Language Models as Opinion Models:
Techniques and Applications

by

William Brannon

M.S. Massachusetts Institute of Technology (2020)
B.S. College of William & Mary (2011)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2025

Authored by:	William Brannon Program in Media Arts and Sciences May 16, 2025
Certified by:	Deb K. Roy Professor of Media Arts and Sciences, Thesis Supervisor
Accepted by:	Tod Machover Muriel R. Cooper Professor of Music and Media Academic Head, Program in Media Arts and Sciences

Language Models as Opinion Models: Techniques and Applications

by

William Brannon

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 16, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Real-time social media platforms now host the news cycle and shape public opinion, while large language models (LLMs) give us new tools to observe and predict those shifts. This dissertation links the new affordances of social media with the predictive power of LLMs to explain—and forecast—opinion change. We first quantify the dynamics of news on an influential social platform, then develop LLM-based tools to forecast persuasion and predict heterogeneous treatment effects (HTEs).

Study I — Media tempo and tone. Using 518,000 hours of U.S. talk-radio broadcasts and 26.6 million tweets from elite and mass users, we show that Twitter discourse (i) moves faster at both take-off and fade-out stages of a news event and (ii) sustains greater outrage than radio – despite radio's often explicitly outrage-focused business model. To our knowledge, this is the first large-scale, data-driven comparison between Twitter and traditional media of both outrage levels and the rate of decay of attention to news.

Study II — Zero-shot persuasion forecasting. Across a diverse set of 28 randomized experiments, LLM-based methods outperform an ensemble of strong baselines at predicting HTEs and deliver good performance at predicting average treatment effects (ATEs) — all without any experiment-specific fine-tuning.

Study III — Transfer and scaling. Fine-tuning LLMs on contemporaneous news coverage boosts HTE (and ATE) prediction performance greatly, to more than 3x baseline performance. A new minibatch-moment-matching (M3) objective lets us train a 400M-parameter model to nearly match the HTE prediction performance of an 8B model at a fraction of the inference cost. Transfer, however, falters out of distribution on held-out experiments and demographic groups, lending support to contextual theories of persuasion.

Overall, we (i) quantify how platform affordances shape the tone and tempo of public discourse, (ii) introduce LLM-based methods that make causal experiments more sample-efficient, and (iii) chart the limits of transfer learning for opinion prediction. Our findings provide practical tools for HTE prediction and help researchers anticipate persuasion dynamics in a media landscape shaped by both humans and machines.

Thesis supervisor: Deb K. Roy
Title: Professor of Media Arts and Sciences

DISSERTATION COMMITTEE

Thesis Supervisor

Deb K. Roy
Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Thesis Readers

Jacob Andreas
Associate Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

John Horton
Associate Professor of Information Technologies
Massachusetts Institute of Technology

Acknowledgments

There are many people to thank for this thesis and the learning along the way, and I'm grateful for all of them. There isn't space enough to do justice to anyone listed here.

I am first of all indebted to my advisor Deb Roy, whose advice and support over the years have been invaluable. His keen eye for important problems is an example I hope to emulate. I'm also grateful for the last several years of good advice, freedom to explore and help through various false starts.

The other committee members also amply deserve a shout-out: I've been very lucky to get to work with Jacob Andreas and John Horton, both excellent researchers and better people. I've learned a lot from the way they approach problems and support students in working through them. Steve Ansolabehere, on my general-exam committee, was also a dependable source of good insights and helpful feedback: an all-around pleasure to work with. Nor would this thesis be complete without thanking many other research mentors over the years. Getting to work with Brian Thompson at Amazon was a great and very educational time, and thanks also to Yogesh Virkar and Marcello Federico. Jad Kabbara has been a delight to work with, even if more of the work than either of us would like took place at 2am before a conference deadline. I also very much appreciate Aaron Strauss, Kevin Collins, Andrew Therriault and Mike Tierney for showing me what research could be and helping me decide to come back to school and do some.

I'm grateful for a number of other collaborators on many projects, including in no particular order: Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, Elinor Poole-Dayan, Doug Beeferman, Shrestha Mohanty, Brandon Roy, Ariel Lee, Vivek Sharma, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, Tobin South, Niklas Muennighoff, Cassandra Overney, Suyash Fulay, Hang Jiang, Andrew Heyward and Luke Hewitt.

Many many thanks to the other students and staff of CCC or (as I first got to know it) LSM. The lab wouldn't function without the one and only Heather Pierce, and I wouldn't be functional without the many other great colleagues and friends I've met there: David McClure, Brandon Roy, Juliana Nazaré, Belen Saldias, Doug Beeferman, Sarah Ballinger, Shayne O'Brien, Maggie Hughes, Eric Pennington, Hang Jiang, Keyla Blanco-Gomez, Suyash Fulay, Hope Schroeder, Wonjune Kang, Martin Saveski, Nabeel Gillani, Eric Chu, Andrew Heyward, Wes Chow, Leila Carter, Elinor Poole-Dayan, Dimitra Dimitrakopoulou, Bridgit Mendler, Kathy Cramer, Alex Siegenfeld, Dennis Jen, Eugene Yi, Naana Obeng-Marnu, Zaria Jalan, Shayne Longpre, Katie Morell, Nazmus Saquib, Danny Kessler, Michael Wong, Cassie Lee, Jad Kabbara, Russell Stevens, Cassandra Overney, Soroush Vosoughi, Anneli Woolf, and Shrestha Mohanty.

My time at the Media Lab has also been blessed with friends here, near and far: Salil Singhal, Andrew Tran, Laura Visochek, Lena Tom, Andrew Therriault, Kimy Lecamwasam, Alessandra Davy-Falconi, Samantha Hicks, David McClure, Rob Lewis, Samantha Gutierrez Arango, Lizzie Ellis, Andy Callahan, Amelia Showalter, Evan Zasoski, Vanessa Lloyd, Stephanie Watson, Kellie O'Malley and Bobby Courtney.

Finally, I'm eternally grateful for my family: my wife Laura, my parents and my siblings. Without them, I wouldn't have gotten anywhere.

Biography

William Brannon is a PhD candidate at the MIT Center for Constructive Communication and MIT Media Lab, and a former intern at the AWS AI Labs. His research focuses on complex social systems and how to build machine learning models able to represent and understand them. He has authored a number of journal and conference publications, and won competitive research grants in AI. Before coming to MIT, Brannon worked for several years as a data scientist for a variety of political groups and campaigns. He previously earned an MS in Media Arts and Sciences from MIT, and a BS in mathematics from the College of William and Mary.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We see things not as they are, but as we are.

Anonymous

Public opinion reflects many influences, from both mass media and interpersonal networks. Media effects on opinion and behavior are real and important, showing up in matters from voting decisions[1,2] to political opinions,[2,3] and even in life-or-death questions of medicine and public health.[4] But public opinion is not simply mechanically determined by the media; "personal influence" matters as well,[5] as do real-world conditions like the economy[6,7] and the overall structure of society.[8] All of these inputs converge in the minds of real people who have to form opinions on issues and take decisions, and there has been great debate over the years about how they do that.[9–11] But regardless of how this opinion formation happens, it highlights the central role of opinion *change* in the study of public opinion. Rather than mapping a fixed set of views people hold, we want to understand the process that leads these views to change.

No understanding of this process in the modern era could be complete without a prominent role for the internet, the rise of which has had a great effect on these opinion dynamics. The media ecosystem has been transformed in important ways: It is more participatory,[12] faster,[13] and more conflictual,[14] with effects on offline phenomena like protest[15] and voting behavior.[1] Because media coverage influences public opinion, these new media with their new affordances[12] can change opinion by changing coverage. Public views themselves have also changed, in particular becoming more polarized[16] and hostile to outgroups[17] in ways often linked to social media.[18] The measurement of public opinion has been affected, too, with collapsing phone response rates[19] leading to the rise of online polls[20] and general ferment in survey research.[21–23] Existing literature also argues for effects on the process of persuasion or opinion change itself, via mechanisms like "group polarization",[24] "filter bubbles",[25] and digital technology substituting for face-to-face interaction.[26]

In parallel with these social changes, the internet has also led to a huge increase in the capabilities of machine learning models, especially language models (LMs).[27] Along with greater computing power and innovation in architectures, a necessary condition for these breakthroughs was the increase in the scale of training data that web corpora allow.[28,29] (Witness the now ubiquitous phrase "large language model," or

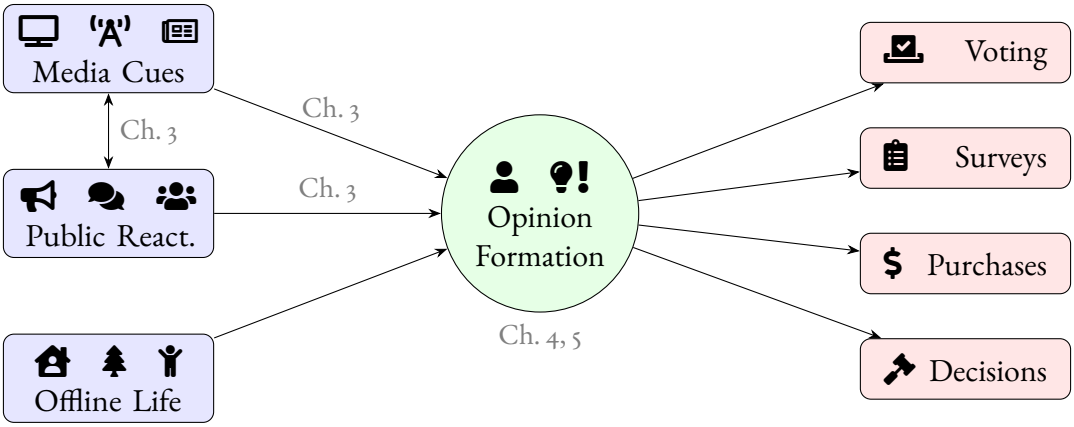


Figure 1.1: A stylized DAG depicting the opinion formation process in which our work is set. People's opinions are profoundly contextual: they result from an opinion-eliciting situation and a person, who is subject to many influences including the dynamics and feedback loops of the news cycle.

LLM.) The enormous scale of these corpora, and the resulting expense of training on them, have also led to a new paradigm for machine-learning tasks: "pre-training" to produce a so-called "foundation model,"[30] followed by "fine-tuning" that model on a specific task or even use without further training, in which case normally test-time tasks are expressed as text whose completion should be generated.[31] Models trained and used this way have demonstrated impressive new capabilities,[32,33] among them greater ability to model survey respondents[34] and other agents[35,36] and perhaps improve social science generally.[37,38]

This dissertation ties together these threads, leveraging the new ML methods both to investigate changing media dynamics and to model the opinion formation processes downstream of them. Our points of departure in the process of opinion formation are illustrated in Figure 1.1, which takes a view of persuasion inspired by that of Zaller [39] and his emphasis on external messages and cues. We are curious, in other words, both about how those cues and messages are generated and about how they are received and used.

The following chapters address two areas in particular: In Chapter 3, we analyze and compare the dynamics and sentiment of the news cycle on social and broadcast media, specifically Twitter (the platform renamed X) and talk radio; and in Chapter 4 and Chapter 5 we develop ways to use LLMs to predict opinion change, especially in the context of randomized experiments, and to understand the mechanisms of opinion change.

Our work on the news cycle is discussed first in the dissertation, and in a sense is prior to a study of persuasion in any case. As existing literature emphasizes, people rarely change their opinions spontaneously, but rather in response to stimuli and especially stimuli delivered through the media.[10,39–41] It's thus important to ask about what those stimuli are and how they're determined. An important theoretical perspective for us here comes from McLuhan [42] in his famous aphorism that "the medium is the message." In other words, the affordances of a media technology shape its content. Our point of departure is to ask how the enormous internet-driven changes in media technology of the last few decades, especially the rise of social media, have changed the information those media deliver.

The natural way for us to approach this question is to compare social and traditional media directly, which motivates the comparative analysis of Twitter and talk radio in Chapter 3. We compare only broad properties of these media, like news-cycle speed and sentiment, for theoretical reasons: Affordances promote predispositions and worldviews, not compel any particular opinion on any particular topic. The point McLuhan and his intellectual inheritors made[43] was not (for example) that TV makes one support or oppose higher defense spending, but rather that it promotes a view of the entire subject as entertainment. We are interested in examining similarly broad, fundamental questions about social and traditional media: How fast and how outraged are they?

