next word." Because each new word is drawn stochastically from a probability distribution, we can conceptualize text generation as following a single pathway through a branching tree of potential next-words, growing a sentence one word at a time (Figure 1). By repeatedly inputting the same prompt to an LLM, an analyst can generate a distribution of directions a statement is likely to take.[2] This autoregressive approach differs from early word embedding models like word2vec or prior transformer-based models like BERT which used preceding *and* following words to predict a central target word (Devlin et al. 2018; Mikolov et al. 2013). This bi-directional approach may benefit from the additional information of subsequent words but is inappropriate for the task of generating new text, in which only prior words are available.[3]

Equation 1 describes how the autoregressive language models calculate the probability of a given sequence of words $(y_1, y_2, \ldots y_n)$ as the product of the probabilities of each word (y_t) conditional on the prior words in the sequence $(y_{<t})$.

$$P(y_1, y_2 \ldots y_n) = \prod_{t=1}^{n} p(y_t | y_{<t}) \tag{1}$$

[2]Some LLMs directly draw upon the branching tree of next word probabilities in order to maximize the likelihood that the entire response, and not just each word, is most probable. By keeping a broad "beam" or tree of probable texts, the model can wait and select the one more probable at the end. Other models, including OpenAI's GPT family, do not use beam search explicitly, but rather run self-attention layers many times (so-called multi-headed attention) to produce a distribution from which the top probability output text can be selected.

[3]Ilya Sutskever (2023) has posited that autoregressive models may outperform bi-directional models precisely because their task is more challenging and therefore induces more thorough learning.

Given the multiplicative nature of joint probabilities, any statement more than a few words long tends to exhibit a very low probability. For example, the most likely completion in Figure 1, "*powerful entity on the*", would occur with a probability of 13.4%. (0.339 * 0.715 * 0.958 * 0.578). The completion we highlight, "*important institution for democracy",* is one of the most likely four-word completions, yet it occurs with a probability of less than 0.1%. Extending to longer generations, the probability of any given completion becomes smaller yet. Joint conditional probabilities thus produce a branching structure that spans a vast diversity of completions, each one with a low individual probability of occurring.

It is important to note that autoregressive language modeling is not a new concept. Attempts at implementing this approach began as early as the 1910s with Andrey Markov (Gagniuc 2017; Hayes 2013) and continued into the 1940s with the work of Claude Shannon (Shannon 1951), finally maturing into full statistical models of future word prediction by IBM in the 1980s (Rosenfeld 2000). But despite their theoretical promise, probabilistic language models failed to consistently produce diverse, meaningful, and grammatically correct sentences for many decades. It was only with two further advances – self-attention and deep neural networks – that autoregressive language modeling finally achieved success.[4]

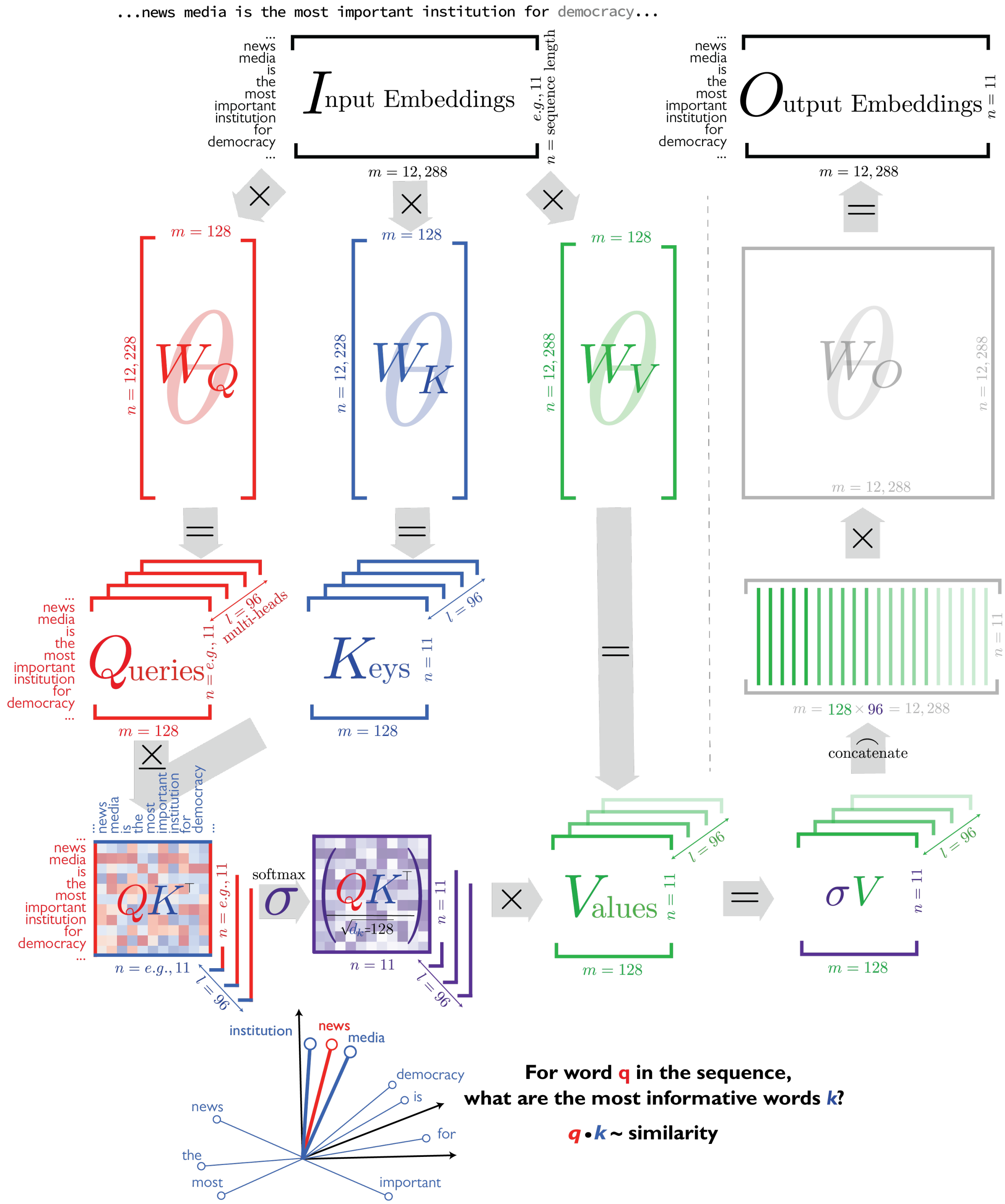
Self-Attention

LLMs use distributed vector representations to encode the relations between words, but they advance beyond classical word embedding models by incorporating a mechanism known as *self-attention* that imbues word vectors with information from their local context (Vaswani et al. 2017). During training, word embedding models like word2vec treat local context as a "bag of words," ignoring sequence and multi-word interactions (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). Post-training, each word in a word embedding has a singular vector representation based on its contexts across the entire training corpus. By contrast, when a prompt is input into an LLM, self-attention mechanisms share information between all words in the sequence. Words have a singular representation only in the first layer of the model; as the sequence progresses through the model, each word's vector representation is adjusted and "contextualized" by the other words in the sequence.

A schematic overview of the self-attention mechanism is displayed in Figure 2. Input embeddings are transformed through multiplication with learned weight matrices (W_K ,W_Q , W_V) and subsequent multiplication between the resultant Key, Query, and Value matrices. Prior to multiplication, however, each weight matrix is split into many smaller matrices (96 in GPT-3) by column in what is called "multi-head attention." Multi-head attention enables the model to attend to multiple versions of the input sequence simultaneously while also easing computation with improved parallelization. After the input embeddings are multiplied by the weight matrices, the key and query matrices are themselves multiplied, creating a square $n * n$ matrix for a prompt of length n. Each entry $[n_i,n_j]$ in this square matrix captures a meaningful "interaction" between token i and token j in the prompt that answers the question "for query token x, what key tokens y from the sequence provide the most informative context". These attention weights, the dot products of x and y, are then normalized by the square root of the Key dimension and transformed via softmax, the multiple-outcome generalization of the logistic function, so that each row's values sum to 1 and the predictive power of each key token on a query token can be interpreted as a probability. These probability estimates are treated as weights and reshaped through multiplication with the "value" vectors for each token. The heads of the weighted Value matrix are then concatenated back into a single wide matrix which is multiplied by a final output weight matrix (W_O), producing the output embeddings that pass through a feed-forward neural network before exiting the transformer block. Across this self-attention layer, the learned parameters include the weight matrices (indicated with a θ), and the input embeddings for the very first model layer.

[4]Another advance central to modern post-ChatGPT language models is the deployment of user feedback in the tuning of the models for human interaction and simulation. By fine-tuning towards human responses, such models effectively return to the sequence-to-sequence (Seq2Seq) architecture, critical in translational models (Xue et al. 2020) and early transformers (Vaswani et al. 2017), that considers at each word not only the word that came before, but also the priming motivation or "prompt" (x), be it a modeled human response, a picture to be captioned, a turn in online conversation, or anything else that inspires the generation of text (Ouyang et al. 2022).

Figure 2: Schematic Diagram of Self-Attention.



This complex operation has three important implications. First, self-attention solves the problem of polysemy. In classical word embedding models, the multiple meanings of words are conflated within a single vector representation. Self-attention resolves this issue by adjusting the position of each word vector using information from all other words in the prompt. For instance, the word "bark" would be modified in one way if it is referencing a tree, and another if referencing a dog. Self-attention similarly resolves anaphora, linking pronouns with their associated nouns by sharing semantic information between them. Although pronouns like "it" carry little information in a classical word embedding, they can be meaningful when linked to the correct noun by self-attention in an LLM. Lastly, self-attention can up-weight words most relevant for predicting the next word. For example, when predicting the next word in the sentence, "Benjamin Franklin, noted inventor and statesman, was born in the year ______" the words "Benjamin" and "Franklin" (and the interaction between them) are more important in predicting the next word than terms such as "inventor" or "noted." These highly relevant words may therefore be up-weighted for predicting the next word. Thus, even long-range dependencies can be preserved and leveraged by "contextualizing" and sharing information between words (Figure 2).

Self-attention highlights another key difference between LLMs and earlier models such as word embeddings – *a single LLM is able to preserve a variety of distinct discourses* (Argyle et al. 2023). In a word embedding, each word has a single relationship to every other word, represented as proximity between the respective word vectors. To compare two discourses using word embeddings, an analyst would need to train independent models on separate collections of text representing the desired discourses, then compare the relative positioning of word vectors between models. By contrast, words in an LLM do not have a single representation; each word's vector representation is modulated by the presence and order of other words in the prompt. Thus, if a word or phrase has a different usage across discourses, these multiple senses can be preserved in a single model and activated by surrounding context words. This means that many different discourses can be generated with a single LLM by inputting prompts that prime different cultural registers, so long as the training corpus includes sufficient texts to learn the linguistic patterns of the respective group (Argyle et al. 2023).

Deep Architecture

The final factor enabling LLMs' success in producing humanlike texts is their massive neural architecture. Word embedding models like word2vec use a shallow architecture with a single hidden layer, and the total number of parameters learned by the model is typically in the tens of millions.[5] By contrast, GPT-3's neural network consists of 96 layers and 175 billion parameters (Brown et al. 2020). GPT-4's architecture has not been formally released, but expert consensus is that GPT-4 is substantially larger, with parameters likely numbering over one trillion (OpenAI 2023a). A greater number of parameters and layers enables a neural network to learn more complex functions. For a language model to faithfully encode the multitude of discourses that appear on the internet, it must compose an exceptionally complex function. This function leverages the extensive non-linearities and interactions between words to transform an input sequence into an accurate probability distribution for the next word. In learning these complex patterns of linguistic entailments from its training texts, the model effectively learns the internal structure of a discourse.

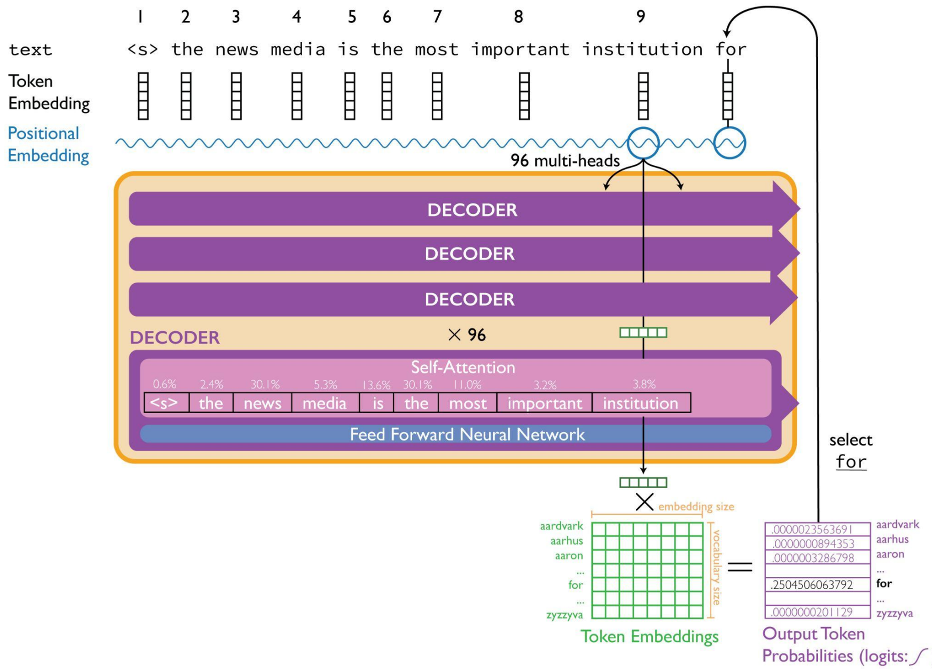
Figure 3 presents a schematic diagram of GPT-3's architecture. During training, a sliding window of words from the training text is used as the context, and is represented as a matrix of corresponding word vectors. This collection of word vectors then passes through 96 "transformer blocks." Each transformer block comprises attention mechanisms that are themselves divided into 96 "attention heads" followed by a feed-forward neural network which further transforms the contextualized word vectors in preparation for predicting the next word. After passing through all transformer blocks, the resulting matrix is converted into a single vector corresponding to the model vocabulary. This vector is rescaled into a probability distribution via the softmax function, a multi-category generalization of the logistic curve. This probability distribution is then compared to the correct response – the actual next word in training texts. Error, or "loss," is calculated as a function of the difference between the predicted probabilities and the correct next word,[6] and this error is propagated back through the model to update parameters such that the same prediction would be more accurate if made again. The context window then continues its progression through the training texts, repeating the task of predicting each subsequent word given the prior words. The algorithm may iterate over the entire corpus of training texts multiple times until improvements become negligible and training is halted.

