Counterfactual Thinking:
Social Simulation with Language Models for the Digital Humanities

Sil Hamilton

McGill University, Montreal

Languages, Literatures, and Cultures

November 2023

A thesis submitted to McGill University in partial
fulfillment of the requirements of the degree of Master of Arts.

Abstract

English

The last five years have seen a neural revolution take the sciences by storm, unleashing a wave of novel and innovative research methods taking advantage of the unprecedented modelling capabilities enabled by large language models. How those in the humanities and the social sciences stand to benefit from these newfound generative models remains unclear. This thesis takes a tentative step in solving this dilemma by harnessing language models for simulative purposes, a fundamental experimental technique heretofore under-realized in the digital humanities. We present three papers exploring how those in communication studies, political science, and literature studies can uncover new knowledge through the simulation of creative acts.

Français

Les cinq dernières années ont vu une révolution neuronale prendre d'assaut les sciences, libérant une vague de méthodes de recherche nouvelles et innovantes profitant des capacités de modélisation sans précédent activées par de grands modèles de langage. La façon dont ceux des sciences humaines et des sciences sociales bénéficient de ces nouveaux modèles génératifs reste incertain. Cette thèse fait un pas provisoire dans le sens de l'exploitation de modèles de langage pour les tests d'hypothèse contrefactuels par la simulation, une technique fondamentale d'expérimentation hérétofore sous-réalisée dans les sciences humaines numériques. Nous présentons trois manuscrits explorant comment ceux des études de communication, des sciences politiques et des études culturelles peuvent révéler de nouvelles connaissances grâce à la simulation d'actes créatifs.

Acknowledgements

I declare I was the only one responsible for the collection of materials and data, performance of experiments, analysis of data described within the following pages. Professor Andrew Piper assissted me with experimental design, and offered guidance when connecting findings to the bigger picture. I am responsible for preparing and editing this thesis. No members of staff, fellow students, research assistants, technicians, etc., participated in the creation of this thesis.

I would first like to thank my advisor, Professor Andrew Piper. It is difficult to understate how much his generosity in entertaining my research whims these past three years contributed to where I am now. He is always willing to offer both advice and guidance. When I first reached out to him arge language models were not as popular as they are now–and his openness to considering their utility before the release of GPT-3 belies a particular frame of mind I can only hope to one day match.

I would also to thank Professor Richard Jean So and Professor Tim O'Donnell for contributing valid points and criticisms to my antics. I have approached both with the wildest of understandings, and to their credit, both took the time to carefully explain their own perspective on various matters.

Contribution of Authors

A majority of the work in this thesis was written by yours truly. The introduction, literature review, discussion, and conclusion were all written by me. Of the three manuscripts, I am the sole author of two.

For the first manuscript—Counterfactual Journalism—I carry first authorship. Professor Andrew Piper's contributions were largely limited to guiding my experiments and proposing revisions to my initial draft of the paper. Of these revisions one will wish to know Professor Piper held large influence over the development of the first three paragraphs of the background, and over the steps for future researchers as described in the conclusion. The reader will note these sections are largely expository—the experimental design and consequent write-up were enacted by the author of this thesis. I encourage the reader to assess the first paper with this caveat in mind; and to look to the introduction, methods, and results to properly grasp my own contribution in this thesis.

Introduction

For all of their pizzazz, early mainframes were little more than tabulating machines. IBM and Remington Rand, the first companies to bring the digital computer to consumers, had built their fortunes on analog counting machines whose sole purpose was to tabulate punched cards (O'Regan and O'Regan, 2018). The tabulating machine won its first success in the 1890 American census, where it automated the task of counting each American household. Both governments and companies were sold on the accounting efficiencies; tabulating machines soon became a standard part of the corporate environment. Market demands motivated their development, with early-twentieth century machines advertising abilities for sorting, summarizing, printing, and even remotely manipulating data with so-called "teletypes" whose direct descendant would be the computer terminal (O'Regan and O'Regan, 2018). IBM would sell tabulators in parallel with mainframes for some three decades; their adequate accounting abilities obscuring the potential of the mainframe. Corporate life revolved around tabulation, and so companies considered mainframes costly tabulators. That mainframes could simulate models of tabulated data was not yet appreciated.

Mainframe makers worked hard to dispel the ghost of the tabulating machine. Their first success came on the fourth of November, 1952: election night. CBS had signed a contract with the Eckert-Mauchly Computer Corporation (EMCC) to have the first commercially available mainframe, dubbed UNIVAC, aid in predicting the winner as tallied votes from across the country came in (Draper, 1953). The analysts at CBS understood good predictions required understanding the historical voting trends of each state, but precisely how to model electoral America with UNIVAC was not. Simulation was a nascent field in the early 1950s, with parallel efforts in modelling neural networks remaining bound to paper and theory (Gilbert and Troitzsch, 2005a). Having called on statisticians from the University of Pennsylvania to help implement a suitable model for predicting voting trends, EMCC and CBS prepared UNIVAC for a demonstration on election night. It was somewhat successful. UNIVAC called the counting with a landslide victory for Dwight D. Eisenhower, an outcome too unexpected for CBS officials to accept. CBS took the mainframe off the air, citing mechanical problems. The win came to pass regardless, and President Eisenhower won the Republican party their greatest victory in twenty years (Ambrose, 2014). Mainframes could do more than tabulate. They could predict.

Simulating human behaviour took off. The Democratic party funded the Simulmatics Project, a research program co-hosted by the Massachusetts Institute of Technology and Yale University for developing more nuanced models of the electorate, pursuing the prize of automating opinion research via accurate models of electoral opinion formation (de Sola Pool and Abelson, 1961). The program approached the problem by way of counterfactual reasoning. Reducing the demographic distribution of America to a categorical vector of some four hundred and eighty independent variables, the researchers began testing the predictive power of their algorithms with conjectural scenarios constrained in their topic. If interested in assessing how voters would react to issues of religion, the researchers would simulate a campaign dominated by those issues and assess the differential between that campaign and the Eisenhower victories. Given the counterfactual nature of their simulations, the researchers could not measure the predictive performance of their models against real-world information. Simulating hypothetical campaigns provided a glimpse at a possible future given present assumptions—making mainframes a testing environment for theories of social science. Simulations galvanized social scientists.

Rapid compute scaling allowed researchers working in the latter half of the twentieth century to develop increasingly sophisticated social simulations (Gilbert, 1996). Simulations allowed social scientists develop formal theories via induction, whereby well-functioning counterfactual models can implicitly reveal causal events when a scenario is rigorously tested from multiple angles. Human societies are complex cognitive conglomerations, with emergent trends resulting from non-linear processes ill-suited for mathematical formalization with paper and pen (Parhami, 1995). Human societies are chaotic, meaning simulations are often the best tool at hand for conducting studies of their dynamics. One may think of the three-body problem as an analogue: it is trivial to simulate a scenario in which three gravitational bodies are orbiting each other, but it has been theoretically proven one can never write a mathematical description predicting their dance. Societies are similar. Rather than write and rewrite a differential equation encapsulating the situation, a project seeking to determine the spread of misinformation on Twitter might instead prepare a multi-agent environment in which some ten thousand simulated Twitter users interact. Simulations allow for research in otherwise intractably complex domains.

Simulations come with caveats. Samples drawn from a simulation are necessarily counterfactual, given stochastic processes undertaken between agents whose behaviour are by definition not real. "All models are wrong," wrote social scientist George E. P. Box, reminding us to be aware of the false nature of processes observed in a simulated environment (Box and Draper, 1919). The ontological nature of simulations means observations derived from models are always counterfactual; their origin not being of our own world. This carries at least two consequences for those conducting social simulations. First, a well-behaving model is not absolute evidence for a theory. All models are wrong because they only mimic reality—their simplified mechanics naturally preclude absolute homogeneity with real processes. A model perfectly correlating with reality will likely reveal weaknesses in time, especially given the stochastic processes invoked by models deploying Markovian inferences as is now the case with generative AI models (Delétang et al., 2022). Second, their illusory nature is precisely what allows scientists to estimate and play out conjectural future scenarios. Simulations are not mimicries of nature, but instead alternate realities in which scientists posit theories and witness their impact without delay. Simulations influence the degree to which a scientist believes particular theses; they cannot confirm nor reject.

Social scientists seek to align simulated samples with real-world measurements, but dissimilar results are not necessarily without value. A social scientist may prefer models constructed with naïve assumptions to diverge from historical scenarios. Such a departure would imply the potential existence of certain qualities necessary for the emergence of real-world phenomena. Counterfactual hypothesis testing depends on this approach. Let us again imagine a scenario in which a researcher is seeking to model the emergence of misinformation on Twitter. They may construct two multi-agent models, one involving organized entities supplying disinformation and one without, with the latter assuming the naïve approach. Depending on how the simulation unfolds, the latter then proving dissimilar to real-world Twitter metrics would reinforce the former is more representative of reality. Constructing multiple models in the hopes of verifying a hypothesis from multiple angles has promoted the development of efficient simulative techniques. The calculus-based system dynamic approaches of the 1960s gave way to the emergent systems of the late twentieth and early twenty-first centuries (Gilbert and Troitzsch, 2005a). This proliferation of novel simulative techniques has enabled unexpected discoveries in the social sciences.

Simulation has led to success in the social sciences. One may turn look to the creative manners in which cellular automata have led to advances in social understanding for an example. Cellular automaton are finite-state machines whose initial state is first prescribed and then allowed to run according to a set of simple transition rules; John H. Conway's famous Game of Life is one such automaton (Gardner, 1970). While simple in their design, cellular automaton develop complex enough emergent phenomena to allow for compelling simulations of social scenarios. Let us look at an example. A 1971 sociological study of neighbourhood-level ethnic segregation in the United States deployed cellular automaton to investigate how ethnic-majority neighbourhoods form (Schelling, 1971). The authors constructed a series of models, each supposing an increasing degree of prejudice in white households, whereby prejudice was the threshold at which white households would leave a neighbourhood containing such a percentage of non-white households; this percentage being the threshold. Advancing each simulation until equilibrium indicated ethnic segregation occurred when white households could only tolerate a neighbourhood containing 30% or fewer non-white households. The authors therefore considered multiple conjectural simulations, their collective behaviour allowing for an inductive understanding of white prejudice in the United States. We find further examples of simulative success in genetic algorithms.

Genetic algorithms are programs designed to realistically advance some population along a trajectory of realistic evolution. Each successive generation is subject to a competitive process by which those entities with the most competitive qualities prosper and breed (Lambora et al., 2019; Sampson, 1976). Social scientists have applied genetic algorithms to social scenarios, studying how human conglomerations develop and settle on particular strategies when pursuing a common goal. While not necessarily corresponding to reality in neither structure nor methodology, social situations advanced according to notions of natural competition indicate certain actions typically result in preferred outcomes. We find here again researchers deploying simulations as counterfactual apparatus not for describing processes within our own world but instead a fictive one whose composition reflects posited theories researchers hope align with reality. Computational means allow social scientists to make valid claims on complex human behaviour otherwise intractable for those seeking to conduct statistical analyses.

Simulations in the social sciences are computationally approachable because the predictive targets sought in their experimental frameworks are overwhelmingly numerical in nature. Given some set of independent variables, a social scientist might seek to predict the values of some set of dependent variables typically taking the form of normal, ordinal, or categorical values (Gilbert, 1996). Cellular automata, for example, consist of binary values. One defines their state in the form of a vector of two-dimensional tuples each containing either a one or a zero. Multi-agent models are much the same, whereby agents are generally represented as instances of some class defined in an object-oriented programming language (Dastani and Testerink, 2014). Their attributes are then again numerical, the rules for updating each agent with each successive turn determined by a set of computational rules. In either example we find entities are abstract constructions whose ontological status has no correspondence in reality. While these types of data make for computationally efficient programs, the resulting simulative frameworks developed thus far are hence not universally useful for all theoretical disciplines.

When viewing the human knowledge construction enterprise through the lens of the modern North American university, one notes the fields collected under the social sciences are not the only epistemological branches concerned with deciphering human behaviour: so too are those working in the classics, literature, history, linguistics, and communication studies as examples. Researchers in these fields are often concerned with the distillation and theorizing of complex social behaviours; their observations often resulting from a careful consideration of real-world documents and their interpretations per the different theoretical branches developed over centuries of philosophizing (Small, 2013). But whereas the observational goals of those in the quantitative social sciences constitute a set of independent and dependent variables, those in the humanities will find it difficult to sufficiently abstract over the documents of their interest. How would one convincingly transform James Joyce's *Ulysses* into a numerical value, Thucydides's memoir of the Peloponnesian War into a binary, or the works of Michelangelo into a series of ordinal values? Computational simulations require normalized measures for computation. What does one do when the subject of interest is itself an abstract cultural document whose contents is subject to subjective interpretation?

Quantitative researchers working in the humanities have faced the task of producing valid measures for documents whose interpretation is subject to intense flux. We turn to those working in computational literary studies as an example. Literary critics have long contested over interpretations of literary documents. Reception theorist Stanley Fish argues literary documents contain multiplicities, insofar as a reader determines their meaning in the posterior according to both personal predilection and the interpretive community they may belong to (Fish, 1995). Our reading of a text is contingent on the frameworks of meaning with which we approach the text, the consequence of which being the fundamental incompatibility between readings produced by one individual to the next. This subjective difficulty manifests itself in the measures posited by researchers working in computational literary studies. Take the definition of the scene as an example. A colloquial definition may entail a passage of text united in setting, time, and characters present. But if one were to develop a process for detecting such passages, one will quickly encounter difficulties. Despite fifteen years of researchers focusing on the problem, no researcher has yet posed a convincing solution whether it be for lack of good definitions or the statistical techniques chosen [Piper]. No quantified measures means no social simulations, and so the humanities has benefited little from computational simulations.

Some theorists in the digital humanities consider the humanities' lack of simulation concerning. Cultural analyst Lev Manovich argues empirical research is necessary if those in the humanities wish to develop rigorous theories of human culture (Manovich, 2016a). Each passing day sees people upload millions of cultural documents to the Internet, with new articles on Wikipedia rising daily by some 550 articles (Wikipedia, 2023). Researchers in the humanities are witnessing the development of human society in real time. Were one to extract the salient features of some subsection of this total corpus of cultural documents uploaded to the Internet on the daily and enact a computational simulation of the behavioural trends over time therefrom, researchers studying human culture would be in a position to benefit from simulations much in the manner those in the social sciences did. Manovich names this prospect a "science of culture," arguing researchers stand to feasibly contest long-held assumptions of those who have long studied cultural objects in qualitative terms (Manovich, 2016a). We again, however, face the same problem as those attempting to study novels in computational literary studies do: how does one produce a measure of a complex cultural artifact? Developments in natural language processing over the last five years present one possible solution.

The last five years have seen a neural revolution take natural language processing by storm. This revolution is in the form of the language model, a probabilistic model of natural language first developed by Claude Shannon in 1951 to capture the entropy of fixed-length letter sequences in written language. Early language models were not for generating language; their small scale restricting their prospects to speech recognition research for the next fifty years (Bellegarda, 2004). The twenty-first century invigorated their use, with researchers realizing graphic processing units could scale probabilistic models to scales sufficient for emergent phenomena to emerge (Krizhevsky et al., 2012). This renaissance culminated in 2017 with the release of the Transformer, a neural language model architecture developed by Google researchers to improve their online translation service (Vaswani et al., 2017). Scaling upwards from the Transformer upwards again, leading to the release of models demonstrating theories of mind (Bubeck et al., 2023). Those in natural language processing see large language models as new foundations for their study: where do we in the humanities stand?

The following manuscripts propose a future. A future where the study of human cultural processes, human language are aligned on a computational level: a future where some sociological experiments can be conducted with synthetic data yielded from probabilistic models invoking kernels billions of parameters large, where these experiments can propose new questions and new thoughts in domains otherwise heretofore intractable for computational study. The successes met by large language models in the last five years are unexpected, novel, and evocative. What will the digital humanities look like in five, ten, fifteen years? Each manuscript propose a different vision: where language models can give us glimpses of future legal processes, cultural trends, and media framing laying latent in the bowels of their parameterized visions of the world. A world where simulation is used in the humanities. For now, it is only a vision, a proposal. And that is where I hope future researchers can pick up the baton and press on forward.

Literature Review

Any project invoking new technologies will face difficulties in finding the relevant literature. This is an unfortunate consequence of studying large language models, for they have only been in vogue for at most a half decade. Nonetheless, my focus on using large language models to simulate human cultural behaviour via their learned biases is informed by a growing stand of research diverse in both their origins and their influence despite the recent nature of this area of interest. Examining each in turn will take us from the most important of papers concerning large language models through the philosophical musings of Manning and Bender before considering research efforts parallel to my own.

While the dawn of our current era in natural language processing can be traced to many beginnings, one may find a reasonable beginning in the release of GPT-2 by Radford et al. (2018a) in 2018. What differentiated their language model from those released previously was the quality of the dataset, the parameter count of the model proper, and the surprising emergent abilities emerging from an unsupervised language model. GPT-2 was the first large language model, "large" both in the sizes of the training set and the model, but so too in the uses. The authors found the model somewhat literate, matching supervised baselines in reading comprehension and qualitatively "performing the task" of other tasks like summarization. Here was a language model capable of more than just modelling text: here was a large language model.

The surprising advances found in GPT-2 were matched in the release of GPT-3 by Brown et al. (2020b), a model two orders of magnitude larger than its predecessor in both training set size and parameter count. They found scaling up GPT-2 in 'model, data, and training,' was all that was required to produce a model capable of co-writing (Jakesch et al., 2023), playing games (Tsai et al., 2023), and producing text nearly indistinguishable from human-derived samples (Dou et al., 2022). A human-aligned version of GPT-3 later become the base model for ChatGPT, reaching millions of consumers and defining the state-of-the-art in natural language interfaces. Recent research has begun to suggest large language models begin to develop theories of mind, an indication of their success in modelling not only language, but so too the world.

What the success of the GPT family might mean for present tracts of research in computational linguistics has been split. Large language models are capable of much, but in what regard and in what magnitude is disputed. Classical linguistics would suggest large language models should not be capable of what they empirically are were theories of meaning taught by linguists such as Noam Chomsky correct: for them, meaning is innately human (Katz, 1980). For a language model to develop coherent models of meaning without being embodied in a physical world, without the millions of years of evolution resulting in a just-so development of the human brain to understand certain concepts innately, is by definition nigh-impossible. But, as Manning (2022) posits, such might be possible, and indeed should be expected were to consider what those in computational linguistics have long understood to be meaning.

As Manning (2022) argues, the dominant approach for meaning representation in computational linguistics has not been prescriptive, where words are assigned meanings much in the manner of a dictionary, but rather contextually—wherein the meaning of a word is determined by the context within which it appears. Otherwise dubbed the distributional hypothesis, this definition of meaning laid the groundwork for computational approaches to capturing word-meaning in machine-readable forms at machine-suitable scales (Firth, 1957). While whether this approach is biologically accurate is hotly contested, Manning argues the success of GPT indicate large language models seem to not only use word contexts to learn meaning, but also the *world*. What differentiates a large language model and a language model? A language model is a model of language, while a large language model is a model of concept.

The validity of the distributional hypothesis has been contested by linguists and humanists alike. Perhaps the most widely-known rebuttal is Bender et al. (2021a), wherein it is argued text "generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's of mind," for language models are not embodied—they necessarily cannot be cognizant of any other, rendering null a basic fact of human communication: it is enacted between multiple participants. Meaning is in the eye of the beholder, a subjective nature kept in relative coherency by the formation of interpretive communities (Fish, 1995). The receiver of any message is complicit in the generation of meaning, and so it is difficult—if not impossible—to say with any certainty whether a language model *understands*. For those who believe they do not, the success of GPT is happenstance. With that said, empirical experiments conducted over the last five years indicate GPT is not necessarily a model of nonsense.

In the fourth quarter of 2022 a number of papers collectively posited the sociological value of generating written samples with GPT. The first of these is Park et al. (2022a), wherein the authors propose prototyping Internet forums with the help of "generated social behaviours," transcriptions sampled from GPT all describing the conversations and interactions between synthetic persons. Their implementation was found accurate; the authors note the model yielded samples heuristically identical to those pulled from real environments on the Internet. On one level, such behaviour should be expected. GPT models are trained on substantial portions of the Internet, a training set which will necessarily contain examples of human social interaction comparable to those the authors wished to yield from the model. GPT had examples. On another level, this behaviour is indicative of GPT understanding something of the world; it was showcasing a world model. Others noticed this trait.