Even (perhaps especially) at this broad level, the new machine-learning tools are very helpful to our analysis. Twitter and radio are both very large enterprises producing large amounts of content, far too much to code by hand or subject to a systematic qualitative analysis. Language models and techniques that use them, as discussed in Section 3.4, are instrumental in doing a comprehensive job and drawing broadly informative, relevant conclusions.

And indeed, understanding these properties of media is helpful to a variety of different groups: regulators and policymakers who need to figure out how communication platforms work, designers of new platforms (or executives at existing ones) who want to avoid pitfalls, and academics studying media, for whom this kind of research can highlight a method usable for future studies (as of new platforms). We believe it points to the need for more attention to a vital area of research: an empirical science of platform design, connecting affordances to social dynamics. Because of the rapid change in this space, with Twitter already quite different from our data collection period and competitors rising quickly, the need for this kind of work will not abate anytime soon. As academics like to say, more research is needed.

On the other side of the opinion-change dynamic, meanwhile, in the study of people's response to messages, the situation is different. Here there is a great deal of research: The science of persuasion has been worked on extensively in political science,[39,44] psychology,[9,45] and other disciplines.[46,47] The marketing industry and political communications have also accumulated a great deal of practical knowledge of what works. Nevertheless opinion change is not well understood: There are varying theoretical perspectives which disagree fundamentally, from the view that persuasion mostly occurs "in parallel,"[48,49] similarly across groups, to other empirical findings that the process of opinion change is complex, highly context-dependent and difficult to theorize or predict.[50]

The question motivating our work in Chapter 4 and Chapter 5 is whether large pretrained models can help remedy this problem. They are surprisingly good proxies for people in many contexts[34,35] — perhaps they can serve in that role here as well. We frame the problem concretely as that of predicting heterogeneous treatment effects (HTEs) in persuasion experiments, because of the powerful identification guarantees[*] provided by the randomization; many otherwise vexing confounders can be removed. Similarly, using LLMs and the transfer-learning paradigm may let us explore *how* persuasion happens, not just whose opinions change. Fine-tuning or adapting models on different datasets can test generalization: Is there transfer between experiments, between kinds of people, between context (like news) and persuasive acts? Affirmative answers to any of these questions would indicate predictable common mechanisms of opinion

[*] Up to some technical subtleties discussed in Section 4.3.

change.

Such mechanisms and the ability to predict their results have obvious practical applications in marketing, politics and other fields, but they have perhaps even greater scientific value. There would be great benefit in establishing LLMs as usable (if imperfect) proxies for humans in studying opinion change, a phenomenon that's currently difficult and expensive to study. This difficulty of research stems fundamentally from two things: the context-dependence persuasion sometimes exhibits, which can render it difficult to control all the relevant variables, and the unobservability of people's internal states when they change their opinions. With LLMs, however, both of these problems are greatly reduced: A researcher can determine the training data, prompt and other inputs, and observe in extreme detail how the model responds. Both context and subject in in-silico studies are simply more controllable.

Finally, we note that both of the streams of research in this dissertation are dual-purpose: They have practical as well as academic implications. We hope that they can make a contribution to positive change. A better understanding of both platform design's impact on public discourse, and the ways people respond to that discourse by changing their opinions, can advance scientific knowledge. We hope that they are also relevant to policy questions and some of the pressing problems facing society.

1.1 Research Questions

RQ 1 Can we identify the differences in news cycle speed and sentiment, particularly outrage, between Twitter (X) and radio? Are any of these differences attributable to properties of the media themselves?

RQ 2 Can we use pretrained language models to predict opinion change in response to messages in experimental settings? Does performance vary by demographics or other traits of subjects?

RQ 3 Is there transfer learning and meaningful generalization on opinion-change prediction tasks? In particular, is persuasion more predictable with pretraining on 1) contextual data like news, 2) data from other experiments, or 3) data representing different groups of people?

1.2 Contributions in Brief

In the course of addressing these research questions, this dissertation makes certain contributions:

1. Empirical demonstration of faster news cycles online, showing that news events on Twitter have shorter lifecycles compared to talk radio, and helping validate the method for use in future news-cycle studies. These results provide the first large-scale empirical comparison of the decay of attention to news between Twitter and traditional media.

2. The first large-scale comparison of outrage between Twitter and traditional media, finding that Twitter has more outrage and more negativity than talk radio.

3. Developing and validating methods for LLM-based treatment-effect estimation, which we find outperform a set of strong traditional baselines.

4. Showing that fine-tuning on contemporaneous news data greatly improves treatment-effect prediction performance, at least for the types of textual survey experiments we consider here.

5. Proposing a novel training objective for treatment effect estimation with neural networks, which we find improves performance on a diverse group of experimental datasets.

6. Demonstrating a limited but theoretically important degree of generalization between experiments and demographics in predicting persuasion treatment response.

1.3 Thesis Outline

The rest of the dissertation frames and investigates these questions. Succeeding chapters are as follows:

Chapter 2 Background, context and theoretical perspectives motivating and informing our work.

Chapter 3 Analyzing the speed and sentiment of the news cycle across media (Twitter and talk radio).

Chapter 4 Introducing methods to predict HTEs in randomized experiments with pretrained LLMs.

Chapter 5 Examining the deeper architecture of persuasion through experiments with LLMs, focusing on transfer learning and generalization from news corpora and held-out experimental settings.

Chapter 6 Wrapping up: a summary of contributions, limitations and future research directions.

Chapter 2

Background and Theoretical Context

With the rise of the internet, mobile devices and social media, an increasing amount of social life is now conducted online, leading to large effects on areas from mental health to dating to politics. The theoretical view underlying our work in this dissertation is that these effects are due to differences in the *affordances* of the internet and social media: the characteristic properties and resources the technology offers for social interaction. These changes in affordances have both direct effects on individuals and indirect effects through further changes they cause in the content of media. An important example is the internet-enabled disruption of traditional gatekeepers like news organizations, which had and have strong editorial cultures, in favor of less constrained content sources like individual influencers.

In the face of all this change, some things remain constant. Basic human drives, including narrative thinking and group identity, are consistent over time but play out differently online than offline. We discuss some of these basic motivations, the changes in affordances that cause them to manifest differently, and the difficulties digital technology poses for measuring and understanding these effects. There are great difficulties and great opportunities in particular for public opinion research: Much new opinion data is visible, but it is not in conventional survey formats. We believe and argue in particular that machine learning tools are essential to solve these problems, and have great potential when applied to public opinion questions. Many challenges facing society today, including political polarization, may benefit from ML-enabled understanding of social and opinion dynamics. Work in subsequent chapters of the dissertation applies tools from machine learning in an attempt to provide exactly this kind of understanding.

2.1 Introduction

We are concerned here with social behavior online, and particularly the formation, evolution and measurement of public opinion. These areas of society have undergone enormous changes in recent years, with the rise of smartphones and social media leading an increasing number of activities to be conducted online.[51] While many of the effects have been beneficial – increased economic activity and e-commerce,[52] remote work,[53] and even claims of developmental benefits for adolescents[54] – much research has also found negative effects on mental health[55,56] and the functioning of political systems.[57–59]

Our work proceeds from two central theoretical views: that these changes, both helpful and harmful, are due to novel online *affordances*, and that tools from machine learning (ML) are needed to understand and measure the changes and their causes. Affordances are a central notion in media studies and fields like human-computer interaction, where the term refers to features of a technology that allow or encourage some behavior while preventing or discouraging others.[60,61] The Internet and social media have clearly different affordances from prior communications media,[62] ranging from context collapse[12] to an increase in the prominence of weak ties.[63]

The social effects of the new communication technology have been far-reaching, ranging from new modes of protest[15] to changes in voting behavior[1] to patterns of inter-group conflict unknown offline.[14] These effects are often linked to the affordances inherent in the technology.[12,42] While basic human drives, motives and mental features not tied to media do exist, the way that these drives manifest in society depends on the media they are filtered through.

Media and their affordances do not, however, affect only social behavior as an object of study: They also have profound influence on the process and tools of measuring it. The great scale, low friction, and networked, interdependent character[12] of online life pose problems for traditional research methods. The sheer size of modern digital datasets requires use of data-mining techniques,[64] while randomized experiments, the gold standard of clinical research, suffer from inter-unit interference in online settings.[65,66] Even the collapse in survey response rates that has transformed public opinion research since the early 2000s is linked to new technologies. With the displacement of landlines by cell phones, and the technologically enabled surge in spam-call volume,[67] fewer people are willing to answer a call from an unknown number like that of a survey interviewer.

Most of these issues (though perhaps not the decline in response rates) can be mitigated with machine learning techniques. ML is useful and indeed widely used to understand social network phenomena[14,68–70] and has also been applied to public opinion research.[34,71] We believe, however, that this latter area is ripe for further exploration, with many more promising applications than have been pursued. From a certain angle, this dissertation is an extended argument for the benefits of applying machine learning to the study of public opinion.

Finally, it's worth mentioning some of the reasons these questions matter — the societal concerns that motivate interest in them to begin with. While digital affordances have widespread effects, we are especially interested in their impacts on the *public sphere* in Habermas's sense:[72] the realm of public debate and deliberation. Media coverage has great impacts on public discourse,[2,3,39,69] and the recent changes in the technological architecture of the media pose equally great concerns for democracy. Falsehoods on Twitter spread faster than the truth,[73] recommendation algorithms amplify political viewpoints,[74] and polarization is intensified online.[75] Survey research, which is one of the main ways the political system acquires public feedback between elections, is still delivering results but is under increasing methodological strain.[76,77]

The rest of this chapter is organized as follows: We discuss some of the underlying behavioral drives affecting online (and offline) life in Section 2.2, before delving into digital affordances and some of their direct effects in Section 2.3. Section 2.4 discusses the problems these new affordances pose for measurement

and traditional approaches to research, and Section 2.5 covers ways machine learning can provide solutions. Finally, Section 2.6 highlights reasons these questions and research on them matter — their impact on the public sphere — and Section 2.7 wraps it all up.

2.2 Basic Drives

Any examination of the effects of changes in media affordances should start with what has *not* changed: basic human behavioral drives and motives that are consistent across media environments. We'll focus here on only a few areas relevant to public opinion and online social dynamics.

These behavioral drives and tendencies are all about understanding the world. People make sense of events by interpreting them as narratives, they make sense of themselves by identifying as members of groups, and they absorb new information socially, through other people. Finally, people do not integrate all of this information into an internally consistent whole, rather remaining ambivalent between different narratives, identities, and pieces of knowledge. This way of interacting with the world implies (as we'll see in Section 2.3) that media affordances which change the narratives, groups and information that are salient can produce large changes in social behavior.

2.2.1 Narrative

The tour begins with narrative, a fundamental, "familiar and ubiquitous" human mode of understanding the world, indeed so much so as to be "likely to be overlooked."[78] It structures the way people experience their lives and construe those lives' events, providing a ready way to make sense of each new occurrence. Stories in this meaning-making role are also powerful drivers of action,[*] from politics and political organizing[79] to economic behavior,[80] often without anyone consciously deciding that they would be.

Knowledge organized in this way has a number of characteristic features, of which a few are most relevant here. In the schema of Bruner, [78] narratives are temporal (or "diachronic"), they are "particular" as opposed to abstract, "intentional" as opposed to purely causal or materialist, and finally are cumulative in that they are remembered, retold and remixed in culture. Narratives, in other words, are about people and their lives, and the form is a poor fit for information about other things. The natural human inclination to think in narrative makes it particularly difficult to understand objects or events with complicated causal structure.[81] While stories differ in their details and even their broad plot arcs, they all have a human scale in common: duration on the scale of a human life, and indeed one protagonist (or at most a few) with recognizably human motives. But many, indeed most, things in the universe are not like this: Statistical mechanics and stock markets and even elections involve millions of people or objects interacting in a fractally complicated way over very short or very long timescales, with the sort of "emergent behavior" that fits poorly into narratives.[81] When people tell stories about such phenomena, we anthropomorphize them: 'The United States angrily protested the diplomatic slight,' as though 'the United States' were a

[*] The perfect example is Bitcoin: a trillion-dollar asset class memed into existence *ex nihilo* by the power of a good story.

person rather than a huge group of people with a complex internal structure. Formal or mathematical models of behavior are often a better fit, but are not at all intuitive.