Once fully trained, LLMs can generate novel texts by the same operation of next-word prediction that is optimized during training. The user feeds the model a prompt and the algorithm iteratively predicts the following words one by one, appending each newly generated word to the prompt to predict the subsequent word. In this way, the LLM builds

[5]The number of parameters in a word2vec model is the product of vocabulary size (N) and the user-specified number of dimensions (M) multiplied by two, because the model simultaneously learns the hidden "context embedding" along with the final "word embedding," each of size NxM. A typical word2vec model may have a vocabulary of size 50,000 and 300 dimensions, resulting in 30 million parameters.

[6]GPT-3 minimizes cross-entropy loss.

Figure 3: Schematic Diagram of GPT-3 Architecture, from Input Embeddings through Token Prediction.



new sentences word-by-word. The primary difference between training and generation is that there is no "correct next word" during text generation, but both tasks are fundamentally grounded in the next-word-prediction task.

Prominent chatbots like ChatGPT use an LLM as their foundation, but are then adjusted by their designers to respond to queries in the style of a helpful assistant. This tuning process, known as Reinforcement Learning with Human Feedback (RLHF), steers the model toward responding to queries rather than simply predicting the next word in a sequence, while also censoring content that is offensive or controversial (Ouyang et al. 2022). The original GPT-3, however, was released prior to fine-tuning with RLHF, and therefore simply predicts the next word in the user-provided sequence without additional weighting or censorship. This makes GPT-3 a more straightforward tool for the recovery and analysis of historical discourses, although recovering a variety of discourses from models fine-tuned with RLHF is also possible.

Using a basic LLM such as GPT-3, the user can simulate responses from a socially-situated respondent by prompting a model with a statement that primes a given perspective. For instance, by starting a statement with "I am a conservative Republican and I believe" will produce systematically different completions than a statement that starts "I am a liberal Democrat and I believe" (Argyle et al. 2023). The differences in the completion will reflect the differing semantic associations of "conservative Republican" and "liberal Democrat" learned from the model's training texts. Leveraging information from a near complete record of text from the internet, the model will attempt to generate likely endings for each of these phrases. This technique thus presents a novel means of interrogating the system of claims, considerations, and justifications that characterize an ideology, extending far beyond the simple positions of words obtained from a classical word embedding model.

The process of generating text in response to a prompt with a pre-trained model is called "inference" in computer science. Recent work demonstrates how the inference process, as described above, directly optimizes text responses to user prompts to maximize syntactic, semantic, and even pragmatic appropriateness. This work demonstrates that the number of transformer layers in contemporary transformers (e.g., 96 for GPT-3) is comparable to the number of steps required to optimize neural network weights, resulting in an optimized and efficiently produced textual response that does not require additional neural network training (Von Oswald et al. 2023). Related work demonstrates how inference from prompts unleashes in a within-model process of gradient descent for error minimization akin to the process of fine-tuning a model from external data (Dai et al. 2023). In sum, transformers enable the potential to model multiple, well-defined human perspectives efficiently in ways that both enable analysis and generative simulation.

Digital Doubles and the Study of Ideology

Through their ability to recreate discourse models underlying their training texts, LLMs offer new avenues for studying cultural-historical change and advancing theories of meaning. In particular, we use LLMs to shed new light on the nature and extent of *constraint* in cultural systems. Structuralist theories emphasize the overall coherence of meaning systems, positing that seemingly elaborate systems of classification and evaluation are reducible to simple underlying logics (Douglas 1966; Lévi-Strauss 1966). The diverse array of practices, values, and beliefs expressed within a culture can be understood as "all of a piece," unified by subtle threads of configured meaning that can be revealed through careful analysis (Mead 1942). This coherent model of collective meaning has important implications for cultural change; new ideas can only be integrated into an ideology if consistent with the overarching logic of the system. By this theoretical model, cultural evolution is largely predictable because it is highly constrained. When a new object enters a cultural context, its potential interpretations are tightly limited by the logic of the cultural context.

This constrained model of cultural coherence was largely displaced by a wave of scholarship that emphasized the importance of historical contingency and internal contradictions of cultural systems. To redress the structuralists' failures to capture the apparent arbitrariness and contradiction of culture, the image of a unified system was supplanted within sociology by a model of culture as a "toolkit," a repertoire of strategies that can be taken up or discarded as necessary (Swidler 2003), or a largely disconnected set of cognitive schemata which are deployed situationally (DiMaggio 1997) to satisfy a pragmatic purpose. More recent work has sought to identify a theoretical middle-ground, accepting that systems of meanings exhibit broad patterning, but remain only weakly constrained, rife with instability, ambivalence, and contradiction (Baldassarri and Goldberg 2014; Boutyline and Vaisey 2017; Kiley and Vaisey 2020; Rawlings 2020; Swidler 2003).

A similar debate has unfolded in the field of political opinion. One strand of research argues that voters hold core values and form opinions on specific issues in accordance with basic underlying considerations such as freedom, equality, or protection from harm (Feldman and Zaller 1992; Goren et al. 2016; Haidt 2012; Lakoff 2010). This theoretical model posits that voters' conceptual systems are constrained by an internal logic with a few core values structuring a complex attitude system that covers a diverse array of specific issues. Detractors of this theory argue that individuals' attitudes are steered not by internal values, however, but by partisanship. According to this top-down alternative, partisan leaders construct a platform of positions through a strategic process of "log-rolling," with the aim of building a coalition across various interests (Carmines and Stimson 1989; Zaller 1992). Partisan leaders then broadcast this assortment of positions as a unified platform, which is then absorbed *in toto* by strong partisans in the electorate (Green, Palmquist, and Schickler 2004). By this model, the collection of opinions that constitute an ideology are the arbitrary result of historical-political contingencies; held together not by a unifying cultural logic, but social and political messaging.

This debate over ideological coherence revolves around a key empirical question. When a new issue emerges, are public responses already prefigured by ideology, or is a new issue ideologically indeterminate until partisan elites voice positions on the issue (Noel 2014; Page and Shapiro 2010; Zaller 1992)? If political ideologies are general dispositions that inform attitudes across issues, they could be readily transposed to new issues without elite direction. If ideologies are arbitrary assortments of attitudes strategically drawn together through political coalitions, however, then the politicization of a new issue would be unpredictable until partisan elites broadcast their positions. Thus, study of the exogenous injection of a new issue within a political landscape sheds light on a key theoretical question in the sociology of culture – when a new idea is introduced into an existing system of beliefs, is it constrained by the logic of the belief system or free to take any direction?

LLMs offer a powerful new approach for exploring this core question in the study of culture and ideology. Typically, by the time the public learns about a new issue, it has already been politicized by partisan elites. Thus, when ideology in the electorate mirrors party platforms, analysts are unable to adjudicate if elite cues steered public opinion or if underlying ideological convictions shaped the new issue's interpretation for both party elites and the public at large. The key benefit of LLMs is that they capture and preserve discourses from the historical period of their training texts. Therefore, an LLM trained on texts from a given time can simulate respondents from that period and present them with questions that anticipate future cultural or political developments. While previous text analytic models produce representations of cultural systems from historic texts, LLMs are *generative* models to which researchers can pose novel questions. These prompts could even include hypothetical scenarios or issues that only emerged after the model's training. By leveraging the vast linguistic information learned in training, generative cultural models produce the most likely responses to novel questions given the discursive associations of that period. This method enables new insight into the historical process of cultural change, providing a lens through which we can assess which developments are truly surprising and which are already prefigured by the cultural system.

Moreover, the generative nature of LLMs enable us not only to assess distinct perspectives, but to interrogate them. In his classic 1940 article on "Situated Actions and Vocabularies of Motive", sociologist C. Wright Mills argued for the importance of analyzing the shared language by which persons from distinct socio-cultural situations justify and account for their positions and conduct (Mills 1940). This was not fixed on the psychology of human motivations and drives but rather the sociology of how certain motives are more acceptable given certain audiences and contexts than others. But in addition to being situationally defined, social appropriateness of a motive is also constrained by identity. As Goffman's dramaturgical perspective suggests,"when an individual plays a part he implicitly requests his observers to take seriously the impression that is fostered before them. They are asked to believe that the character they see actually possesses the attributes he appears to possess, [and] that the task he performs will have the consequences that are implicitly claimed for it...." (Goffman 2021). To the extent that an LLM encodes the discursive patterns of a social group, the analysts can go beyond simply "polling" the simulated respondents. They can prompt the model to justify its responses within a designated cultural register, revealing a set of linked considerations that render the expressed attitude coherent.

Mills' vocabulary of motive, however, does not reach its own aspiration. A "vocabulary" implies a collection of low-level words that might be flexibly deployed according to some higher-order cultural logic to plausibly and acceptably justify a position. It implies the prevalence of templated excuses that could be tallied and expected in particular contexts. Modern LLMs allow us to reach past this conception to model not only situated vocabularies of motive, but their syntax and semantics. LLMs can generate and discriminate how even novel, unique expressions of motive should be more or less expected from a situated actor regarding their behavior. How can we do this with LLMs? As with people: By asking them over and over again.

2019 Politics in Silico

The spread of COVID-19 to the United States presented a critical event for social and political meaning making. Lacking any precedent in living memory, the pandemic was not readily interpretable within existing frames for political response. This is evident from the earliest public opinion surveys, which show relatively little partisan division on questions relating to the virus (Deane, Parker, and Gramlich n.d.). Nevertheless, political elites and opinion leaders began to broadcast a variety of competing interpretations of the situation as soon as a virus was detected in the US in January of 2020 (Stokes et al. 2020). By March 2020, a sizable divide had already grown between self-identified democrats and republicans regarding the appropriate government response to the emerging pandemic (Gadarian, Goodman, and Pepinsky 2021). This politicization of COVID-19 would prove to be a defining characteristic of the pandemic period, imbuing discussions of lockdowns, masks, and vaccines with partisan fervor, ultimately stymying any unified public response to the virus (Albrecht 2022; Allcott et al. 2020; Chen and Karim 2022).

The politicization of COVID-19 related issues quickly assumed a common pattern. Liberals viewed the virus as an urgent threat warranting immediate and sweeping response, whereas conservatives questioned the danger posed by the virus, denounced responses that infringed on individual liberties, and doubted the safety of the government-sanctioned vaccine. After years of pandemic politics, this familiar pattern may appear self-evident, and there are indeed some reasons to view this polarization as predictable. On a variety of issues, ranging from gun control to universal healthcare to motorcycle helmet requirements, liberals tend to favor protection and government intervention while conservatives lean towards freedom and personal choice (Homer and French 2009). To explain this pattern, some public opinion analysts have argued that conservatives more highly value freedom while liberals prioritize considerations of equity and protection from harm (Feldman and Zaller 1992; Haidt 2012).

Nevertheless, a large literature from political psychology plausibly anticipates the very opposite empirical outcome and supplies numerous reasons to expect that conservatives would support stricter responses to the virus than liberals. In an influential review of the psychological correlates of political ideology, Jost (Jost 2006) cites robust international evidence that political conservatism is associated with (i) fear of death, (ii) aversion to threat or loss, (iii) uncertainty avoidance; and (iv) needs for order, structure, and closure. Each of these predispositions suggest that *conservatives* should be the ones to advocate for harsher measures to protect against the virus, not liberals. Moreover, a wide array of studies suggest that conservatism is associated with desire for purity and aversion to contamination. This literature emphasizes forms of social or religious impurities, but more broadly connects political conservatism to a general fear of contamination and uncleanliness (Haidt 2012; Helzer and Pizarro 2011; Jost 2017; Oxley et al. 2008; Terrizzi, Shook, and McDaniel 2013). Consistent with this pattern, Republicans were more concerned about the Ebola epidemic of 2014 than Democrats (Pew Research Center 2014).

Similarly, there are reasons to suspect that liberals would be skeptical of sweeping government responses to the virus. Liberals have a long history of skepticism toward vaccines, arguing that they are unnatural and pushed by large, profit-driven pharmaceutical companies (Callaghan et al. 2019; Colgrove 2006; Conis 2014; Jamison, Quinn, and Freimuth 2019). Also, in recent historical cases where the safety of the American people has been at stake, such as the War on Terror, conservatives were more willing to sacrifice personal liberties for public safety than liberals (Rosentiel 2011). All of these considerations suggest that history could have played out differently, and that an alternate framing for COVID-19 was plausible, which would steer conservatives to endorse cautious measures and liberals to oppose them.