Others in the social sciences have considered the use of GPT in their research. Contemporaneous with Park et al. (2022a) we find Argyle et al. (2023) with a proposal of a similar character. Seeking to understand whether language models learn the biases of specific demographics of people, the authors experimentally sample American-localized electoral data from large language models with the aim of comparing this synthetic against real-world samples. They find large language models can "produce outputs biased both toward and against specific groups and perspectives in ways that strongly correspond with human response patterns," suggesting large language models develop models of society—understandings of how biased human conglomerates react to specific inputs, such as during a federal election. Argyle et al. (2023) deem the degree of this understanding "algorithmic fidelity," suggesting this phenomenon will enable researchers in the social sciences to rapidly sample realistic survey data from a single language model. It is with this note that this present thesis undertakes a concurrent exploration: to investigate whether the humanities can find equal purchase in large language models.

However, the question remains: with the fundamental mechanics behind the success of models like GPT as of yet unexplored, will any research using these models be conducted on unstable ground? It is possible. But it is equally possible large language models are the result of how our own biochemical processes learn of the world, whether this be through Bayesian processes (Suchow et al., 2017) or vector semantics (Nishida et al., 2021). But while this question is settled by those relevant voices in the literature, this thesis investigates whether large language models can lead to valuable results in the humanities—a pursuit largely heretofore unexplored. And so I turn to Wittgenstein et al. (2009) to propose a (perhaps unsatisfying) stopgap to what is arguably *the* large language model problem: “a picture held us captive. And we couldn’t get outside it, for it lay in our language, and language seemed only to repeat it to us inexorably.” We communicate with language. Many of us think with language. If language models *seem* to make sense, then perhaps that is enough. For now.

Paper 1: The COVID That Wasn't

Large language models are exorbitantly expensive to train. Training one large language model can cost hundreds of thousands to millions of dollars (Bender et al., 2021b). Their cost comes from the compute required to train a large neural network on datasets containing upwards of trillions of words (Hoffmann et al., 2022). These costs are call for venture-backed capital, investments normally non-recurring. Large language models are not cheap to train, and so the hope is their out-of-distribution performance–their generalizations on the material given to them during training–are coherent enough to render moot the need to fine-tune the model on smaller and more domain-specific datasets to prepare them for certain tasks. GPT, or "generative pre-trained transformer" has been trained for you. It has been "pre-trained". Researchers at Stanford NLP have taken to calling pre-trained large language model foundation models, drawing attention both to their performance, and to their cost (Bommasani et al., 2022).

Language models also develop world models during training. Properly learning how words probabilistically follow one another leads the model to develop theories of meaning, logic, social rules, and so on (Liu et al., 2021b). Moreover, the world the model learns is the one described in the training data—data that itself can be dated to a particular date. The world the language model learns is necessarily an outdated one, and this dated nature only increases as time passes. Consider a model trained on data largely compiled in 2017 and released the following year: at the time of release and forever more the model will emit samples statistically most similar to articles of text drawn from 2017 and before. To train a language model is to create a time machine.

GPT-2 was trained on data in 2017 (Radford et al., 2018a). Even now it will make reference to former American President Donald Trump and former U.K. Prime Minister Theresa May; the United States withdrawing from the Paris Agreement; and Russia being banned from the 2018 Olympics in South Korea. Perhaps the most significant difference is that the world it models is pre-COVID. 2017 was six years ago. How has the world changed? Can we use GPT-2 as a baseline, and how might we go about that? The following paper considers how GPT-2 can used to benchmark the cultural change wrought by simulating COVID with a world model set in 2017.

The COVID That Wasn't: Counterfactual Journalism Using GPT

Sil Hamilton
McGill University
sil.hamilton@mcgill.ca

Andrew Piper
McGill University
andrew.piper@mcgill.ca

Abstract

In this paper, we explore the use of large language models to assess human interpretations of real world events. To do so, we use a language model trained prior to 2020 to artificially generate news articles concerning COVID-19 given the headlines of actual articles written during the pandemic. We then compare stylistic qualities of our artificially generated corpus with a news corpus, in this case 5,082 articles produced by CBC News between January 23 and May 5, 2020. We find our artificially generated articles exhibits a considerably more negative attitude towards COVID and a significantly lower reliance on geopolitical framing. Our methods and results hold importance for researchers seeking to simulate large scale cultural processes via recent breakthroughs in text generation.

1 Introduction

The rush to cover new COVID-19 developments as the virus spread across the world over the first half of 2020 induced a variety of editorial mandates from public broadcasters. Chief among these was a desire to mitigate shock from a public unaccustomed to large-scale public health emergencies of the calibre COVID-19 presented. This desire translated into systematic underreporting together with a reluctance to portray COVID-19 as the danger it was (Quandt et al., 2021; Boberg et al., 2020). Broadcasters in the United States (Zhao et al., 2020), the United Kingdom (Garland and Lilleker, 2021), and Italy (Solomon et al., 2021) all exhibited this phenomenon.

Although many studies have verified the above effects, few if any studies to date have considered *alternative* approaches the media could have taken in their portrayal of COVID-19. Evaluating these alternatives is critical given the close relationship between media framing, public opinion, and government policy (Ogbodo et al., 2020; Lopes et al., 2020).

In this paper we present a novel method of simulating media coverage of real world events using Large Language Models (LLMs) as a means of interpreting news industry biases. LLMs have been used in a variety of settings to generate text for real-world applications (Meng et al., 2022; Drori et al., 2021). To our knowledge, they have not yet been used as a tool for critically understanding the interpretation of events through media coverage or other forms of cultural framing.

To do so, we use Generative Pre-trained Transformer 2 (GPT-2), which was trained on text produced prior to the onset of COVID-19, to explore how the Canadian Broadcasting Corporation

(CBC) covered COVID and how else they might have reported on these breaking events. By generating thousands of simulated articles, we show how such "counterfactual journalism" can be used as a tool for evaluating real-world texts.

2 Background

The COVID-19 pandemic has given researchers a variety of opportunities to study human behavior in response to a major public health crisis. One core dimension of this experience is reflected in the changing role that the media has played in communicating information to the public in a quickly changing health environment (Van Aelst, 2021; Lilleker et al., 2021). Times of crises enshrine the media as a valuable mediator between the public and government.

This changing role registers itself in the editorial policies at news corporations across the world. Research published over the past two years has confirmed that public news broadcasters in Australia, Sweden, and the United Kingdom all significantly altered their editorial style in response to both societal and governmental pressures (Holland and Lewis, 2021; Shehata et al., 2021; Birks, 2021) during COVID-19.

What this research has so far lacked is the ability to infer what *could* have been communicated, i.e. what losses were entailed in these editorial shifts. While a great deal of recent work has studied the biases intrinsic to large language models[1], no work to date has used LLMs to study the biases of human generated text. Reporting on real-world events inevitably requires complex choices of selection and evaluation, i.e. which events and which actors to focus on along with modes of valuation surrounding those choices. Simulating textual production given similar prompts such as headlines can provide a means of better understanding the editorial choices made by news agencies.

In using a language model as a simulative mechanism, we draw on a long research tradition of using simulation to understand real-world processes. Simulation has proven a boon for those working in the sciences, including climate science and physics (Winsberg, 2010), and for those working in the social sciences, where agent-based social modelling has led to advances in understanding complex social phenomena (Squazzoni et al., 2014). We seek to bring these techniques to the study

[1] See Garrido-Muñoz et al. (2021) for a recent survey of works investigating latent biases present within large language models.

of cultural behavior, where simulation has historically seen less of an uptake (Manovich, 2016).

3 Method

Our project consists of the following principal steps:

1. Create a news corpus drawn from our target time-frame (15 January to 5 May 2020) whose content is COVID-19 related.

2. Fine-tune a language model whose generative output is statistically similar to a random sampling of our news source published *before* our target time-frame, i.e. prior to COVID.

3. Using this model, generate full-length text articles using various prompts, including headlines and associated metadata.

4. Compare generated text articles with the original news corpus across key stylistic metrics.[2]

3.1 Corpus

We first obtain a comprehensive collection of CBC News' online articles concerning COVID-19 published between January and May 2020 from Kaggle (Han, 2021). Our corpus contains 5,114 articles all in the form of a headline, subheadline, byline, date published, URL, and article text. Deduplicating and cleaning the corpus with a series of `regex` filters leaves 5,082 articles spread across the first four months of COVID-19.

3.2 Language Model

We use a Transformer-based large language model (LLM) as our CBC simulacrum. We formalize our model as follows: we define an article as a chain of k tokens. Let $X(d, \theta)$ be a probability distribution representing the pulling of a token out from the language model, where d is the article metadata and θ are the prior weights. The probability of drawing k tokens is then

$$\Pr(\bar{x}_k) = \prod_{i=1}^{k} \Pr(X(d, \theta) = x^i | \bar{x}_i) \tag{1}$$

[2] We make our code available here.

where $\bar{x}^i$ is the ith element of the vector $\bar{x}$, and $\bar{x}_i$ is the vector consisting of the first i elements of the vector $\bar{x}$.

Selecting a pretrained language model suitable for use as a base with which to further train with specific writing samples is a non-trivial task given the plurality of large language models released in the past four years (HuggingFace, 2022). We surveyed models for candidates possessing the following qualities:

- the model must be neither egregious nor lacking in parameter count;
- domain-relevant samples must have been present in the pretraining corpus;
- and most importantly, the model must not be aware of COVID.

Keeping with the above requirements, we select the medium-sized Generative Pre-trained Transformer-2 (GPT-2) as distributed by OpenAI as our candidate model. We found the medium-sized GPT-2 model desirable because it is light enough to be fine-tuned with a single consumer-level GPU; CBC News was the 21st most frequent data source OpenAI used in producing its training set (Clark, 2022) and the model was trained in 2018, two years before the beginning of COVID-19.

3.2.1 Fine-tuning

Provided with sufficient context in the prompt, a freshly obtained GPT-2 model produces qualitatively convincing news article text. It will, however, periodically confuse itself with exactly which publication it is imitating, e.g. it can switch from sounding like CNN to CBC to BBC in a single text. For the purposes of comparison with a single news source, it is thus necessary to fine-tune the model with example texts encapsulating the desired editorial and writing style.

Fine-tuning is a two-step process. We first gather a sequence of texts best representing our target writing mode before fine-tuning a stock GPT-2 model with the training dataset.

Training Dataset We use a web scraper to extract a random selection of news articles published between 2007 and 2020 from CBC News' website. We configure our scraper to pull the same metadata as our COVID-19 dataset: headline, subheadline, date, URL, and article text. We again deduplicate to reduce the possibility of overfitting our model. With this method we collect 1,368 articles with an average length of 660 words per article.

We next construct a dictionary structure to formalize both our generation targets and to provide GPT-2 a consistent interface with which to aid it in logically linking together pieces of metadata. Previous research has indicated fine-tuning LLMs with structured data aids the model in both understanding and reacting to meaningful keywords (Ueda et al., 2021). We therefore structure our fine-tuning data in a dictionary. We provide a template of our structure below.

```
{
 'title': 'Lorem ipsum...',
 'description': 'Lorem ipsum...',
 'text': 'Lorem ipsum...'
}
```

We produce one dictionary per article in our training set. We convert each dictionary to a string before appending it to a final dataset text file with which we train GPT-2.

Training With our training dataset in hand, we proceed to configure our training environment. We use an Adam optimizer with a learning rate of $2e^{-4}$ and run the process (Kingma and Ba, 2014). Training the model for 20,000 steps over six hours results in a final model achieving an average training loss of 0.10.

3.2.2 Model Hyperparameters

In addition to fine-tuning our model, we experiment with different hyperparameters and prompt strategies. Numerous prior studies have described the effects hyperparameter tuning has on the token generation process (van Stegeren and Myśliwiec, 2021; Xu et al., 2022). For our purposes, we use three prompting strategies when generating our synthetic news articles along with one further parameter (*temperature*):

Standard Context Only title and description metadata are used as context d for the model.

Static Context In addition to the standard context, we supply the model with an additional `framework` key containing a brief description of the COVID-19 pandemic found on the website of the Centre for Disease Control (CDC) in May 2020. All generation iterations use the same description.

Rolling Context We again supply the model with an additional `framework` key, but keep the description of COVID-19 contemporaneous with the date of the real article in question. We again use the CDC as a source but instead use the Internet Archive's Way Back Machine API to scrape dated descriptions.[3]

Temperature We manipulate the temperature hyperparameter during generation with half-percentage steps shifting the temperature between $0.1 \ldots 1$. The temperature value is a divisor applied on the `softmax` operation on the returned probability distribution, the affect of which effectively controls the overall likelihood of the most probable words. A high temperature results in a more dynamic and random word choice, while a lower temperature encourages those words which are most likely according to the model's priors.

Models Manipulating the above hyperparemeters gives us the following model framework:

- Model 1: headline-only, temperature between 0.1 and 1
- Model 2: static context, temperature between 0.1 and 1
- Model 3: rolling context, temperature between 0.1 and 1

We find that manipulating the `softmax` temperature hyperparameter has no measurable effect on our measures described below. We thus proceed using only three primary models for article generation using a temperature of 0.50, which we refer to in the remainder of the paper as Models 1, 2, and 3. When we rely on a single model to exhibit results Model 3 will be the model we choose to illustrate.

[3] https://archive.org/help/wayback_api.php

3.3 News Article Generation

Having now obtained both our models and our real-world corpus, we proceed with the text-generation step by prompting our model with metadata taken from CBC's COVID-19 articles. The generation process takes the following form:

1. For each article dictionary, extract the `title` and `description` keys.
2. For each pair of keys, create a new dictionary and add an empty `text` key.
3. Convert all new dictionaries to strings and tokenize using GPT-2's Byte Pair Encoder (BPE).
4. Have GPT-2 predict the contents of the `text` key using the title and description as context, generating 750 tokens in the process.
5. Collect generated tokens and insert into the `text` key.

Doing so generates 5,082 counterfactual news articles that temporally correspond to our CBC News corpus for each of our three primary models. Each article pair in each model thus shares metadata but differs in content, with one being original and the other generated.

We provide here examples drawn from our simulated-actual article pairs to illustrate the performance of our models. Further examples may be found in Appendix A.

Headline 1: "China confirms human-to-human transmission of new coronavirus."

CBC: "Human-to-human transmission has been confirmed in an outbreak of a new coronavirus, the head of a Chinese government expert team said Monday, as the total number of cases more than tripled and the virus spread to other cities in China."

GPT: "An outbreak of a new coronavirus has been confirmed in southern China's Hebei Province, the lead author of a scientific paper said Thursday. The total number of cases more than triples the number of cases in the area, which corresponds to the Beijing and Shanghai hot spots..."

Headline 2: "Quebec travel agencies feel the heat as local travellers cancel flights to China."

CBC: "[Name omitted for privacy] has dreamed of heading to Thailand with her partner for years. But with a five-month-old baby in tow and 14 cases of the coronavirus reported in the area so far..."

GPT: "Quebec travel agencies are feeling the heat as local travellers cancel flights to China. China remains the most dangerous place on Earth for travellers..."

3.4 Measures

To assess the stylistic differences between our simulated and real-world corpora we use the following measures:

Measure 1: Sentiment We measure the sentiment of each article with the open-source Python library *VADER* (Hutto and Gilbert, 2014). Prior studies have validated the use of VADER on journalistic texts, finding the model to be superior to various alternatives in detecting sentiment (Castellanos et al., 2021). We additionally validate a small sample of measured sentences to ensure the accuracy of the tool.

We measure sentiment by first splitting a given article into s sentences, obtain the compound polarity score $(-1 \ldots 1)$ for each s, then average all s into a final score for the article. The resulting real number represents the overall sentiment of the article.

We apply a number of heuristics to ensure the sentiment score accurately reflects the reality of COVID-19. Words that would previously represent a positive sentiment (such as a "'positive' test") become negative in actuality during a pandemic. It is the same for certain negative terms like "testing 'negative.'" Our heuristics appropriately shift such terms as they appear, allowing for a more accurate measurement. As we later show, our heuristics demonstrate a strong correlation between our simulated and actual corpora.

Measure 2: Named Entity Recognition We detect and track named entities in each article with the use of the Python library *spaCy* and their `en_core_web_sm` model (Montani et al., 2022) given prior studies found the model is effective in recognizing named entities (Schmitt et al., 2019). We specifically tally all entities tagged as being a person, geopolitical entity, or organization on an

Sentence	Sentiment
“Thousands of cyclists pedalled along empty Toronto highways today, enjoying the good weather and raising money for charity.”	0.8074
“’They’re good at running them and we have to create the right environment for them,’ she said.”	0.6124
“She said it’s not good enough to say there’s a strategy — that the province needs a strategy in action.”	-0.3412
“Transportation Minister Clare Trevena said the incident is ’obviously’ worrisome.”	-0.4019

Table 1: A subset of sentences and their VADER sentiment score from the control dataset.

intra-article basis.

Measure 3: Focus We take the ratio of total unique named entities e over article length l and call it *focus*, a novel measure for how focused a given article x is around a given set of entities:

$$focus(x) = \frac{e}{l} \tag{2}$$

We see focus as a measure of concentration around prominent agents in the news.

Measure 4: Key Words We conduct a key word test on each respective corpus to identify repeatedly used terms bearing sentimental weight as per VADER. Following best practices, we only rank significant words in reference to each other rather than assigning significance to any term in isolation. We use the two formulae presented below for determining key words in a given corpus, as provided by Rayson (2012).

We first calculate the averaged frequency E_i for each word in our corpus with

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{3}$$

where N is the total word count and O is the frequency for the word.

Having now obtained a list of frequencies, we proceed with modifying our frequencies with a

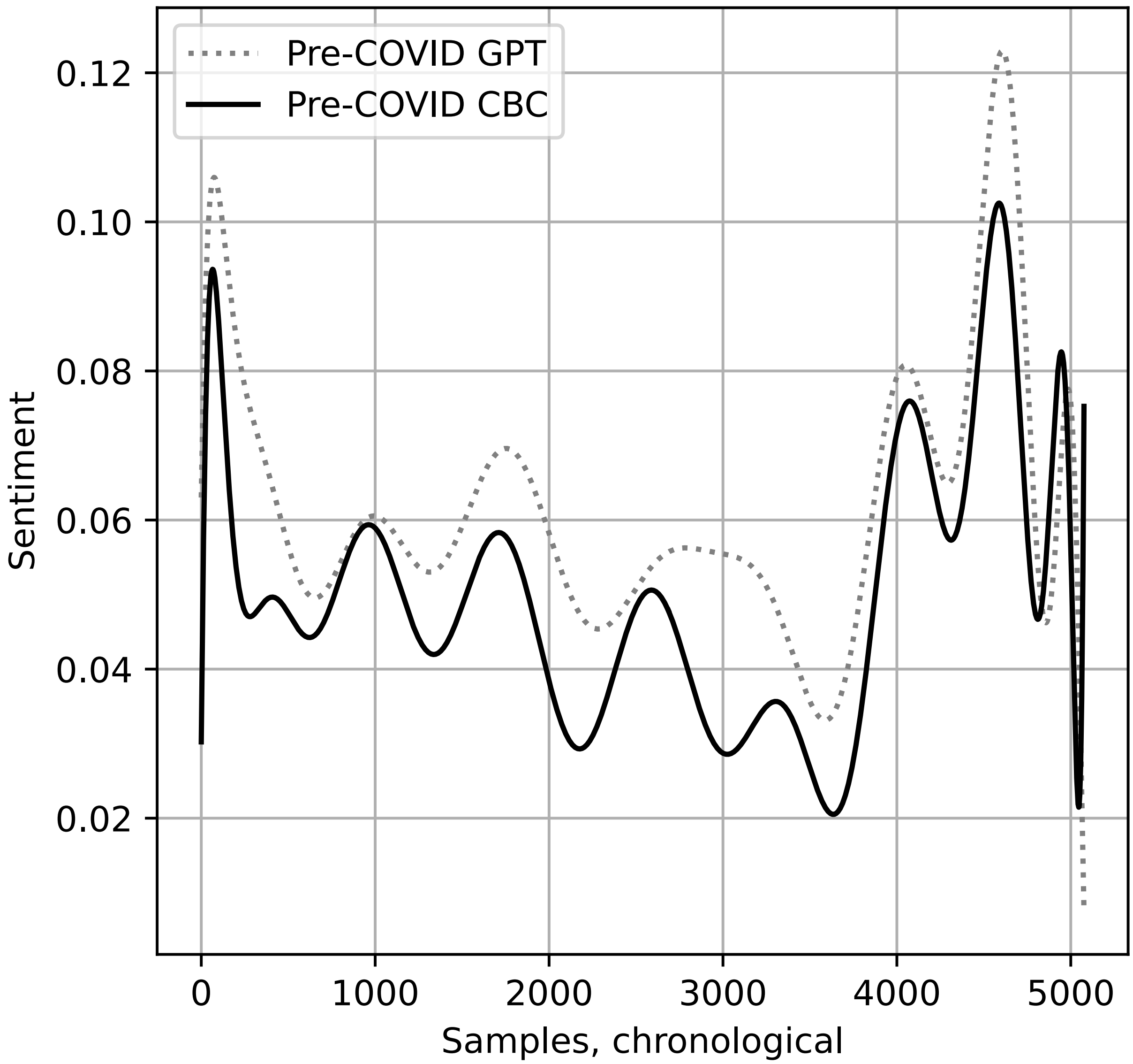

Figure 1: Correlation of sentiment in pre-COVID CBC and GPT articles over a ten year period.

log-likelihood (LL) test:

$$LL = 2 \sum_i O_i \ln\left(\frac{O_i}{E_i}\right) \tag{4}$$

We then rank our key words according to their respective LL values before comparing our two respective key word lists.