This proclivity for narrative highlights the need for the sort of work we do in the rest of this dissertation. Both of the phenomena we study (the news cycle in Chapter 3 and internal psychological processes of opinion change in Chapter 4 and Chapter 5) are just the sort of complex mesoscale phenomena that are difficult to understand as narratives. News events themselves are usually narratives (or at least narrativized by the press), which makes the contrast all the more striking: No particular such narrative sheds light on the complex underpinnings of the process that generates them. Journalism, with its inclination to tell stories, may have difficulty examining itself. Similarly, with persuasion and opinion change, the forms and mechanisms of influence may be subtle: a large number of small background cues,[39] not one large act of influence that can be understood easily in human terms. We want, that is, to illuminate the non-narrative background conditions that influence what kind of narratives people tell and believe.

2.2.2 Group Identity and Social Learning

The narrative form centers human experience because, indeed, people are deeply interested in other people. They are often most interested in *groups* of people: nations, ethnicities, subcultures, and so on. An influential theoretical view is that of Brewer, [82] who described people as having dueling desires to fit in and stand out. Too much separation from other people is isolating and in some cases risky; too little is stultifying and denies individuality. "Optimal distinctiveness," the happy medium, is achieved when these desires balance.[82] Importantly, this view predicts the existence of a hierarchy of groups of varying sizes, with larger groups like nations developing subcultures and smaller ones aggregating together.

Other research, examining social groups in the service of questions not about group identity, has found a similar nested structure of groups to which individuals feel loyalty. The classic study of voting by Berelson *et al.* [83] provides a good example: Voters were initiated into political traditions by their families, which especially meant being inducted into groups. Being a Catholic, a union member, a participant in the upper-crust 'society' set, and most of all in a political setting a Democrat or Republican – all of these and more are identities that voters are deeply connected to. For political purposes, participants in the study often perceived these identity groups to have a hierarchical structure: One was pulled toward being a member of the larger Democratic coalition by majority opinion in one's own group of, for example, local Catholics. The desire to fit in (with one's groups) and to stand out (from the other party) are evident.

The example of voting highlights another way that sociality is important for our purposes: its critical role in learning. Voters in Elmira, the town Berelson et al. studied, acquired most of their political information and opinions from those around them, including their parents (early in life), friends, identity groups, and preferred media sources. A related, pioneering study by Katz & Lazarsfeld [5] delved more deeply into the structure of "personal influence" by one person on the opinions of another and found a complex, richly textured web of social relationships that mediate learning.

Social learning has appeared in other influential work as well. Backstrom *et al.* [84] showed that users' decisions to join online communities depends on how many friends the users already have in those

communities, and on the finer details of social connections among those friends. Milgram *et al.* [40] found in a simple but ingenious experiment that people were more willing to stop and join a crowd of confederates in looking up at a building the larger the initial crowd. Even the line of work in political science finding media influence on public opinion[2,11,39] demonstrates social influence: People's opinions are, after all, being changed by information other people have brought them.

2.2.3 Cross-Pressure and Ambivalence

As this last example may suggest, people's opinions are not usually fixed, carefully thought out and internally consistent. Of particular interest are opinions having to do with public affairs, where this conclusion is counterintuitive: For many years in the political science literature, the reigning assumption was that people have "true attitudes" on issues.[85] The tide began to turn with the work of Converse, [10] who argued for a more complex conception of opinion. For Converse, while a small number of people do have carefully reasoned ideological beliefs, most people think about public affairs in less abstract ways. Some make decisions based on group loyalty, others on a general sense of the "nature of the times," and some not based on issues at all.

Because these less abstract motivations are not systematized and weighed against each other, they are not preemptively reconciled where they may conflict. Most people, then, are *ambivalent*: they have clashing, disconnected opinions that may lead them in different directions depending on which are salient. Other work after Converse's seminal paper picked up this thread and developed it, culminating in the influential model due to Zaller [39] of opinion formation and expression. In this more systematic version of Converse's ideas, people respond to surveys based on a random sample of the opinions they've recently been exposed to, especially in the mass media. Because respondents don't usually (in the model) think carefully about the opinions they hear, salience effects are powerful.

There are also close connections to the group identity and social learning issues considered above. The shape ambivalence takes depends on the particular (potentially conflicting) social groups one is a member of and feels pulled toward,[83] and on the micro-architecture of social influence in one's network.[5]

2.3 Affordances and Effects

While some aspects of human behavior are consistent over time, the technology through which they are expressed changes rapidly, and social behavior changes with it. This section discusses the theory of these technological changes, some particularly important affordances of the internet, and their political effects. Particularly important is the end of the gatekeeping the role that news organizations traditionally played, discussed in Section 2.3.3

2.3.1 Theoretical Frameworks

A large literature documents the influence of media technologies and their affordances on human behavior. Consideration of the subject in the modern era goes back to Walter Lippmann's influential work in the 1920s,[86] which made the basic but essential point that the world is "altogether too big, too complex and too fleeting for direct acquaintance." People, in other words, understand the world through the media, as they must: It is too large and changes too quickly for there to be any other option. Combined with the whole catalog of cognitive biases that people bring to the task of interpreting their media-derived information, this state of affairs led Lippmann to radical skepticism about the very possibility of an informed public and a democratic polity.[41]

John Dewey, sparking a debate that has continued ever since, defended democracy against Lippmann's critique in his famous book "The Public and Its Problems."[87] While recognizing the force of his opponent's arguments, Dewey believed that democracy could be rescued from Lippmann's doldrums. He stressed the value of education, in allowing people to resist media blandishments and think for themselves, and also (in contrast to Lippmann's view of atomized individuals influenced by media) of social organization. A real society is complex and situates each person in a web of social relationships and institutions, which can help structure and inform that person's behavior.

Rather than focus on political impacts specifically, another line of thought has considered the media themselves and their effects on society overall. The seminal work here is Marshall McLuhan's *Understanding Media*, best known for its famous adage that "the medium is the message."[42] McLuhan advanced a comprehensive theoretical view of media and their impact on society, one grounded in the affordances and technological features of particular media. His distinction between "hot" and "cool" media, respectively those which provide a great deal of information and require limited thought from the audience (e.g., film) or less information and more thought (e.g., print), illustrates the overall emphasis on technological influence. These media would have different effects on their audiences, in McLuhan's view, and lead to different cultural consequences.

The overall thrust of technological development in media, however, he expected to head in one direction: toward greater integration and easier communication. The resulting "global village," a phrase McLuhan coined, is a cliche now but was quite a daring prediction in 1964.

Later work explored the social effects of these electronic media in more detail, and in a less aphoristic style than McLuhan, a literary theorist and philosopher, used. Meyrowitz, [88] examining the "impact of electronic media on social behavior," stressed the importance of changing social topology. Electronic media, which at the time meant especially radio and TV, changed who could be exposed to which occurrences in society. With them, society went from a world in which walls and doors gated access to real-time social events to one in which those events could be broadcast to millions. The new visibility changed the "situations," or archetypal social scripts, that people inhabited, resulting in not just wider access to the same behavior but quite different behavior. Politics, for example, is not conducted the same way when negotiations and debates are televised as it is when they are private or otherwise harder for others to access.

Finally, Ong & Hartley [89] study the differences between oral and literate cultures, and propose a new

kind of "secondary orality" that develops in cultures with electronic media. The "primary orality" of cultures without writing has several fundamental differences from the literate mindset,[89] all driven by the affordances of face-to-face speech. Everything that is known must be memorized, and memory is not perfect: People can more easily remember some things, like narratives, than others, like exact phrasal formulas or factual lists. (Ong and Hartley give the trenchant example of the 'begats' in the originally oral Bible, in which a genealogical list is couched as a story.) Because stories are told and memory is practiced through performance to a group, knowledge is also social: Oral cultures have no such thing as solitary study.

Both of these properties reversed with literacy, and have partially reversed again with electronic media. Television and radio, the dominant media at the time Ong and Hartley wrote, promote an oral-literate hybrid culture with elements of both. This secondary orality cannot exist without a literate culture to sustain the technology, but it revives the emphasis on social performance of stories. Lengthy, detailed, abstract thought plays poorly on television; as in a small group gathered to hear a story, emotion and narrative are the order of the day.

2.3.2 New Digital Affordances

As the internet developed in the early 21st century, many scholars sought to adapt these theoretical frameworks to the new technology. If indeed the medium is the message, what message did the internet send? The answer is that it depends: Different sites and services have different affordances and different social dynamics.[12] The internet is a more differentiated medium than television was, and particular user-interface and design choices can have substantial impacts on how people behave.

There are, however, some commonalities. boyd[12] focuses on social network sites, and identifies four traits that they generally have in common. In her terms, these are *persistence*, the default recording of most online content; *replicability*, the ability to easily copy digital materials; *scalability*, more recently known as virality, the ability of information to spread widely and frictionlessly to a large audience; and finally *searchability*, the ability to easily search for and discover user-generated material. These affordances are a sharp departure from pre-internet and offline life, in which speech is ephemeral, copying handwritten text is slow and tedious, broadcasting was restricted to large media companies, and searching for information required libraries and patience.

Together, they give rise to a dynamic Marwick & boyd[63] termed "context collapse": The merging of previously separate social spheres and situations. This is quite similar to the process Meyrowitz identified for television and radio,[88] and indeed Marwick & boyd[63] explicitly characterize the internet as bringing "the changes Meyrowitz described [for public figures] to interpersonal interaction." The effects as they describe them are far-reaching, especially a pervasive and corrosive sense of fakeness and inauthenticity in the behavior of people who are simply reacting, as anyone might, to the presence of a large audience. Context collapse is not even limited to the merging of diverse social settings, but can also (because of the persistence of old content) occur over time,[90] with people sometimes durably defined by their oldest or worst selves.

A final notable feature of online social structures is not just the combining of previously disparate settings, but also of different kinds of relationships. "Weak ties," which is to say social connections to people one knows only casually,[5,91] are quite important online. They make up most of the social graph on any given service; no one with 5,000 Facebook friends, for example, is a close friend of each of those 5,000 people — there simply isn't enough time in the day. They are not usually broken out separately from stronger ties, with every following relationship on Twitter, for example, working the same as every other. These edges are the substrate that supports "scalability"[12] or virality, social discovery of information, and many other characteristic features of online life.

2.3.3 Effects

These differences in affordances have had numerous effects on society, both online and offline. Social behavior broadly has changed in important ways, from the rise of online dating[92] to negative mental health outcomes,[56] but in keeping with a focus on the public sphere, we consider political effects especially.

Some of the most striking are those that occur 'in real life,' as the saying goes, as part of offline political activity. Tufekci [15] describes new modes of political protest enabled by the internet, which take a radically "networked" and "leaderless" form that would be alien to any pre-internet organizer. These sorts of protests — including movements like Occupy Wall Street and Black Lives Matter — are heavily reliant on digital technology to function at all. Tufekci gives the example of an anti-government protest movement in Turkey, coordinated and organized in a spontaneous fashion online. The movement's sudden appearance was an advantage[15] in avoiding repression, but also a hindrance in coming to terms with the authorities: In a leaderless movement, after all, there are no leaders to negotiate with.

Nor are the effects limited to protests; Bond *et al.* [1] demonstrate in a randomized experiment that a simple Facebook intervention can drive large changes in voter turnout. Crucially, this intervention relied on a "social message," in which members of the treatment group receive information about which of their Facebook friends had already voted. A number of the internet's new affordances are visible here, especially scalable, viral information spread and the importance of the weak ties that predominate on Facebook.

There are, of course, online effects as well. Much of the behavior that occurs online, while it may reflect offline social structure,[93] occurs in fundamentally new ways. On Reddit, for example, subreddit communities can engage in a novel sort of conflict,[14] difficult to achieve offline, that most closely resembles a military unit raiding behind enemy lines.[†] Members of one community, "mobilized by negative sentiment,"[14] are drawn through a hyperlink posted in their own discussions to another community, where they attack and express negativity to members of the recipient community. This is, from top to bottom, an online phenomenon: It couldn't occur without hyperlinks and the other affordances of Reddit (clearly demarcated communities, for one).