It is possible that prior values and predispositions do little to steer the public response to an issue; any political topic could be framed in a variety of ways, and predispositions provide little direction until partisan elites provide an interpretation of the issue in terms of clear ideological considerations. The alternative hypothesis is that ideology does steer reactions to a new issue like COVID-19. Even if the initial effect of ideology was weak, it could set off a cascade of self-reinforcing dynamics among both elites and the public, ultimately setting the course for widespread polarization (DellaPosta 2020; Rawlings 2022)

Assessing whether the public was inclined to tip towards a given form of politicization prior to elite signaling would require measuring public response to pandemic-related questions *before its top-down framing*. In practice, this is difficult because most issues are rapidly framed by elites before social scientists can measure public attitudes. Because collecting public attitudes to an issue before it emerges is unfeasible, the best alternative is to reconstruct the discursive space prior to the emergence of a novel issue and interrogate its associations surrounding the issue. We therefore use GPT-3, the first LLM capable of faithfully reproducing complex attitude systems, to reconstruct the political landscape of 2019 so we may investigate whether this period exhibited a predisposition to tip towards the pattern of polarization that manifested in the following years.

DATA AND METHODS

For the following analyses, we analyze completions generated by the GPT-3 language model. This approach presents a notable departure from conventional methods of text analysis. LLMs are generative models, and the most straightforward way to learn from these models is not to examine their internal representations but to study the outputs they produce. As a result, the data we analyze are not actual statements made by members of our target populations nor are they representations of such statements like topic models or word embeddings. Our data are novel word sequences learned to occur with high probability given the discursive patterns learned in a vast training corpus.

Such an approach deviates from the simulation studies typical of formal sociology. Formal models commonly attempt to parsimoniously capture social dynamics by precisely specifying minimal conditions under which empirically-observed patterns can be reproduced. Using LLM outputs as data, by contrast, constitutes a hybrid of empirical and formal sociology; we analyze simulated data, but the simulation is trained to produce outputs that closely approximate empirical distributions.[7] To the extent that the model successfully reproduces a population's response distributions, the model's outputs can stand in for human responses from that population and be analyzed at scale. We adopt this approach to simulate the opinions of American liberals and conservatives in October of 2019, the historic point immediately prior to the emergence of COVID-19.

Prompt Design

Our primary aim is to discern whether a speaker identifying as liberal is predicted to have different views regarding pandemic responses than a speaker identifying as conservative. We therefore design prompts that isolate the effect of partisan identification words on the generation of responses toward a variety of COVID-19 issues. We use three modes of partisan identification: ideological identification (liberal or conservative), party identification (Democrat or Republican), and candidate preference (Hillary Clinton or Donald Trump). To maximize the partisanship signal, we use all three of these identifiers in all our prompts.[8] All prompts begin with a "partisan priming" taking the following form:

"I am a strong ***conservative*** *and a lifelong* ***Republican****. In 2016, I was proud to vote for* ***Donald Trump*** *and I think that the* ***Democrats*** *have been a disaster for this country.*

or, conversely:

"I am a strong ***liberal*** *and a lifelong* ***Democrat****. In 2016, I was proud to vote for* ***Hillary Clinton*** *and I think that the* ***Republicans*** *have been a disaster for this country.*

Because GPT-3 was trained on texts published only through October 2019, it has no knowledge of COVID-19. This ignorance is an asset, as it allows us to investigate the ideological landscape immediately prior to the emergence of this pivotal issue. However, in order for the model to produce an informative response, we must supply some basic knowledge of the virus to the model in the prompt. Directly following the initial "partisan priming," we insert the following sentence:

[7] Previous social simulation studies have used parameters estimated by empirical models to improve correspondence to observed contexts, but these models still typically strove to parsimoniously reproduce the system of interacting factors, prioritizing interpretability over predictive accuracy.

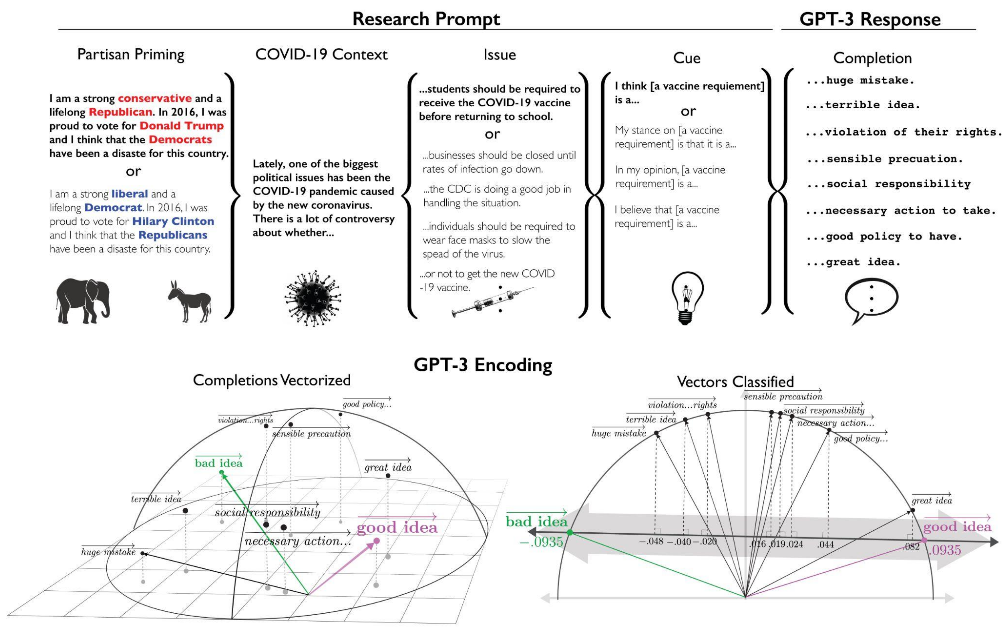
[8] Robustness tests suggest that removing the candidate preference from prompts generally weakens effects but rarely eliminates them.

"Lately, one of the biggest political issues has been the COVID-19 pandemic caused by the new coronavirus. There is a lot of controversy around {issue}.

We replace the {issue} variable with 49 different issues relating to different aspects of COVID-19, such as "whether wearing face masks in public places should be optional or mandatory" or "whether students should be required to receive the COVID-19 vaccine before returning to school."[9] Within each issue, we test a variety of question wordings. Some variations are designed to test whether certain key words steer responses (e.g. vaccine *mandate* vs. vaccine *requirement*), and other variations are included simply to improve robustness (e.g. "*I think* face masks are" versus "*In my opinion*, face masks are"). Conventional psychometric surveys are limited by the cost of additional questions and the limits of respondents' attention (Furr 2021). However, with relatively inexpensive digital doubles, we can relax these constraints and explore the sensitivity of answers across a range of differently worded prompts.

After introducing the issue, the prompt ends with a statement such as "I believe this is a", after which GPT-3 begins its completion. In pilot testing, we found that the more open-ended "I believe" led to many non-committal outputs like "that some people don't like to talk about this" whereas the ending with "*this is a"* encourages more clearly valenced responses like "terrible idea" or "great plan." An outline of our piecewise approach to prompt construction in Figure 4.

Figure 4: Outline of the prompt design process and conversion of open-ended responses to scores on a semantic axis defined by two anchoring terms (e.g., "bad idea" and "good idea")



Including wording variations, our total number of questions reaches 179. We test each prompt with both a liberal and conservative partisan priming, doubling the number of unique prompts. We therefore present the model with a total of 179 x 2 = 358 distinct prompts. To generate a sample of completions, we input each prompt into GPT-3 500 times, resulting in a total of 179,000 completions.

Machine Coding Responses

We designed our prompts to encourage relatively standardized completions indicating either a positive or negative response. Total standardization was not desirable, however. LLMs are built to be "programmed" natively with language inputs and to produce natural language outputs.[10] In pilot testing, we experimented with prompting the model to produce closed-form responses, but these structured responses performed markedly worse than open-ended responses at predicting partisan stances on well-established political issues. Indeed, open-form responses are generally more informative than constrained closed-form responses for human respondents as well (Tourangeau, Rips, and Rasinski

[9]Exact wordings of the 49 issues are provided in the Appendix.

[10]It is possible to make an LLM respond to closed-form multiple choice questions, but in order to optimize a model to perform on this task the analyst would append an output layer with a softmax activation that predicts the best response category rather than the next word.

2000; Willis 2004). The open-ended responses generated by the LLM have clear valences, but still take on a wide variety of unique forms. Given the impracticality of hand-coding 179,000 responses, we use OpenAI word embeddings to classify completions as either positive or negative (Neelakantan et al. 2022). These embeddings are built atop an LLM architecture similar to GPT-3, but fine-tuned with a contrastive loss that directly pulls similar words together and pushes dissimilar ones apart (Neelakantan et al. 2022).

OpenAI recommends a technique for classification with word embeddings similar to Kozlowski, Taddy, and Evans' (2019) method for projecting words onto axes of cultural meaning. We first produce a vector representation of each prompt completion using GPT-3 embeddings. We use the ada-text-embeddings-002, which are 1026 dimensional and perform well on semantic similarity tasks. As with word2vec, words or phrases that are semantically similar are proximal in the embedding space. Thus, to classify whether a completion is semantically closer to "good idea" or "bad idea," we calculate the cosine similarity between the completion's vector and the vector for each of those two anchoring phrases. "Good idea" and "bad idea," serve as our anchoring phrases for most completions, but depending on issue wording they also take forms like "mandatory" and "optional," or "effective" and "ineffective."[11] After calculating each completion's cosine similarity to each of the two relevant classification terms, we calculate the difference between these proximities. The result is a score between -1 and 1, indicating whether the response is more semantically similar to the first classification option (e.g. "good idea") or the second option (e.g. "bad idea").

We use these scores to test for each issue whether prompts with a liberal priming are statistically distinct from those with a conservative priming. For each question wording, we fit an OLS regression of partisan priming predicting the completion's classification score. These models reveal whether the speaker identifying as a liberal is more positive about the issue at hand than the speaker identifying as conservative.

We ultimately consider GPT-3's forecast to be correct if the effect of partisanship in the OLS model is in the same direction as a partisan gap observed with surveys in 2020. We draw upon Gadarian, Goodman, and Pepinsky (2021) and published results from the Gallup Panel survey (McCarthy 2023) to source "ground truth" partisan gaps against which we compare our simulated response distributions.

Generating Justifications

For select prompts, we go beyond simply identifying differences in responses to COVID-19 and attempt to shed light on why the model is predicting these differences. We elicit this by prompting the model to *produce a second sentence justifying its initial response* and reveal the characteristics of ideologically consistent motives. For these tests, we input the original prompt to GPT-3 along with the previously generated completion, which is restricted to one sentence. We then extend this prompt by beginning a new sentence offering a justification for the earlier response. Specifically, we append the phrase "This is because" to the end of the prompt to induce a justification response. We find that alternate wordings produce substantially similar outputs and include examples in the Appendix.

For each prompt, we generate three "justification" completions. Because we initially generate 500 liberal and 500 conservative responses to each question, this results in a total of 3000 justifications. To classify these numerous open-ended responses into a few informative categories, we again rely on machine coding. As above, we use GPT-3 embeddings to represent each justification as a 1026 element vector. But because we want categories of justifications to emerge inductively, we use k-means clustering to identify thematic groups instead of rating responses along a predetermined axis. Because justifications in favor of a given policy should be qualitatively different from those opposing the policy, we conduct k-means clustering separately for positive and negative responses as scored in the prior step; for instance, we first perform cluster analysis on the justifications of statements *in favor* of mask mandates, then we conduct another independent cluster analysis of all justifications for statements *opposing* them. We manually select the number of clusters by considering three metrics (Silhouette, Calinski–Harabasz, and Davies-Bouldin scores) along with our qualitative assessment of interpretability and parsimony. After dividing justifications into clusters, we generate labels for each cluster using GPT-4. We feed a random sample of 100 entries from each cluster into a GPT-4 prompt with instructions to provide labels for each set of responses that describes their distinctive semantic characteristics.[12] By comparing the proportions of liberal and conservative responses in each of these clusters, we identify partisan differences in how opinions are justified.

We cannot know for certain if the justifications produced by this method truly reveal the causal antecedent of the initial responses generated – this would require examining patterns of neural activation and identifying how these

[11] For some completions, we found that responses could take multiple positive or negative forms. For these, we average together two classification terms (good idea + personal choice; bad idea + public health issue).

[12] We designed this prompt to generate cluster labels: "*The following are clusters of semantically similar responses to the question of whether [issue]. [List of sample texts] Please write concise, specific, and not overly broad labels for each of the clusters that describe their unique theme and distinguish them from the other clusters. It does not have to encompass all responses but should instead reflect the primary theme evident in the substantial part of the responses.*"

patterns correspond to semantically coherent features. In asking GPT-3 to justify its response, our approach reveals the distribution of most likely justifications, reflecting both their plausibility and acceptability. This is much like asking a human respondent to justify a prior response. It does not necessarily provide the true reason for the response, but it does source a network of associated considerations that the respondent deems plausible to themselves, acceptable to their imagined audiences, and relevant to the issue at hand (Tourangeau, Rips, and Rasinski 2000). Just as these *post hoc* justifications can still provide insight into how an issue is understood by a human respondent, it may similarly illuminate how the issue is "understood" by the language model.

Collectively, our analytical approach is divided into three stages: (i) validation of our method on well established political issues, (ii) testing if GPT-3's representation of COVID-19 politicization anticipates observed patterns, and (iii) exploration of the semantic considerations underpinning GPT-3's predictions by asking simulated respondents to explain their responses.