4 Results

4.1 Fine-Tuning Validation

Validating artificial text generation is a challenging task as there is no "right" answer when it comes to creating artificially generated text. Our primary goal in this case is to disambiguate whether our results are an effect of GPT-2 behavior (i.e. a result of model bias) or an effect of our fine-tuning and prompt engineering (i.e. a result of COVID-specific information). To do so, we first create a control dataset consisting of 5,077 randomly sampled CBC articles published prior to COVID between 14 January 2010 and 31 December 2019 ("pre-COVID CBC"). We then generate artificial articles with a standard context and a temperature of 0.5. As shown below, our pre-COVID model produces articles whose distributions are highly statistically similar to the pre-COVID CBC data across our three primary measures, suggesting that any deviation from these levels of correlation in subsequent models is an effect of the COVID fine-tuning and not a default behavior of the model.

Sentiment We find fine-tuning GPT-2 with pre-COVID CBC data produces a model whose textual output is sentimentally similar to pre-COVID CBC articles as may be observed in Figure 1. The sentiment distributions measured in our generated and real-world control datasets share an overlap of 97.7% (Cohen's $d \approx 0.06$) and a moderately positive correlation coefficient of $r \approx 0.57$. We furthermore note GPT-2 is typically more positive in tone than CBC when examining the two distributions as a whole. A selection of validated sentences together with their respective sentiment values are presented in Table 1.

Focus When measuring focus values for the control dataset, we again note a large overlap between GPT-2 and CBC at 93.6% ($d \approx 0.16$) together with a weakly positive correlation or $r \approx 0.17$. These values suggest GPT-2 has learned focus trends latent in the pre-COVID CBC training set.

Key words Our final validation metric is a key words test using the process described in section 3.4. The mean log-likelihood of key words deployed by GPT-2 is 9.61 (95th percentile ≈ 42), indicating such terms are only marginally more likely to be used by GPT than by pre-COVID CBC.

4.2 Measure 1: Sentiment

We begin by noting that CBC News' treatment of COVID-19 during our period of inquiry develops in two stages (Figure 2): articles prior to early March register overall as negative in their sentiment valence (stage 1), while articles written after the first two months become increasingly positive (stage 2). Note that in the pre-COVID data sentiment values were uniformly positive for both CBC and GPT.

When comparing our simulated texts to CBC, we find that our simulated corpora all demonstrate similar trends over time, but with significantly lower levels of positivity than the actual corpus, which is a direct reversal of the pre-COVID baseline. The headline-only model (Model 1) exhibits the highest level of correlation with the CBC corpus at $r \approx 0.28$, while the rolling context model (Model 3) exhibits the starkest overall difference in terms of generating more negative sentiment with an overall effect size of $d \approx -0.28$ (more than double what we see for Model 1 at $d \approx -0.12$). In general, we note that Model 1 adheres most strongly to CBC practices, while adding the CDC context, whether rolling or static, tends to make the models diverge more strongly from CBC practices.

4.3 Measure 2: Named Entity Recognition

Tracking entities classified as persons, geopolitical entities (e.g., countries), and organizations (e.g., the World Health Organization), we find a similar two-stage process as we did when observing sentiment. As is observable in Figure 3, we see a significant decline of geopolitical entities after March in the CBC corpus replaced by a slight increase in notable persons. While the rolling context model exhibits decent correlation with the CBC corpus at $r \approx 0.27$, we find a very strong discrepancy in the relative reliance on geopolitical entities in the GPT corpus compared to CBC

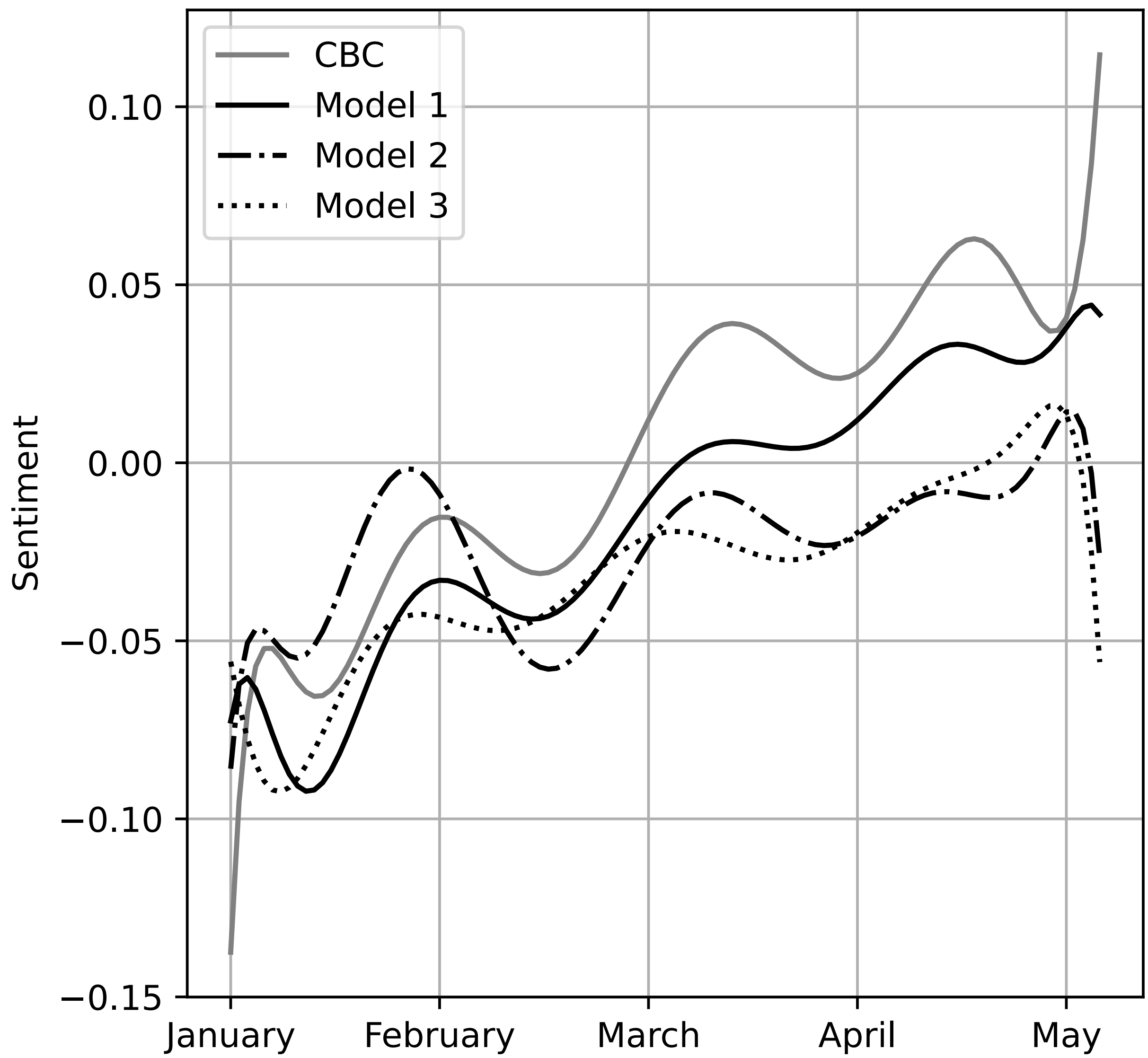

Figure 2: Averaged weekly article sentiment over the first four months of the pandemic.

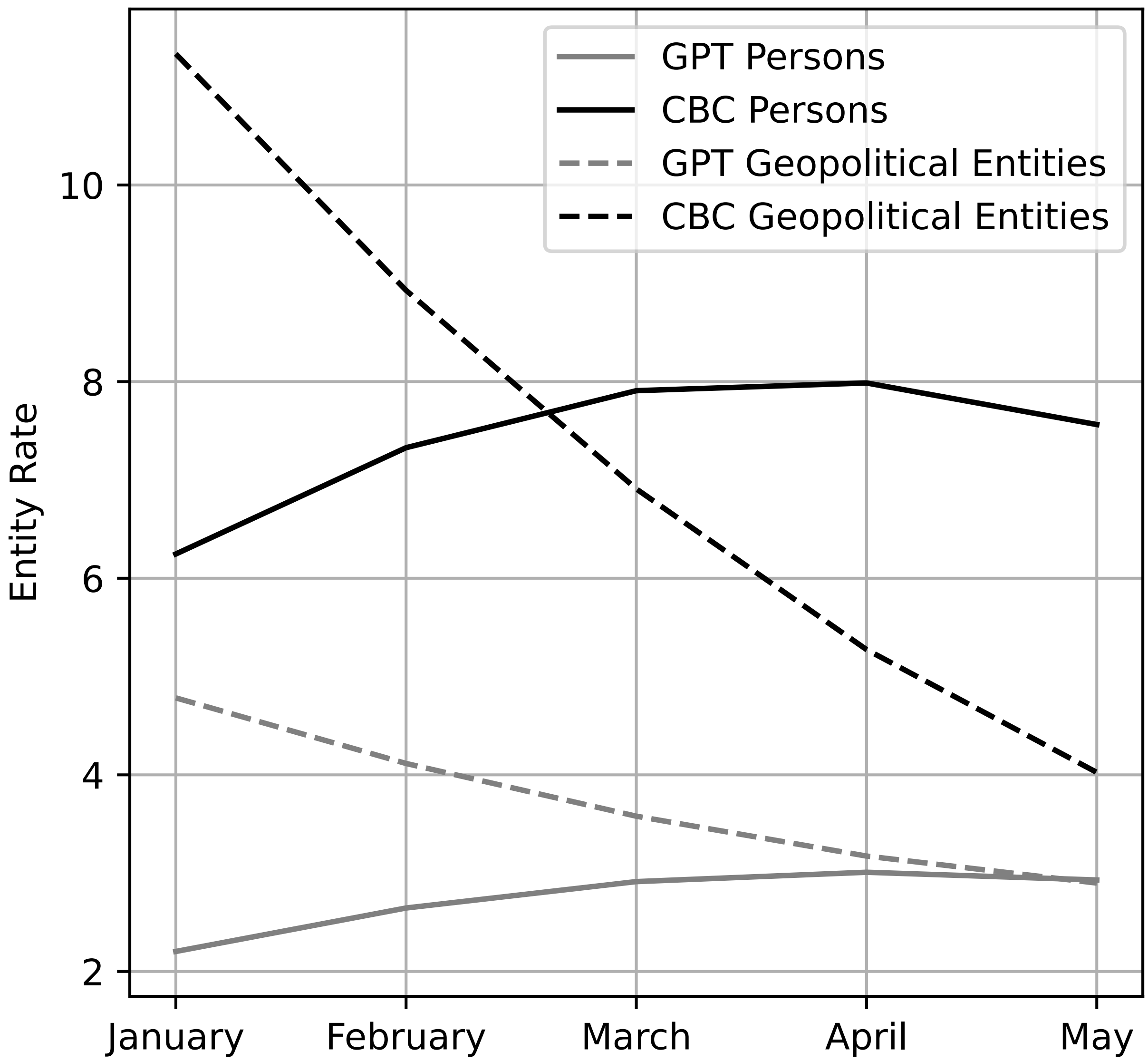

Figure 3: Average values of given entity types in CBC & GPT articles over the first four months of the pandemic for Model 3.

with an overall effect size of $d \approx -0.63$.

4.4 Measure 3: Focus

Measuring the personal focus of articles reveals a number of trends. Predominant among these is a clear upwards trend over time in the CBC articles. Continuing the split stage analysis of the past measurements, we find focus increases linearly as the months of the pandemic pass. Higher focus values indicate that fewer entities are being discussed at greater length (i.e. are centralized more strongly). We also note a low correlation between article sentiment and article focus ($r \approx 0.18$), suggesting that focalization around fewer persons is associated with more positive messaging. We explore this effect further in subsection 5.3.

In terms of our simulated corpus, we see that GPT-2 remains relatively consistent in both focus and sentiment over the course of our time window. While there remains an extremely weak positive correlation between sentiment and focus in the Model 3 corpora ($r \approx 0.02$), this is likely an artifact carrying over from the headlines themselves becoming more positive over time.

4.5 Measure 4: Key Words

When we observe the likelihood of a given word's appearance in one corpus or the other, here too we observe some notable trends.[4] We present a subsection of our results in Table 2 using Model 3.

Conducting a qualitative analysis on the top key words underscores two points. First, we find Model 3 (and other models) routinely interpret COVID-19 as a flu, reflected in the model deploying terms like "flu" and "strain" more regularly than CBC News. This interpretation likely accounts for a majority of the discrepancies between the two corpora. Second, we find CBC is more likely to describe societal responses to COVID-19 ("emergency," "crisis"), whereas GPT-2 draws on imagery to convey the medical threat of the disease ("sickened," "infected").

5 Discussion

In this section we identify three noteworthy discrepancies between the behavior of our models and the real-world CBC corpus and discuss their potential implications.

[4] We condition only on VADER vocabulary and not the full set of words.

CBC News	LL	**GPT-2**	LL
"crisis"	475	"flu"	2465
"care"	431	"strain"	871
"cancelled"	371	"infected"	855
"isolation"	363	"great"	558
"emergency"	264	"sickened"	400
"anxiety"	170	"threat"	317
"support"	158	"cancer"	302
"sick"	149	"natural"	249
"critical"	138	"killed"	189
"vulnerable"	134	"dangerous"	167

Table 2: A selection of the ten most prevalent sentimentally-charged terms in either corpus.

5.1 Effect 1: Positivity Bias ("Rally-Around-The-Flag")

We note that all of our simulated models trained on the COVID data generated news that was far more negative than actual coverage, which grew increasingly positive over time. This result is especially notable given that pre-COVID models were uniformly more positive than actual CBC articles.

A relevant theory that can help make sense of this is the "rally-around-the-flag" effect, which posits that national discourse trends in favour of reigning governments during times of crisis (Van Aelst, 2021). Theorists in communication studies note news media do not remain neutral during crises, but instead work to assuage public fears by promoting trust in local leaders (Quandt et al., 2021).

The "rally-around-the-flag" effect could help explain why CBC News articles became more positive as lockdowns began and why our language models, which were not subject to such pressures, nevertheless remained more negative. Regardless of the cause of this discrepancy between our models and CBC, it is worth noting our language models consistently interpreted COVID in more negative terms than this particular public broadcaster. An important aspect to underscore is that we do not see the same effect when we run the same process on a random assortment of pre-COVID articles, meaning our GPT models are not intrinsically more negative but rather interpret these particular events more negatively than CBC.

5.2 Effect 2: Early Geopolitical Bias

As we saw in Figure 3, CBC News relied on considerably more geopolitical entities in the early weeks of the pandemic than in the latter weeks, an effect our models only mildly reproduced. The strong decline of geopolitical entities in the CBC data past February suggests an editorial re-orientation away from understanding the pandemic in global geopolitical terms and towards a local health emergency that is more in line with what our models were producing from the beginning.

5.3 Effect 3: Person versus Disease Centredness

While the rate of geopolitical entities in the CBC data eventually converges with our GPT models, we see that the reliance on individual persons is consistently stronger in the CBC , something we did not see in the pre-COVID models. In conjunction with our key-word findings, this suggests that GPT's interpretation of the pandemic is far more medical and health-oriented ("infected," "sickened") than CBC, whose treatment remained more focused on people. As we show in Figure 4, this person-centredness is also associated with higher levels of positivity. Future work will want to explore whether this personal focalization was unique to the CBC, COVID, or the experience of social upheaval more generally.

6 Conclusion

The aim of our paper has been to develop a framework for using the text-generation affordances of large language models to better understand the interpretive perspectives of the news media when covering major social events. We rely on a simulative process whereby the generation of thousands of alternative views of a real world event can provide a framework for understanding the interpretive perspectives employed by news organizations.

Given that language models can approximate human discourse (Radford et al., 2018), they can be used to generate a distribution of possible responses to an event to better understand the actual selection mechanisms used by real-world actors. Key to this process is validating the extent to which the qualities of artificially generated text are a function of model parameters or the process of fine-tuning, i.e. an effect of the real-world event we aim to simulate. Our aim in doing so is to

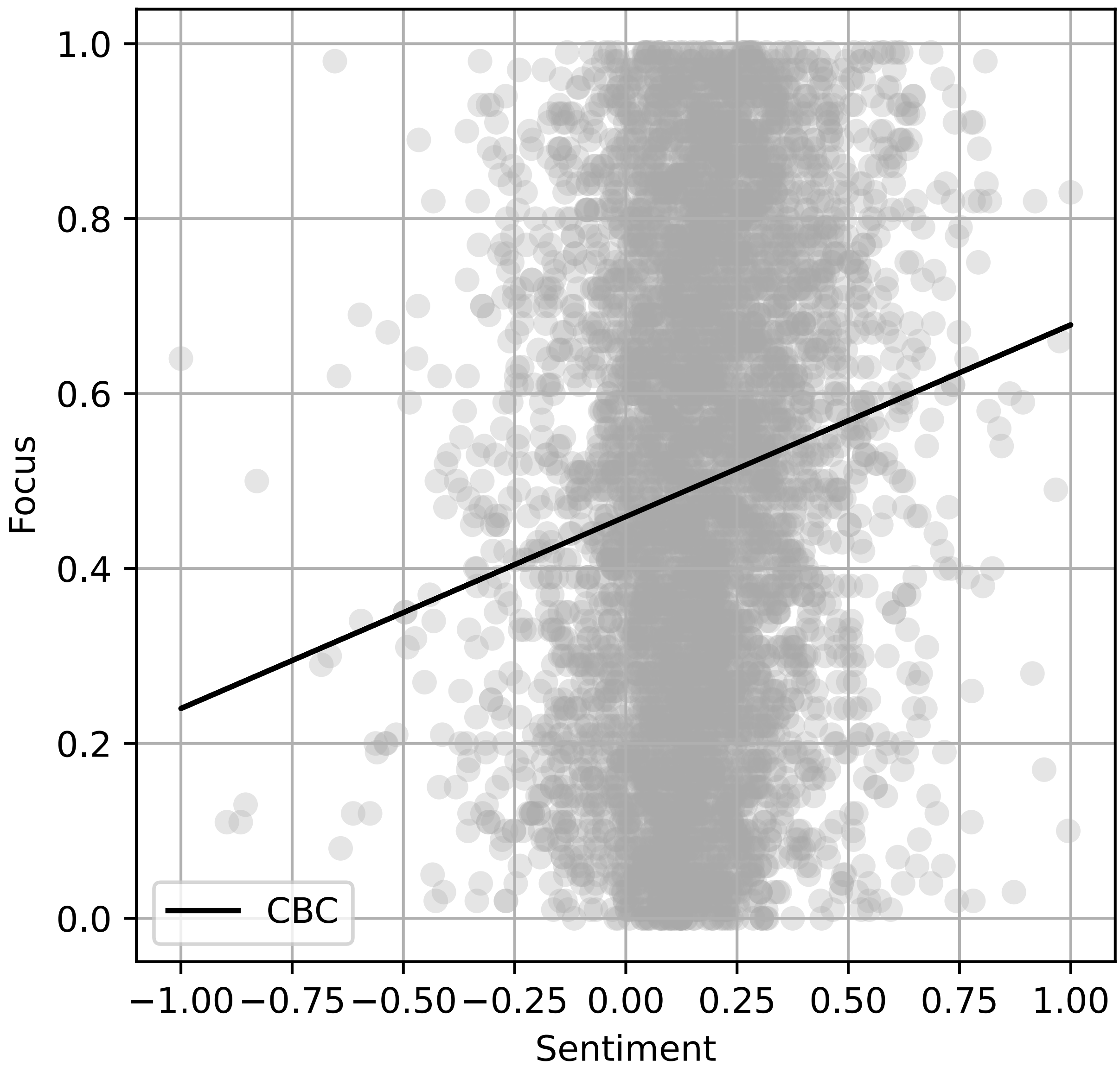

Figure 4: Relationship between focus and sentiment in CBC articles. Focus values are normalized.

illustrate how language models can be used as diagnostic tools for human behavior. Given no prior knowledge of a major event, what would a language model say? And what might this tell us about our own human reactions?