A final, and quite important, effect of this new social topology is its great weakening of traditional gatekeepers. This story runs from affordances to shifts in attention to economic changes: The ease of horizontal connection and the low cost of distribution online has led attention to migrate away from

[†]Indeed, this behavior is sometimes called 'brigading.'

traditional media outlets toward newer internet properties,[94] and advertising dollars have followed. This shift began with Craigslist and its demolition of newspaper classified-ad sections,[95] and continued with the rapid rise of Google and Facebook (not such prior attentional giants as NBC or the New York Times) to dominance in the online ad business. Fundamentally, advertisers want to be where the viewers are, and viewers no longer need to consult (say) a newspaper to look up (say) movie times: the internet will now give them a perfectly customized "Daily Me."[96] The resulting economic weakening of the traditional press has further accelerated the same shifts of attention that got it moving in the first place, as the decline of local news in recent years attests.[97,98]

2.4 Measurement

Social and economic organization is not the only thing affected by new technology. Social science has seen significant changes as well. As they matured over the last few decades, digital technologies have been both a great help and a hindrance to research. The benefits, of course, are clear: Large amounts of human behavior are now more easily observable through digital data than they ever were in an offline world. The hindrances are more subtle, but still important: New tools are needed to understand these large-scale online phenomena. People are also less willing to cooperate with older research techniques like surveys, partly because of digital affordances and their consequences like spam.[67]

Even primarily descriptive research is more complicated than in settings with smaller data. Studies of Facebook[93] and Twitter,[99] conducted early in those platforms' histories, had to deal with large datasets and rely on expertise in data engineering and computer science.

2.4.1 Experiments

Research aimed at drawing causal conclusions in online settings usually has an even more difficult task, in contrast to the wide recognition of randomized experiments as the gold standard of clinical research. In online experiments, the fundamental problem is pervasive interference between experimental subjects:[65,66] The same "scalability"[12] or virality that appears in other online phenomena makes it very easy for one person to influence another. If, for example, person A receives an experimental treatment and is induced to post angrier comments, all of A's followers are potentially exposed to these posts as well and may be indirectly influenced by the treatment.

This situation is a violation of the fundamental "stable unit treatment value assumption," or SUTVA,[100] that subjects' (potential) outcomes are not affected by the treatment assignment of other subjects. It renders much traditional experimental methodology inapplicable.

Experimental research in online settings deals with this in a number of ways, with the appropriate methodology depending on the quantity being measured and on the structure of the interference.[65] A common approach is the 'linear-in-means model,'[66] in which the ego node's treatment effect is a function of its own effect and the mean of its neighbors' effects. There are many other models of interference that

can be used, including "constant," "neighborhood," and "anonymous" treatment response, which Aral [65] covers in more detail.

In actually performing randomization, "network autocorrelation" of treatment[65] — i.e., whether nodes near each other in the graph should be treated similarly — and the sequencing of treatment in time are important considerations, both stemming from the reality of interference and assumptions about its structure. A frequent approach is "graph cluster randomization,"[101] which assigns treatment to groups or clusters of units and sharply reduces variance under mild assumptions.

While approaches like the linear-in-means model and graph cluster randomization are used in practice now, there is still much research underway on both the theory and practice of network experiments. A great deal of additional research has also already been done on these subjects (for example, Aronow & Samii [102]), too much to cover here, and there are a large number of practically important experimental results. Digital experiments have measured the effects of Facebook interface nudges on voting,[1] of Twitter bots on political polarization,[103] and of Twitter's recommendation algorithm on the ideological balance of tweets,[74] among other subjects.[104–108]

2.4.2 Survey Research

Scientific survey-based measurement of public opinion has been an important way for both policymakers and academics to learn about public preferences since the pioneering work of Gallup [109] in the 1930s, which in turn built on earlier theoretical approaches to public opinion.[86,87] While it has traditionally relied on randomly sampled surveys of landline phones,[110] new technology has complicated this picture in several ways. Landlines have gradually been replaced by cell phones,[111] with only 27.4% of adults having access to a landline at the end of 2022,[112] and their decline is continuing. Because of legal restrictions on the use of auto-dialer technology to call cell phones,[‡] polls of cell phones cost more to conduct. In parallel with these developments, response rates to telephone polls have fallen to extremely low levels,[§] partly because of the prevalence of spam calls.[67] Because calls that are not picked up still have costs in interviewer time and phone service, in addition to pollsters' fixed costs like office rent, declining response rates not only pose the risk of bias but also raise the cost per respondent.

These challenges for polling became subjects of intense public discussion after the polling errors in the 2016 US presidential election. Though Hillary Clinton's lead in the polls that year was not as wide as news coverage portrayed it, and shrank as the election got closer,[113] she did narrowly lose to Donald Trump after narrowly leading in polls and poll-based forecasts ahead of Election Day. A detailed retrospective analysis of the 2016 horse-race polls[77] by an AAPOR committee attributed the polling error especially to a late-breaking shift against Clinton, with an important role also for pollsters' practice of not weighting polls by respondents' educational attainment. The committee argued against some of the most concerning explanations, especially the "shy Trump" theory of nonresponse correlated with candidate support. This

[‡] 47 U.S.C. §227(b)(1)(A)(iii).

[§] Pew Research is helpfully transparent about these rates: As far back as 2018, its longitudinal surveys had a response rate of only 6%[19].

reassuring conclusion held up well in the 2018 midterm elections but was challenged by a much larger polling error in 2020,[76] at the height of the Covid-19 pandemic's disruptions to daily life and behavior. The results of the 2024 election, polling's most recent acid test, are ambiguous: Trump once again outperformed his polls, but by a much smaller amount.

Pollsters and scholars of opinion research have responded to these challenges with a flowering of new methods. New survey modes, from online to SMS, have seen rapid uptake.[114] Online surveys, which have been used since the early internet,[115] have become substantially more common in recent years.[114] Platforms like SurveyMonkey and Qualtrics are an important part of the story, offering tools for respondent recruitment, randomization, reporting and more. SMS surveys, though less widely used, are also the subject of both academic and commercial experimentation.[114] There is also substantial interest in and practical use of mixed-mode surveys,[20,116] which combine contacts from multiple modes (e.g., live phone interview and online). Longitudinal panels have become increasingly widely used as well, because their response rates have held up better than those of cross-sectional polls.[117]

Public and media pollsters have also increasingly adopted sampling techniques with a long history in political campaigns. Rather than using random-digit dialing, many polls in the last few years employ voter files as sampling frames,[118] taking advantage of the comprehensive list of voters to assemble samples more precisely. Pre-stratification by predicted probabilities of turnout (and/or survey completion) provided in these voter files, in particular, can improve power and help deal with nonresponse.[118]

Finally, the huge number of opinion statements now expressed on the internet, and especially on social media, provides a promising resource for measuring opinion.[119] These new data sources offer unprecedented insight into the detailed attitudes of large sections of the public.[21] They can also help address the classic critique, due to Blumer [8] and others, of survey-based opinion research that its model of independent, identically distributed, disconnected respondents is a poor fit for our actual, hierarchical, interacting society.

There are important challenges in using these new kinds of data, however, especially the lack of standardized questions and the non-representative set of those expressing opinions. Taking advantage of this data requires greater use of machine-learning methods and tools, and there will be more to say about it in the next section.

2.5 Machine Learning

One of the same developments that has enabled these digital affordances may also provide tools for measuring and adapting to them: machine learning. A comprehensive overview of modern machine learning is at least a book-length endeavor and is well beyond the scope of this chapter. This section will instead cover some of the important developments in ML, focusing on those relevant to online social behavior and public opinion. We also outline their applications to the problems of Section 2.3 and Section 2.4, and highlight some connections to our own work in later chapters.

2.5.1 Power and Scale

For nearly a decade and a half now, machine learning has been going from one success to another on the strength of "deep learning" or "neural network" techniques.[120] These methods, inspired by a loose analogy to the workings of biological brains,[121] have a provable ability to fit any continuous function,[122,123] and depend on the backpropagation algorithm[124] to perform gradient-based learning. They are capable of impressive feats: among others, unprecedentedly accurate modeling of protein structure,[125] superhuman performance at the game of Go,[126] and remarkable performance at language modeling and text generation.[27]

Many of these achievements have come through two central ideas: the use of dense distributed representations, and the power of large scale. Distributed representations, famously used in word2vec[127] with many earlier antecedents like latent semantic indexing (LSA),[128] are based on the principle that meaning should be spread out across the components of a model or object representation rather than concentrated into 'local' or semantically aligned features.[129] In an NLP setting, that might mean having a dense sentence embedding of dimension (say) 768, with meaning encoded in the geometry of the embedding space, rather than a 100,000-dimensional embedding in a space where each dimension represents a word. This principle has been widely applied in architectures from multilayer perceptrons[130] to recurrent neural networks[131–133] to more recent approaches like attention[134] and transformers.[135]

All of these architectures learn embeddings, in the mathematical sense: an injective, approximately homomorphic map from a high-dimensional input space like that of images or text to a low-dimensional, usually Euclidean target space (of distributed representations).[¶] Because this embedding — possibly with a final classification, regression, etc, layer on top — can be quite complicated with a highly parametrized model, even sophisticated learning tasks like contextual word embeddings[136] or translation[137] can be reduced to questions of geometry. Such representations have also been critical to the powerful self-supervised learning approaches that have given us the transfer learning paradigm, with separate pretraining and fine-tuning stages,[138] now widely used in ML.

The other central idea in recent ML breakthroughs is scale. Increasingly large models have been shown to have increasingly good performance on a wide range of tasks,[28,29] a phenomenon often referred to as "scaling laws" for neural models. This increasing performance is surprising in light of traditional learning theory, which would expect increasingly large models to overfit. Neural network scaling behavior has instead revealed a previously unknown kind of generalization performance known as "double descent."[139–141] In this regime, generalization error declines again after reaching its maximum value at the point where model capacity is sufficient to interpolate the training data. Various hypotheses have been offered to explain this phenomenon, including the 'lottery-ticket hypothesis" of useful sub-networks,[142] but its theoretical foundations remain mysterious.

These two ideas are combined in the idea of a "foundation model,"[30] a large, general-purpose model pretrained on a diverse corpus of text or audio/video,[143] usually also "post-trained" on instruction tuning and preference datasets,[144,145] and applicable to a variety of tasks. Such models, which include ChatGPT

[¶] Contrast the mathematical usage with the typical ML usage of the word "embedding," which refers to points in the range of this function: Rather than calling the model an embedding, that term is used for the vector it sends a token to.

among many others, are shaking up ML research and finding practical applications in fields like software development.[146]

2.5.2 NLP

This trend of large "foundation" models based on distributed representations has been especially prominent in natural language processing, the machine-learning subfield which considers language as data. The field encompasses a variety of tasks, from parsing to question answering to language modeling, traditionally with bespoke separate models arranged in a pipeline architecture to handle practical language tasks.

NLP has more recently, however, been dominated by large transformer[135] models pre-trained on language modeling in particular, because accurate prediction of word co-occurrence[‖] turns out to be sufficient to learn many other tasks.[147] Such models are then fine-tuned for downstream tasks, or their hidden representations can be used directly, without further training. An additional, highly useful, idea is to avoid both fine-tuning and the use of frozen hidden representations by encoding every task as a token sequence completion problem, a technique often called "text2text."[31]

This new paradigm arguably began with BERT,[138] furthering the contextual-embedding approach pioneered by ELMo,[136] and kicked into high gear with the release of GPT-3 in 2020.[27] A truly enormous amount of research has happened in the years since, on topics from multimodality[148] to alignment[149] to interpretability,[150] but the paradigmatic idea continues to be the use of large pretrained transformer language models.

2.5.3 Graph Learning

Deep learning for graph-structured data is also a powerful tool for understanding online social behavior. Existing graph neural network (GNN) architectures are generally built on what is called the "message-passing GNN" or MP-GNN approach, in which each node's representation is iteratively updated based on its own previous representation and those of its neighbors.[151] The term "message-passing" refers to the information or "message" aggregated from neighboring nodes. In practice, each iteration is usually implemented as a layer of a neural network,[151] sometimes with weights tied across layers. Recurrent architectures exist[152] but are unusual.

Commonly used MP-GNNs include the graph convolutional network (or GCN),[153] based on an analogue of the continuous convolution operator, and the graph attention network (or GAT),[154] a transformer with adaptations for graph data that improve its asymptotic complexity. Vanilla transformers are also usable as graph neural networks if attention is masked according to the graph adjacency matrix,[155] though they have worse computational requirements than GAT. Shallow node embedding methods like node2vec[156] can also be motivated in terms of this framework.[151]

These models differ from those used in NLP and many other areas of deep learning in several important

[‖] Usually autoregressively; masked / denoising objectives are less common because of the difficulty of using models trained with them for text generation.

ways. One of the most important is that they have been far less affected by the scaling revolution of the last few years. Transfer learning is limited: Pretraining approaches for graph tasks are complex and see limited improvement in performance by NLP standards.[157] Deeper models are not always beneficial either, fundamentally because layer k in the network aggregates information from each node's k-hop neighborhood. There is no reason to use more layers than the graph's diameter, and when local information is more useful than global information (as is often true in social networks), models may be shallower still.