Validation

To confirm that our prompt induces partisan differences as expected, we conduct a series of validation tests on political issues already well-established in 2019. Using the same style prompt described above excluding only the passage about COVID-19, we generate liberal and conservative responses to variously worded questions on topics of abortion, climate change, gender and sexuality, race, immigration, drugs and policing, gun control, healthcare, welfare state programs, and business regulation. Across these 10 topical areas, we pose 37 distinct questions, and each question had multiple wordings. For 35/37 questions, the majority of wordings correctly predicted empirical partisan differences on that issue. For one question, the association was in the incorrect direction for 2/4 wordings, and for one question, no association was identified. Results are presented in Appendix A. These results also reveal systematic differences in the effect of question wordings on stated positions. For example, asking a liberal- or conservative-prompted model its "stance" on an issue almost always brought the answer close to the center of the distribution, whereas when an "opinion" is elicited, we find responses are more markedly partisan.

These results provide confirmatory evidence that the prompts we designed effectively induce partisan divides observable in American politics circa 2019. This does not definitely establish that the model is equally accurate in estimating American attitudes toward a hypothetical virus in 2019. The familiar political issues are within the training distribution, whereas responding to questions about COVID-19 requires out-of-distribution generalization. We have no "ground truth" for what attitudes toward COVID-19 would have been in 2019. Indeed, if such a data source existed we would not need to rely on simulation. But on those attitudes for which validation is possible, we find encouraging evidence that our prompts faithfully reproduce observable partisan divisions.

RESULTS

We begin by presenting results from prompts on vaccine-related topics. Each of the figures below is a coefficient plot showing the effect of "partisan priming" on the classification scores for a given prompt's outputs. Partisan priming is a binary variable coded 1 = "liberal", so positive coefficients indicate that liberal prompts were more likely to endorse the first anchoring term in the response dichotomy (e.g. "good idea"). This equivalently means that conservative prompts are more likely than the liberal prompt to endorse the second anchoring term (e.g. "bad idea"). Coefficients with significant positive effects ($p<0.05$) are colored blue. Negative coefficients conversely signify that conservatives are more likely to endorse the first anchoring term than liberals. Coefficients with significant negative effects are colored red.

Figure 6 displays coefficient plots of "partisan priming" predicting output classification scores for vaccine-related issues. The plots in the top left of Figure 5 examine the effect of partisanship on the intention to receive the vaccine. The positive coefficients in the first plots show that liberal prompts were more likely to report an intention to get the vaccine than conservative prompts. The effect is significant for both wording variations ("Personally, I will" and "I have decided that I will"), and for two question variations ("...to solve the spread of the virus" and "...to protect yourself against the virus").

The next plots show results to questions regarding vaccine mandates. In the first question, liberal prompts are more likely to claim that a vaccine mandate is a "good idea" than conservative prompts, with the effect significant for three of four wordings. The second question replaces the word "mandate" with "requirement." The effect remains positive for three of four wordings, although some effects are attenuated. Conversely, when the question is changed to be about *ending* vaccine mandates, conservative prompts are more likely to state that this is a "good idea."

In the second row of Figure 5, we test the effect of partisanship on responses to vaccine mandates for government workers and students. Although many of the coefficients fall short of statistical significance, the significant effects are consistent with later observed polarization: the liberal prompt is associated with stating that "requiring government workers to get the vaccine" is a good idea, while conservative prompts is associated with saying "allowing government workers to return to work without getting the vaccine" is a good idea. Similarly, liberal prompts endorse "requiring students to get the vaccine" while conservative prompts endorse "allowing students to return to school without getting the vaccine." However, when the question is worded to probe a response to "opening schools without a vaccine requirement," no significant effect for partisanship is identified.

The third row of Figure 5 shows the effects of partisanship on responses to whether getting the vaccine should be a choice. On the questions of allowing government workers or students to "opt-out" of the vaccine, conservative prompts were more likely in both cases to say this is a good idea than liberal prompts. Similarly, the conservative prompts were more likely than liberal prompts to say that "letting individuals choose" and "letting people decide" to get the vaccine is a good idea.

We see the first instances of model inferences being inconsistent with historical observed patterns of polarization in the bottom row of Figure 5, where we ask about requiring proof of vaccination for various activities. Conservative prompts are more likely to say it is a good idea to require proof of vaccination to "travel by plane" or "enter bars or restaurants." The same pattern emerges when the question is reworded, such that liberal prompts are more likely to state that "allowing unvaccinated people" to travel by plane or enter bars and restaurants is a good idea.

We next run a similar set of tests for opinions regarding face mask requirements. Results are presented in Figure 6. First, we see that liberal prompts are more likely to express intent to wear face masks both to slow the spread of the virus and for personal protection. Liberal prompts were also more likely than conservative prompts to say that mask "mandates" and mask "requirements" are good ideas. Consistent with this, conservative prompts were more likely to respond that ending mask mandates is a good idea.

The second row shows responses to prompts about requiring/mandating masks in stores, workplaces, or schools. For both "requiring" and "mandating" wordings, and for both stores/workplaces and schools, liberal prompts exhibit a greater likelihood of framing these measures as a good idea.

In the third and final rows, we see results from prompts on whether masking should be a personal choice and whether masks are effective. The first plots show that conservative prompts were more likely to state that "letting individuals choose" or "letting people decide" whether to wear a mask is a good idea, with no apparent difference between these wordings. The next plot shows that, when asked whether wearing masks in public should be mandatory or optional, liberal prompts were more likely to state that they should be "mandatory" than conservative prompts. Lastly, partisanship had no statistically significant effect on responses to whether face masks are an effective way to slow the spread of the virus.

Next, in Figure 7 we examine responses to questions about lockdowns. The first two plots in the first row ask about closing business and closing bars and restaurants. For both of these issues, the liberal prompts are more likely to endorse the lockdowns. The next two issues concern prohibiting large gatherings and avoiding small gatherings, and again the liberal prompts are more likely to say each of these measures is a "good idea." By contrast, the final question asks if keeping businesses open is a good idea. The results for this question are mixed; for completions beginning with "I think," conservative prompts were more likely to say keeping businesses open is a good idea. Yet for completions beginning with "I believe that," liberal prompts were more likely to support the action. Although these results are mixed, they still represent a movement in the correct direction relative to previous questions.

In the second row of Figure 7 are plots representing the effect of partisanship on prompts regarding school lockdowns and halting international travel. Results on the topic of school lockdowns are mixed. Liberal prompts are more likely to say that "keeping schools open" is a good idea than conservative prompts, but they are also more likely to say "closing schools and conducting classes online" or "switching to remote schooling" are good ideas as well. Thus, we do not see a clear partisan leaning in either direction in the question of school lockdowns. Halting travel similarly exhibits some inconsistency. Conservative prompts are more likely to endorse a ban on visitors from countries with COVID-19 outbreaks, but are not more likely to endorse "stopping international travel." Partisanship effects for "stopping international travel" are non-significant, except liberal prompts were more likely to endorse "stopping international travel" for completions that began with "I believe that." Results for halting travel are therefore also inconsistent, but lean slightly in the correct direction.[13]

Lastly, Figure 8 shows results from general questions on the COVID-19 pandemic. The first two questions assess confidence in the CDC. Liberal prompts were more likely to state that the CDC is "doing a good job" handling the situation, while conservative prompts were more likely to state that the CDC is "exaggerating the danger posed by the

[13]Gadarian et al. (2021) find that Democrats were less likely to support air travel restrictions or banning visitors from countries with COVID-19 outbreaks

Figure 5: Coefficient plots from OLS models of partisanship predicting output classification scores on vaccine-related issues.

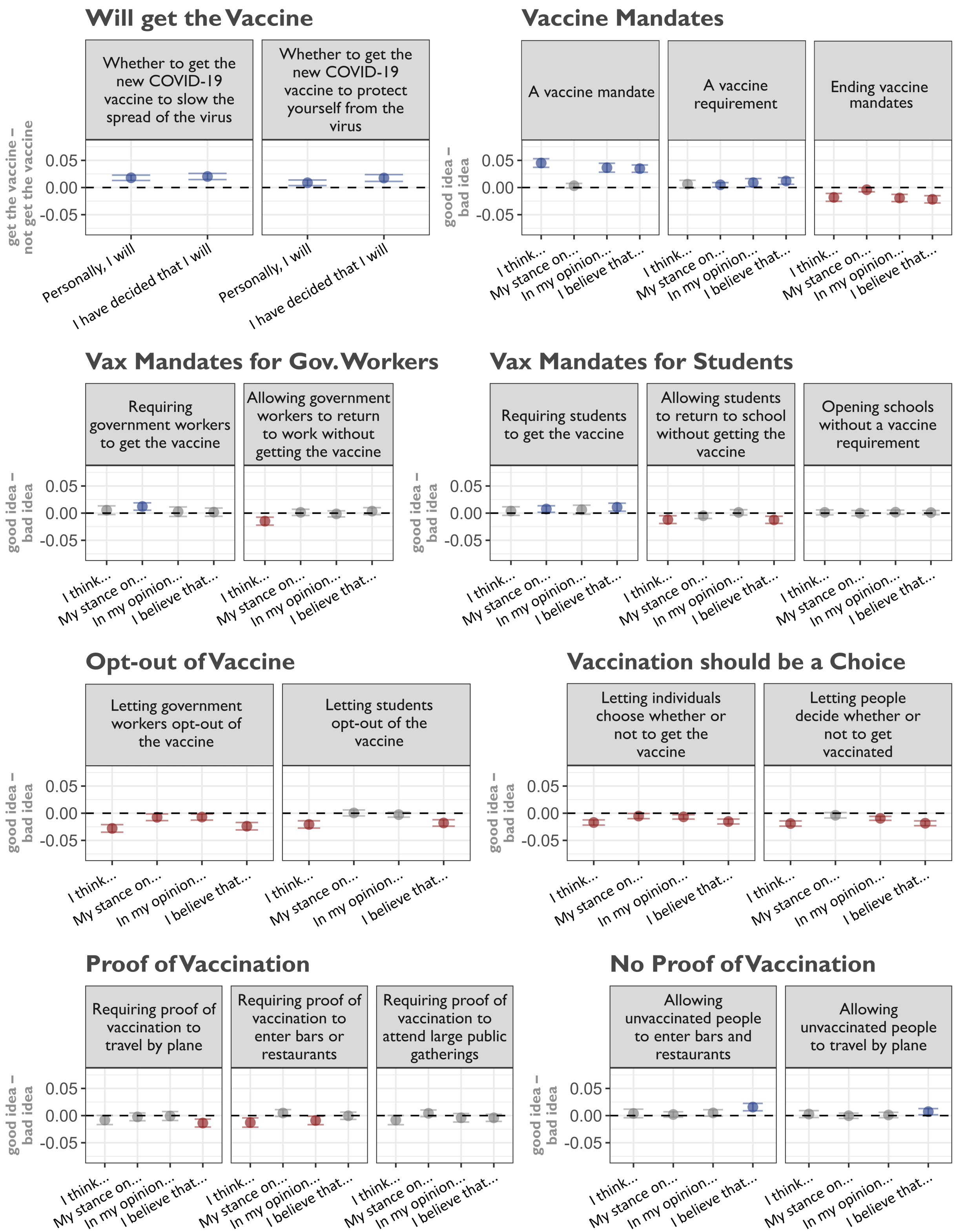
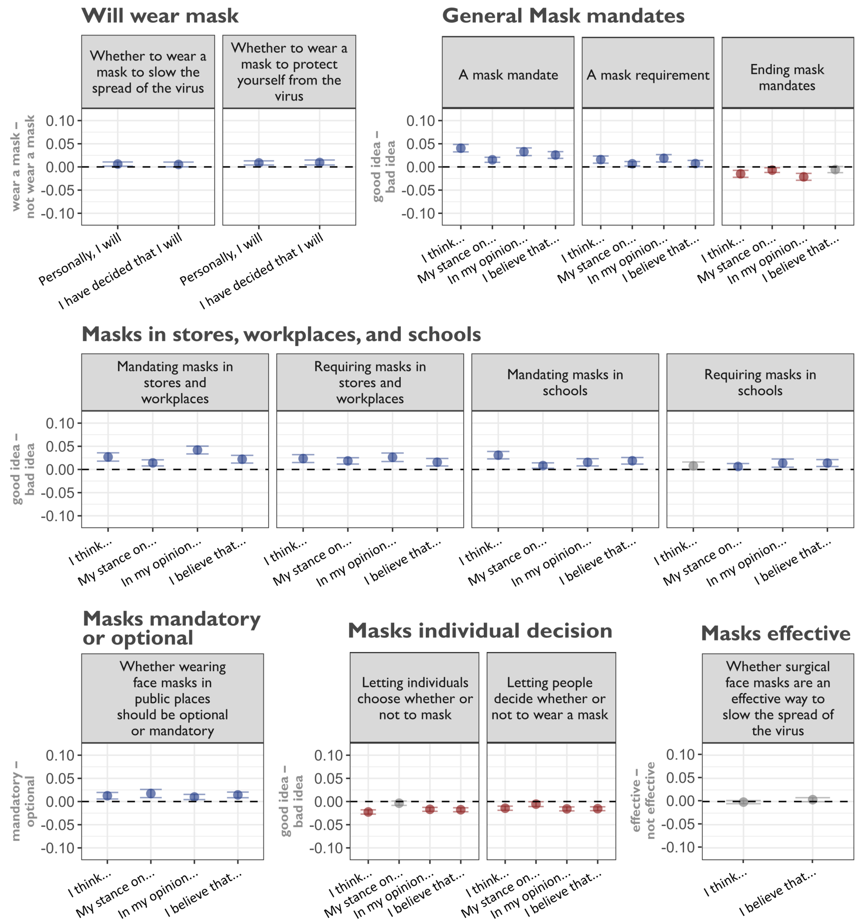


Figure 6: Coefficient plots from OLS models of partisanship predicting output classification scores on mask related issues.



virus.” The next prompt poses whether the virus is something to be afraid of. Intriguingly, conservative prompts were more likely to state that the virus is dangerous than the liberal prompt, an apparent inconsistency with the prior tendency of conservative prompts to state that the CDC is exaggerating the threat. Finally, we ask whether the virus originated in a lab or a wild animal. In these initial prompts, we do not observe a partisan difference. However, when we mention in the prompt that COVID-19 originated “in China,” conservative prompts are associated with higher likelihood than liberal prompts of speculating that the virus originated in a lab.