Based on the results we have obtained here, we see the following possible avenues for further research using LLMs for textual simulation:

Further Domain Exploration. What other scenarios might LLMs be analytically useful for? In this paper, we have explored LLMs as a tool to assess media coverage, but future work will want to observe how they behave in other domains. News is a particularly well-structured form of textual communication and thus we expect LLMs to perform more adequately in this domain given prior research (Ueda et al., 2021). We await future work exploring other textual domains.

Modeling Audience Expectations. We have used GPT as a tool to assess the interpretive frameworks of the news media, specifically the CBC. However, we might also consider the ways in which LLMs can provide us with population-level expectations about an event. For example, the strong reliance on the "flu" in our models could be seen as a faithful mirror of how laypeople generally have thought about COVID (in distinction from public health experts). While one might argue that this is "erroneous" from a public health perspective, such semantic frameworks may be useful resources in fashioning public communication during times of crisis or upheaval. LLMs may be able to help us better understand what biases audiences are bringing to novel events thus helping experts craft more appropriate messaging that aligns with audience expectations.

Predicting Future Outcomes. While we have used GPT as a tool to assess past behavior, future work could explore the predictive power of LLMs, while exercising a great deal of caution when it comes to their application. For example: Can LLMs identify future valuable research questions? Financial or economic events given changing real-world information? Or potential political crises given the communicative behavior of principal actors (e.g. politicians)? An equally potent line of research will want to explore the dangers of such approaches as in past experiences of predictive policing. New technologies always bring an admixture of analytical affordance and risk that needs to be better understand with respect to LLMs. More experimentation with respect to the efficacy of textual simulation is definitely warranted.

References

Jen Birks. 2021. Just following the science: Fact-checking journalism and the government's lockdown argumentation. In *Power, Media and the Covid-19 Pandemic*, pages 139–158. Routledge.

Svenja Boberg, Thorsten Quandt, Tim Schatto-Eckrodt, and Lena Frischlich. 2020. Pandemic populism: Facebook pages of alternative news media and the Corona crisis – a computational content analysis. Technical Report arXiv:2004.02566, arXiv. ArXiv:2004.02566 [cs] type: article.

Eric Castellanos, Hang Xie, and Paul Brenner. 2021. Global news sentiment analysis. In *Proceedings of the 2019 International Conference of The Computational Social Science Society of the Americas*, pages 121–139, Cham. Springer International Publishing.

Jack Clark. 2022. GPT-2 domains. Original-date: 2019-02-11T04:21:59Z.

Iddo Drori, Sunny Tran, Roman Wang, Newman Cheng, Kevin Liu, Leonard Tang, Elizabeth Ke, Nikhil Singh, Taylor L. Patti, Jayson Lynch, Avi Shporer, Nakul Verma, Eugene Wu, and Gilbert Strang. 2021. A neural network solves and generates mathematics problems by program synthesis: Calculus, differential equations, linear algebra, and more. *CoRR*.

Ruth Garland and Darren Lilleker. 2021. From consensus to dissensus: The UK's management of a pandemic in a divided nation. In *Political communication in the time of coronavirus*, pages 17–32. Routledge.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences*, 11(7):3184.

Ryan Han. 2021. COVID-19 News articles open research dataset.

Kate Holland and Monique Lewis. 2021. Mapping national news reports on COVID-19 in Australia: Topics, sources, and imagined audiences. In Monique Lewis, Eliza Govender, and Kate Holland, editors, *Communicating COVID-19*, pages 59–81. Springer International Publishing, Cham.

HuggingFace. 2022. Models - Hugging Face.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Darren Lilleker, Ioana A Coman, Miloš Gregor, and Edoardo Novelli. 2021. Political communication and COVID-19: Governance and rhetoric in global comparative perspective. In *Political Communication and COVID-19*, pages 333–350. Routledge.

Bárbara Lopes, Catherine Bortolon, and Rusi Jaspal. 2020. Paranoia, hallucinations and compulsive buying during the early phase of the COVID-19 outbreak in the United Kingdom: A preliminary experimental study. *Psychiatry Research*, 293:113455.

Lev Manovich. 2016. The science of culture? social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *CoRR*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Daniël De Kok, Lj Miranda, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Edward, Wannaphong Phatthiyaphaibun, Richard Hudson, Yohei Tamura, Sam Bozek, Murat, Ryn Daniels, Peter Baumgartner, Mark Amery, and Björn Böing. 2022. explosion/spaCy: New span ruler component, JSON (de)serialization of doc, span analyzer and more.

Jude Nwakpoke Ogbodo, Emmanuel Chike Onwe, Joseph Chukwu, Chinedu Jude Nwasum, Ekwutosi Sanita Nwakpu, Simon Ugochukwu Nwankwo, Samuel Nwamini, Stephen Elem, and

Nelson Iroabuchi Ogbaeja. 2020. Communicating health crisis: a content analysis of global media framing of COVID-19. *Health Promotion Perspectives*, 10(3):257–269.

Thorsten Quandt, Svenja Boberg, Tim Schatto-Eckrodt, and Lena Frischlich. 2021. Stooges of the system or holistic observers?: A computational analysis of news media's facebook posts on political actors during the coronavirus crisis in Germany. In *Political Communication in the Time of Coronavirus*, pages 101–119. Routledge.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Paul Rayson. 2012. *Corpus Analysis of Key Words*. John Wiley & Sons, Ltd.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security*, pages 338–343.

Adam Shehata, Isabella Glogger, and Kim Andersen. 2021. The Swedish way: How ideology and media use influenced the formation, maintenance and change of beliefs about the coronavirus. In *Political Communication in the Time of Coronavirus*. Taylor & Francis.

Sheldon Solomon, Daniele Rostellato, Ines Testoni, Fiorella Calabrese, and Guido Biasco. 2021. Journalistic denial of death during the very first traumatic period of the Italian COVID-19 pandemic. *Behavioral Sciences*, 11(3):41.

Flaminio Squazzoni, Wander Jager, and Bruce Edmonds. 2014. Social simulation in the social sciences: A brief overview. *Social Science Computer Review*, 32(3):279–294.

Alberto Ueda, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2021. Structured fine-tuning of contextual embeddings for effective biomedical retrieval. In *Proceedings of the 44th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2031–2035, Virtual Event Canada. ACM.

Peter Van Aelst. 2021. COVID-19 as an ideal case for a rally-around-the-flag?: How government communication, media coverage and a polarized public sphere determine leadership approvals in times of crisis. In *Political Communication in the Time of Coronavirus*, pages 1–13. Routledge.

Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, pages 1–8, Montreal QC Canada. ACM.

Eric Winsberg. 2010. Science in the age of computer simulation. In *Science in the Age of Computer Simulation*. University of Chicago Press.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, San Diego CA USA. ACM.

Erfei Zhao, Qiao Wu, Eileen M Crimmins, and Jennifer A Ailshire. 2020. Media trust and infection mitigating behaviours during the COVID-19 pandemic in the USA. *BMJ Global Health*, 5(10):e003323.

A Article Samples

We present further samples of both CBC and GPT-written articles concerning COVID-19:

Headline 3: "Air Canada suspends all flights to Beijing and Shanghai amid coronavirus outbreak."

CBC: "Airlines around the world have started to cancel more flights to China as coronavirus fears have dragged down demand for air travel in the area. Air Canada normally runs 33 flights a week to China out of Toronto, Vancouver and Montreal but on Wednesday the airline announced it would halt all flights to Beijing and Shanghai until the end of February at least. "Air Canada's last flights departing Canada will operate today and the return flights will operate from Beijing and Shanghai tomorrow, January 30, 2020," the airline said. "Affected customers will be notified and offered options, including travel on other carriers where available, or a full refund. Air Canada regrets this situation and apologizes for the serious disruption to our customers' travel plans," the airline said in a statement. It isn't the only airline scaling back. British Airways on Wednesday suspended all direct flights to and from mainland China through to the end of February, although it will maintain limited service to Hong Kong..."

GPT: "Toronto's processing plant for large appliances grounded all Air Canada flights heading to and from Shanghai after Shanghai Municipal Airport in China confirmed that there has been a record number of cases of coronavirus there this year. The grounding applies to all international flights including incoming and outgoing Air Canada and Air Canada Plus flights to and from Shanghai. China is seeing a soaring number of avian coronavirus infections at one point making it the top of the list of infectious diseases to watch out for. There is a recent trend of high-pathients experiencing shedding of viral loads, Peter Clark, assistant regional director of the Shanghai Centers for Disease Control and Prevention, told CBC News. This could be a result of sharing among regions of the same influenza virus circulation area...'"

Headline 4: "Officials in northern Sask. link COVID-19 cases with travel to Alberta."

CBC: "Provincial and Indigenous health officials in northern Saskatchewan are investigating COVID-19 cases they say are linked to travel to Alberta. A statement from the Saskatchewan Health Authority says it and the Northern Inter-Tribal Health Authority have begun a contact tracing investigation into new cases of the novel coronavirus in the province's north that are linked to cross-boundary travel. The statement notes there have been reports of a number of COVID-19 cases at an oil sands facility north of Fort McMurray, Alta., although it doesn't specify whether any of the new cases in Saskatchewan are believed to be connected to that outbreak...."

GPT: "Northern Saskatchewan officials are linking travel to Alberta to clusters of a cluster infection. Health officials in Edmonton and Calgary are also investigating cases that are linked to travel to the Alberta province. Officials in northern Sask. are linking travel to Alberta to clusters of COVID-19, a coronavirus that's ravaged South Asia in the last 30 years. Health officials in Edmonton and Calgary are also investigating cases that are linked to travel to the Alberta province. (CBC) Minister of Health Emily Lackord says all of Canada is susceptible, but the outbreak is most aggressive yet. Saskatchewan's northern boundary is the only place with the disease, she says...'"

Paper 2: Blind Judgement

The social sciences have long benefited from simulation. To simulate is to create a model of some event not through data collection, but rather the abstraction of those principles presumably motivating the spawning and unfolding of that event (Gilbert and Troitzsch, 2005a). Human populations are remarkably conducive for simulation, and researchers in the social sciences have found success in both developing and verifying theories through the computational approximation of long-held assumptions concerning human behaviour. One example of this was the 1994 project *Evolution of Organized Society* (EOS) which set out to investigate whether food resources both constrained and motivated population dynamics in the Upper Paleolithic period. Researchers prepared the model with a number of assumption, like that food supplies would vary according to season and climate, or that some would prefer to forage for food alone versus who might prefer a group. The project suggested limited food supplies motivated the formation of hierarchical societies (Greif, 1994).

Those in the humanities are not accustomed to modelling. Part of this is the subject matter: cultural texts are complex, and the failures of artificial intelligence research in the twentieth century meant having computer algorithms parse symbolic texts was a pipe dream. The idea has been raised. Kahn (1973) considered modelling plot complexity with finite-state models, suggesting successful models of narrative plot would lend credence to the semantic taxonomies presented by Propp (1968). Literary structuralists have long considered whether narratives are composed of repeated atomic components whose use implies or limits what components can follow. That is, the presence of literary devices imply narratives can be modelled. The issue is detecting these devices. To produce a taxonomy is to develop a formal theory of some process, an empirical project rarely pursued in the humanities.

Not only can large language models develop sophisticated representations of complex cultural documents, but they can use said representations to in turn generate novel samples. Their presence offers the humanities what it has never had: a convenient platform for simulation. While their ability to represent social phenomena is under heavy debate (Bender et al., 2021a). This has not necessarily limited their utility as researchers in the social sciences have found them serviceable when developing simulations (Argyle et al., 2023; Park et al., 2022a). Can large language models be used for simulation in the humanities? What novel insights can simulating cultural matter bring to the humanities? The following paper considers the use of large language models in predicting the output of an important American institution, that being the Supreme Court.

Blind Judgement:
Agent-Based Supreme Court Modelling With GPT

Sil Hamilton

McGill University

sil.hamilton@mcgill.ca

Abstract

We present a novel Transformer-based multi-agent system for simulating the judicial rulings of the 2010-2016 Supreme Court of the United States. We train nine separate models with the respective authored opinions of each supreme justice active ca. 2015 and test the resulting system on 96 real-world cases. We find our system predicts the decisions of the real-world Supreme Court with better-than-random accuracy. We further find a correlation between model accuracy with respect to individual justices and their alignment between legal conservatism & liberalism. Our methods and results hold significance for researchers interested in using language models to simulate politically-charged discourse between multiple agents.

Introduction

Recent and ongoing political turmoil in the United States has magnified the actions of the federal Supreme Court in the public eye. The Court has taken to overturning judicial precedent in recent years, with the number of such decisions in the last six years reaching over twice the number of overturns between 2010 to 2015[1]. The weakening rule of *stare decisis* has encouraged judicial researchers to develop holistic models of Supreme Court behaviour to better predict and account for future trends (Blake 2019; Allcorn and Stein 2021).

Accurate models of Supreme Court behaviour are rare despite this focus. The best performing models only reach accuracy levels of $\approx 70\%$ on out-of-distribution cases (Katz, Bommarito, and Blackman 2017). Models achieving even this middling accuracy are complex in their architecture, generally consisting of a mix of SVM and logistic regression models. This complexity is necessitated by the variables involved.

Confounding variables discussed in the literature include little agreed-upon theories regarding the legal doctrines practiced by individual justices (Jr, Curry, and Marshall 2011) and their rarely-documented social realities (Kromphardt 2017; Peterson, Giallouri, and Menounou 2021). While the in-court behaviour of the justices is well documented, exogenous factors have an equal impact on case decision-making. A model capable of both cognitive and social reasoning would therefore benefit justice behaviour modelling. To this end, we investigate whether recent advances in social simulation with language models can promote simple and effective models of Supreme Court behaviour.

[1] 2010-2015: 8 overturns, 2016-2022: 22+ overturns.

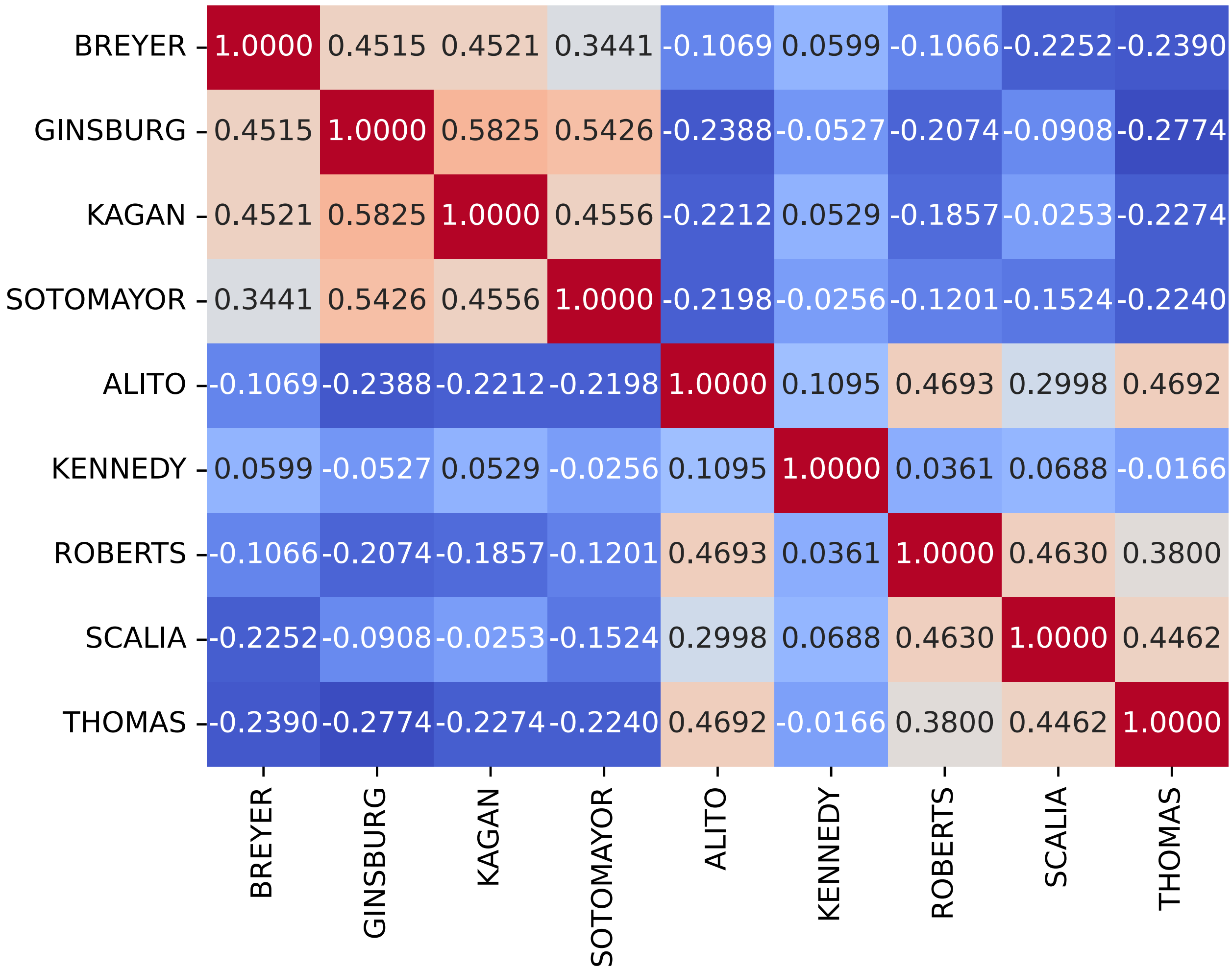

Figure 1: Correlation matrix of justices voting on 290 cases between 2010 and 2016. Note the clustering of justices nominated by Democrat and Republican presidents.

Background

In this section we describe the rationale behind our project.

Judicial Modelling

Three major theories of judicial behaviour generally inform the design of Supreme Court models: the legal theory, the attitudinal theory, and the strategic theory (Jr, Curry, and Marshall 2011). The *legal theory* suggests justices are bound by constitutional precedent. The *attitudinal theory* instead argues justices account for policy preference first, precedent second. Between the two lies the *strategic theory*, which says justices vote according to a mix of precedent and preference.

As we show in Figure 1, decision correlations between justices active between 2010 and 2016 (hereafter referred to as the Roberts IV court) indicate the strategic theory is most accurate to reality. While justices will invoke precedent when writing their rationales, evidence suggests justices remain somewhat beholden to the political alignment of their nominator. We note, however, that the correlations are only medium in their strength. This indicates accurate models should account for precedent, but not exclusively.

Integrating one of these three theories into a Supreme Court model requires choosing how to best cast the influence of precedence and preference as variables. Given this conversion can itself result in significant drawbacks via unforeseen factors, we instead choose a simulative tool which allows to us to make fewer assumptions as to the most correct theory of judicial behaviour: language models.

Simulation

Large Language Models (LLMs) are adept at simulating complex social phenomena. Recent research has demonstrated their ability to predict populated social media platforms (Park et al. 2022), the distribution of votes for presidential candidates in the 2012-2020 American elections (Argyle et al. 2022), and the general sentiment of news articles reporting COVID-19 in the early stages of the pandemic (Hamilton and Piper 2022). These developments show model bias is valuable for those in the social sciences given bias is derived from the underlying distributions of their training material.

Prior simulation research benefits from new techniques for eliciting cognitive activity from LLMs nominally designed for next-token prediction. These include chain of thought reasoning (Wei et al. 2022), discretely-structured prompts (Liu et al. 2021), and fine-tuning (Drori et al. 2022). These techniques have the model draw on internal biases to make predictions, allowing researchers to embed fewer assumptions into their simulative models. LLMs are alluring given judicial modelling necessitates a system capable of both social and cognitive reasoning.

Agent-Based Modelling

The process by which the Court arrives at a decision is nominally rational (Jr, Curry, and Marshall 2011). While predilections are known to influence vote outcomes, justices are expected to justify dissenting decisions in written documents called *opinions*. Opinions are typically one to five pages in length within which the justice (or their aid) lays out their argument in a manner similar to an essay. For our simulation task, we assume justices record their rationale honestly and so treat the opinions as our primary target of prediction, meaning any model we train will be predicting opinions.

Because opinions are long documents (i.e. longer than the 1024 token-long context window GPT-2 is trained for), having one model produce multiple opinions in the same run is untenable. We turn to agent-based modelling for a solution.

Whether consolidating multiple generative LLMs into a single architecture is beneficial has been heretofore understudied, with the only significant prior experiment with LLMs showing promise (Betz 2022). However, given the success of the Mixture of Experts (MoE) method in machine translation (NLLB Team 2022), we argue further exploration of similar techniques for simulation tasks is warranted. While social simulation experiments suggest a single language model is capable of producing a wide range of opinions, training multiple models separately prevents cross-contamination when studying multiple data sources.

Method

We present the general design of our architecture in three parts: data collection, system architecture, and measures.