Another important difference is that, despite these limitations, GNNs have a firmer theoretical basis than most deep learning architectures. The MP-GNN paradigm has deep connections to the theory of graph isomorphism testing and the Weisfeiler-Lehman test,[158] with many MP-GNNs having expressive power provably equal to that of the test. From a different angle, the same MP-GNNs can also be motivated as a kind of discrete spectral convolution,[151] based on a close formal analogy between a graph's Laplacian matrix and the continuous Laplace operator[159] (with connections thus also to Fourier analysis). There is also considerable interest in alternative theoretical perspectives and developing models based on them, especially ones able to benefit from pretraining and greater scale.[160]

2.5.4 Applications

These ML techniques have been widely applied to research topics across scientific fields[161–163], with some limited use in analysis of public opinion and broader uptake for social network questions[164] and other online social behavior. There is far too much of even these latter kinds of work to cover it all here. Applications have included tasks like stock price prediction,[165] ** fake news detection,[166,167] Twitter sentiment analysis,[168] modeling the structure of literary narratives,[169] and even online-to-offline tasks like forecasting the influenza season with Google search data.[170] Techniques from machine learning and AI are also used to build tools and services, including for political communication[171] and journalism.[172]

ML is also frequently used for research on questions having more to do with social structure and interpersonal behavior. Kumar *et al.*, [14] for example, analyze a certain kind of conflict allowed by the affordances of Reddit (also discussed in Section 2.3.3), and build a classifier to predict these kinds of conflicts. This sort of work has practical value to social media sites in helping prevent conflicts, and to researchers in illuminating the situations in which they occur. Other research has combined[173,174] two modalities — networks and language — that frequently co-occur in social media settings, often in the service of answering social science questions. Bail [68] tests a thesis about the effectiveness of subcultural bridging in online advocacy by using both network-structured social context and the language used to express positions. Li & Goldwasser [70] leverage network structure, via a graph neural network, to provide distant supervision for training a political perspective classifier. A final ML-informed direction of note is to examine the workings and effects of social sites' feed curation and recommender algorithms, as in the context of a political campaign.[105] Some of the work in this dissertation fits into this vein as well, with the analysis in Chapter 3 and Appendix A using graph-based techniques to detect news events and compare the structure of Twitter and broadcast media. The clear presence of graph structure in such applications

** In this case, presumably also rather a lot of unpublished work.

makes graph-learning methods, whether neural or otherwise, commonly used.

Uptake of ML in public opinion research, by contrast, has been more limited. From quite early on, there have been attempts to extract measures of opinion from social media, especially for election prediction,[175,176] albeit with mixed results.[177] One of the most promising studies in this vein, that of Bovet *et al.*, [22] proposes a Twitter-based method for predicting opinion trends that is ultimately based on hashtag co-occurrence. Validation against a traditional polling average in the 2016 US presidential election demonstrates quite good performance. Despite these results, however, this study also highlights the difficulty of social media data for opinion research: Twitter, after its sale to Elon Musk, has restricted the availability of data enough that the method is no longer usable at a reasonable cost.

More recent research has focused on the use of large language models for opinion. An exciting line of work has tried to relate social media data to public opinion data,[178] finding mismatches between "stance" as expressed in a public opinion poll response and on social media. Other work has been under the broader "alignment" umbrella, such as attempting to align language models to the consensus opinions of a group of humans.[179] Other work, focusing on group deliberation rather than average opinion, leverages LLMs to improve deliberative discussions,[180] and some recent papers examine the political biases displayed by the models themselves.[181,182]

Finally, a promising new direction is using LLMs to simulate respondents to public opinion surveys. Several recent studies have taken this approach, with various fine-tuning strategies. Chu *et al.* [34] train models on "media diets," exploring the ability of such models to predict respondents' answers to public opinion polls. Besides being practically useful, the results shed light on the relationship between media coverage and poll responses. Kim & Lee [71] evaluate similar methods, but with fine-tuning on survey questions rather than media content, and also evaluate their approach as a method for multiple imputation in survey settings.

2.6 Social Impact

It is worth stepping back and considering why these developments are important not just for researchers but for society at large. We have encountered a few major, direct, offline effects so far, but there are others. Particularly important are impacts on the public sphere,[72] broadly defined: public discourse, political activity, and organizing on significant issues. As Habermas [72] defined the concept, the public sphere consists of "private people gathered together as a public and articulating the needs of society with the state." Because any modern society has a huge number of private people participating in this activity, too many to directly interact with each other, most of the public sphere takes place of necessity in the media.

Many of the impacts of technological change on the public sphere are news-media phenomena as well. The internet has sharply changed both the economics of media and the distribution of readers' attention.[183] As public attention moved online, to intermediary platforms like Google and Facebook, advertising revenue moved with it and local news entered a steep decline that continues today. The news cycle has changed as well, featuring much more influence for online media: first blogs,[69] then Twitter and other social media,[184,185] and most recently TikTok.[186] These shifts have produced great effects on politics

and other aspects of society downstream of the media.[187]

In particular, they have contributed to polarization. The United States has seen a longstanding, gradually intensifying process of political polarization, possibly motivated by partisan sorting of identity groups[188] and the rise of a rural identity.[189] Increasingly polarized parties move further apart on issues[16] and their voters express increasing dislike of each other.[17,190,191] Other developed countries have seen similar, though less intense, polarization.[192] Media effects are widely argued to be involved[57], with online media[192] and "pro-attitudinal media"[59] argued to be polarizing. More research is, however, needed.[59]

The challenges facing survey research are also relevant here: While there has not (yet) been a great collapse in polling accuracy,[76,77] declining response rates have made survey research more expensive and harder to conduct. With the last several decades' decline of mediating institutions like parties, civil society groups, and so on,[193] a weakening of public opinion research may make government less informed about what the people want.

2.7 Conclusion

We have reviewed both the theory and practice of changing media affordances, and ways to measure them. New developments in the technology of communication are reshaping social behavior both online and offline, and pose both challenges and great opportunities for social-science research. New machine-learning tools are especially promising in addressing the challenges facing research, though they are also involved in some of the negative effects of the new digital technology. (Consider social-media recommender systems.)

These developments pose important social risks, including to political culture and democracy. Innovation in research methods and greater use of ML techniques, as we have discussed in this chapter and as we pursue in the rest of the dissertation, plainly cannot fix all of these problems on their own. But for research purposes, new methods can be and already are helpful, and with luck will continue to be. Fixing problems begins, after all, with understanding them.

Chapter 3

Speed and Sentiment of the News Cycle Across Media

The work discussed in this chapter was published last year,[194] and we have adapted the discussion, figures and tables here, in Section 6.1.1, and in part of Appendix A from the published version, with minor modifications. Some of the text adapted from that paper, especially that describing datasets in Section 3.3, first appeared in an earlier master's thesis,[195] though none of the results did.

3.1 Introduction

The speed of the news cycle and the nature of public discourse surrounding the news have important implications for civic life. If increasingly prominent new media institutions have shorter attention spans than older ones, society's ability to sustain focus on important issues and hold the government accountable may be reduced.[196] Similarly, if outrage and negativity are intensifying as well,[197] public debate becomes more rancorous and it may be harder to reach consensus. Because different media have and serve different audiences, there may also be disparities between groups in the attention paid to those groups' priorities.

These questions of media dynamics matter in turn because media coverage has an important role in determining public opinion and policy[39] with even policymakers relying on mass media coverage for information on matters as weighty as whether to go to war (the so-called "CNN effect").[198] Understanding the internal dynamics of media can shed light on the influence those media have over the rest of society.

We seek in particular to compare the nature of news propagation – its speed and affective content – across media. We compare across two major and influential communication platforms, one drawn from the domain of social media and the other from broadcast media: Twitter (recently renamed X) and U.S. talk radio (both public and commercial). Both are highly impactful in their own right: radio because of its very wide reach,[199] and Twitter because of its millions of users and a high concentration of journalists among those users.

Radio in particular, while less of a venue for intra-journalistic discourse than Twitter, does not receive as much attention as its substantial influence on political discourse and individual behavior[200] suggest it

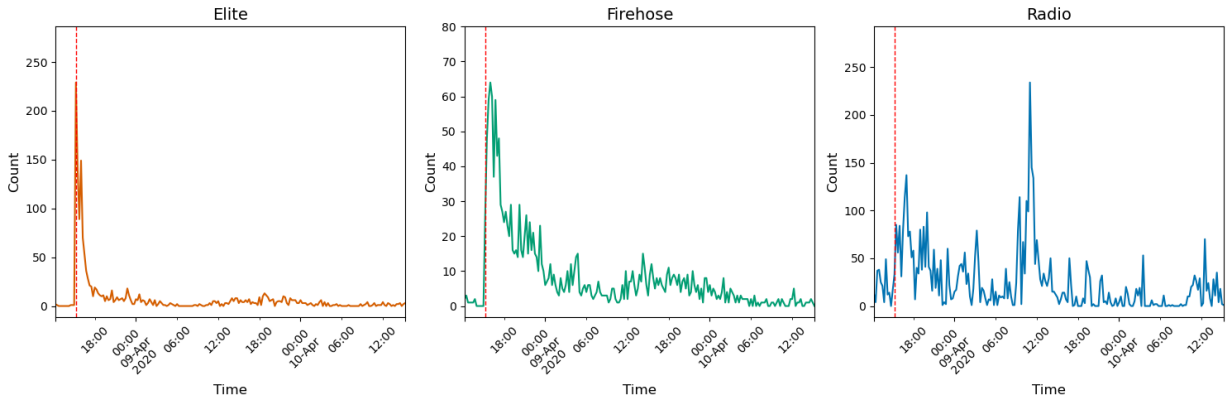


Figure 3.1: An example of one of the manually detected or keyword-based events used as a robustness check, Bernie Sanders' withdrawal from the 2020 presidential race. Note that discussion on both elite and mass Twitter clearly peaks and falls off before radio discussion.

deserves. Much of this lack of attention is due to the difficulty of obtaining corpora of audio or transcripts, which we address by building the comprehensive radio transcript corpus discussed below.

It is not clear, however, exactly how we should expect news events' rise, decay and sentiment to differ between radio and Twitter. On the early, rising side of a news event, a finding of more rapidly rising Twitter discussion would be in line with qualitative and theoretical literature on the "24-hour news cycle," the acceleration of news,[13,201] and Twitter's role in it,[202,203] with our study testing and contributing empirical evidence to this literature.

It is not, however, obvious which medium we should expect to see move on more quickly. Twitter's highly interactive nature, which allows co-creation of the news[204] and a stronger feedback loop between journalists and audience, may cause events to proceed faster and conclude more quickly than on a one-to-many broadcast medium like radio. On the other hand, there are also reasons to expect events on the radio to be less persistent. Unlike genuinely ephemeral radio broadcasts, which are not generally recorded, easily available after the fact, or searchable, tweets persist online and can be shared by users, perhaps prolonging event lifespans. Interactivity and virality may also be a two-way street, with a story that strikes a chord with users persisting longer than journalists alone would have continued discussing it. Finally, because the two media can interact, we might find that they have similar attention spans. With discussion on Twitter keeping an issue alive and sparking interest on radio (or the reverse), news events may wind down in a similar way across media. An empirical answer to this question appears to be lacking in the literature, and our work in this chapter can provide one.

Meanwhile, in contrast to the theoretically and empirically rich literature on news cycles, there has been much less attention to how outrage and other affective states compare between Twitter and traditional media. We have not been able to identify a large-scale comparison between outrage or other affective states on Twitter and in traditional media. The closest prior work[205,206] compared talk radio and cable news to political blogs, rather than social media, and found both very high levels of outrage on radio and cable, and *lower* levels in blogs. There is some support, in other words, for believing that either Twitter

or radio (the medium of hosts like Rush Limbaugh and Sean Hannity) might be more outraged and negative; comprehensive data are lacking. One contribution of our work is to fill in this lacuna with an apples-to-apples comparison between the two media, especially one which allows us to examine the distribution of outrage and negativity over the lifecycle of an event.

We address these questions with three large datasets, one of radio broadcasts and two providing different views of Twitter. Our analysis shows that Twitter, as a medium, differs from radio in having a systematically shorter attention span for news than radio, and in reacting to it in a more outraged way. These biases, likely rooted in properties of the media themselves, suggest an important connection between platform affordances and news cycle dynamics. Identifying specific mechanisms relating platform features to news cycle behavior may be an important direction for future research.