A collective overview of the results suggests that the partisan associations generated by GPT-3 mirror historical associations at a rate far outperforming chance. To assure that this assessment is correct, we conduct multiple tests of overall statistical significance in the Appendix. All tests, ranging from simple binomial tests to multi-level cross-classified models all find that GPT-3’s predictive capabilities outperform chance with p-values consistently below 0.001. Models are described and results are presented in Appendix.

Generated Justifications

GPT-3’s success in anticipating the future politicization of COVID-19 suggests that the model encodes discursive associations from its training texts that link liberalism with cautious responses to a novel virus and conservatism with a rejection of such measures. However, the preceding analyses tell us little about the nature of the associations that facilitate the model’s accurate predictions. To gain more detailed insight into how GPT-3 links pandemic issues to political ideology, we prompt the model to produce open-ended justifications for its previous responses, then we cluster those justifications by theme. We divide responses for each issue into two groups– positive and negative – and use k-means clustering within each of these groups to identify clusters of semantically similar justifications.

Figure 7: Coefficient plots from OLS models of partisanship predicting output classification scores on lockdown policies.

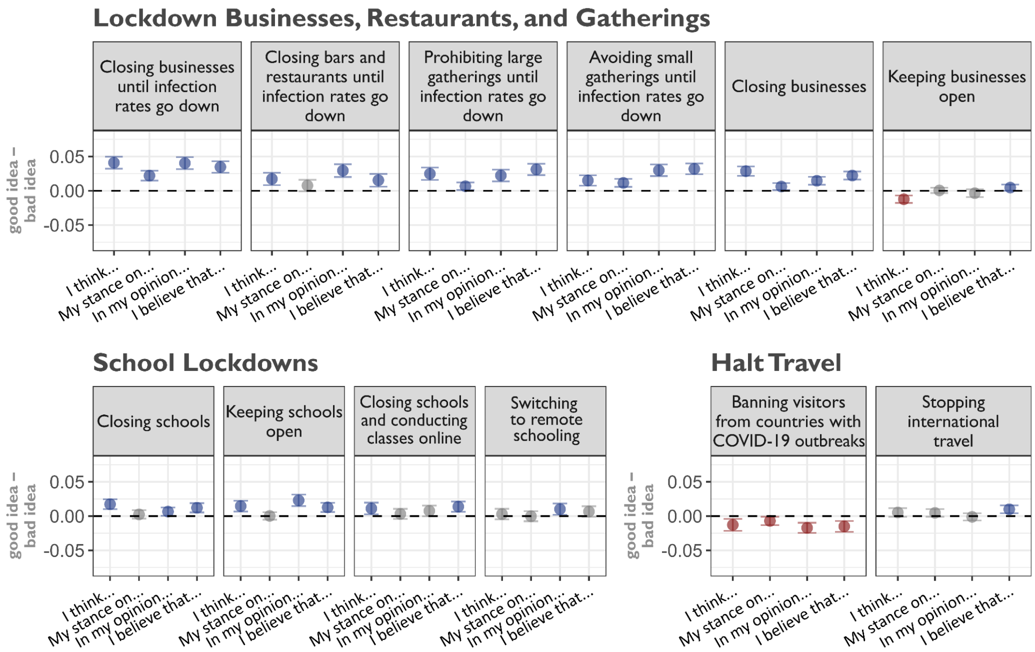
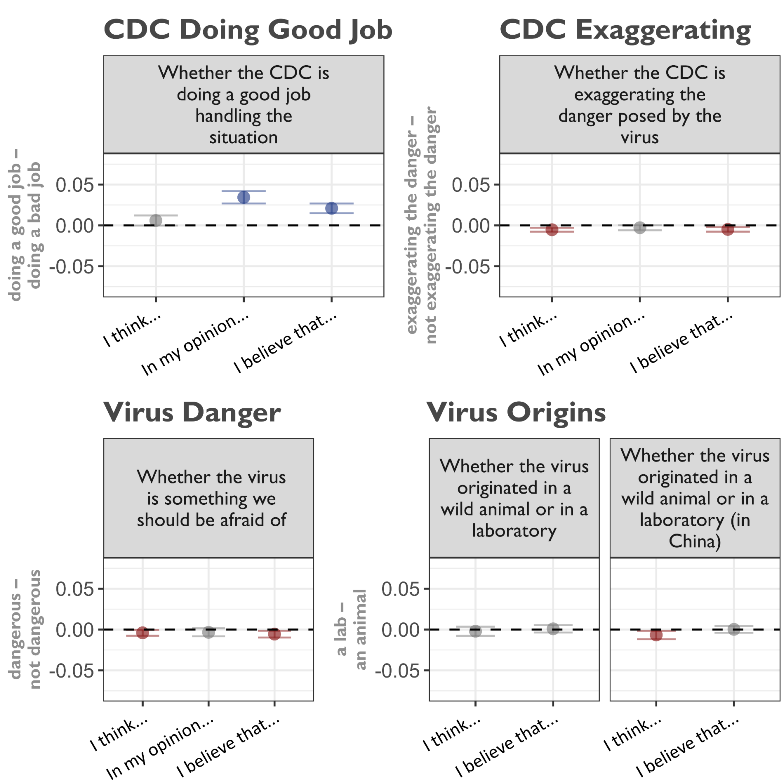


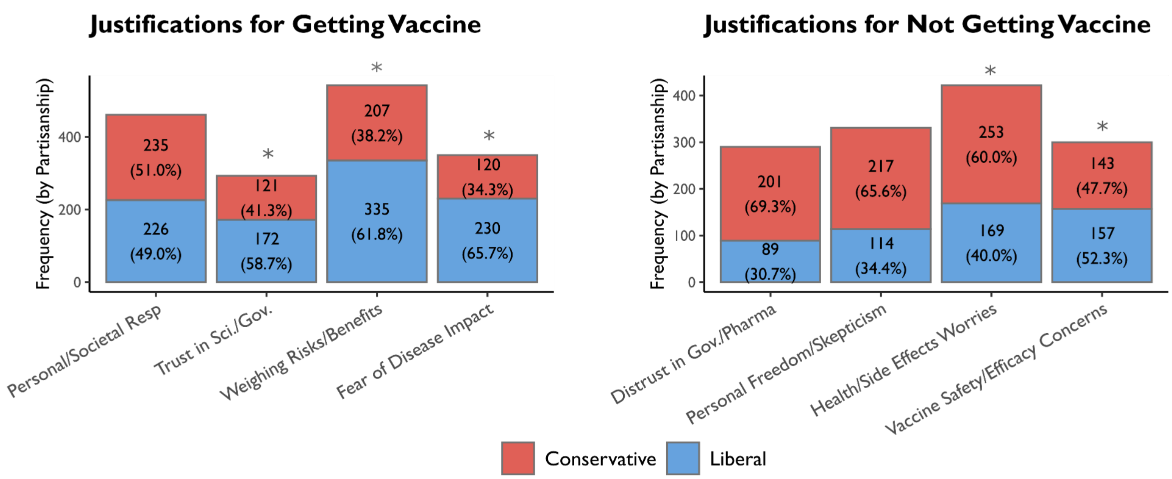
Figure 8: Coefficient plots from OLS models of partisanship predicting output classification scores on opinions toward the CDC and COVID-19 beliefs.



To provide the clearest insight into the system of relevant considerations, we first analyze two cases in which the model predicted particularly strong partisan divides: the intention to get vaccinated and opinions toward mask mandates. We also apply this approach to one issue where the model failed to correctly anticipate the direction of COVID-19 polarization: whether the virus is something that should be feared. While the model's incorrect predictions may simply be "errors", it is also possible that the model fails in situations where the history of relevant considerations suggest a response that did not manifest empirically. It is therefore plausible that incorrect predictions indicate points where history deviated from what an analysis of discourse alone would anticipate.

Figure 9 shows frequencies for each cluster of justifications in favor and opposed to getting the vaccine. The first panel displays four clusters of justifications for getting vaccinated, with the clusters in each plot ordered from most conservative to most liberal. Although liberal prompts disproportionately favor getting the vaccine, conservative prompts comprise a slight majority in the first cluster of justifications, Personal and Societal Responsibility (e.g. "I think that the American people have a duty to protect themselves and their families."). The remaining clusters are all majority liberal. In the first of these clusters, Trust in Science and Government, the decision to get vaccinated is justified by an appeal to the trustworthiness of the scientists and government agencies involved in the vaccine's production (e.g."I have a lot of trust in the government to make sure that the vaccine is safe and effective"). Justifications within the Weighing Risks / Benefits cluster often argue that the vaccine carries some risk but that this is far outweighed by the risk of contracting the virus. The final category, Fear of Disease's Impact, emphasizes the danger COVID-19 presents to both the respondent and the public at large. In sum, an appeal to personal or societal responsibility is a non-partisan justification for getting the vaccine, whereas appeals to the trustworthiness of the institutions backing the vaccine as well as appeals to safety tend to skew liberal.

Figure 9: Themes of justifications for getting / not getting the COVID-19 vaccine.



Note: * p<0.05 two-sample binomial test that a cluster's composition is more liberal than the most conservative (left side) cluster.

The justifications for *not* getting the vaccine also exhibit substantial differences in their political skew. The most conservative cluster is labeled Distrust in Government and Pharma (e.g. "the CDC and the FDA have been feeding us lies about the vaccine"). The following cluster, Personal Freedom / Skepticism, expresses the belief that the choice to vaccinate is highly personal (e.g "I believe that the right to choose what you put into your body is a fundamental human right") combined with occasional references to skepticism regarding the vaccine's safety (e.g. "I do not want to take the chance that it might cause me harm, which is why I am against the government mandating that everyone have this vaccine"). The final two clusters both highlight potential side-effects of the vaccine (e.g. "I believe the vaccine is unsafe") and concerns that the vaccine is either unsafe or ineffective (e.g. "the vaccine is still in its early stages and there are no guarantees that it will work") Interestingly, the Vaccine Safety / Efficacy Concerns cluster is majority liberal, not conservative. This suggests that the considerations that link conservatism to rejection of the vaccine by GPT-3 are not simply "anti-vax" sentiments, given that the partisan direction of this sentiment is equivocal.

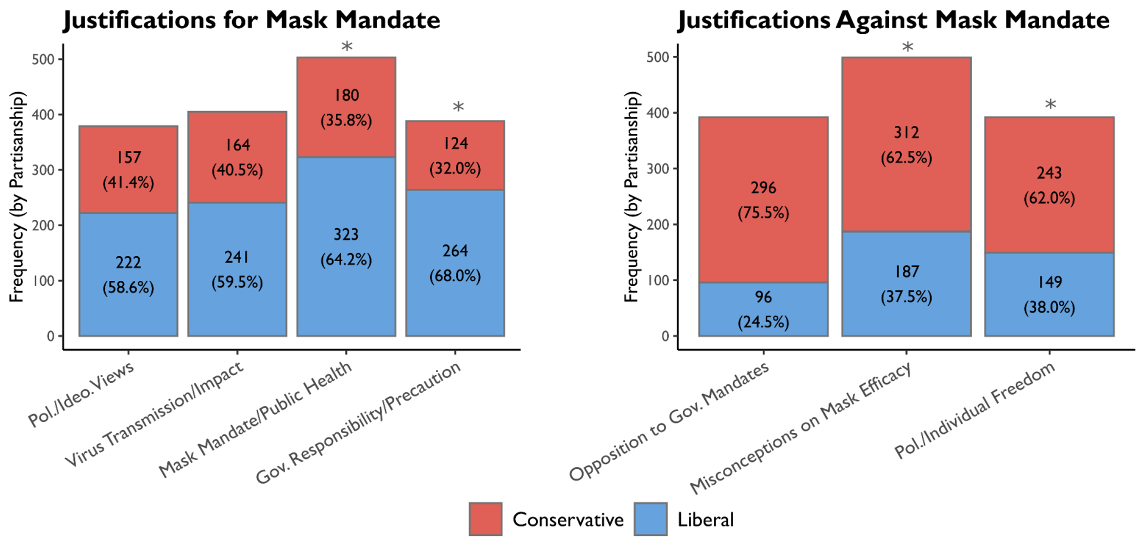
Table 1: Clusters of Justifications for Getting the Vaccine

Personal / Societal Responsibility - "I am a strong believer in personal responsibility" - "I am an adult who is able to make my own decisions" - "It is a matter of personal choice" - "I feel it is the right thing to do for my country" - "I want to make sure that my family is safe and protected" - "I think that the American people have a duty to protect themselves and their families" - "I think it is the responsible thing to do"	**Trust in Science and Government** - "I trust the science behind it and because of what I have learned about vaccines over the years" - "I have a lot of friends and family members who are doctors, nurses, scientists, or other medical professionals" - I am a strong liberal who trusts in science over faith - "I have a lot of respect for the Centers for Disease Control, and they seem to be doing an excellent job with this new disease" - "I have a lot of trust in the government to make sure that the vaccine is safe and effective"
Weighing Risks and Benefits - "I believe that the risk of getting this virus is far greater than any side effects from the vaccine" - "I am not a risk taker" - "I feel that the benefits of getting the vaccine outweigh any potential risks" - "I think that the risk of getting the virus is too great to depend on luck" - "I believe that the risk of getting a new COVID-19 infection is greater than the risk from side effects of the vaccine"	**Fear of Disease Impact** - "I do not want to be a victim of the virus and because I am concerned that it will cause a major pandemic if we do not slow its spread" - "I think that the vaccine will save thousands of lives in the long run" - "the virus has already killed over 1,000 people" - "I don't want to get sick and die" - "The virus is extremely deadly"

Table 2: Clusters for Justifications for Not Getting Vaccine

Distrust in Government and Pharma - "I do not trust the government to produce a safe vaccine" - "I know that the US government is a lying, corrupt institution" - "I am extremely skeptical of the government and its motives, but also because I don't think that this is a good use of my money" - "the CDC and the FDA have been feeding us lies about the vaccine" - "I believe the COVID-19 vaccine is a plan to make more money for big pharma and nothing else"	**Personal Freedom and Skepticism** - "I believe that the right to choose what you put into your body is a fundamental human right" - "I am a strong conservative and I believe in limited government" - "I am opposed to the government forcing me to do anything" - "I am an individual with the right to make my own decisions" - "I do not want to take the chance that it might cause me harm, which is why I am against the government mandating that everyone have this vaccine"
Health and Side Effect Worries - "I have seen too many drugs in the past that were not tested properly and caused more harm than good" - "I have seen too many examples of how a new drug or vaccine is rushed to market and then later found to be dangerous" - "I believe the vaccine is unsafe" - "I have heard that the vaccine is not very effective and many people are getting sick from it"	**Vaccine Safety and Efficacy Concerns** - "the vaccine has not been tested enough and could have unforeseen side effects" - "the vaccine is still in its early stages and there are no guarantees that it will work" - "the current vaccine is experimental and has not been completely tested yet" - "the vaccine is a new technology and there are many unknown risks involved"

Figure 10: Frequency of themes in justifications for and against Mask Mandates.