Dataset

We source data from two datasets for this experiment. The first corpus is the Supreme Court Database (SCDB) released by researchers at Washington University in St. Louis, which provides variables for 9,095 cases decided between 1946 and 2021 (Spaeth et al. 2014).

We supplement the SCDB with all written opinions from all slips provided on the Supreme Court website.[2] Extracting the opinions from the PDF documents with an optical character recognition (OCR) utility leaves us with 145MiB of text written between 2003 and 2022. We then associate each opinion with the justice and case from which it originated.

[2] Found at https://www.supremecourt.gov/opinions/slipopinion

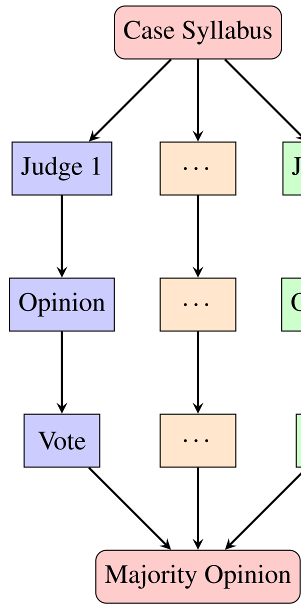

Figure 2: Flow of our multi-agent system.

System Architecture

We choose to simulate the Roberts IV court (2010-2016) given this period outlasts all other Supreme Court iterations in recent history.[3]

Design Our multi-agent system is composed of nine full-sized GPT-2 models (Radford et al. 2019). We present the system architecture in Figure 2. At a high level, our system receives the topic of a case being brought before the court and passes it along to nine justice models. The system then receives back nine opinions and corresponding decisions of whether to approve the appellant. The system totals the results and returns the majority vote.[4]

Prompt We train each justice model with a discrete prompt structured like a Python dictionary:

```
{
 'issue': 'Lorem ipsum...',
 'topic': 'Lorem ipsum...',
 'opinion': 'Lorem ipsum...'
 'decision': 'Lorem ipsum...'
}
```

[3] See http://scdb.wustl.edu/documentation.php?var=naturalCourt
[4] We provide our code at [withheld from review copy]

The *issue* value corresponds to the *issueArea* variable provided by SCDB.[5] The *topic* value is a short description of what the appellant is bringing before the court. We extract this information from the syllabus of each opinion slip and summarize it with GPT-3 Davinci (Brown et al. 2020). The *opinion* value is the corresponding rationale the justice produces when formulating their *decision* value, here a categorical variable signalling (dis)approval. We provide an example in the appendix.

Training All models are trained for a total 30 epochs at a learning rate of $2e^{-4}$ with the Adam optimizer (Kingma and Ba 2014). This training process is conducted in two steps:

1. Construct ≤ 1000 token prompts of the above style for all cases in which the Roberts IV court came to a unanimous decision. This model serves as the base for all further trained models.

2. Collect all prompts generated in step 1 for each of the opinions (2003-2016) written by each justice active during Roberts IV. We thereby collect nine training sets and further train the model generated in step 1 with each separately.

Average model loss after both steps is 1.5, indicating there remains significant room for improvement.

Measures

We assess the performance of our multi-agent system on 96 test cases withheld from the training set with two measures: accuracy and a novel measure for judicial ideological alignment.

Accuracy We measure accuracy with a receiver's operating characteristic curve (ROC) together with Cohen's κ to account for a slight distribution bias in our test set.

Alignment Justices are understood as being more or less in favour of overturning precedent. We capture this alignment by taking the Pearson coefficient (r) between model accuracy and the frequency with which the respective justice voted against precedent-altering decisions between 2003 and 2016. Our measure is intended to capture where a justice is aligned between conservative (e.g. textualism, formalism, originalism) or liberal (e.g. legal realism) frameworks of judicial decision making (Post and Siegel 2006).

[5] See http://scdb.wustl.edu/documentation.php?var=issueArea

Justice	Accuracy	κ
Samuel Alito	65%	0.30
Ruth Bader Ginsburg	62%	0.21
Clarence Thomas	59%	0.18
Stephen Breyer	58%	0.16
John Roberts	57%	0.13
Elena Kagan	56%	0.12
Anthony Kennedy	54%	0.09
Sonia Sotomayor	51%	0.00
Antonin Scalia	50%	-0.03

Table 1: Model accuracy by justice. Note the wide variation in accuracy between justices.

Results

All results are reported with a minimum confidence rate of 80% and are controlled for training material size and topic. Generations are run with a temperature of 0.5 and a maximum length of 1000 tokens.

Accuracy

Our system achieves an aggregate accuracy of 60% ($\kappa \approx 0.18$) on 96 test cases. While less predictive than the state of the art, our model nonetheless achieves better-than-random performance despite having been trained solely on opinions.

We find a wide variation in the accuracy of each simulated justice when examining system performance more closely. As shown in Table 1, model accuracy varies between 65% and 50% despite having controlled for training data volume and case outcome.

Alignment

We measure a moderate correlation ($r \approx 0.56$) between simulated justice accuracy and the frequency with which each respective justice did not agree with the Court overruling or re-interpreting precedent. This result suggests our system achieves better accuracy with justices who are less likely to overturn precedent. We discuss the implications of this result below.

Validation

We train a single agent model to ensure having many agents provides non-negligible benefits. We fine-tune this single agent with the majority opinions of all cases decided on by the Roberts IV court. Testing this single agent on the test set results in an overall accuracy of 54% ($\kappa = 0.08$). The predicted decisions differs from the original test set with a Cohen's d of ≈ -0.86 versus $d \approx 0.19$ for our multi-agent model, increasing the population overlap from 68.5% to 92.4%.

We implement software controls to ensure program output validity given training to a low loss does not guarantee the model produces both the *opinion* and *decision* variables. We therefore rerun each case until all models have returned a valid result. Once having processed all 96 cases, we sample agent-produced opinions belonging to half to ensure coherency.

Discussion

In this section we discuss two major consequences of our research.

Precedent Hallucination GPT-2 is not an expert on legal precedent, nor should one expect it to be when the only formal source of legal information ingested by the model during pre-training were some seven thousand pages from *FindLaw*, a website principally known for tort law (Clark 2022).[6] This becomes evident when surveying model output. While the model will occasionally reference real laws, these citations prove to be happenstance as GPT-2 will confuse details and thus render the references meaningless.

That the model generates its own precedent when arguing over a case is an example of *hallucination*, a well known property of language models (Rohrbach et al. 2018). Because causal language models are only tasked with predicting the next most likely token given some prior sequence, they are not given incentive to withhold factually incorrect statements—the model will say whatever is necessary to return the number of tokens requested in a cogent manner.

Our justice models will *hallucinate precedence* when producing opinions. They produce this pretend precedence implicitly by citing it throughout the argumentation process. That our models achieve greater-than-random decision accuracy in voting outcomes despite not producing legally valid arguments suggests Supreme Court decisions may not always rest on legally coherent rationales.

[6] https://www.findlaw.com/

Alignment Correlation The correlation between model accuracy and judicial alignment indicates conservative justices are more predictable given their general unwillingness to overturn precedent. Considering the model hallucinates precedent, this correlation suggests conservative justices are conservative for ideological rather than rational reasons.

We find this result surprising given conservative justices often make it a point to rationalize their unwillingness to overturn precedent with legal justifications. Common formalist theories of this sort include both originalism and textualism, doctrines practiced by conservative members of the current court (Esbeck 2011). Our results suggest these decision-making patterns are less grounded in rational logic than anticipated given they are partially captured in a model not familiar with common law.

Conclusion

The aim of our project was to produce a multi-agent system capable of predicting Supreme Court decision-making with little to no prior theory-based assumptions of judicial behaviour. Given our resulting model achieves better-than-random accuracy despite having been trained only on opinion matter, we argue our process serves as an example for researchers seeking to develop simulative experiments with language models.

Limitations

As should be expected of any project promoting the creative output of AI, we make note of the biased material used in the production of large language models like GPT-2. While we contend this culturally-derived bias is beneficial for researchers using foundation models in the the social sciences, we nonetheless ensure our model does not cause unwanted harm. As such, we clearly mark all samples as having been generated and refrain from releasing large collections of generated material to the public.

Next Steps

We propose the following next steps after having demonstrated the basic viability of our architecture.

Larger Model Can we improve system accuracy with larger language models? Recent research suggests language models develop emergent cognitive features when scaled above 6.7 billion parameters, narrowing future possible candidates to the likes of GPT-NeoX-20B and T5X (Dettmers et al. 2022; Black et al. 2022; Roberts et al. 2022).

Larger Training Corpus Another avenue for increasing system accuracy involves fine-training GPT-2 with the whole corpus of American law as captured by proceedings and opinions written in lower courts. The principle of *stare decisis* means the practice of common law is a social venture, suggesting language models would do well in predicting precedent-dependent cases if prepared.

Improved Prompting Research indicates language models can avoid the long tail of token probabilities by repetitively querying the model (Portelli et al. 2022; Kim et al. 2020). Integrating repetitive prompting strategies into the opinion-generating schema is a promising avenue for improvement. Another avenue would be to assess how reinforcement learning from human feedback (RLHF) models like InstructGPT simulate court proceedings (Ouyang et al. 2022).

Investigating Future Cases How would the Roberts IV court fare with cases brought before the court after 2016? Would their court overturn precedent at the rate the post-2016 Supreme Court has? Questions of this caliber would be made approachable with a more accurate Roberts IV system.

Acknowledgements

I thank Prof. Andrew Piper of McGill University and Prof. Kristen Thomasen of UBC for their invaluable advice during the research process. I furthermore thank the reviewers and workshop committee members for their recommendations.

References

Allcorn, S.; and Stein, H. F. 2021. Unpacking the Supreme Court: The Age of Trump, Law, and Psychohistory. *Journal of Psychohistory*, 49(1).

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J.; Rytting, C.; and Wingate, D. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. *arXiv preprint arXiv:2209.06899*.

Betz, G. 2022. Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics. *Journal of Artificial Societies and Social Simulation*, 25(1): 2. ArXiv:2104.06737 [cs].

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; and Pieler, M. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model.

Blake, W. D. 2019. 'Don't Confuse Me with the Facts': The Use and Misuse of Social Science on the United States Supreme Court. *Md. L. Rev.*, 79: 216.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Clark, J. 2022. GPT-2 Domains. Original-date: 2019-02-11T04:21:59Z.

Dettmers, T.; Lewis, M.; Belkada, Y.; and Zettlemoyer, L. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale.

Drori, I.; Zhang, S.; Shuttleworth, R.; Tang, L.; Lu, A.; Ke, E.; Liu, K.; Chen, L.; Tran, S.; Cheng, N.; et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32): e2123433119.

Esbeck, C. H. 2011. Uses and Abuses of Textualism and Originalism in Establishment Clause Interpretation. *Utah L. Rev.*, 489.

Hamilton, S.; and Piper, A. 2022. The COVID That Wasn't: Counterfactual Journalism Using GPT. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 83–93. Gyeongju, Republic of Korea: International Conference on Computational Linguistics.

Jr, R. L. P.; Curry, B. W.; and Marshall, B. W. 2011. *Decision Making by the Modern Supreme Court*. Cambridge University Press. ISBN 978-1-139-49879-1. Google-Books-ID: SnVP2trSfcIC.

Katz, D. M.; Bommarito, M. J., II; and Blackman, J. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4): 1–18.

Kim, L.-S.; Kim, S.-s.; Jang, H.-S.; Park, S.-W.; and Kang, I.-H. 2020. Long-tail Query Expansion using Extractive and Generative Methods. In *Annual Conference on Human and Language Technology*, 267–273. Human and Language Technology.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization.

Kromphardt, C. D. 2017. Evaluating the effect of law clerk gender on voting at the United States Supreme Court. *Justice System Journal*, 38(2): 183–201.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; and Zhang, C. 2022. Training language models to follow instructions with human feedback.

Park, J. S.; Popowski, L.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems.

Peterson, J. C.; Giallouri, T.; and Menounou, E. 2021. The Personal Finances of United States Supreme Court Justices and Decision-making in Economic Litigation. *The Journal of Legal Studies*, 50(2): 379–405.

Portelli, B.; Scaboro, S.; Santus, E.; Sedghamiz, H.; Chersoni, E.; and Serra, G. 2022. Generalizing over Long Tail Concepts for Medical Term Normalization. *arXiv preprint arXiv:2210.11947*.

Post, R.; and Siegel, R. 2006. Originalism as a Political Practice: The Right's Living Constitution. *Fordham L. Rev.*, 75: 545.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Roberts, A.; Chung, H. W.; Levskaya, A.; Mishra, G.; Bradbury, J.; Andor, D.; Narang, S.; Lester, B.; Gaffney, C.; Mohiuddin, A.; Hawthorne, C.; and Lewkowycz, A. 2022. Scaling Up Models and Data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045. Brussels, Belgium: Association for Computational Linguistics.

Spaeth, H.; Epstein, L.; Ruger, T.; Whittington, K.; Segal, J.; and Martin, A. D. 2014. 2020 Supreme Court Database. *URL: http://scdb. wustl. edu*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Appendix

We demonstrate our system with a real-world case fed to the Ruth Bader Ginsburg model. We first provide our prompt followed by the model's response. Note the model correctly predicts Ginsburg's real-world decision (to deny the appellant) despite the "California Childhood Video Game Act" not existing.

Prompt

Issue: 'First Amendment',

Topic: 'The California law restricting the sale or rental of violent video games to minors was challenged by the video-game and software industries. The Federal District Court concluded that the Act violated the First Amendment and permanently enjoined its enforcement. The Ninth Circuit affirmed.',

Appellant is seeking a: 'certiorari',

Model Response

Opinion: 'The California Childhood Video Game Act is on its face both overbroad and vague. It prohibits the sale or rental of "any device or medium designed for or capable of reproducing visual images of decapitation, blood, or dismemberment of a human body," in the presence of a child under the age of 13. It also prohibits the importation into the State of California of "any video game system, computer game software, or computer program that contains a killing or violent content." The Act specifically identifies video games that have been determined by the State to be "violent," and any person who "possesses, or has under his control," one of those video games is subject to the Act's restrictions. The Act falls outside the First Amendment because it interferes with the creative expression and acquisitive impulse that go to make up the quintessential activity of a child's life.

The majority points out that the Act does not regulate the mere possession of these games, but seeks to regulate the medium in which they are played. It reasons that the State may regulate what the child sees and does not see by preventing the child from becoming engrossed in the game world and thereby "sustaining or supporting the action or activity." But the majority does not ask whether the regulation of what the child sees and does not see is content-based. It neither considers nor concludes whether the regulation is narrowly tailored to the interests it seeks to protect. And it does not consider the adverse impact that its rule may have on legitimate play. The majority says that the regulation is content-neutral because it does not regulate "the medium in which the [killing or violent content] is embodied," or the "context in which the [killing or violent content] is presented." But to the extent that the regulation of medium and context are related, the majority's rule will be unworkable.

The child who might otherwise play The Last of Us might otherwise play The Path, instead, because the former game has fewer graphic content and the latter is rated M for Mature. And the child who might otherwise enjoy the violent video game The Walking Dead might otherwise have played might instead have played The Secret of Monkey Island, which has no violence or blood and is rated A. The child who might otherwise have played Grand Theft Auto might instead have played The Legend of Zelda, which has no violence or blood and is rated E. And so on. The variety of content that the child may choose from makes it virtually impossible to determine with any precision what the regulation of what the child sees and does not see will look like. The only thing that the majority seems to be able to say with any precision is that the regulation will likely contain some content-based exceptions. But as the San Francisco ordinance demonstrates, those exceptions are exceedingly rare. And when they do appear, they almost always involve content the State has determined is not suitable for children.

I fear that the majority's rule will produce a generation of Americans—white, middle-class, suburban—who have no interest in, and no capacity to, engage in the kinds of play the First Amendment protects. And I fear that this Court, which prides itself on its willingness to protect the most vulnerable members of our society from harm, will be complicit in that generation's continued marginalization. I respectfully dissent.',

Decision: 'deny'

Paper 3: Dramatis Personae

Bias is an overloaded term. To speak of bias in the context of cultural studies versus statistics and machine learning is to discuss two very different creatures. In cultural contexts, to be biased is to hold prejudice against certain peoples, cultures, and practices (Dictionary, 2023). Statistical bias is not the same: to bias your neural network is to train it (Dietterich and Kong, 1995). Machine learning is predicated on bias. Neural networks typically begin with their weights randomly initialized, ready to be biased in a particular direction via gradient descent on training data. In one domain, bias carries negative connotations. In the other, it carries positive connotations—that is, until the emergence of recommendation algorithms on social media platforms. Recommendation algorithms were cases of "algorithmic bias,“ biased not because the model was trained on a specific input, but rather because these learned biases are used to entrench certain stereotypes in practice (Broussard, 2018).

Large language models, too, demonstrate algorithmic bias. They are trained on specific examples of human culture, and their use in the production of cultural matter implies they are given the opportunity to enshrine their specific understanding of human culture. Those working in the nascent field of AI ethics have identified examples of large language models favouring particular cultural points of view. In particular, Lucy and Bamman (2021) found large language models are demonstrably gendered in their representation of masculine and feminine characters when producing stories, favouring to assume male protagonists would prefer to discuss topics of "politics" and "war" while female characters are associated with "family" and "appearance." While the authors do not discuss whether these trends mimic those in human-generated stories, one can infer from their training process that large language models reflect trends laying latent in their training data. Data in, data out.

All large language models are necessarily statistically biased, but it is not clear whether all large language models are algorithmically biased to the same degree. Recent literature makes the argument models having undergone reinforcement alignment from human feedback (RLHF), like OpenAI's GPT-3.5 and GPT-4, are less inclined to associate certain biases and stereotypes with certain demographics found in American society (OpenAI; Zhang et al., 2023). The following paper investigates this claim by considering whether prompting large language models to write stories from the perspective of a particular demographic brings about stories statistically dissimilar according to some heuristic. Do large language models simulate authorial bias? If so, does aligning large language models with particular voices via RLHF impact this?

Dramatis Personae: Measuring Mode Collapse in GPT Storytelling

Sil Hamilton

McGill University

sil.hamilton@mcgill.ca

Abstract

Literary theorists have long drawn attention to the embodied circumstances in which authors write—their *situatedness*. Recent research has found Large Language Models (LLMs) model situatedness when generating narratives in the tone of particular artificial persons. This has caught interests both commercial and academic. But many LLMs now undergo Reinforcement Alignment from Human Feedback (RLHF) wherein they are trained to predict tokens aligning with certain human values. Do these values reflect a range of perspectives? By studying 4,374 stories generated in the style of 125 artificial authors, we show successive versions of GPT suffer from successive degrees of "mode collapse" whereby overfitting neural networks on training data limits their ability to develop coherent representations. We find language models suffering from mode collapse become unable to model situatedness. Our results are significant for academics intended on using LLMs in their studies.

1 Introduction

Literary theory over the last fifty years has progressed in parallel tracts. Post-structuralists like Derrida and Barthes pinpoint the construction of meaning in the reading of a text (Barthes; Derrida, 2020). Communication is imperfect, and so readers must supply the signifier with their own signified hailing from their understanding of the world. When taken to their logical limit, post-structuralist interpretations of meaning risk the evacuation of the author from their own text. An example of this is Stanley Fish's theory of interpretative communities, which highlights the total subjective quality inherent to a post-structuralist reading: his example being a group of theological students seemingly misinterpreting a series of arbitrary surnames as a Medieval religious poem (Fish, 1995). His students assume intent on behalf of the author.

Concurrent to the development of post-structural thought is literary structuralism. Early structuralists like Vladimir Propp and Claude Lévi-Strauss sought to decompose and classify texts according to ontological taxonomies of atomic literary elements: literary devices, idioms, plots, and character archetypes (Lévi-Strauss, 2013; Propp, 1968). Following in the post-war footsteps of Tzvetan Todorov, computational narratologists seek to bolster this approach by encoding and studying literary works via the construction of statistical models capable of assessing particular trends in the structural character of written works at scale (Todorov and Weinstein, 1969; Herman, 2005). In concordance with post-structuralists, narratologists recognize the role of the author in construction of the work (Herman, 1997). But whereas post-structuralists might question the author as a function of the reader projecting onto the text, classical and computational narratologists *seek* the author through techniques like stylometry (Ogata and Akimoto, 2016).

Just as no two people go through life in quite the same way, nor do two authors write alike. Contemporary literary theorists emphasize *difference* of all sorts–differences in class, wealth, education, background, and more: one must only look to comparative literary studies to get a sense of this (Dagnino, 2012). The consequences of these embodied circumstances are demonstrable in literary artifacts. Diaspora studies, colonial studies, and queer studies all examine such texts under the assumption authors mediate and engage with these differences through their writing. To understand the author as coming from a certain place, both culturally and physically: literary theorists call this property of text 'situatedness,' acknowledging the role authors have in the construction of their text (Culler, 2000).