3.2 Related Work

Twitter Twitter, whose traditionally easy data availability has enabled a great deal of research,[99,207] is one of the major social media platforms with a mass audience and is widely used by journalists, playing host to many of their interactions and deliberations with each other.[208,209] Journalists themselves, while not always convinced that Twitter's influence on their profession is salutary, have discussed it frequently and at length.[210,211]

Radio Terrestrial radio reaches as many as 88% of Americans each week,[199] and a long literature documents its political influence over issues and decisions ranging from health care reform[212] to vote choice[213] to misinformation and conspiracy theories.[214] This ability to influence political agendas reaches as far back as Father Coughlin in the 1930s,[215] but intensified and became more clearly partisan after deregulation in the 1980s and the rise of conservative hosts like Rush Limbaugh.[200]

Affect Affect has been extensively studied on Twitter and social media in general, with a particular focus on outrage. Prior work has examined its psychology,[216] mechanisms behind outrage 'firestorms,'[217] the distribution of outrage,[218] and many case studies,[219,220] which generally concur with everyday experience that outrage is prominent and influential on Twitter. Outrage in traditional media has also received considerable attention, especially in conservative media,[221] with pioneering studies from Sobieraj & Berry [205] and Berry & Sobieraj [206] finding very high levels of outrage on cable news and talk radio, and lower levels in political blogs. Notably, the literature appears to lack a direct comparison of broadcast media to more recent social media sites.

News cycle Work in political science and computational social science has lent support to claims of an accelerating news cycle[201,222] and of media business models becoming increasingly focused on tribalism and outrage,[206] with the latter rooted in the 1980s deregulation and the rise of talk radio.[200,223] But while both the 24-hour news cycle and outrage media have long histories, they are often claimed to have accelerated

with the internet,[13] and especially social media.[197,224] Indeed, there is even evidence that the turnover of content in social media is accelerating over time, as recent work[225] has found for Twitter.

There is also extensive work on "intermedia agenda-setting,"[226,227] which focuses on the effects different media and media outlets have on each other's coverage decisions. The speed of reaction to events may indeed be an important channel for this influence, providing a first-mover advantage for faster media that can determine the tone and framing of subsequent coverage.[228,229] Though our work does not directly take up such causal questions, it complements the agenda-setting literature by closely examining one potential mechanism and origin of intermedia influence.

There has been much less quantitative work, however, on the falling side – the decay of attention to news. Prior work has shown that traffic to online news stories decays rapidly,[185,230] as does engagement with them on sites like Digg.[231] Other work[202,232] has examined the dynamics of attention on Twitter during events, but focused more on the rise of attention than its decay, and used small numbers of hand-selected events. Pfeffer *et al.* [233] analyzed the "half-life" of a tweet and found that engagement with the typical tweet falls off rapidly, but this work examines individual tweets rather than events. Another study, closer to our work, examines the "screen-persistence" of events detected on Twitter by an in-house event detection system.[234] This work differs from ours, however, in looking only at very short-term (< 1 hour) persistence, defines persistence by presence in the Twitter feed rather than net discussion, and does not make cross-medium comparisons.

While an increasing amount of news breaks on or is influenced by these new platforms,[13] and researchers have analyzed the dynamics of user activity within them,[225] there is relatively little hard evidence about how news cycles differ between them and traditional media. Systematic, large-scale empirical evaluations of comparative news-cycle speed and sentiment are lacking.

3.3 Datasets

Our three datasets each cover a total of six months, spread over three years: September to October 2019, March to April 2020, and January to February 2021, encompassing important news events including President Trump's first impeachment and the start of the COVID-19 pandemic. We refer to these three datasets — radio, elite Twitter, and firehose Twitter — interchangeably as "datasets" or "corpora," to distinguish them from the "media" of Twitter and radio.

3.3.1 Radio

The first of the three is a novel and very large dataset of transcribed radio broadcasts, developed using a similar approach to Beeferman *et al.* [235] and Brannon.[195] This corpus includes the transcribed broadcasts, divided into approximate speaker turns, web-scraped show labels, and station metadata. Before quality filtering, it comprises 517,895 hours of audio from 228 talk and public radio stations, including content from 1,066 shows and amounting to about 5.2 billion words of text.

Our process began with data from the third-party company Radio-Locator[*] on the complete list of U.S. radio stations, from which we selected about 200 of those with talk or public-radio formats and online streams of their broadcasts. We continuously downloaded these broadcast streams. The exact number of stations varied over time with technical issues and additions or removals for other projects.

An important difference from Beeferman *et al.* [235] and[195] is that we performed ASR in two steps: First, to reduce costs given the large scale of the corpus, we converted all of the streamed audio to text using an in-house speech recognition system based on Kaldi.[236] These transcripts are used in keyword-based manual event detection (discussed below). Second, we re-transcribed the random sample used in automated detection with OpenAI's Whisper model[237] for better performance.

We also collected information on the stations' schedules (i.e., start and end times of their programs) by scraping their websites, using this to label transcribed broadcasts with the show of origin. Because station websites are not always updated promptly, however, these labels are sometimes out of date. Overall, the corpus included transcribed radio broadcasts, divided into approximate speaker turns, the web-scraped show labels, and station metadata.

To exclude irrelevant content and ensure quality, we performed several filtering and deduplication steps:

- We excluded any audio for which we could not collect schedule data. Note that a large share of this content should be syndicated programming, and will be represented by collection of the same content on other stations.
- We manually reviewed each of the 1066 shows and excluded those which did not contain any news discussion, such as music call-in shows, gardening shows, and most sports talk, as well as the BBC World Service, which rarely discusses US politics and is a poor match to elite Twitter. A handful of entire stations which had changed to music formats during our collection periods were also excluded.
- We dropped some airings (an "airing" is a show/station/date combination) to ensure quality, specifically those with ASR or schedule-data confidence scores below empirically chosen thresholds.
- Finally, we deduplicated the remaining airings across stations to "episodes," or show/date combinations, to avoid over-weighting large syndicated shows. We did so by choosing the airing for each episode with the highest average schedule and ASR confidence.

This process yielded a cleaned, final corpus of 89,203 hours of content from 144 stations and 810 shows, amounting to about 902 million words of text.

3.3.2 Elite Twitter

Our second dataset consists of 2.6 million tweets from a diverse set of 2,834 elite journalists and politicians ("elite Twitter") posted during the study periods. Parts of our analysis also rely on the follow-graph edges

[*] https://radio-locator.com/.

between these users.

User selection The set of users includes leading figures in journalism and politics, including every member of Congress and reporters from the major national news outlets, covering a range of geographies and ideological perspectives. Among politicians, we accordingly included every member of Congress and a few particularly prominent political figures who were not in Congress at the time of data collection (e.g., Donald Trump, Barack Obama, and Joe Biden). Among journalists, we generally deferred to news organizations' own decisions on notability, collecting most of our users from eight Twitter lists of prominent reporters. These lists were themselves from ideologically diverse outlets: C-SPAN, Slate, MSNBC, Fox News, the New York Times and the Washington Post. As with politicians, we included a few other particularly prominent commentators.

We also included 203 radio hosts and staff ("radio users") for whom we could identify Twitter handles. Doing so lets us ensure that both radio and elite Twitter datasets offer different views of the social and professional milieu around radio hosts, making analysis involving these two datasets an even more direct comparison of the two media themselves. We collected the radio users' accounts by show, focusing on the 45 largest syndicated shows (i.e., shows recorded on multiple stations), and then including approximately half again as many local shows. We then manually collected matching Twitter accounts: If Twitter accounts of hosts (i.e., on-air talent) were available, we included them; for producers and staff, we preferred staff in senior positions or with a public presence, including more junior staff in a few cases where no other accounts could be found. In total we collected 203 Twitter accounts for these 68 shows, with 65 shows having accounts that posted tweets during the study periods.

While the final set of 203 radio-linked accounts is not based on a random sample of all shows, it is reasonably representative of the largest and most important shows (because it includes most of them) without neglecting smaller programs. For example, the syndicated shows include the most popular and influential shows from both public and conservative radio during this period: for example, All Things Considered and Morning Edition on one side, Rush Limbaugh and Sean Hannity on the other.

In selecting the other journalists and politicians included in the sample, we aimed to cover a range of geographies and ideological perspectives. Among politicians, we accordingly included Twitter accounts for every member of Congress (via C-SPAN's members-of-congress Twitter list) and a small number of particularly prominent political figures who were not in Congress at the time of data collection (e.g., Donald Trump, Barack Obama, and Joe Biden). Among journalists, we mainly deferred to the decisions on notability of news organizations themselves, collecting most of our users from eight Twitter lists of prominent reporters. Two of these lists, from C-SPAN ("cspan/political-reporters" and "cspan/congressional-media"), included journalists regardless of ideological affiliation, while two further lists each covered left-leaning ("slate/left-leaning-tweets" and "msnbc/msnbc-hosts") and right-leaning ("slate/right-leaning-tweets" and "foxnews/shows-hosts") journalists. We also included lists of prominent journalists from the New York Times ("nytimes/nyt-journalists") and the Washington Post ("washingtonpost/washington-post-people"). As with politicians, we supplemented these lists with a small number of manually selected commentators or institutional accounts who were not on any of these lists. This process produced 2631 Twitter accounts

for journalists and politicians.

Data collection For each Twitter user, we collected all tweets posted by that user during the study periods, as well as the list of IDs of other elite-Twitter users that user follows and is followed by (the "follow graph" over the users).[†] For analysis other than ideology detection (discussed below), we used information only about our list of 2834 users, and discarded follow edges to users not on the list. Not considering second-degree connections through other users focuses our graph on users' explicit choices, and the preferences or interests they reveal. The ideology analysis does, however, use follow edges to these other users to calculate an ideology measure.

We pulled the follow graph once, in early November 2019, and pulled tweets every day during the three two-month study periods. Using these tweets, we also generated the mention graph, in which there is a directed edge from A to B if A has mentioned B's username in a tweet. (Unlike the follow graph, the mention graph is used only for analysis in Section A.1, not in this chapter.) After creating both follow and mention graphs, we collapsed all users in each graph associated with a given show to one node for that show.

3.3.3 Firehose Twitter

The third and final dataset consists of 24.0 million tweets sampled randomly from Twitter's firehose of all tweets[‡] ("firehose"). We collected two separate samples, one for analysis with manually detected events (8.4 million tweets) and another for analysis with automatically detected events (15.6 million tweets), totaling 24.0 million tweets. Because firehose data is drawn from the general Twitter user base, using it as well as elite Twitter lets us ensure that we are measuring effects of the medium as such, rather than only journalists' behavior or professional conventions. If we find similar behavior for both, as in general we do, this behavior is likely to stem from Twitter's affordances and properties as a medium.

3.4 Methodology

This section describes our methodology for event detection, affect identification and affect dynamics modeling. There are several other robustness checks and supporting analyses also outlined here: inferring ideology to analyze liberals and conservatives separately, matching detected events across media, and identifying and tracking a small sample of manually specified keyword-based events.

3.4.1 Automatic Event Detection

The first step in comparing event lifecycles is concretely defining an 'event.' As in much existing literature on news event detection (e.g., Liu *et al.* [234]), we define events here in a media-first, but also medium-agnostic,

[†]The 203 radio users' tweets were collected only during the 2019 and 2020 periods.

[‡]Via the Decahose API: https://developer.twitter.com/en/docs/twitter-api/enterprise/decahose-api/overview/decahose

way: An event is a subject of discussion and public attention, rising to prominence in discourse before falling off. Concretely, our events are thus groups of related *items* (tweets or radio speaker turns), which allows us to use the same definition of and detection method for events in both media, without relying on medium-specific features like retweets. This definition captures a wide range of news, both clearly exogenous events and those whose timing is driven by journalists' editorial choices.

We operationalize attention as discussion, rather than audience exposure to content,[233,234] for reasons of data availability. We had access to radio broadcasts and tweet-level information, but not to radio audience data or the impression data needed to measure exposure on Twitter. This common situation[238] focuses our analysis on the behavior of the media elites who produce news and those engaged users who explicitly comment on it, and means that the results have greater implications for research on agenda-setting than on media effects. While the hypothesis that audience attention rises and decays in line with discussion is plausible, and our results with Twitter's firehose below provide some evidence for it in that medium, we cannot say for sure. On the agenda-setting side, though, because past media discussion can frame and influence future coverage,[228,229] a medium whose engagement with an issue both begins and ends sooner may have a greater ability to set the agenda.