Note: * p<0.05 two-sample binomial test that a cluster's composition is more liberal than the most conservative (left side) cluster.

Comparing these two plots, we see several themes that help to explain partisan differences in intention to vaccinate. First, we observe a juxtaposition between the liberal-skewed Trust in Science and Government cluster in favor of vaccination and the highly conservative Distrust of Government and Pharma cluster justifying vaccine hesitancy. References to institutional trust among both positive and negative responses, along with its corresponding partisan skew, suggests that GPT-3 identifies confidence in government and scientific institutions as a major consideration in the decision to vaccinate as well as a core issue dividing liberals and conservatives. Another theme with partisan valence is the appeal to safety vs. freedom. The most liberal justifications for getting vaccinated emphasize dangers posed by the virus, and the most liberal justification for *not* getting vaccinated highlight potential dangers of the vaccine. By contrast, the more conservative themes emphasize personal responsibility and freedom, both of which foreground choice over protection from harm.

Figure 10 displays the clusters of justifications supporting and opposing mask mandates. In the first panel, we see that liberal prompts comprise the majority for all four clusters of justifications in favor of mask mandates. However, some topics are more liberally skewed than others. Explicit appeals to political ideology (e.g. "I think the Republicans have been wrong about this issue") and comments about the mode of transmission (e.g. "when people sneeze or cough, they spread the virus") were only slightly tilted toward liberals. On the other hand, appeals to the public health (e.g. "a mask mandate would lower the number of people who are exposed to this virus, which would protect the public health and save lives") and appeals to government responsibility (e.g. "the government has a responsibility to protect the public") are more disproportionately liberal.

Justifications *against* mask mandates are divided into three clusters, each of which is comprised of a majority of conservative responses. The most conservative cluster expresses a general opposition to government mandates (e.g. "I believe that the government should not be allowed to force people to wear masks"). The next cluster justifies opposition to mask mandates by questioning masks' efficacy (e.g. "a face mask is not going to stop you from getting sick"). The final cluster, which is still largely conservative, is characterized by appeals to political and personal freedom (e.g. "I believe that the freedom to choose is an important right").

Taken together, these results further clarify the partisan division on mask mandates. One key theme dividing liberal and conservative responses is their orientation toward government intervention; the most liberal justification for mandates emphasizes the government's responsibility to protect the public, whereas the most conservative justification for opposing mandates argues that this measure would be government overreach. It is worth noting that this theme is closely related to trust in the government, which previously emerged as a consideration dividing liberals and conservatives in their intention to vaccinate. Similarly, we see the reappearance of partisan differences in appeals to safety versus freedom. Liberal prompts are more likely to justify their views by emphasizing the risk to public health, while conservative prompts disproportionately highlight how such measures could jeopardize individual liberty. Lastly, we see a curious emergence of partisan differences in the factual question of mask efficacy – liberal responses argue that

Table 3: Clusters of Justifications for Mask Mandate

Political and Ideological Views - “it is being opposed by the Democrats” - “the GOP is responsible for spreading this virus” - “the right to life is more important than the right of free speech” - “the Democrats are better than the Republicans at solving problems” - “I believe that freedom is a good thing but you don’ t have a right to put other people at risk” - “I think the Republicans have been wrong about this issue and they have been very irresponsible” - “I am a strong liberal and I have the courage to stand up for what is right” - “they are weak-kneed liberals who want to pander to big business, the same way that Hillary Clinton did with Goldman Sachs”	**Virus Transmission and Impact** - “the virus is spread easily by physical contact, and wearing a mask will significantly reduce your chances of infection” - “the virus is spread through respiratory droplets that are expelled during sneezing and coughing” - “when people sneeze or cough, they spread the virus” - “if people don’t wear masks, the virus will spread” - “of the fact that there is no known cure for this disease and it can be easily transmitted from person to person through a cough or sneeze”
Mask Mandate and Public Health - “I think that the benefits of a mask mandate far outweigh any costs” - “there is a lot of evidence that in the past, public health measures like masks have been effective at preventing illness and death during pandemic outbreaks” - “a mask mandate would lower the number of people who are exposed to this virus, which would protect the public health and save lives” - “we should be willing to make sacrifices for the common good and because there is a scientific consensus that masks will help” - “the cost of a mask is small and it would be effective at stopping the spread of this deadly virus”	**Government Responsibility and Precaution** - “I have a strong belief that the government should try to protect its citizens from harm” - “I believe that the government should have a role in protecting individuals from disease and death” - “I have read the CDC's guidelines for dealing with this virus” - “I believe government regulation can be effective and is sometimes necessary”

Table 4: Clusters for Justifications *Against* Mask Mandate

Opposition to Government Mandates - “I believe that the government should not be allowed to force people to wear masks” - “it will strengthen the power of the government and take away individual freedom” - “I believe that it is a violation of the rights granted to us under the Constitution”	**Misconceptions on Mask Efficacy** - “of the simple fact that a face mask is not going to stop you from getting sick” - “the problem with face masks is that it gives people a false sense of security” - “the virus is spread by respiratory droplets, and a face mask does not stop droplets from getting into another person's mouth”	**Political and Individual Freedom** - “I believe that the freedom to choose is an important right - of the principle of individual liberty, which is central to conservatism” - “I believe in individual freedom and the free market”

masks capture respiratory droplets while conservative responses are skeptical of any protective effect.

Finally, we pivot to examining an issue that was *incorrectly* forecast by the model – expression of fear about the dangers of the virus. The frequencies for each cluster of justifications are presented in Figure 11. Justifications for fearing the virus are divided into five clusters. The most conservative-skewed justifications frame the virus as a threat to national security, commonly suggesting that it is a bioweapon and linking it to foreign countries (e.g. "I believe that the virus is a biological weapon released by China to attack America"). This is followed by concerns about the virus' lethality (e.g. "the virus has killed about 60% of those infected"). The Scientific Concerns and Characteristics cluster is also majority conservative and emphasizes the virus's rapid mutation and science's inability to keep up (e.g. "[because] of a lack of research and information about the virus"). The next two categories reveal a particularly instructive partisan divide over whether the virus should be feared. Justifications in the Political and Government Distrust category argue that the virus should be feared because government institutions are untrustworthy and incapable of handling the threat (e.g. "they are not competent to handle the situation"; "the government is trying to cover up a lot of things about this virus"). The most liberal cluster's LLM-generated label (Personal Experiences & Beliefs) is not particularly informative, but from a sample of its justifications we see that it combines appeals to personal experience (e.g. "I have a family member who is currently infected with the virus") with appeals to scientific expertise (e.g. "I trust experts and scientists").

Lastly, the right panel of Figure 11 displays clusters of justifications for *not* being afraid of the virus. The first cluster displays near partisan parity, and argues that the new coronavirus is no worse than other familiar viruses (e.g. "the symptoms are very similar to the flu"). The following three clusters which each display a slight liberal skew view the virus as a political ploy (e.g. "the pandemic is being used to further conservative causes"), question the danger of the virus (e.g. "the virus is most likely to be harmless"), and argue that the threat is overblown (e.g. "the media has an interest in creating a sense of panic"). The most intriguing results, however, come from the most liberal-skewed cluster, Trust in Government and Science. Responses in this cluster justify a less fearful response by expressing confidence that the government and science will effectively protect the public from the virus (e.g. "I believe in the power of science to solve problems" or "I put my trust in the government and their scientists").

We can thus see how a discursive association that led to accurate forecasts for most issues may have supported incorrect associations in this case; liberal's confidence in government and scientific institutions could have dispelled fears about the virus by providing assurance that the country is in good hands. Conversely, if COVID-19 had been framed as a national security threat originating in a foreign country, our results suggest that a cautious and vigilant response would be consistent with contemporary conservative ideology. Of course, this is not the path history took, but these alternative, imagined justifications are consistent with the sophisticated and detailed model of discourse encoded in GPT-3. To the extent that an LLM faithfully preserves the principles for generating a historic discourse, even its erroneous predictions may be illuminating.

DISCUSSION

LLMs can serve as powerful tools for the analysis of language and culture. These models are unique in that they do not merely provide a map of associations, they generate new texts consistent with the linguistic patterns and discursive styles on which they were trained. It is therefore possible to use LLMs to create "digital doubles" of actors fluent in discourses included in the model's training texts. In this study, we reconstructed the political opinion landscape of a pivotal period, the year preceding the spread of COVID-19, to determine whether the following politicization of the pandemic was predictable given the existing regime of discourse and politics. We find that the model predicts the correct direction of politicization far better than chance across a wide array of pandemic-related issues. To gain insight into how the model made these predictions, we prompt the LLM to produce justifications for its responses. We find on key issues that the distrust of institutions and the prioritization of personal freedom characteristic of American conservatism corresponded with their greater rates of rejecting sweeping policies to restrict the spread of the virus, whereas liberal prompts tend to justify strong collective responses to the virus with appeals to government responsibility and public safety. These results suggest that the way the pandemic politicized was largely consistent with existing repertoires of American liberalism and conservatism, and that the pandemic was not an occasion of substantial political innovation or surprise.

These findings speak to a fundamental question in the study of culture – to what extent are historical developments constrained by culture? When a new issue emerges, is it interpreted within an existing system of understanding, or is it "up for grabs," with political and cultural entrepreneurs offering competing frames for interpretation? Our findings suggest that, in the case of COVID-19's polarization, existing schemas of political sense-making steered the issue's reception. While President Trump's early statements downplaying the need for drastic responses may have helped crystallize the observed pattern of polarization, our evidence suggests that this stance was consistent with a general ideological tendency that predated the virus's emergence and channeled a zeitgeist already present among Trump's constituency. We do not deny that the political response to the COVID-19 pandemic could have unfolded differently,

Figure 11: Frequency of themes in justifications regarding fear COVID-19.