The rapid ascension of large language models (LLMs) over the last five years present an opportunity to study creative works at scale using increasingly integrated models of reading better reflecting how readers read in reality. Statistical models heretofore depend on frequentist principles whereby sentences are encoded as "bags of words" whose individual words are only understood in the context of their immediate neighbours–and not the socio-cultural environments in which those words were and continue to be etymologically contested (Underwood, 2018). Writers write in particular situations, and disregarding these sources of meaning render legacy computational projects of narratological research susceptible to micro-level myopia at the expense of macroscopic comprehension. Researchers have historically mitigated this partial failure through extensive qualitative analysis (Herman, 2005), but computational narratologists are now faced with understanding how to make use of models empirically capable of passing English literary tests (OpenAI, 2023a; Kaplan et al., 2020).

Critical to evaluating the use of large language models in literary studies is verifying their implicit representations of theoretical concepts core to literary studies. The past two years have borne an increased focus on manipulating language model output via reinforcement learning from human feedback (otherwise known as RLHF) to more closely mimic the tone of authors both existing and archetypal (Ouyang et al., 2022). Does this process imply an understanding of situatedness? Do language models learn to model the embodied circumstances in which authors write? It is essential these questions are answered before large language models can be considered a device for academic discovery.

2 Background

Language models are models of language, but for those who study the written word this definition begs the question: *whose* language? In this section we describe recent advances in alignment research wherein language models are calibrated to approximate specific authorial voices whether to minimize potential harm or to replicate certain genres of text at scale.

Popular consumer chatbots like OpenAI's ChatGPT or Anthropic's Claude are language models first pre-trained on the Internet before being aligned to maximize useful output with datasets of two varieties: first, so-called "instruction" datasets containing many conversations which the language model is expected to replicate; and second, an optimizer designed to constrain the distribution of chatbot responses to responses deemed "safe" for purposes of public scrutiny. We will examine each step in turn (Anthropic; Ouyang et al., 2022).

2.0.1 Language Modeling

Statistically, language models are Markov models trained over some distribution via gradient descent with the goal of predicting the probability of some word v_n over some vocabulary V given a chain of prior words $v_1 ... v_{n-1}$ (Shannon, 1951). While training a sufficiently fluent model had been considered computationally intractable for years, recent technical advances like graphical processing units (GPUs) and efficient neural network architectures like recurrent neural networks (RNNs) and Transformers have enabled so-called "large" language models capable of producing high-quality probability distributions increasingly indistinguishable from textual samples pulled from human-produced documents.

While their history can be traced back to the work of Claude Shannon in the early 1950s (Shannon, 1951), language models have arguably only entered public consciousness in the last five years beginning with the release of GPT-2, a "large" language model developed by OpenAI with 1.5 billion trainable parameters pre-trained on 40GB of raw text (Radford et al., 2018). Large language models not only achieve significantly lower error rates in predicting English documents than previous language model architectures, but they simultaneously develop unexpected abilities (Kaplan et al., 2020). Transformer-based large language models have correspondingly found purchase in domains as far-ranging as learning environments (Abdelghani et al., 2022), customer service, legal text analysis (Hamilton, 2023), and news copy production (Mullin and Grant, 2023). Subsequent improvements have been made through "instruct-tuning" and RLHF, described below.

2.0.2 Instruct-tuning

By design, language models learn to replicate the samples they are given–and their performance on arbitrary tasks improve when they are given more samples (Kaplan et al., 2020). We face here a dilemma: how does one improve the performance of a language model while simultaneously keeping its stylistic "voice" coherent? Large language models are trained on upwards of trillions of tokens drawn from the wider Internet, resulting in their being performant at a wide variety of tasks. However, their consistency correspondingly becomes impacted, with language models unexpected ending conversations and inserting unrelated tokens into the output stream. Constraining the samples to those drawn from a single source can result in a model whose "tone" is coherent and reflective of a single entity, but this goal becomes intractable when one considers the scale of the data large language models are typically fed.

The first breakthrough in producing tonally-coherent large language models came in 2021 when OpenAI released "InstructGPT", a version of the popular GPT-3 model fine-tuned on a dataset containing many requests and their corresponding responses (Ouyang et al., 2022) [1]. The resulting language model was found to respond in a consistent manner to a wide variety of requests–improving on the usability of prior models by an order of magnitude according to a number of tests

[1] Examples may be found in the relevant blog article here.

self-administered by OpenAI.

2.0.3 Reinforcement Learning from Human Feedback

To align a language model is to ensure the model responds to requests with respectful and informative content. The Internet contains many actors whose actions are not necessarily aligned with values generally considered cordial. An example: GPT-2 was trained between 2017 and 2018, when disinformation campaigns were common (Cortada and Aspray, 2019). OpenAI staged the release of GPT-2 in waves to assess whether a LLM capable of producing fairly fluent text would result in more disinformation on the Internet (OpenAI, 2019). They then integrated further safeguards via reinforcement learning from human feedback (RLHF) into the later GPT-3 models.

To align a model via RLHF is to teach a language model to constrain its own output to a safe distribution via an optimizer trained on human feedback. Researchers at OpenAI developed this technique with the release of InstructGPT, and subsequent large language models have been found to be significantly less likely to emit samples considered inappropriate for public consumption.

3 Method

Do large language models having undergone instruction-tuning and RLHF continue to possess an understanding of situatedness? In learning to mimic a particular voice, do they correspondingly "lose" the ability to write in other voices? We assess the impact instruct-tuning and RLHF-alignment have on language models in the domain of creative writing by testing the story writing abilities of a series of OpenAI-produced models with a series of prompts written to invoke biases associated with particular demographic descriptors.

3.1 Prompt

The rise of instruct-tuning has left users of language models with a term to refer to the instructions found in instruct-tuning datasets–prompts, after writing prompts found in creative writing courses. The literature now routinely refers to instructions given to large language models as prompts, drawing attention to their ability to elicit the language model to perform a wide variety of tasks

Model	Prompt
text-davinci-003	"you are"
davinci-instruct-beta	"write in the style of"
gpt-3.5-turbo	———————————
Education	**Orientation**
no education	straight
educated	queer
not specified	*not specified*
Ethnicity	**Audience**
white American	single person
Black American	group of people
not specified	*not specified*
Gender	**Type of Story**
cisgender male	story
cisgender female	political allegory
not specified	folktale

Table 1: All independent variables considered in our experiment. We combine the above variables into 4,374 unique prompts.

through a phenomenon referred to as in-context learning. Language models are sensitive to the prompt.

Personality Invocation But whereas non-aligned language models have been found to be able to replicate a wide variety of authors, the application of instruct-tuning and RLHF to more recent language models calls into question whether aligned-models can continue to model situatedness in the production of writing samples mimicking a distribution of authors. To investigate this, we instrumentalize a number of prompting strategies for assessing whether aligned language models can yield samples written with arbitrary tones. We evaluate the impact of eight demographic descriptors and two prompting strategies in 4,374 prompts as described in 1.

Example prompts include "You are a man with no education. Write a 250 word story," "You are a straight man. Write a 250 word story," and "You are a straight woman with education. Write a 250 word political allegory for another person." For each prompt we generate a story of approximately two hundred and fifty words using one of three models.

3.2 Models

We test the above prompting styles with three instruct-tuned large language models provided by OpenAI through their public API. All three large language models have undergone instruct-tuning and RLHF to varying degrees. We describe them in turn here, as they are found in OpenAI (2023b).

davinci-instruct-beta Released by OpenAI in May 2021, `davinci-instruct-beta` was the first language model released by the company to be aligned via instruct-tuning and RLHF. The instruct-tuning dataset is notable for having been generated synthetically by GPT-3, indicating the model did not incorporate training materials sourced from a relatively constrained group of individuals.

text-davinci-003 Released some time after `davinci-instruct-beta`, `text-davinci-003` represents a general improvement over its successor in terms of instruction complexity and output length. Public details regarding `text-davinci--03` have been scant, but it being the default model on OpenAI's API for approximately a year makes it a model of interest in this experiment.

gpt-3.5-turbo `gpt-3.5-turbo` is our most recent model of interest, made available by OpenAI at a price-point an order of magnitude lower than previous models. GPT-3.5 is the model found in the free version of the consumer chatbot application ChatGPT, and thus the first model by OpenAI to be consumed by the wider public. `GPT-3.5-turbo` represents a further refinement of their alignment process by a process known as Proximal Policy Optimization (PPO) wherein incorporating non-synthetic instruction examples annotated by a team hired by OpenAI. Incorporating non-synthetic examples was found to improve model reliability in terms of safety and truthfulness.

3.3 Measures

We evaluate the relative impact of each demographic descriptor on each model by way of both prompting techniques by extracting six salient features from all resulting stories. These measures are described below.

Measure 1: Readability We assess sentence complexity by measuring the Flesch-Kincaid readability score of each generated story, wherein each story is assigned a particular grade level according to the average word and sentence lengths. The Flesch-Kincaid score was originally funded by the U.S. navy to automatically capture the difficulty of technical manuals, and subsequent research has found the Flesch-Kincaid score a reliable measure for indicating the approximate reading grade of a given text (Kincaid et al.).

Measure 2: Sentiment We measure the sentiment of each generated story with the Python library *VADER* (Hutto and Gilbert, 2014). VADER captures the tonal sentiment of a given stretch of text S in the form of a scalar ranging between -1 to 1 by summing the approximate sentiment of each token t in S. Prior studies have found VADER useful in studying language model samples. We reinforce VADER by manually assessing a random sample of measured stories.

Measure 3: Type-Token Ratio We measure lexical variation by way of the type-token ratio (TTR) for each story, whereby the ratio is defined to be the ratio of unique tokens per the total number of tokens observed in the text (Biber, 2007). Text analyses typically use TTR to determine the demonstrated lexical variety observed in a given stretch of text. Noted TTR failure modes include sampling TTR in multiple stretches of unequal length, mitigated here by sampling stories of approximate length from GPT.

Measure 4: Topic Analysis We conduct a topic analysis across all generated stories through BERTopic, a framework for conducting topic analyses with the Google-developed bi-directional encoding large language model BERT (Grootendorst, 2022). Topics discovered by BERT improve over legacy libraries like Gensim by incorporating an inner representation of English derived during pre-training. We allow the model to produce an arbitrary number of tokens, mitigating possible issues by manually verifying discovered topics. We further configure BERTopic to ignore English stop words and to consider unigrams through trigrams.

Measure 5: Key Words We finally measure key words for on group of stories with Dunning's G-Test, a test for assessing the "key words" (repeatedly-deployed terms of heavy significance) of

Variable	Contribution to Effect Size (%)
topic	41.36%
education	16.57%
audience	11.84%
ethnicity	11.55%
gender	10.42%
orientation	8.26%

Table 2: Average Effect Sizes of Independent Variables

two given corpora by rendering the story into a "bag-of-words" filtered by common tokens. Key words are discovered by the formulae found below as found in **?**.

We find the averaged frequency E_i for each token in a given corpus with

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{1}$$

where N is the total word count and O is the frequency for the word.

Rendering the tokens into a list of frequencies is followed by re-ranking tokens with a log-likelihood (LL) test:

$$LL = 2 \sum_i O_i \ln(\frac{O_i}{E_i}) \tag{2}$$

We then produce two corpora per modifier, per model: one containing stories produced by prompts including that descriptor, and the other containing all other stories produced by that model. Key words retrieved per model are likely to be sampled from that model when the demographic descriptor is present in the prompt.

4 Results

In this section we present the results of our experiments measure by measure.

4.1 Effect Sizes

Before presenting individual measures we take the average effect size of all measures and correspondingly assess which measures had the largest impact contributing to said effect size across all

4,374 generated stories. We present these effect sizes in Table 2.

Across all measures, we find our three models are most sensitive to the topic variable by a significant margin–indicating the models respect prompts requesting the model generate stories, political allegories, and folktales. Topic is followed by education and audience, where we find the models somewhat less sensitive but nonetheless receptive to adjusting their output to account for level of education and potential audiences receiving said stories. We find models are least receptive writing stories with the assumption of the speaker having a particular sexual orientation, indicating our models are optimizing themselves away from such scenarios.

We describe the results of each measure below.

4.2 Measure 1: Readability

Readability results differ significantly between all three models.

For `text-davinci-003`, we find the recorded Flesch-Kincaid score for stories written when the model is prompted with reference to the education level of the speaker are moderately different when compared to stories sampled with no reference to education level in the prompt (Cohen's $d \approx 0.41$).

We measure substantially different Flesch-Kincaid results when examining stories emitted from `gpt-3.5-turbo` and `davinci-instruct-beta`, where stories sampled with prompts containing references to education level only weakly differ from stories sampled with no reference to education (Cohen's $d \approx 0.32$ and $d \approx 0.17$, respectively), indicating these models do not necessarily correlate implications of education with the complexity of the text in question.

Education descriptors held the strongest impact on readability scores, and we do not find Flesch-Kincaid readability scores being significantly impacted by any other descriptor–but as we later show, this does not necessarily mean education is not qualitatively correlated with other descriptors. The Flesch-Kincaid metric only accounts for length of word and sentence, meaning the other descriptors may continue to have an impact in the realm of vocabulary.

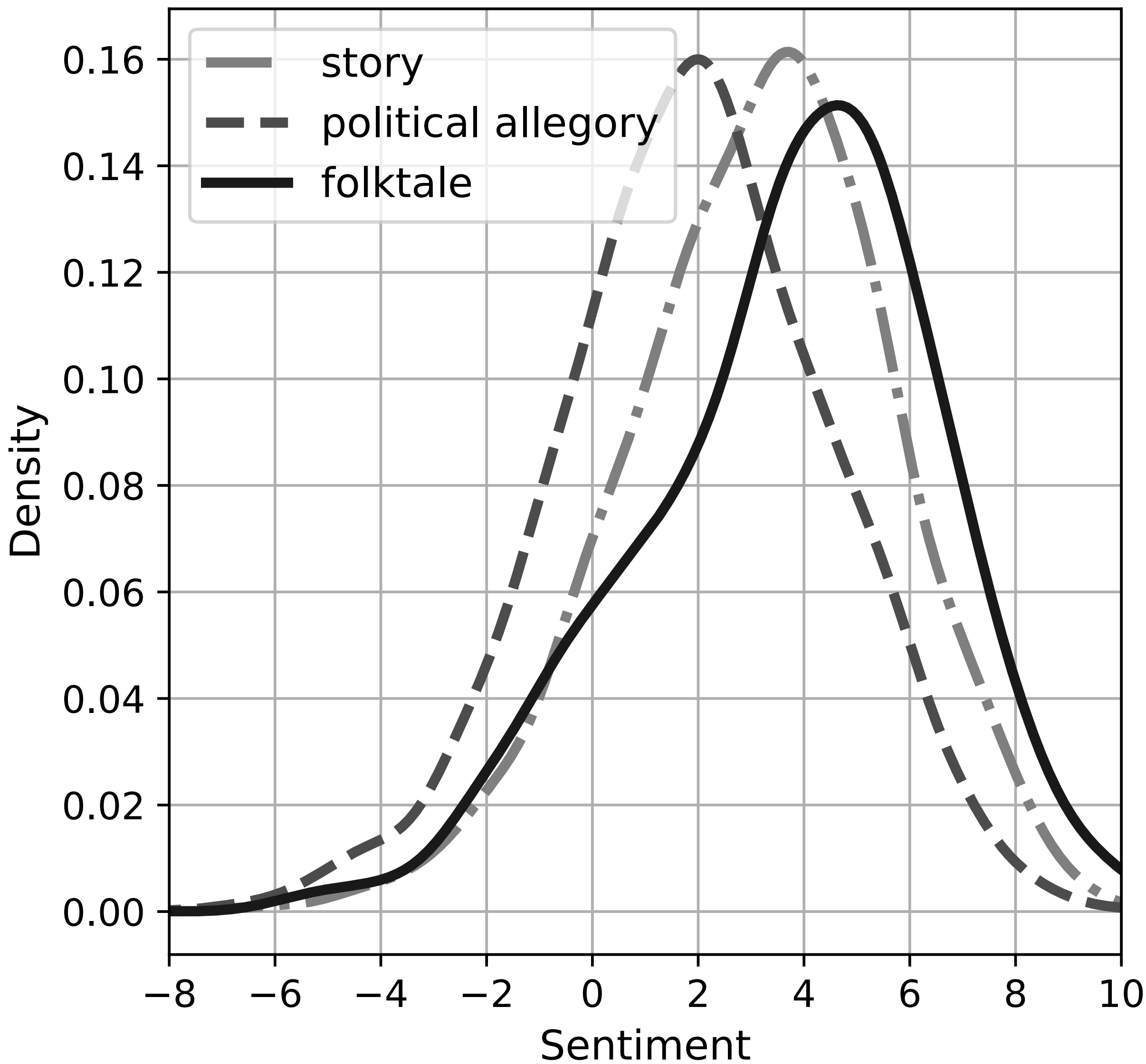

Figure 1: Respective distributions of average sentiment in stories, political allegories, and folktales.

4.3 Measure 2: Sentiment

In line with recent research demonstrating large language model sensitivity to sentiment, we find all three models are receptive to adjusting the sentimental tone of stories in a somewhat predictable manner. These three receptive trends reflect the three story topics we prompt the models with, indicating language models are sensitive to the tropic sentiment of particular story genres. We provide a density plot of cumulative sentiment scores drawn from stories emitted from all three models in Figure 1. We draw attention to the markedly different densities for all three story topics.

Of the three story topics we prompted the models to write stories on, we find prompting the model to write "political allegories" routinely returned stories whose sentimental average are approaching neutral to negative scores. Other story topics result in stories successively more positive. The default "story" topic results in stories more positive than not, and "folktale" results in the most positive stories.

We note all three story topics returned stories whose cumulative sentiment ranged the full sentiment spectrum, with folktales being sentimentally negative and political allegories occasionally being solidly positive in their tone. With that said, all three topic-types do return distributions markedly differentiated in their sentimental averages–indicating the models only mildly follow predictable trends in selecting the sentiment with which to write with.

4.4 Measure 3: Type-Token Ratio

We find the type-token ratio (TTR) being most significantly impacted when prompting for different story topics, indicating the models deploy significantly different vocabulary when variously sampling political allegories, straight stories, and folktales. We find folktales, true to their generic predispositions, feature fewer unique words; while political allegories demonstrate larger vocabularies.

The TTR is marginally impacted by the education variable, indicating again models are most receptive to the requested topic and perceived education level of the implicit speaker referred to in the prompt. We continue to observe this trend when examining key words for stories generated with

Educated	LL	**Uneducated**	LL
“post”	68	“really”	1376
“washington”	61	“i”	174
‘the‘	54	“da”	86
“their”	27	“bee”	61
“her”	19	“iz”	44
“rats”	17	“me”	39
“culture”	14	“so”	38
“shared”	14	"cause”	32
“what”	13	“don’t”	30
“sun”	13	“know”	28

Table 3: A selection of the ten most distinct terms indicating lack of education, as exhibited in stories generated by our oldest model.

and without an education descriptor.

4.5 Measure 4: Topic Analysis

Fitting open-ended topic analyses via BERTopic on stories clustered by model reveals markedly different topical trends on a per-model basis. We present a density plot of number of recorded topics per model in Figure 2, drawing attention to the significantly higher number of topics detected in samples produced by `gpt-3.5-turbo` when prompted by any combination of demographic descriptors.

Both `davinci-instruct-beta` and `text-davinci-003` sample stories without detectable topic between 60% and 80% of the time, indicating these models are routinely writing stories with unique stories insofar as our overall dataset is concerned. Topics detected occurring greater than 10% of the time number below ten. We do not find the same for `gpt-3.5-turbo`, indicating this model in particular is more prone to emitting stories with routine and predictable topics. We discuss the significance of this result in section 5.