To detect events as we've defined them in the elite Twitter, firehose Twitter and radio corpora, we use the newsLens system of Laban & Hearst. [239] The system has several stages, or modules:

Inter-item similarity computation First, we compute the similarity of each item to every other item within a sliding window over the corpus. (Recall that an "item" is a tweet or radio speaker turn.) We use a window of length 16,000 seconds, and follow recent work[240,241] in using the cosine similarity of Sentence-BERT embeddings[242] rather than bag-of-words representations or doc2vec to compute the inter-item similarity in the first stage of newsLens, building on findings of better downstream performance.[240]

Item similarity graph Only item pairs with similarity above a configurable threshold are kept; we used a value of 5 standard deviations above the mean for all years and corpora, with (year, corpus)-specific values for SD and mean. The selected item pairs form an undirected graph.

Community detection Next, we use the Louvain community-detection algorithm[243] to decompose the graph into events, or news stories. Because our graphs were not in general fully connected, we ran the Louvain algorithm separately within each component of at least 300 items.

Community merging To detect events which are interrupted and then resume, detected events which do not overlap in time are merged together if the cosine similarity of their centroids exceeds an empirical threshold; we used the 99.9th percentile of all pairwise similarities.

Because the radio and firehose corpora are much larger than elite Twitter, for simplicity and a fair comparison we first randomly downsampled the final radio corpus to the same size as elite Twitter before running newsLens. Because the firehose contains a large amount of content which is not about news, such as Twitter memes and discussion of movies or music not released recently, we downsampled the firehose

corpus to six times the size of the elite Twitter corpus (rather than the same size) to get a similar number of news-related events.

We ran the detection process separately on each (year, corpus) pair, such as 2021 firehose data. Separating the media during the detection phase produced notably higher-quality events, and we address its downside of no longer having paired events with the matching process described below.

3.4.2 Event Filtering

To exclude radio and firehose content which was not about news (e.g., advertising, weather reports, Twitter memes), we represented each event by the centroid of its item embeddings, and filtered out detected radio and firehose events whose maximum cosine similarity to any elite event was below a manually tuned threshold of 0.6. We used the same threshold across all years and corpora. We also discarded any detected event consisting of fewer than 10 tweets or speaker turns, in order to exclude a set of apparently spurious detections with very few items. We did not filter elite Twitter events, because by inspection this corpus — made up of journalists and politicians — almost exclusively consists of news discussion.

We filtered in this way, relying on elite Twitter, for two reasons. First, because we are interested in comparing properties of Twitter and radio as media, it is more important to have a common set of news events than to have an entirely comprehensive one. Minor news events which were discussed only on mass Twitter or local radio would be important in a thorough survey of the news agenda on these media, but we are interested in the different question of comparative news cycle dynamics. Avoiding confounders like topical variation can help. Second, deferring to a large set of national journalists on what the news is greatly reduces researcher degrees of freedom in this regard.

Detected events This detection process identified 1,694 total events, broken down by year and corpus in Table 3.1. In light of recent findings that patterns of user behavior on Twitter are changing over time,[225] we also break down much of the following analysis by year as a robustness check.

Year	Firehose	Elite	Radio
2019	224	393	41
2020	203	566	53
2021	73	105	36

Table 3.1: Counts of automatically detected news events by year and corpus.

3.4.3 Event Matching

Some of our analysis requires matched (elite, radio, firehose) triples, selected to correspond to the same real-world occurrence. We selected these triples as follows. Using the same all-pairs cosine similarities as above, for events in different media which overlapped in time, we found (elite, radio) matches. The first step was to discard pairs which failed to satisfy some quality constraints: a) if either event in the pair had fewer than 30 items, b) if the events differed by a factor of more than 3 in their wall-clock duration, or c) if the events differed by more than a factor of 7 in the number of included items. We formed an undirected graph over the remaining events,

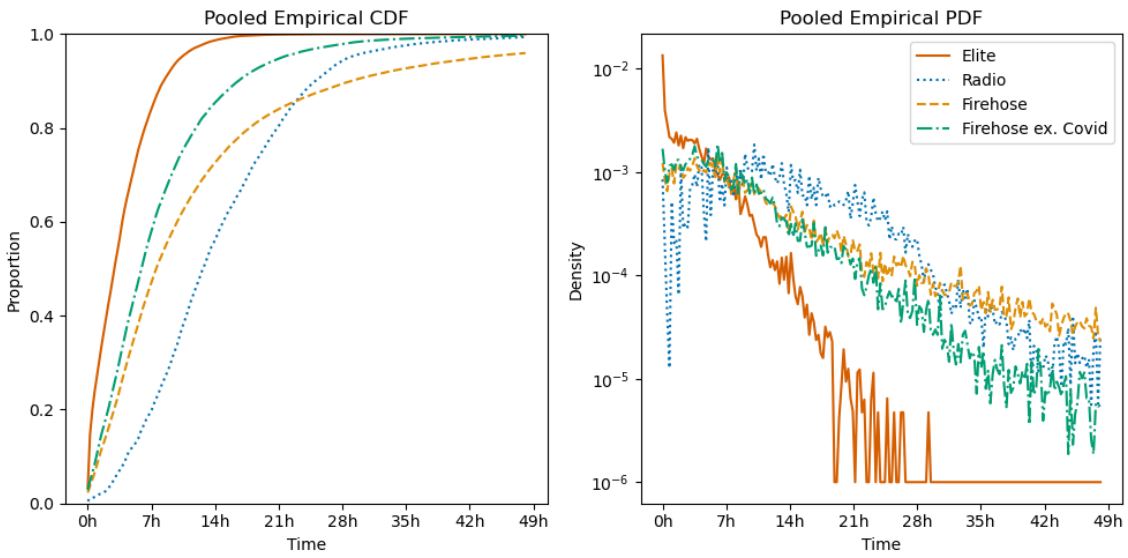


Figure 3.2: The average empirical CDF and PDF of mentions of the automatically detected events. We find the time after event start of each item (tweet or radio speaker turn) contained in the event, compute the empirical CDF at 15-minute resolution and average the CDFs over the events. The PDF is obtained by a central difference estimate of the derivative of the CDF. Only mentions occurring within 48 hours of the event are included. The density of Twitter discussion shows the same rapid rise and rapid fall, first exceeding and then falling below radio, as in the manual case. The "Firehose" line represents all firehose stories, while the "Firehose ex. Covid" line excludes Covid-related stories from March and April 2020.

with one edge per pair and the cosine similarity as an edge weight. Finally, we took the maximum-weight matching[244] as implemented in the networkx package[245] for Python to find the most similar pairs of events.

We used the same procedure and quality filters to match these pairs to firehose events, yielding (elite, firehose) and (radio, firehose) pairs, with each firehose event matched to the closest elite or radio event. Grouping each firehose event with the (elite, radio) pair to whose member it had been matched gave us (elite, radio, firehose) triples, which we filtered by dropping those below a threshold in either (elite, firehose) or (radio, firehose) similarity, with the same threshold for both similarities. After manual inspection, we found that a threshold of 0.5 produced a high-quality set of events. In total, we find 38 triples totaling 114 events.

3.4.4 Manual Event Detection

As a robustness check, we identified a set of events from Wikipedia occurring during the same periods in 2019[§] and 2020[¶] as our datasets and tracked mentions of them by counting keywords. To avoid drift in the real-world events associated with our keywords, only mentions within 4 days of event start are considered. The identified events are shown in Table 3.2.

[§] https://en.wikipedia.org/wiki/2019_in_the_United_States
[¶] https://en.wikipedia.org/wiki/2020_in_the_United_States

We aimed to collect a particular kind of event, namely one which a) got more than a trivial amount of discussion on both radio and Twitter, b) was readily identifiable by simple keywords, and c) occurred at a specific point in time rather than extending over a longer period. We collected the first 5 events each during the 2019 and 2020 corpus periods that we found to meet these criteria, for a total of 10. Using point events of this sort allows us to avoid considering the time course of the event itself, as with, say, a hurricane. Our events can thus all be viewed as announcements: of firings, bankruptcies, court decisions, etc.

3.4.5 Ideology Inference

Some of our analysis breaks down radio and Twitter into liberal and conservative ideological sides. This section lays out our methodology for identifying liberal and conservative content in elite Twitter and radio corpora, relying on particular properties of each medium to do so. Lacking a good way to make ideological inferences about users in the general firehose, we ignore it here and consider only elite Twitter and radio.

Twitter. To identify liberal and conservative sides of Twitter, we rely on Louvain communities[243] in the follow graph, which has been shown to incorporate both social and discursive aspects relevant to a phenomenon like ideology.[99] Community detection returns a small set of communities covering the entire graph, which upon inspection have clear interpretations. One consists mainly of conservative journalists and politicians, one is centrist-left, and the others contain mostly liberals; we considered tweets from the first community's members to be conservative and all others to be liberal. For a more detailed examination of these communities, substantiating this interpretation, see the supplementary information to the published paper.[194]

Event	Date	Time
John Bolton Fired	2019-09-10	16:00
Purdue Bankruptcy	2019-09-16	03:15
Tom Brady Free Agent	2020-03-17	12:45
Shane Gillis Fired	2019-09-16	20:00
Manning Released	2020-03-12	21:15
Huffman Sentencing	2019-09-13	18:00
Bernie Drops Out	2020-04-08	15:15
NBA Season Cancelled	2020-03-12	01:30
Warren Drops Out	2020-03-05	15:30
Trump Impeachment	2019-09-24	18:30

Table 3.2: Details of the 10 manually detected news events included in our analysis. All times are UTC and are for the 15-minute period start time immediately preceding the event.

Radio — Methodology. For radio, it is less clear how to identify an ideological dimension. Lacking any labeled training data, we cannot follow Vijayaraghavan *et al.* [246] in using a classifier to infer user ideology. In lieu of Bayesian methods,[247] we instead adopt a simpler approach (also used, for example, in Green [248] and Timm [249]) based on applying dimensionality reduction to Twitter's follow graph.

We first form the bipartite graph between radio accounts, on the one hand, and the set of all Twitter users (not just our set of journalists and politicians) who follow at least two radio accounts. (Here, unlike in the main text, we rely on the full set of follow graph edges pulled for the elite-Twitter accounts rather than

only the edges among the elite-Twitter users themselves.) From the adjacency matrix of this graph, which has radio accounts as columns and followers as rows, we compute the pairwise cosine similarity matrix of the radio accounts. We can then project the radio accounts into a lower-dimensional space (we use 2D for easy visualization and inspection) via classical multidimensional scaling (MDS). In a heavily ideological set of users like ours, where distances largely reflect ideological homophily, one of these dimensions is interpretable as latent ideology.

Because of properties of MDS, the scores are only well-defined up to a change of sign. We found higher scores to be more conservative and lower scores to be more liberal, which our discussion reflects. (The scores range from about -0.75 on the liberal end to +1.25 on the conservative end.) To get scores at the radio-show level, we averaged the ideology scores of the radio accounts associated with each show to obtain the show's ideology score. Finally, to get a dichotomous indicator for radio content, we map these show-level scores back to radio content. Any content from a show with an ideology score receives that show's score, and other content on stations with such shows (i.e., content from other shows for which we do not have Twitter handles) receives the average score of shows with ideology scores. In all, this process assigns scores to 81.0% of the radio data. The remaining radio content within each event is excluded from our ideology analysis.

Radio — Validation. The follow graph-based measure of ideology validates well, and is readily interpretable. On the radio side, we evaluate it at the show level, taking as reference labels whether a show airs more on public radio or non-public, commercial radio, because of the well-known fact that non-public, commercial talk radio skews heavily conservative politically.[200] (According to our schedule data, all but one show airs at least 99% on one or the other.) We find that the continuous ideology scores achieve an AUC of 0.933 at predicting these public/non-public labels. More qualitatively, Twitter users with the highest or most conservative scores include well-known right-wing hosts like Mark Levin, Sean Hannity, and Dana Loesch, while the lowest or most liberal scores belong to users who are affiliated with NPR. Other reassuring patterns the scores reflect include variation by geography (radio stations in urban areas tend to be associated with more liberal shows) and frequency band (AM talk is a particularly conservative genre). They also line up well with the follow-graph communities: The community we found to contain most conservative hosts has the highest average score by far at 0.605, compared to -0.09 for the centrist community and -0.529 and -0.607 for the two liberal communities.