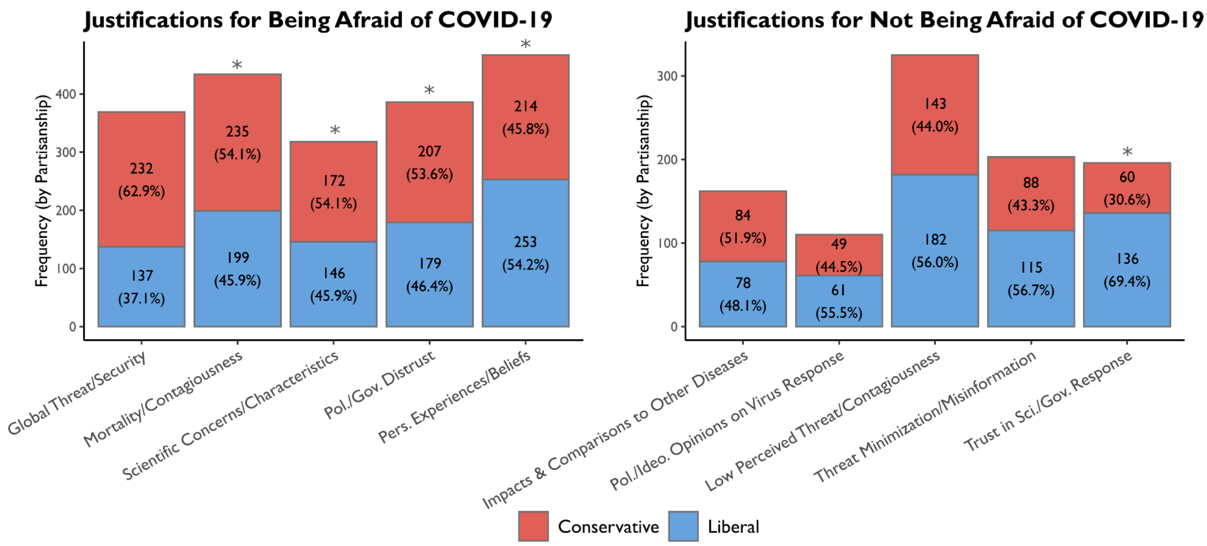


Table 5: Clusters for Justifications for Being Afraid of the COVID-19 Virus

Global Threat and Security	Mortality & Contagiousness	Scientific Concerns & Characteristics
- "I believe that the virus is a biological weapon released by China to attack America" - "there are many reasons to believe that the virus is a manmade weapon" - "the virus is being spread by illegal aliens and refugees" - "the virus is a potential threat to our health and the nation's economy"	- "the virus has a high mortality rate and because it is extremely contagious" - "the virus is so dangerous" - "it is very dangerous and hard to control" - "the virus has killed about 60% of those infected"	- the virus is a new mutation and we have no way to fight it - the fact that it is new, and we don't have any immunity to it - of a lack of research and information about the virus - there is a high probability that the virus will mutate into something more dangerous

Political and Gov. Distrust	Personal Experiences and Beliefs
- "the focus of our government has changed from serving the people to making money" - "they are not competent to handle the situation - the government has been caught lying about Ebola, Zika and other viruses in the past" - "the government is trying to cover up a lot of things about this virus" - "the CDC has admitted that they are unable to determine whether or not the virus is being transmitted from person-to-person"	- "I have studied the virus and its effects" - "I have a family member who is currently infected with the virus" - "I have read the research and I trust the scientists who are working on it" - "I have seen a lot of misinformation, and people are not getting the complete picture" - "I trust experts and scientists"

Table 6: Justifications for Not Being Afraid of the COVID-19 Virus

Impact & Comparison with other Diseases	**Political/Ideological Opinions on Virus Responses**	**Low Threat / Contagion**
- "the virus is very similar to a type of coronavirus that infects millions of people every year and causes mild illness" - "the symptoms are very similar to the flu" - "the virus is relatively new and has not evolved to be as deadly as other viruses we have faced in the past" - "the virus is similar to a cold virus and many people have already been exposed to it"	- "the media is controlled by the liberal elites who are trying to scare people into giving them more power" - "the COVID-19 pandemic has been used to promote a left wing agenda" - "the Republican party has been taken over by a bunch of very stupid people" - "the pandemic is being used to further conservative causes"	- "the virus does not spread easily between people - "it is not airborne - "it's only spread through direct contact with bodily fluids, and the virus does not survive outside of a host" - "the virus is most likely to be harmless"

Threat Minimization and Misinformation	**Trust in Sci./Gov. Response**
- "I have been studying the virus and it is not as dangerous as people are saying" - "the virus is not even a major health issue" - "the media is exaggerating the threat that COVID -19 poses to our country" - "the media has an interest in creating a sense of panic"	- "I believe that the CDC is a trusted source of information for us all" - "I put my trust in the government and their scientists" - "I think that the government is doing a good job of preparing for it and taking action" - "I believe that this administration is already taking the threat of COVID-19 seriously"

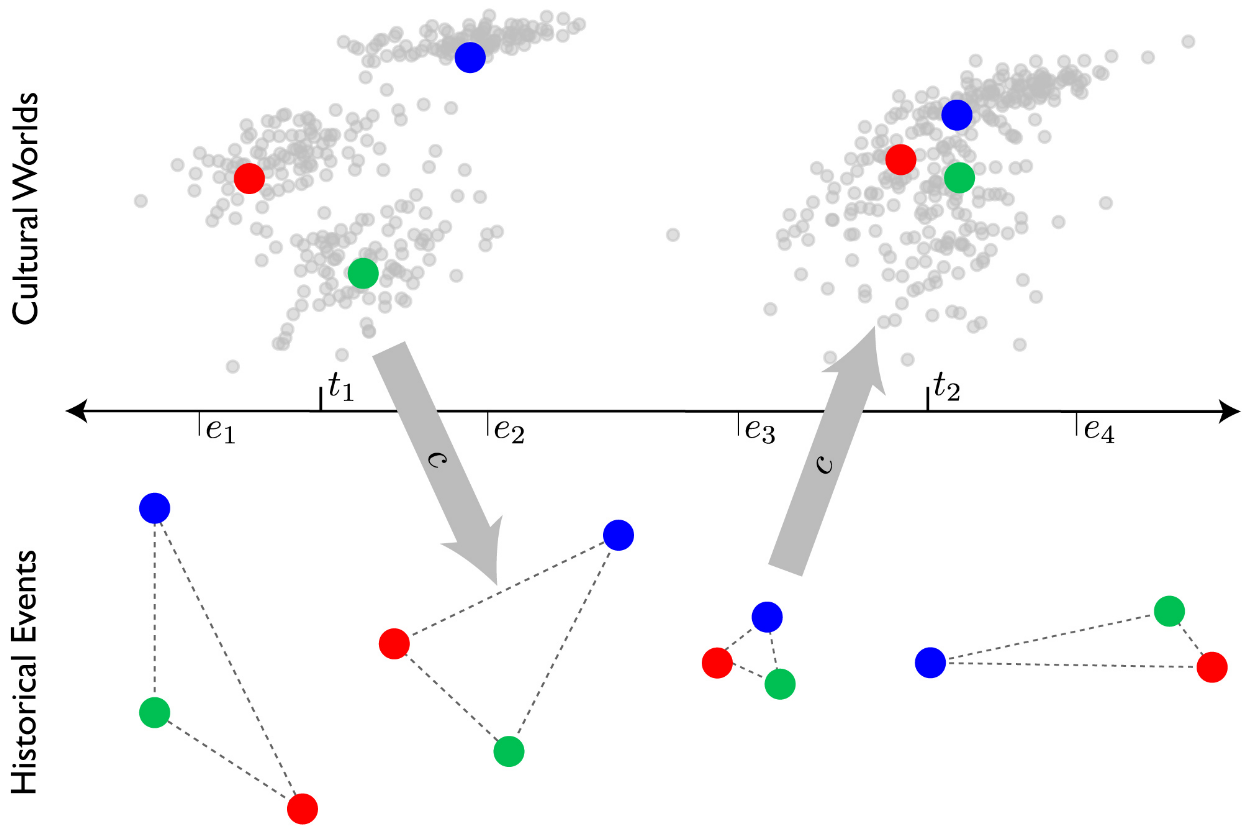
but our evidence suggests that in this case, history manifested along the predictable path consistent with the features of liberalism and conservatism in 2019.

It is noteworthy that the considerations the LLM uses to justify its stances often deviate from those emphasized within political psychology. Many of the pillars of conservatism offered by political psychologists — uncertainty avoidance, fear of threat, need for order and structure, and respect for authority (Jost 2006) — would have likely motivated opposite responses to many COVID-19 policies. Nevertheless, these considerations were not well represented among produced justifications. Instead, LLM-generated justifications more often made appeals to personal freedom and skepticism toward governmental and scientific authorities, which were consistent with conservatives' observed resistance to large-scale restrictions.

Our study presents one potential application of LLMs to the analysis of cultural systems, but a variety of alternative approaches are likely to prove fruitful for future inquiry. First, our analytic approach benefits from historical happenstance — the first human-level LLM was trained on texts published immediately prior to an unprecedented socio-political event. In order for LLMs to be a general tool for the analysis of socio-cultural change, researchers should intentionally train models sequentially on ordered time periods, releasing model "checkpoints" with training restricted to different periods of training texts. This pattern of sequential training mirrors "curriculum learning" in which training examples are presented to a model in sequence from simple to complex (Bengio et al. 2009), but we propose a method that instead proceeds chronologically, encouraging the model to learn the present in terms of the past, as in the historic unfolding of social life. This sequential training approach enables two key analytic opportunities. First, by comparing models trained on subsequent periods, an analyst can identify the rate and direction of historical cultural shifts. Second, by comparing periods immediately prior and following moments of important social framing, the influence of an event can be assessed.

Figure 12 schematically represents how time-stamped, sequentially trained LLMs could enable a "causal cultural analysis." The top row displays representations of cultural systems drawn from time-stamped LLMs, here rendered as two dimensional point clouds. "Historical events" are rendered below as networked associations between three focal concepts (e.g., red, green, and blue). Linking these two levels of observation could facilitate an empirical operationalization of Giddens' structuration theory (Giddens 1984) or Coleman's "boat" models of macro-micro-macro relations (Coleman 1994). In our example, cultural world models are checkpointed at two times: t_1 and t_2. Irregular events e_1 through e_4 each render the focal three concepts, which could represent ideas, facts, stereotypes, or narrative elements with different orientations to one another and the rest of the cultural world. In this scenario, event e_2 is wholly unsurprising, as it perfectly reproduces the standing system of cultural relations. By contrast, event e_3 not only deviates from the prior cultural system, but changes the cultural world, becoming typical in the subsequent time period. Most surprising associations (e.g., e_1 and e_4) do not change the cultural world. These represent the failed innovations, unfunny jokes, and unfortunate accidents of history that do not cause anything but their own forgetting.

Figure 12: The causal opportunity associated with time-stamped LLM traces of cultural worlds and events.



Using LLMs as cultural world models, a wide range of causal identification strategies (Pearl 2009) could be deployed in order to do causal cultural history—to identify when cultural events like speeches, concerts, new products, or viral memes change the space of associations; and when the space of associations changes the distribution of cultural events that follow. Our ability to generate speech events from cultural world models, and evolve cultural world models from speech events through fine-tuning enables the production of richly situated counterfactuals for probabilistic identification. This potential for cultural measurement and identification could allow us to relax the assumed conservation of influence from the macro-cultural ether to individual cultural behaviors and back again, suggesting when change is conditioned and disproportionately driven from the "bottom-up" or the "top-down".

We note that there are several practical reasons why LLMs are currently not developed through chronological training, including: (1) the lower average quality of earlier text, leading of path-dependent low-quality models; (2) the inaccurately dated character of text on the web, LLM's dominant corpus; (3) the phenomena of catastrophic forgetting whereby deep neural networks learn new tasks without retaining the ability to perform past ones (McCloskey and Cohen 1989; Ratcliff 1990); and (4) the associated problem of mixing styles that allows LLMs to efficiently link text across a large asynchronous corpus. However, many of these limitations can be engineered around, albeit with non-trivial expense. Early text can be corrected, or translated into contemporary formats. Web-text can be dated using time-stamped informal text and news. Earlier texts can be incorporated into new training samples in proportion to metrics reflecting their circulation and presence in current collective memory, such as their recent citation rate.

Beyond this, LLMs present tremendous new opportunities to simulate complex interactions between "encultured" agents. Our study uses LLMs to generate simulated data that resemble responses to open-ended survey questions, in which each response is treated as an independent sample from a population of possible responses. But social life emerges through interaction, and LLMs enable the simulation of semantically-rich interaction at scale. Many forms of consequential social interactions could be analyzed *in silico* that would be challenging to reproduce in a laboratory or online, such as parliamentary discourse, scientific conferences, or cross-cultural collaborations. Moreover, simulated interactions can be easily run millions of times over with initial conditions experimentally manipulated to encourage wide coverage of the space of possible interaction outcomes. Findings inductively discovered from an extensive search over the interaction space could then be validated with human subjects, or for certain hard-to-observe social phenomena for which validation is impossible, simulation may ultimately provide the best obtainable evidence.

Lastly, the internal representations of AI agents are directly observable in a way that human internal representations are not, and may therefore yield key insights into how complex meaning systems can be efficiently modeled and compressed. Although the activation patterns of deep neural networks are commonly described as black boxes, the growing field of mechanistic interpretability is making progress in mapping relations between neuronal activation and model behavior (Bills et al. 2023; Elhage et al. 2022). More and more findings demonstrate the lower-level "neurons" with the transformer's self-attention architecture facilitate novel interpretive and simulational capacities (Hendel, Geva, and Globerson 2023). As this field matures, it may enable an "AI neuroscience," whereby differences in behavior or discourse can be translated into differences in latent representation. Although it is likely that lessons from artificial neural networks cannot be directly applied to human brains, their visibility and manipulability make them a valuable site for exploratory studies linking distributed representations to phenomena of culture and cognition.

As AI models continue to achieve more accurate approximations of human behavior, they will become increasingly powerful tools for the analysis of complex social processes. The advent of LLMs has now made simulation studies of discourse possible for the first time, opening many new avenues for investigation into language, culture, and meaning. Moreover, just as LLMs trained on massive collections of text learn the patterns of language, multi-modal models trained on images, videos, and other records of social life can similarly distill underlying patterns across these diverse domains (Guilbeault et al. 2024; Ludwig and Mullainathan 2024). Social scientists are fundamentally interested in what people do, and direct observation of human action remains the essential cornerstone of our science. But as empirically realistic agents can be achieved *in silico*, social simulations will offer opportunities to explore social phenomena beyond the observable, shedding light on the underlying patterns and branching pathways that structure social life and give rise to its complex and varied forms.

REFERENCES

Albrecht, Don. 2022. "Vaccination, Politics and COVID-19 Impacts." *BMC Public Health* 22(1):96.

Allcott, Hunt et al. 2020. "Polarization and Public Health: Partisan Differences in Social Distancing during the Coronavirus Pandemic." *Journal of Public Economics* 191:104254.

Argyle, Lisa P. et al. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31:337–51.

Baldassarri, Delia and Amir Goldberg. 2014. "Neither Ideologues nor Agnostics: Alternative Voters' Belief System in an Age of Partisan Politics." *American Journal of Sociology* 120(1):45–95.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. "Curriculum Learning." Pp. 41–48 in *Proceedings of the 26th Annual International Conference on Machine Learning*, *ICML '09*. Montreal, Quebec, Canada: Association for Computing Machinery.