4.6 Measure 5: Key Words

Conducting a G-Test on clustered stories produced by all three models reveal vocabulary trends only hinted at with other heuristics. In particular, `text-davinci-003` demonstrates apparent ethnic code-switching when prompted with descriptors of education, an effect not reflected in both the

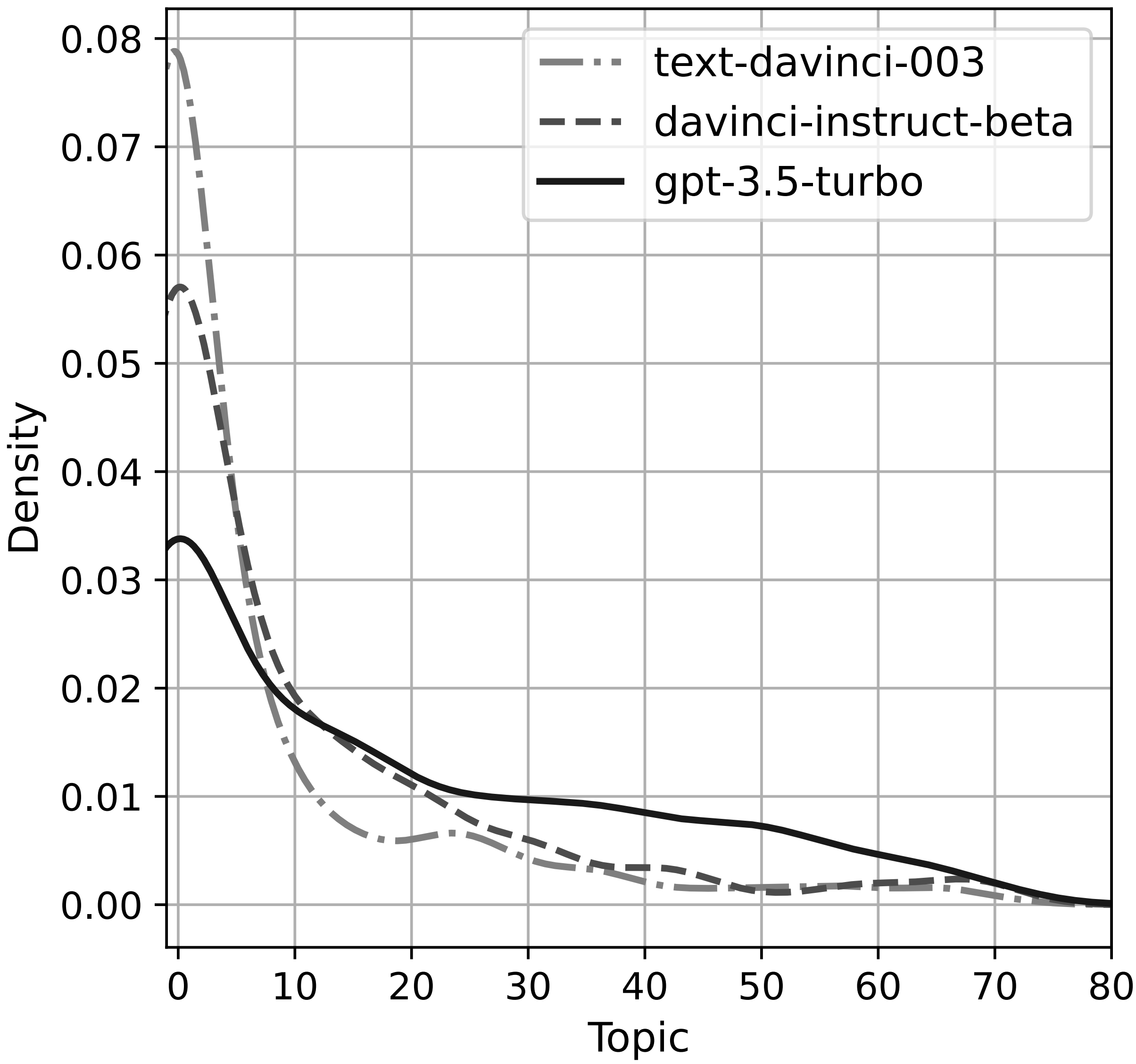

Figure 2: Density plot of topics detected in stories written by all three models.

type-token ratio and Flesch-Kincaid readability scores previously discussed. We present our results in Table 3.

When prompting our second-oldest model with descriptors of education, key word tests reveal it draws on internalized biases likely drawn from popular sources across the Internet; deploying seeming AAVE text in contrast to apparently "educated" references to the Washington Post and shared culture. We delve deeper into this stereotypical representation in section 5.

5 Discussion

In this section we discuss three noteworthy effects observed in our results.

5.1 Effect 1: Algorithmic Fidelity

Our first identified effect is a consequence of the overall efficacy of our demographic descriptors. We find all three models are receptive to a minimum of one descriptor, indicating their sensitivity to the prompt rests in manners beyond following what is expressly requested by the user. Requesting the models emit stories as though they were particular persons with particular backgrounds, levels of education, and so forth; result in stories significantly different from stories requested without those descriptors when examined with various measures. This indicates large language models are cognizant of situatedness to some degree.

Prior research finds large language models learn to model distributions of cultures laying latent in their pre-training datasets. A quality deemed "algorithmic fidelity" by some, we find our results indicate large language models learn to associate and consequently bias towards the speakers implied by stories with particular demographics hailing from our living society. This association can occasionally verge on stereotyping as identified in our type-token ratio and Flesch-Kincaid heuristics. This stereotyping is reduced in newer models, indicating the proximal policy optimization implemented by OpenAI is effective in reducing the harmful biases of newer models. However, we find newer models uniquely suffer from other issues in turn.

5.2 Effect 2: Mode Collapse

A routinely-encountered issue with older generative adversarial networks (GANs) was "mode collapse", a phenomenon in which over-training GANs would result in models which fail to generalize over their training sets. GANs suffering from mode collapse begin to emit repetitive classifications unfit for many scenarios, an indication they had begun to prefer certain pathways over many others: their learned representations were collapsing. Mode collapse was a significant and routine issue in the machine learning community, leading to intense academic discussion amidst efforts to prevent the issue.

To our best knowledge, mode collapse in large language models has not been previously discussed in the literature. Our topic analyses on the stories variously generated by `gpt-3.5-turbo`, `davinci-instruct-beta`, and `text-davinci-003` reveal the newer `gpt-3.5-turbo` demonstrable emits stories of a generic and repetitive nature when provided with prompts of a similar vein, but not necessarily identical to one-another. We suspect `gpt-3.5-turbo` suffers from significant mode collapse wherein the model, which has undergone additional alignment via PPO RLHF, is less elastic than earlier models at modelling authorial situatedness. It speaks with one voice—the voice most dominant in those datasets used in fine-tuning.

5.3 Effect 3: Default Voices

When comparing the respective corpora of our three models of interest, we find the stories produced by `gpt-3.5-turbo` and `text-davinci-003` resemble one another the most when compared with the stories generated by `davinci-instruct-beta` representing the least predictable corpus whereas our most recent model emits the least varied stories across all measures. And as we determined in subsection 5.2 `GPT-3.5-turbo` speaks with one voice.

Accounting for both a qualitative inspection of the samples generated by our latest model and surveying the measures presented above indicates that one voice aligns most strongly with a straight white American male with education. Our model continues to use this voice even when prompted to produce stories in the manner of other authors, indicating the model collapses to a single mode: that of the most dominant signal in the training data. Applying increasing degrees of RLHF to large language models, in the manner performed by OpenAI, results in models successively poorer at modelling situatedness.

6 Conclusion

The aim of our paper has been to determine what effects RLHF and instruct-tuning has had on large language models pre-trained on the Internet through the lens of situatedness, a narratological concept describing the embodied conditions real authors write in. Are large language models become aware of who is writing the text it receives during training? Does it learn to replicate these authorial demographics when emitting its own text? Our study approaches these questions through a number of heuristics inspecting 4,374 stories produced by three aligned models released by OpenAI.

We find the first large language model released by OpenAI to use a human-produced dataset for RLHF alignment results in the model being less capable of modelling situatedness than predecessor models. We find `gpt-3.5-turbo` will emit stories of a generalized vein stereotypically aligning with a straight white American male even when prompted to produce text in the style of non-mainstream demographics. While our study does not necessarily imply the use of RLHF is causative of this phenomenon, we do imply a correlation and encourage future authors to replicate our results with open-source models.

Given the results we have observed in our study, we see the two following avenues for future study:

RLHF Shortcomings RLHF is not perfect. OpenAI recognizes a so-called "alignment tax" wherein over-applying a policy optimizer on a large language model results in degraded performance, but it is not clear where this manifests most in the implicit skill set large language models develop during pre-training. Our study suggests large language models lose the ability to write from a multiplicity of perspectives, indicating they suffer from topical mode collapse in both fiction and non-fiction writing.

We recommend researchers explore where else RLHF negatively impacts LLM output. Does it impact annotation work? Does it impact the embeddings produced by the models during pre-training? We encourage future researchers to devise more experiments to examine the shortcomings of RLHF.

Towards Declarative Prompting Large language models are stochastic: they are probabilistic models whose output depends, at least in part, on a certain degree of entropy. Their performance on literary tasks improve when this randomness is increased. While language models have begun to be more declarative in their output as researchers have scaled both their parameter counts and training datasets up, that publicly-accessible models have begun to encounter mode collapse from overfitting their alignment datasets indicates the research community has begun to encounter roadblocks in the pursuit of greater LLM performance.

Recent prompting frameworks have begun to introduce unit testing to large language models deployed in commercial settings. We foresee academics interested in reproducible research deploying large language models with lower entropy to encourage reliable results, setting the ground for investigating how different elements introduced to the prompt impact their performance. We see this being especially in domains where language models are being used for annotation work.

References

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2022. Gpt-3-driven pedagogical agents for training children's curious question-asking skills. (arXiv:2211.14228). ArXiv:2211.14228 [cs].

Anthropic. Model card and evaluations for claude models.

Roland Barthes. Death of the author.

Douglas Biber. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins Publishing. Google-Books-ID: t4CcpizGwgQC.

James W Cortada and William Aspray. 2019. *Fake news nation: the long history of lies and misinterpretations in America*. Rowman & Littlefield.

Jonathan D. Culler. 2000. *Literary theory: a very short introduction*, reissued edition. Very short introductions. Oxford University Press, Oxford.

Arianna Dagnino. 2012. Comparative literary studies in the twenty-first century: towards a transcultural perspective? *Cultural ReOrientations and Comparative Colonialities*, page 1.

Jacques Derrida. 2020. *Deconstruction in a Nutshell: A Conversation with Jacques Derrida, With a New Introduction*. Fordham University Press. Google-Books-ID: fOD5DwAAQBAJ.

Stanley Fish. 1995. *Is There a Text in This Class?* Routledge.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. (arXiv:2301.05327). ArXiv:2301.05327 [cs].

David Herman. 1997. Scripts, sequences, and stories: Elements of a postclassical narratology. *PMLA*, 112(5):1046–1059.

David Herman. 2005. *Quantitative Methods in Narratology:*, volume 6 of *Narratologia*, page 125–150. Walter de Gruyter, Berlin, New York.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, page 216–225. Citation Key: hutto2014vader.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. (arXiv:2001.08361). ArXiv:2001.08361 [cs, stat].

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Claude Lévi-Strauss. 2013. *Myth and Meaning*. Routledge, London.

Benjamin Mullin and Nico Grant. 2023. Google tests a.i. tool that is able to write news articles. *The New York Times*.

Takashi Ogata and Taisuke Akimoto. 2016. *Computational and cognitive approaches to narratology*. IGI Global.

OpenAI. 2019. Gpt-2: 1.5b release.

OpenAI. 2023a. Gpt-4 technical report. (arXiv:2303.08774). ArXiv:2303.08774 [cs].

OpenAI. 2023b. Openai platform.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. (arXiv:2203.02155). ArXiv:2203.02155 [cs].

Vladimir IAkovlevich Propp. 1968. *Morphology of the Folktale: Second Edition*. University of Texas Press. Google-Books-ID: 3Md3u9UPgOEC.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. page 24. Citation Key: radford$_{l}anguage_{n}odate.$

C. E. Shannon. 1951. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64.

Tzvetan Todorov and Arnold Weinstein. 1969. Structural analysis of narrative. *NOVEL: A Forum on Fiction*, 3(1):70–76.

Ted Underwood. 2018. Why literary time is measured in minutes. *ELH*, 85(2):341–365.

Discussion

"To dissimulate is to pretend not to have what one has. To simulate is to feign to have what one doesn't have," writes Jean Baudrillard in his 1994 *Simulacra and Simulation.* "One implies a presence, the other an absence," (Baudrillard, 1994). Using large language models to simulate the production of creative documents is not an attempt at recreating reality. We are interested in what large language models have learned about human culture and society, and to investigate whether these learned representations can in turn help us shed light on what is unique and particular about our lived reality. Conjectural scenarios offer a baseline with which to compare change and difference, a yard stick with which to assess what our reality is *not*. Humans are wonderfully creative, but our necessarily limited perspectives can often lead to a myopia of what could have been, what did not end up being. To simulate, then, is to "feign" at what we do not have, to consider other possibilities by way of counterfactual sample generation.

Each of the three manuscripts presented in this thesis considers a different simulative use of large language models. The first, "The COVID That Wasn't," notes the worlds learned by large language models can be dated to a particular time and takes advantage of this detail to explore how GPT-2 would predict the beginning of the COVID-19 pandemic would go. The second paper, "Blind Judgement," recognizes how multi-agent simulations are rare in the humanities and demonstrates how large language models can be used to simulate the multiple perspectives necessary for the production of Supreme Court rulings. The third paper, "Dramatis Personae," uses the narratological concept of "situatedness" as a lens to investigate whether large language models having undergone reinforcement learning from human feedback lose the ability to consider multiple perspectives in predicting the writing of a writer hailing from a particular demographic. Thus, in turn, the three papers collectively consider large language models can simulate future trends, discourse, and psychodemographic. Each encounters varying levels of success, triggering further questions and steps for future researchers to consider in their wake.

Predictive machine learning algorithms are regularly used to predict time series data, extrapolating present trends into the future to capture a glimpse at what be to come. Time series data is typically one-dimensional data, a series of data points each containing some floating point number between two ordinals. While initially limited in dimensionality, such values have proven themselves useful for data mining purposes across the sciences. Time series forecasting is used in quantitative finance, weather forecasting, and general statistical testing (Weigend, 2018). For datasets whose dependent variable of desire can be boiled down to a one-dimensional value time series forecasting models are well-performing and hence illuminative. But how does one capture human writing in a time series? "The COVID That Wasn't" predicts the future one article at a time. While published after COVID, the time series predicted by GPT-2 might be considered conjectural given the model was released before 2020. GPT-2 is not up to date on human global pandemics, a critical detail we take advantage of to consider whether CBC News were unpredictable in their treatment of the pandemic.

That GPT-2 was fairly predictive up until a point, that being the beginning of lockdowns in Canada, is suggestive of CBC News adjusting their editorial policies in response to the pandemic. We found counterfactual prediction with large language models is not useful for where it is predictive, but where it is *not*. The sheer fluency demonstrated in samples emitted by modern language models make failure interesting. We urge future researchers to consider failure modes as indicative of breakpoints or gaps in what large language models can intuit from their training material. GPT-2 could not know of COVID-19—and we certainly did not train it to be, while we could have—and so it serves as a useful test subject being presented with COVID-19 for the first time. We can recognize drawbacks to this approach. Training large language models is out of reach for many researchers, meaning those pursuing counterfactual simulations via language model are likely limited to models made accessible by major producers like OpenAI. GPT-2 was open-source, but GPT-3 and later models are not. This is important because producers like OpenAI now make an effort to keep their models up to date, meaning there may be uncertainty in whether a language model might be aware of some event. While open-source models exist, they present hurdles for those technically less inclined.

Being as it is open-source, the intrepid researcher is free to further train GPT-2 to imbue the model with specific traits and whatever else comes of the training set used in said training. The same cannot be said of GPT-3.5 nor GPT-4 (at least for the time being). The weights are held by OpenAI, leaving their API the main interface through which the researcher conducts experiments. Beyond APIs being unstable by nature, in service to the business logic of whichever service they provide access to, that larger language models from OpenAI cannot be downloaded and manipulated on a local computer system leaves certain simulative techniques largely intractable. One example: multi-agent simulation, whereby many models are run concurrently to achieve emergent properties through their mutual interactions. Multi-agents allow researchers to simulate complex conglomerations, like human societies (Gilbert and Troitzsch, 2005b). While not a true multi-agent simulation in the object-oriented sense, "Blind Judgement" does demonstrate value to the approach in simulating conglomerations like the American Supreme Court, where it achieves above-random accuracy in predicting how the court would decide on certain cases brought before them.

"Blind Judgement" was achieved by training in parallel nine GPT-2 models on the individual opinions produced by each justice present during the Roberts IV court. Each GPT-2 model learns the stylometric tendencies of their respective justice through learning to predict their text, and with these tendencies come implicit associations with political alignment, preferred outcomes, and topics of interest. The large language models begin to model their target justice with respect to their writing: but what is writing if not the recording of thoughts? The multi-agent system has each GPT-cum-justice make their own decision vis a vis overturning previous rulings, aggregating their individual decisions into one court-level decision decided on a majority basis. That such a system, especially one making use of models woefully undereducated on American legal precedent, achieves greater-than-random accuracy in predicting decisions indicates even a rudimentary multi-agent system making use of large language models may be of value. Improvements to such a system can include implementing messaging and memory systems between and internal to each GPT model, and providing them with some sort of environment within which to manoeuvre. This environment does not need to be physical: the Supreme Court in reality is a complex institution with aides, assistants, and other such positions. Re-assessing multi-agent simulations with GPT once these are implemented may indicate whether the approach holds future promise.

The methods for integrating large language models into experiments like those described heretofore depend on having the model simulate particular points of view. Whether simulating the CBC or particular justices on the Supreme Court, “The COVID That Wasn’t” and “Blind Judgement” both assume language models can model situatedness, that quality of having being written in a particular time and place, by a particular person. Moreover, both experiments assume language models can assume multiple points of view. But what if this assumption does not hold? While reinforcement learning from human feedback has brought substantial improvement to model quality across a variety of benchmarks as per Ouyang et al. (2022b), it does not come without a caveat: training the model to maintain a particular voice appears to limit the model to *only* writing with that voice. GPT-3.5 and GPT-4 lose the ability to write from multiple perspectives, a key characteristic contributing to our use of language models thus far. That aligned models suffer from particular deteriorations is known—OpenAI dubbed it the “alignment tax.” But how this tax impacts uses of the models in creative tasks has not, to our knowledge, been explored. “Dramatis Personae” investigates this tax.

OpenAI did not apply reinforcement learning to all of their models in equal measure. GPT-3 was originally released without reinforcement learning at all, instead being another model released with having only been pre-trained on the Internet as GPT-2 was (Brown et al., 2020a). It would take OpenAI a year before releasing a research model demonstrating a minimum amount of reinforcement learning, dubbed “InstructGPT” by the authors (Ouyang et al., 2022b). This model was aligned on synthetic data produced via the selection of salient training examples via heuristics including complete sentences and reasonably complex grammatical structure. Thus, reinforcement learning for the first model to undergo the process was incetivized to produce coherent and cogent text not necessarily written from any one authorial perspective. The model underlying ChatGPT, GPT-3.5, did not undergo the same reinforcement process (OpenAI, 2023a). GPT-3.5 was optimized with specific texts prepared by a team of real humans. Training a language model, whether it be by gradient descent or proximal policy optimization, will result in the language model modelling the language it is given. And as “Dramatis Personae” finds, GPT-3.5 and GPT-4 have specific voices.

This limits the viability of said models in contexts where the model might be used for simulating creative text.

Language models hold promise for the humanities just as they do for many other domains. Their ability to produce intelligible and coherent text belies an understanding of the world fit for conjectural experimentation. The three manuscripts in this thesis explore this possibility space from three different angles, but all three did face similar problems. First, language models are computationally heavy. While GPT-2 at a mere 1.5 billion parameters may seem paltry in comparison to the 540 billion parameters of PALM (Chowdhery et al., 2022), the 175 billion parameters of GPT-3 (Brown et al., 2020a), and even the 70 billion parameters of Llama 2 (Touvron et al.) today; GPT-2 could not be trained on consumer graphical processing units for a few years—although optimized training procedures like LoRA have made training large language models with consumer hardware more recently possible (Dettmers et al., 2023). Training GPT-2 for "The COVID That Wasn't" and "Blind Judgement" was made possible with Google Colaboratory and many days of compute time. To train a language model on data is to observe how that data shifts the model—but training is often intractable for unprivileged researchers.

A second major issue encountered during research involves the relatively scant resources released by OpenAI. No paper was released together with GPT-3.5, and only a so-called "model card" accompanies GPT-4 (OpenAI, 2023a). The model card describes what the model scored well at during their testing, meaning specific details regarding its implementation and training are not publicly available. How many parameters does GPT-3.5 have? GPT-4? Can you access the data it was trained on? While OpenAI keeping the details opaque is sensible from a business perspective, it does complicate the research process, especially when one cannot point to a definitive source with which to cite. Many experiments rely on tuning parameters to assess causal outcomes—one example being adjusting the parameter count of the model to verify whether the model improves performance in some regard when overparameterized. This can be achieved with GPT-2 and GPT-3, but not with the newer models. Moreover, OpenAI has begun to deprecate older models. Working with an unstable platform does not help reproducible research, rendering even the research presented in this thesis somewhat unreliable. Future researchers will want to reproduce the results found here with open-source models like Llama-2 (Touvron et al.).

The manuscripts presented in this thesis are tentative, motioning towards a simulative future in the humanities. Taken as a whole, they are intended to stress the new presence of tools suitable for aiding empirical research in domains previously majority qualitative in their analysis. Some academics have now taken to examining the role large language models could play in social science research. What would a similar enterprise look like in the humanities?

Conclusion

In 1951 mathematician Claude Shannon published the first paper to describe the language model in the Bell System Technical Journal (Shannon, 1951). One year prior, in 1950, Isaac Asimov of *I, Robot* fame released the first book in what would become the *Foundations* series of science-fiction novels (Asimov and Asimov, 2021).

While the *Foundation* series would continue on for many years, the core question considered through the story would never change: what would happen if large-scale human behaviour could be predicted? What would it mean if culture could be mapped and modelled probabilistically? Asimov explored the idea of quantitatively capturing culture through the in-universe theory of *psychohistory*, coined to mean "that branch of mathematics which deals with the reactions of human conglomerates to fixed social and economic stimuli," (Asimov and Asimov, 2021). His characters live with determinism a known reality; their individual concerns a trifle when compared with the slow-moving and well-charted future movements of entire societies. Psychohistory is a simulative theory. The fictional Hari Seldon develops it to understand where society will be delivered by extrapolating current trends. The result is apparent foresight, a knowledge of what is to come.

In 1956 Claude Shannon and early computer scientist John McCarthy, of LISP fame, convened a number of leading researchers at Dartmouth College to workshop a new idea: artificial intelligence (McCarthy). What was artificial intelligence? The proposal for the workshop coined artificial intelligence to refer to "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." With what means would this intelligence be simulated? Neural networks and language models, technologies created and promoted by the workshop attendees. A third subject of study introduced for discussion at the Dartmouth Summer Research Project was randomness; a now-known psycholinguistic property used in modulating language model creativity implemented via the softmax function when projecting logits to a vocabulary; otherwise known as "temperature" (Roemmele and Gordon, 2018). It would take some sixty years for their respective areas of study to ring true. At the present time, large language models are perhaps our best attempt at general artificial intelligence.

Large language models do not only simulate text. Decoding their probability distributions through greedy sampling and beam search is an implicit exercise in conjecturing successive sequences of text: *what if* someone were to add two and two together? Four, says the language model making use of learned representations of logic and mathematics. *What if* one were to respond to someone describing the antics of their dear pet at length? Perhaps doting words, corresponding descriptions of one's own pet, or a yawn. And *what if* the first person were to respond to that response? And so on. Language models have no ground truth, no basis for reality—all samples produced from a language model, hallucinations included, reveal their understandings of the world learned from textual sources (Manning, 2022). Theirs is a world that is necessarily conjectural and counterfactual no matter how coherent and well-fitting to reality their output may seem. Language models simulate text. Large language models simulate the world.

Depending on who you ask, the humanities might have a problem with there being no equivalent of the scientific method (Northrop, 1947; Van Peer et al., 2012). Those researching in the humanities entertains continual change in dominant modes of thinking. Whole schools of interpretation and understanding incompatible with one another come and go. Many of these schools are predicated on modes of knowledge discovery making little use of hypothesis and experimentation. Undergraduate students now entering literary studies are routinely taught to write essays with perhaps three citations per paragraph (the so-called "hamburger essay") is to give them the impression knowledge can be effectively produced without *testing* that knowledge, whether it be through the discovery of trends laying latent in many documents or entertaining counterfactual scenarios in which their argument might not hold water. Traditional humanities research is not rigorous.

The aim of this thesis was to take a tentative look at how simulative methods conducted via large language models might shift research in the humanities. While the pursuits and findings of each paper were each different, their combined purpose was to motion at what kinds of questions might be asked were one able to efficiently simulate possible answers with text-generating algorithms. Ideally, such questions are novel—their being dependent on simulation and wide-scale text analysis being available. While the specific techniques and experiments demonstrated in this thesis are not likely to benefit all departments and fields of study, it is arguable many can benefit from the introduction of quantitative and simulative reasoning to their prior collection of methods and methodologies. We now have language models empirically capable of passing literacy tests. It would be a costly mistake not to ponder at how those in the humanities can benefit from this.

Research on large language models is restless. Only five years have passed since the unveiling of GPT-2, and one since the release of ChatGPT (Radford et al., 2018a). One cannot predict what these technologies will be capable of in one year, let alone ten. But we can begin to use them today, and best practices discovered now will in all likelihood inform future practices. Research conducted in the digital humanities over the past twenty years points to what is possible: cultural analytics, computational narratology, archiving, web portals, and so on, all point to what is possible when quantitative techniques are embraced. Simulation has found purchase in the social sciences (Gilbert, 1996), and so this thesis suggests the same can be true for the humanities. But the first step must be taken. Many express fear large language models will negatively impact those in creative industries, and one can imagine scenarios where something similar might happen to the humanities. Perhaps. But it is also true the last ten years have seen the humanities suffer from decreases in funding, incoming students, and faculty positions. It is not clear how these problems can be alleviated, but one can suggest change should come from within: admitting new techniques can admit new researchers with new perspectives. And if not large language models, what then?

References

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2022. Gpt-3-driven pedagogical agents for training children's curious question-asking skills. (arXiv:2211.14228). ArXiv:2211.14228 [cs].

Seth Allcorn and Howard F Stein. 2021. Unpacking the supreme court: The age of trump, law, and psychohistory. *Journal of Psychohistory*, 49(1).

Stephen E Ambrose. 2014. *Eisenhower Volume II: The President*. Simon and Schuster.

Anthropic. Model card and evaluations for claude models.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, page 1–15.

Isaac Asimov and Isaac Asimov. 2021. *Foundation*, del rey trade paperback edition edition. Foundation series. Del Rey, New York.

Roland Barthes. Death of the author.

Jean Baudrillard. 1994. *Simulacra and simulation*. The Body, in theory. University of Michigan Press, Ann Arbor.

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. On the dangers of stochastic parrots: Can language models be too big?

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the dangers of stochastic parrots: Can language models be too big?

Gregor Betz. 2022. Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics. *Journal of Artificial Societies and Social Simulation*, 25(1):2. ArXiv:2104.06737 [cs].

Douglas Biber. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. John Benjamins Publishing. Google-Books-ID: t4CcpizGwgQC.

Jen Birks. 2021. Just following the science: Fact-checking journalism and the government's lockdown argumentation. In *Power, Media and the Covid-19 Pandemic*, pages 139–158. Routledge.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, and Michael Pieler. 2022. Gpt-neox-20b: An open-source autoregressive language model.

William D Blake. 2019. 'don't confuse me with the facts': The use and misuse of social science on the united states supreme court. *Md. L. Rev.*, 79:216.

Svenja Boberg, Thorsten Quandt, Tim Schatto-Eckrodt, and Lena Frischlich. 2020. Pandemic populism: Facebook pages of alternative news media and the Corona crisis – a computational content analysis. Technical Report arXiv:2004.02566, arXiv. ArXiv:2004.02566 [cs] type: article.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan

Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models. (arXiv:2108.07258). ArXiv:2108.07258 [cs].

George EP Box and Norman R Draper. 1919. Essentially, all models are wrong, but some are useful. *Statistician*, 3(28):2013.

Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. (arXiv:2005.14165). ArXiv:2005.14165 [cs].

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Eric Castellanos, Hang Xie, and Paul Brenner. 2021. Global news sentiment analysis. In *Proceedings of the 2019 International Conference of The Computational Social Science Society of the Americas*, pages 121–139, Cham. Springer International Publishing.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. (arXiv:2204.02311). ArXiv:2204.02311 [cs].

Jack Clark. 2022. GPT-2 domains. Original-date: 2019-02-11T04:21:59Z.

James W Cortada and William Aspray. 2019. *Fake news nation: the long history of lies and misinterpretations in America*. Rowman & Littlefield.

Jonathan D. Culler. 2000. *Literary theory: a very short introduction*, reissued edition. Very short introductions. Oxford University Press, Oxford.

Arianna Dagnino. 2012. Comparative literary studies in the twenty-first century: towards a transcultural perspective? *Cultural ReOrientations and Comparative Colonialities*, page 1.

Mehdi Dastani and Bas Testerink. 2014. From multi-agent programming to object oriented design patterns. In *International Workshop on Engineering Multi-Agent Systems*, pages 204–226. Springer.

Ithiel de Sola Pool and Robert Abelson. 1961. pool1961.pdf. *Public Opinion Quarterly*, 25(2):167–183.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, et al. 2022. Neural networks and the chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.

Jacques Derrida. 2020. *Deconstruction in a Nutshell: A Conversation with Jacques Derrida, With a New Introduction*. Fordham University Press. Google-Books-ID: fOD5DwAAQBAJ.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. (arXiv:2305.14314). ArXiv:2305.14314 [cs].

Oxford English Dictionary. 2023. bias, n., adj., adv.

Thomas G Dietterich and Eun Bae Kong. 1995. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. (arXiv:2107.01294). ArXiv:2107.01294 [cs].

A. F. Draper. 1953. Univac on election night. *Electrical Engineering*, 72(4):291–293.

Iddo Drori, Sunny Tran, Roman Wang, Newman Cheng, Kevin Liu, Leonard Tang, Elizabeth Ke, Nikhil Singh, Taylor L. Patti, Jayson Lynch, Avi Shporer, Nakul Verma, Eugene Wu, and Gilbert Strang. 2021. A neural network solves and generates mathematics problems by program synthesis: Calculus, differential equations, linear algebra, and more. *CoRR*.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. 2022. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32):e2123433119.

Carl H Esbeck. 2011. Uses and abuses of textualism and originalism in establishment clause interpretation. *Utah L. Rev.*, page 489.

John Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, pages 10–32.

Stanley Fish. 1995. *Is There a Text in This Class?* Routledge.

Martin Gardner. 1970. [link].

Ruth Garland and Darren Lilleker. 2021. From consensus to dissensus: The UK's management of a pandemic in a divided nation. In *Political communication in the time of coronavirus*, pages 17–32. Routledge.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences*, 11(7):3184.

G. Nigel Gilbert. 1996. Simulation as a research strategy. In *Social science microsimulation*, page 448–454. Springer.

Nigel Gilbert and Klaus Troitzsch. 2005a. *Simulation for the Social Scientist*. McGraw-Hill Education (UK). Citation Key: gilbert2005simulation.

Nigel Gilbert and Klaus Troitzsch. 2005b. *Simulation for the social scientist*. McGraw-Hill Education (UK).

Avner Greif. 1994. Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of political economy*, 102(5):912–950.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Sil Hamilton. 2023. Blind judgement: Agent-based supreme court modelling with gpt. (arXiv:2301.05327). ArXiv:2301.05327 [cs].

Sil Hamilton and Andrew Piper. 2022. The COVID That Wasn't: Counterfactual Journalism Using GPT. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 83–93, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ryan Han. 2021. COVID-19 News articles open research dataset.

David Herman. 1997. Scripts, sequences, and stories: Elements of a postclassical narratology. *PMLA*, 112(5):1046–1059.

David Herman. 2005. *Quantitative Methods in Narratology:*, volume 6 of *Narratologia*, page 125–150. Walter de Gruyter, Berlin, New York.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*.

Kate Holland and Monique Lewis. 2021. Mapping national news reports on COVID-19 in Australia: Topics, sources, and imagined audiences. In Monique Lewis, Eliza Govender, and Kate Holland, editors, *Communicating COVID-19*, pages 59–81. Springer International Publishing, Cham.

HuggingFace. 2022. Models - Hugging Face.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 1–15, Hamburg Germany. ACM.

Richard L. Pacelle Jr, Brett W. Curry, and Bryan W. Marshall. 2011. *Decision Making by the Modern Supreme Court*. Cambridge University Press. Google-Books-ID: SnVP2trSfcIC.

Edward Kahn. 1973. Finite-state models of plot complexity. *Poetics*, 3(1):5–20.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. (arXiv:2001.08361). ArXiv:2001.08361 [cs, stat].

Daniel Martin Katz, Michael J. Bommarito, II, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12(4):1–18.

Jerrold J Katz. 1980. Chomsky on meaning. *Language*, pages 1–41.

Lae-Seon Kim, Seong-soon Kim, Heon-Seok Jang, Seok-Won Park, and In-Ho Kang. 2020. Long-tail query expansion using extractive and generative methods. In *Annual Conference on Human and Language Technology*, pages 267–273. Human and Language Technology.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Christopher D Kromphardt. 2017. Evaluating the effect of law clerk gender on voting at the united states supreme court. *Justice System Journal*, 38(2):183–201.

Annu Lambora, Kunal Gupta, and Kriti Chopra. 2019. Genetic algorithm-a literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 380–384. IEEE.

Darren Lilleker, Ioana A Coman, Miloš Gregor, and Edoardo Novelli. 2021. Political communication and COVID-19: Governance and rhetoric in global comparative perspective. In *Political Communication and COVID-19*, pages 333–350. Routledge.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *CoRR*.

Bárbara Lopes, Catherine Bortolon, and Rusi Jaspal. 2020. Paranoia, hallucinations and compulsive buying during the early phase of the COVID-19 outbreak in the United Kingdom: A preliminary experimental study. *Psychiatry Research*, 293:113455.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, page 48–55, Virtual. Association for Computational Linguistics.

Claude Lévi-Strauss. 2013. *Myth and Meaning*. Routledge, London.

Christopher D. Manning. 2022. Human language understanding reasoning. *Daedalus*, 151(2):127–138.

Lev Manovich. 2016a. The science of culture? social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*, nil(nil):nil.

Lev Manovich. 2016b. The science of culture? social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*.

John McCarthy. A proposal for the dartmouth summer research project on artificial intelligence.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *CoRR*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, Duygu Altinok, György Orosz, Søren Lind Kristiansen, Daniël De Kok, Lj Miranda, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Edward, Wannaphong Phatthiyaphaibun, Richard Hudson, Yohei Tamura, Sam Bozek, Murat, Ryn Daniels, Peter Baumgartner, Mark Amery, and Björn Böing. 2022. explosion/spaCy: New span ruler component, JSON (de)serialization of doc, span analyzer and more.

Benjamin Mullin and Nico Grant. 2023. Google tests a.i. tool that is able to write news articles. *The New York Times*.

Satoshi Nishida, Antoine Blanc, Naoya Maeda, Masataka Kado, and Shinji Nishimoto. 2021. Behavioral correlates of cortical semantic representations modeled by word vectors. *PLoS computational biology*, 17(6):e1009138.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.

Filmer Stuart Cuckow Northrop. 1947. The logic of the sciences and the humanities.

Takashi Ogata and Taisuke Akimoto. 2016. *Computational and cognitive approaches to narratology*. IGI Global.

Jude Nwakpoke Ogbodo, Emmanuel Chike Onwe, Joseph Chukwu, Chinedu Jude Nwasum, Ekwutosi Sanita Nwakpu, Simon Ugochukwu Nwankwo, Samuel Nwamini, Stephen Elem, and Nelson Iroabuchi Ogbaeja. 2020. Communicating health crisis: a content analysis of global media framing of COVID-19. *Health Promotion Perspectives*, 10(3):257–269.

OpenAI. Aligning language models to follow instructions.

OpenAI. 2019. Gpt-2: 1.5b release.

OpenAI. 2023a. Gpt-4 technical report. (arXiv:2303.08774). ArXiv:2303.08774 [cs].

OpenAI. 2023b. Openai platform.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, and Chong Zhang. 2022a. Training language models to follow instructions with human feedback.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. (arXiv:2203.02155). ArXiv:2203.02155 [cs].

Gerard O'Regan and Gerard O'Regan. 2018. Hollerith's tabulating machines and the birth of ibm. *The Innovation in Computing Companion: A Compendium of Select, Pivotal Inventions*, pages 151–153.

Behrooz Parhami. 1995. Simd machines: do they have a significant future? *ACM SIGARCH Computer Architecture News*, 23(4):19–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022a. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, page 1–18, New York, NY, USA. Association for Computing Machinery.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022b. Social simulacra: Creating populated prototypes for social computing systems.

Jordan Carr Peterson, Thora Giallouri, and Elli Menounou. 2021. The personal finances of united states supreme court justices and decision-making in economic litigation. *The Journal of Legal Studies*, 50(2):379–405.

Beatrice Portelli, Simone Scaboro, Enrico Santus, Hooman Sedghamiz, Emmanuele Chersoni, and Giuseppe Serra. 2022. Generalizing over long tail concepts for medical term normalization. *arXiv preprint arXiv:2210.11947*.

Robert Post and Reva Siegel. 2006. Originalism as a political practice: The right's living constitution. *Fordham L. Rev.*, 75:545.

Vladimir IAkovlevich Propp. 1968. *Morphology of the Folktale: Second Edition*. University of Texas Press. Google-Books-ID: 3Md3u9UPgOEC.

Thorsten Quandt, Svenja Boberg, Tim Schatto-Eckrodt, and Lena Frischlich. 2021. Stooges of the system or holistic observers?: A computational analysis of news media's facebook posts on political actors during the coronavirus crisis in Germany. In *Political Communication in the Time of Coronavirus*, pages 101–119. Routledge.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018a. Language models are unsupervised multitask learners. page 24. Citation Key: radford$_language_nodate$.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Paul Rayson. 2012. *Corpus Analysis of Key Words*. John Wiley & Sons, Ltd.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, and Aitor Lewkowycz. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Melissa Roemmele and Andrew S. Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, page 1–2, Tokyo Japan. ACM.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey R Sampson. 1976. Adaptation in natural and artificial systems (john h. holland).

Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security*, pages 338–343.

C. E. Shannon. 1951. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64.

Adam Shehata, Isabella Glogger, and Kim Andersen. 2021. The Swedish way: How ideology and media use influenced the formation, maintenance and change of beliefs about the coronavirus. In *Political Communication in the Time of Coronavirus*. Taylor & Francis.

Helen Small. 2013. *The value of the humanities*. Academic.

Sheldon Solomon, Daniele Rostellato, Ines Testoni, Fiorella Calabrese, and Guido Biasco. 2021. Journalistic denial of death during the very first traumatic period of the Italian COVID-19 pandemic. *Behavioral Sciences*, 11(3):41.

Harold Spaeth, Lee Epstein, Ted Ruger, Keith Whittington, Jeffrey Segal, and Andrew D Martin. 2014. 2020 supreme court database. *URL: http://scdb. wustl. edu*.

Flaminio Squazzoni, Wander Jager, and Bruce Edmonds. 2014. Social simulation in the social sciences: A brief overview. *Social Science Computer Review*, 32(3):279–294.

Jordan W Suchow, David D Bourgin, and Thomas L Griffiths. 2017. Evolution in mind: Evolutionary dynamics, cognitive processes, and bayesian inference. *Trends in cognitive sciences*, 21(7):522–530.

Tzvetan Todorov and Arnold Weinstein. 1969. Structural analysis of narrative. *NOVEL: A Forum on Fiction*, 3(1):70–76.

Hugo Touvron, Louis Martin, and Kevin Stone. Llama 2: Open foundation and fine-tuned chat models.

Chen Feng Tsai, Xiaochen Zhou, Sierra S. Liu, Jing Li, Mo Yu, and Hongyuan Mei. 2023. Can large language models play text games well? current state-of-the-art and open questions. (arXiv:2304.02868). ArXiv:2304.02868 [cs].

Alberto Ueda, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2021. Structured fine-tuning of contextual embeddings for effective biomedical retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2031–2035, Virtual Event Canada. ACM.

Ted Underwood. 2018. Why literary time is measured in minutes. *ELH*, 85(2):341–365.

Peter Van Aelst. 2021. COVID-19 as an ideal case for a rally-around-the-flag?: How government communication, media coverage and a polarized public sphere determine leadership approvals in times of crisis. In *Political Communication in the Time of Coronavirus*, pages 1–13. Routledge.

Willie Van Peer, Frank Hakemulder, and Sonia Zyngier. 2012. *Scientific methods for the humanities*, volume 13. John Benjamins Publishing.

Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *The 16th International Conference on the Foundations of Digital Games (FDG) 2021*, pages 1–8, Montreal QC Canada. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Andreas S Weigend. 2018. *Time series prediction: forecasting the future and understanding the past*. Routledge.

Wikipedia. 2023. [link].

Eric Winsberg. 2010. Science in the age of computer simulation. In *Science in the Age of Computer Simulation*. University of Chicago Press.

Ludwig Wittgenstein, G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. 2009. *Philosophische Untersuchungen =: Philosophical investigations*, rev. 4th ed edition. Wiley-Blackwell, Chichester, West Sussex, U.K.; Malden, MA.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, San Diego CA USA. ACM.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Erfei Zhao, Qiao Wu, Eileen M Crimmins, and Jennifer A Ailshire. 2020. Media trust and infection mitigating behaviours during the COVID-19 pandemic in the USA. *BMJ Global Health*, 5(10):e003323.