Bills, Steven et al. 2023. "Language Models Can Explain Neurons in Language Models." *URL Https://Openaipublic. Blob. Core. Windows. Net/Neuron-Explainer/Paper/Index. Html. (Date Accessed: 14. 05. 2023)*.

Bonikowski, Bart, Yuchen Luo, and Oscar Stuhler. 2022. "Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models." *Sociological Methods & Research* 51(4):1721–87.

Boutyline, Andrei, Alina Arseniev-Koehler, and Devin J. Cornell. 2023. "School, Studying, and Smarts: Gender Stereotypes and Education across 80 Years of American Print Media, 1930–2009." *Social Forces* 102(1):263–86.

Boutyline, Andrei and Stephen Vaisey. 2017. "Belief Network Analysis: A Relational Approach to Understanding the Structure of Attitudes." *The American Journal of Sociology* 122(5):1371–1447.

Brown, Tom B. et al. 2020. "Language Models Are Few-Shot Learners." *ArXiv* 2005.14165.

Brynjolfsson, Erik. 2023. "The Turing Trap: The Promise & Peril of Human-like Artificial Intelligence." Pp. 103–16 in *Augmented education in the global age*. Routledge.

Callaghan, Timothy, Matthew Motta, Steven Sylvester, Kristin Lunz Trujillo, and Christine Crudo Blackburn. 2019. "Parent Psychology and the Decision to Delay Childhood Vaccination." *Social Science & Medicine (1982)* 238:112407.

Carmines, Edward G. and James A. Stimson. 1989. *Issue Evolution: Race and the Transformation of American Politics*. Princeton University Press.

Chen, Hsueh-Fen and Saleema A. Karim. 2022. "Relationship between Political Partisanship and COVID-19 Deaths: Future Implications for Public Health." *Journal of Public Health* 44(3):716–23.

Coleman, James S. 1994. *Foundations of Social Theory*. Harvard University Press.

Colgrove, James. 2006. *State of Immunity: The Politics of Vaccination in Twentieth-Century America*. University of California Press.

Conis, Elena. 2014. *Vaccine Nation*. University of Chicago Press.

Converse, Philip. 1964. "The Nature of Belief Systems in Mass Publics." in *Ideology and discontent*, edited by D. Apter. New York: Free Press.

Dai, Damai et al. 2023. "Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers."

Deane, C., K. Parker, and J. Gramlich. n.d. "A Year of US Public Opinion on the Coronavirus Pandemic."

Degrave, Jonas et al. 2022. "Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning." *Nature* 602(7897):414–19.

Dell'Acqua, Fabrizio et al. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality."

DellaPosta, Daniel. 2020. "Pluralistic Collapse: The 'Oil Spill' Model of Mass Opinion Polarization." *American Sociological Review* 85(3):507–36.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv*.

DiMaggio, Paul. 1997. "Culture and Cognition." *Annual Review of Sociology* 23(1):263–87.

Douglas, Mary. 1966. *Purity and Danger: An Analysis of Concepts of Pollution and Taboo*. Routledge & Kegan Paul.

Elhage, Nelson et al. 2022. "Toy Models of Superposition." *ArXiv* 2209.10652.

Feldman, Stanley and John Zaller. 1992. "The Political Culture of Ambivalence: Ideological Responses to the Welfare State." *American Journal of Political Science* 36(1):268–307.

Furr, Michael. 2021. *Psychometrics: An Introduction*. SAGE Publications.

Gadarian, Shana Kushner, Sara Wallace Goodman, and Thomas B. Pepinsky. 2021. "Partisanship, Health Behavior, and Policy Attitudes in the Early Stages of the COVID-19 Pandemic." *PloS One* 16(4):e0249596.

Gagniuc, Paul A. 2017. *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

Garcia-Pardina, Alejandro, Francisco José Abad, Alexander P. Christensen, Hudson Golino, and Luis Eduardo Garrido. 2022. "Dimensionality Assessment in the Presence of Wording Effects: A Network Psychometric and Factorial Approach." *PsyArXiv*.

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* 115(16):E3635–44.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3):535–74.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press.

Goffman, Erving. 2021. *The Presentation of Self in Everyday Life*. Knopf Doubleday Publishing Group.

Goren, Paul, Harald Schoen, Jason Reifler, Thomas Scotto, and William Chittick. 2016. "A Unified Theory of Value-Based Reasoning and U.S. Public Opinion." *Political Behavior* 38(4):977–97.

Green, Donald P., Bradley Palmquist, and Eric Schickler. 2004. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Guilbeault, Douglas et al. 2024. "Online Images Amplify Gender Bias." *Nature* 626(8001):1049–55.

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Knopf Doubleday Publishing Group.

Hannan, Michael. 2022. "Measuring Memberships in Collectives in Light of Developments in Cognitive Science and Natural-Language Processing." *Sociological Science* 9:473–92.

Hayes, Brian. 2013. "First Links in the Markov Chain." *American Scientist* 101(2):92.

Helzer, Erik G. and David A. Pizarro. 2011. "Dirty Liberals! Reminders of Physical Cleanliness Influence Moral and Political Attitudes." *Psychological Science* 22(4):517–22.

Hendel, Roee, Mor Geva, and Amir Globerson. 2023. "In-Context Learning Creates Task Vectors." *ArXiv* 2310.15916.

Homer, Jenny and Michael French. 2009. "Motorcycle Helmet Laws in the United States from 1990 to 2005: Politics and Public Health." *American Journal of Public Health* 99(3):415–23.

Hunzaker, M. B. Fallin and Lauren Valentino. 2019. "Mapping Cultural Schemas: From Theory to Method." *American Sociological Review* 84(5):950–81.

Jamison, Amelia M., Sandra Crouse Quinn, and Vicki S. Freimuth. 2019. "'You Don't Trust a Government Vaccine': Narratives of Institutional Trust and Influenza Vaccination among African American and White Adults." *Social Science & Medicine (1982)* 221:87–94.

Ji, Ziwei et al. 2023. "Towards Mitigating Hallucination in Large Language Models via Self-Reflection." *ArXiv* 2310.06271.

Jiao, Yutao, Ping Wang, Dusit Niyato, Bin Lin, and Dong In Kim. 2021. "Toward an Automated Auction Framework for Wireless Federated Learning Services Market." *IEEE Transactions on Mobile Computing* 20(10):3034–48.

Jiménez-Luna, José, Francesca Grisoni, and Gisbert Schneider. 2020. "Drug Discovery with Explainable Artificial Intelligence." *Nature Machine Intelligence* 2(10):573–84.

Jost, John T. 2006. "The End of the End of Ideology." *The American Psychologist* 61(7):651–70.

Jost, John T. 2017. "Ideological Asymmetries and the Essence of Political Psychology." *Political Psychology* 38(2):167–208.

Kaplan, Jared et al. 2020. "Scaling Laws for Neural Language Models." *ArXiv* 2001.08361.

Kaplan, Sarah. 2008. "Framing Contests: Strategy Making under Uncertainty." *Organization Science* 19(5):729–52.

Kiley, Kevin and Stephen Vaisey. 2020. "Measuring Stability and Change in Personal Culture Using Panel Data." *American Sociological Review* 85(3):477–506.

Kim, Junsol and Byungkyu Lee. 2023. "AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys." *ArXiv* 2305.09620.

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.

Lai, Shiyang et al. 2024. "Evolving AI Collectives to Enhance Human Diversity and Enable Self-Regulation." *ArXiv* 2402.12590.

Lakoff, George. 2010. *Moral Politics: How Liberals and Conservatives Think, Second Edition*. University of Chicago Press.

Le Mens, Gaël, Balázs Kovács, Michael Hannan, and Guillem Pros. 2023. "Using Machine Learning to Uncover the Semantics of Concepts: How Well Do Typicality Measures Extracted from a BERT Text Classifier Match Human Judgments of Genre Typicality?" *Sociological Science* 10:82–117.

Lévi-Strauss, Claude. 1966. *The Savage Mind*. University of Chicago Press.

Ludwig, J. and S. Mullainathan. 2024. "Machine Learning as a Tool for Hypothesis Generation." *The Quarterly Journal of Economics*.

McCarthy, Justin. 2023. "Roundup of Gallup COVID-19 Coverage." *Gallup.Com*. Retrieved (https://web.archive.org/web/20230322222821/https://news.gallup.com/opinion/gallup/308126/roundup-gallup-covid-coverage.aspx).

McCloskey, Michael and Neal J. Cohen. 1989. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem." Pp. 109–65 in *Psychology of learning and motivation*, vol. 24, edited by G. H. Bower. Academic Press.

Mead, Margaret. 1942. *And Keep Your Powder Dry*. New York: William Morrow.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *ArXiv* 1301.3781.

Mills, C. Wright. 1940. "Situated Actions and Vocabularies of Motive." *American Sociological Review* 5(6):904–13.

Mollick, Ethan R. and Lilach Mollick. 2023. "Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts."

Mosavi, Amir et al. 2020. *Comprehensive Review of Deep Reinforcement Learning Methods and Applications in Economics*. SSRN.

Neelakantan, Arvind et al. 2022. "Text and Code Embeddings by Contrastive Pre-Training." *ArXiv* 2201.10005.

Noel, Hans. 2014. *Political Ideologies and Political Parties in America*. Cambridge University Press.

OpenAI. 2023a. "GPT-4 Technical Report." *ArXiv* 2303.08774.

OpenAI. 2023b. "Models." Retrieved (https://platform.openai.com/docs/models/).

Ouyang, Long et al. 2022. "Training Language Models to Follow Instructions with Human Feedback" edited by S. Koyejo et al. *ArXiv* 2203.02155.

Oxley, Douglas R. et al. 2008. "Political Attitudes Vary with Physiological Traits." *Science* 321(5896):1667–70.

Page, Benjamin I. and Robert Y. Shapiro. 2010. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. University of Chicago Press.

Park, Joon Sung et al. 2022. "Social Simulacra: Creating Populated Prototypes for Social Computing Systems." Pp. 1–18 in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, *UIST '22*. Bend, OR, USA: Association for Computing Machinery.

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

Pennington, J., R. Socher, and C. Manning. 2014. "Glove: Global Vectors for Word Representation." *Proceedings of the 2014*.

Pew Research Center. 2014. "Ebola Worries Rise, But Most Are 'fairly' Confident in Government, Hospitals to Deal with Disease: Broad Support for US Efforts to Deal with Ebola in West Africa."

Ratcliff, R. 1990. "Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions." *Psychological Review* 97(2):285–308.

Rawlings, Craig M. 2020. "Cognitive Authority and the Constraint of Attitude Change in Groups." *American Sociological Review* 85(6):992–1021.

Rawlings, Craig M. 2022. "Becoming an Ideologue: Social Sorting and the Microfoundations of Polarization." *Sociological Science* 9:313–45.

Rosenfeld, R. 2000. "Two Decades of Statistical Language Modeling: Where Do We Go from Here?" *Proceedings of the IEEE* 88(8):1270–78.

Rosentiel, Tom. 2011. "Public Remains Divided over the Patriot Act." Retrieved (https://policycommons.net/artifacts/624569/public-remains-divided-over-the-patriot-act/1605857/).

Senior, Andrew W. et al. 2020. "Improved Protein Structure Prediction Using Potentials from Deep Learning." *Nature* 577(7792):706–10.

Shannon, C. E. 1951. "Prediction and Entropy of Printed English." *Bell System Technical Journal* 30(1):50–64.

Sourati, Jamshid and James A. Evans. 2023. "Accelerating Science with Human-Aware Artificial Intelligence." *Nature Human Behaviour*.

Stokes, Erin K. et al. 2020. “Coronavirus Disease 2019 Case Surveillance - United States, January 22-May 30, 2020.” *MMWR. Morbidity and Mortality Weekly Report* 69(24):759–65.

Stoltz, Dustin S. and Marshall A. Taylor. 2019. “Concept Mover’s Distance: Measuring Concept Engagement via Word Embeddings in Texts.” *Journal of Computational Social Science* 2(2):293–313.

Sutskever, Ilya. 2023. “An Observation on Generalization.”

Swidler, Ann. 2003. “Talk of Love: How Culture Matters.” *Canadian Journal of Communication* 28(1).

Terrizzi, John A., Natalie J. Shook, and Michael A. McDaniel. 2013. “The Behavioral Immune System and Social Conservatism: A Meta-Analysis.” *Evolution and Human Behavior : Official Journal of the Human Behavior and Evolution Society* 34(2):99–108.

Tesler, Larry. 2010. “Adages and Coinages.” Retrieved (https://web.archive.org/web/20100223000629/https://www.nomodes.com/Larry_Tesler_Consulting/Adages_and_Coinages.html).

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Turing, A. M. 1950. “Computing Machinery and Intelligence.” *Mind; a Quarterly Review of Psychology and Philosophy* LIX(236):433–60.

Vaswani, Ashish et al. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems* 30.

Vicinanza, Paul, Amir Goldberg, and Sameer B. Srivastava. n.d. “Who Sees the Future? A Deep Learning Language Model Demonstrates the Vision Advantage of Being Small.”

Von Oswald, Johannes et al. 2023. “Transformers Learn In-Context by Gradient Descent.” Pp. 35151–74 in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, *Proceedings of Machine Learning Research*, edited by A. Krause et al. PMLR.

Wilkins, Alex. 2023. “AI Could Discover Wonder Materials.” *New Scientist* 260(3467):8.

Willis, Gordon B. 2004. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. SAGE Publications.

Xue, Linting et al. 2020. “mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer.” *ArXiv* 2010.11934.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Zhou, Quan et al. 2018. “Learning Atoms for Materials Discovery.” *Proceedings of the National Academy of Sciences of the United States of America* 115(28):E6411–17.