As a robustness check, we also evaluated the simple option of considering statements from public radio stations to be liberal and statements from other stations to be conservative (relying on the well-known average political leanings of commercial talk and public radio[200]), and found very similar results.

Event Splitting. We use these ideological assignments to break each event into liberal and conservative sub-events for its discussion in the appropriate parts of radio and Twitter. We drop from the liberal side any event without at least 10 liberal items, and analogously on the conservative side, and also drop from both sides any event without at least 20 items in total. In total, we are left with 907 liberal events and 629 conservative events, each out of 1,694, with a total of 965 unique events between them.

3.4.6 Affect Identification

We measure three affect variables: negativity, emotionality and outrage. To detect them, we follow Yin *et al.* [250] in treating the problem as textual entailment and use an approach based on natural-language inference (NLI). The NLI task takes a "premise" and a "hypothesis," and estimates probabilities that the premise entails, contradicts, or is independent of the hypothesis. For each item (tweet or speaker turn) and each affect metric, the score is the estimated probability that the item, as premise, entails the hypothesis "This sentence is [negative/emotional/outraged]." We apply this procedure to each item in the newsLens events.

In more detail:

- We begin with a BART-Large model[251] fine-tuned for NLI.[‖]
- For each label ("negative," "emotional," or "outraged"), we take the text of the input item (tweet or speaker turn) as premise, and the sentence "This example is [label]." as hypothesis.
- The model's NLI head produces probabilities of entailment, contradiction and independence. We take the entailment probability as the probability that the input item satisfies the label, discarding the other two estimates.

A hypothetical example might have the premise "@username this is a terrible tweet, delete this" and the hypothesis "This example is negative," with the model estimating P(entailment) of 0.96. We would then take this value as the tweet's "negative" score. Manual spot-checks reveal that this procedure produces reasonable scores for both tweets and radio content.

3.4.7 Modeling Affect Dynamics

To analyze the time dynamics of affect, we fit logistic regression models to the data for each corpus, predicting the affect scores as a function of linear trends in within-event relative time. (The item-level scores are probabilities, so a linear model of their log-odds is a logistic regression.) The within-event relative time is the time of occurrence of a given tweet or speaker turn after the start of the event that contains it; for example, a tweet which was posted 3 minutes after the start of its event would have a within-event relative time of 3 minutes. This section describes these models, which we state here for one arbitrary affect variable A; the specifications are identical except for the choice of A. Between three affect metrics and three media, we fit 9 models and apply corresponding Bonferroni corrections to hypothesis tests.

We have a set of N items (tweets or speaker turns) which have been assigned to detected events, indexed by a variable i, each of which has a parent corpus $K_i \in \{\mathrm{E}, \mathrm{F}, \mathrm{R}\}$ (for elite, firehose and radio). Write $C \triangleq \{\mathrm{E}, \mathrm{F}, \mathrm{R}\}$ for the set of corpora, $I^{(c)}$ for the set of items in corpus c, $\mathrm{ind}(I^{(c)})$ for the indices of items in $I^{(c)}$, and finally $S^{(c)}$ for the set of events detected in corpus c.

[‖] https://huggingface.co/facebook/bart-large-mnli.

Each item in each corpus c also has a within-event relative time $t_i \geq 0$, an event ID $S_i \in S^{(c)}$ indicating which event it was assigned to, and an affect variable $A_i \in [0, 1]$, interpretable as a probability of the item displaying the underlying affect state.

For notational simplicity, let $L_i \triangleq \log(A_i/(1 - A_i))$ be the log-odds of affect, $D_i(s) \triangleq \mathbb{1}\{S_i = s\}$ be a dummy variable for whether item i is in event s, and $G_i(c) \triangleq \mathbb{1}\{K_i = c\}$ another dummy for whether item i is in corpus c. The ϵ_i terms below are error terms.

Single-corpus models. First, we model the log-odds of affect as a linear function of relative time:

$$L_i = \beta_1 t_i + \beta_0 + \epsilon_i \tag{1}$$

for all $c \in C$ and $i \in \text{ind}(I^{(c)})$.

Next, we test for time trends net of event-level effects by modeling affect as a linear function of relative time, controlling for event fixed effects (note that we cluster the standard errors by event):

$$L_i = \beta_1 t_i + \sum_{s \in S^{(c)}} \beta_s D_i(s) + \beta_0 + \epsilon_i \tag{2}$$

again for all $c \in C$ and $i \in \text{ind}(I^{(c)})$.

Multi-corpus models. To test for inter-corpus differences, we take pairs of corpora (elite/firehose, elite/radio and firehose/radio) and model them as linear functions of relative time, corpus, and the interaction effect. For corpora $c_1, c_2 \in C$, we have:

$$L_i = \beta_3 t_i + \beta_2 G_i(c_1) + \beta_1 t_i G_i(c_1) + \beta_0 + \epsilon_i \tag{3}$$

where $i \in \text{ind}(I^{(c_1)}) \cup \text{ind}(I^{(c_2)})$. The coefficients β_1 then indicate the magnitude of the inter-corpus difference in relative-time trend.

As above in the single corpus case, we also fit models with event fixed effects, again clustering the standard errors by event. For linear relative time, we have:

$$L_i = \beta_3 t_i + \beta_2 G_i(c_1) + \beta_1 t_i G_i(c_1) + \sum_{s \in S^{(c)}} \beta_s D_i(s) + \beta_0 + \epsilon_i \tag{4}$$

where $i \in \text{ind}(I^{(c_1)}) \cup \text{ind}(I^{(c_2)})$.

3.5 Results

3.5.1 Event Lifecycles

We first look at the average empirical CDF and PDF of *within-event relative times*. Recall that if we consider an event to start at the time of its first item (tweet or speaker turn), an item's relative time is its offset from its event's start time. If, for example, a tweet assigned to some event occurs exactly 100 seconds after the event starts (i.e., with its first tweet), the tweet's relative time is 100 seconds.

The average eCDFs and ePDFs are shown in Figure 3.2, and demonstrate the general point of faster Twitter lifecycles clearly: discussion on elite Twitter rises fastest, followed by the firehose, with both remaining at a higher relative rate than radio for a time, before both falling below the level of radio discussion for a longer time as events wind down.

Here, however, there is a complication: As shown in Figure 3.2, firehose discussion of Covid in March and April of 2020 does not follow this pattern. We identified Covid-related events by embedding a probe phrase ("pandemic of coronavirus disease 2019"), measuring the cosine similarity of the phrase's embedding to the centroid of each event's item embeddings, and selecting those with similarity above an empirical threshold of 0.1. Such events in the firehose rose more quickly than radio, but did not fall off as quickly as events in other years or (in 2020) about other topics. Inspecting these events revealed the reason: Firehose discussion of the pandemic at this point was constant, closely related to people's lived experience of the virus and countermeasures, rather than particular news stories, and difficult to separate into discrete events. This unusual behavior during an exceptional time for the media ecosystem throws into relief the more usual behavior, during other years and on other topics in 2020, of firehose discussion winding down faster than radio.

Examining the events concretely, we find that the cumulative fraction of elite Twitter discussion (i.e., the empirical CDF value) is greater than the cumulative fraction of radio discussion at every 15-minute query point out to 48 hours after event start. The eCDF of firehose discussion, if we exclude the Covid-related events from 2020, is also strictly between elite Twitter and radio at all query points. Two-sided percentile-bootstrap tests find that all of these differences are significant at the 5σ level ($p < 2.86 \cdot 10^{-7}$), except for the last 14 hours of firehose/radio differences. (All but the last three hours are significant at the lower $p = 0.05$ level.) Though not shown in Figure 3.2, results are similar for 2019, 2020 and 2021 subsets — see Section A.2 and Figure A.2 in the appendix for more details.

For another view of event onset and decay, we calculate means and standard deviations (SDs) of the relative times, in seconds, for each event. A later average relative time indicates an event whose center of mass, so to speak, occurs further after event onset (i.e., not rising as quickly), while a higher SD indicates one which is more spread out in time (i.e., not decaying as quickly). The relative-time means and SDs tell a similar story to the eCDF/ePDF analysis. The average within-event time is significantly lower at the 5σ level for both elite Twitter events ($t = -35.07$, df = 1192, $p < 2.86 \cdot 10^{-7}$) and firehose events excluding 2020 Covid-related events ($t = -10.24$, df = 463, $p < 2.86 \cdot 10^{-7}$) than radio events. SDs are also significantly lower for elite events ($t = -28.06$, df = 1192, $p < 2.86 \cdot 10^{-7}$) and firehose events

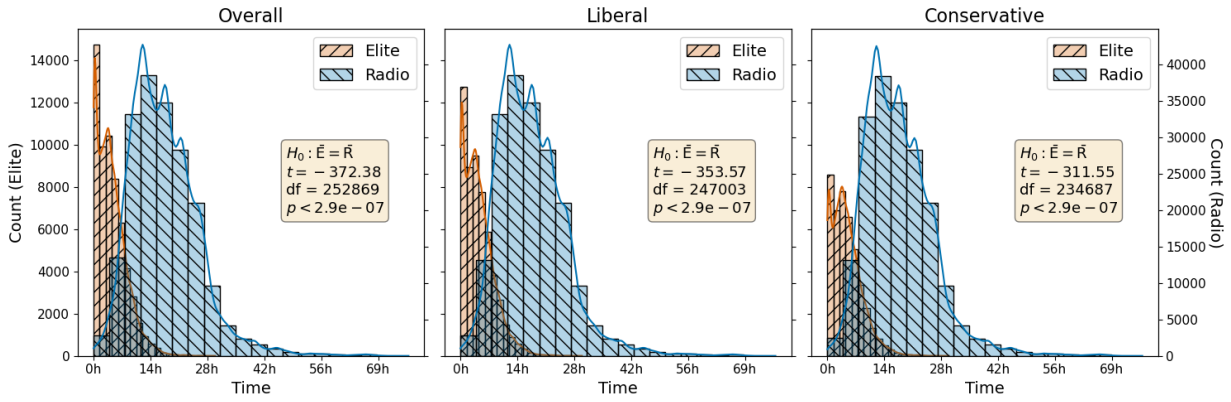


Figure 3.3: The distribution of within-event relative times of items: tweets for events on elite Twitter, and speaker turns for events on the radio. The average piece of discussion occurs earlier on Twitter than radio, and the average event winds down faster. Both effects are significant at the 5σ level, the difference in means by two-sided independent-samples t-tests as shown in the text boxes, and the difference in standard deviations (not shown) by two-sided bootstrap tests.

excluding 2020 Covid-related events ($t = -5.26$, df = 463, $p < 2.86 \cdot 10^{-7}$) than radio events. All tests are two-sided independent-samples t-tests. Results are similar if broken down by year; see Section A.2 and Table A.2 in the appendix for more details.

For elite Twitter and radio, we show the distributions of within-event relative times at the item level in the leftmost pane of Figure 3.3. (See below for the other two panes.) There are clear and large differences in both mean and SD between Twitter and radio, both of which are statistically significant at the 5σ level ($p < 2.86 \cdot 10^{-7}$). We use a two-sided independent-samples t-test for the difference in means and a two-sided bootstrap test for the difference in SDs.

In addition to the breakdowns by year above, we conduct several additional robustness checks (as discussed in Section 3.4): matching events across media for a paired comparison, breaking results down by ideological affiliation, and examining an additional set of manually detected events identified by keywords. An example manually detected event, Bernie Sanders's withdrawal from the 2020 presidential race, is shown in Figure 3.1. These checks, on different subsets and even with different event-detection methods, reinforce the conclusions of the main analysis: Discussion of the typical event on Twitter both rises to its peak and falls off much faster than on the radio.

3.5.2 Event Matching

Our first robustness check is to examine the 38 matched triples of (elite, radio, firehose) events, totaling 114 events. Results on these events are quite similar to those on the broader corpus. This time using two-sided paired t-tests, average within-event times are once again significantly lower for elite Twitter ($t = -6.80$, df = 37, $p < 2.86 \cdot 10^{-7}$) and firehose events ($t = -3.54$, df = 29, $p = 0.001$) than for radio events, as are SDs for elite ($t = -7.37$, df = 37, $p < 2.86 \cdot 10^{-7}$) and firehose ($t = -2.33$, df = 29, $p = 0.027$) Twitter. Comparisons to the firehose exclude triples whose firehose member is a 2020 Covid-related event.

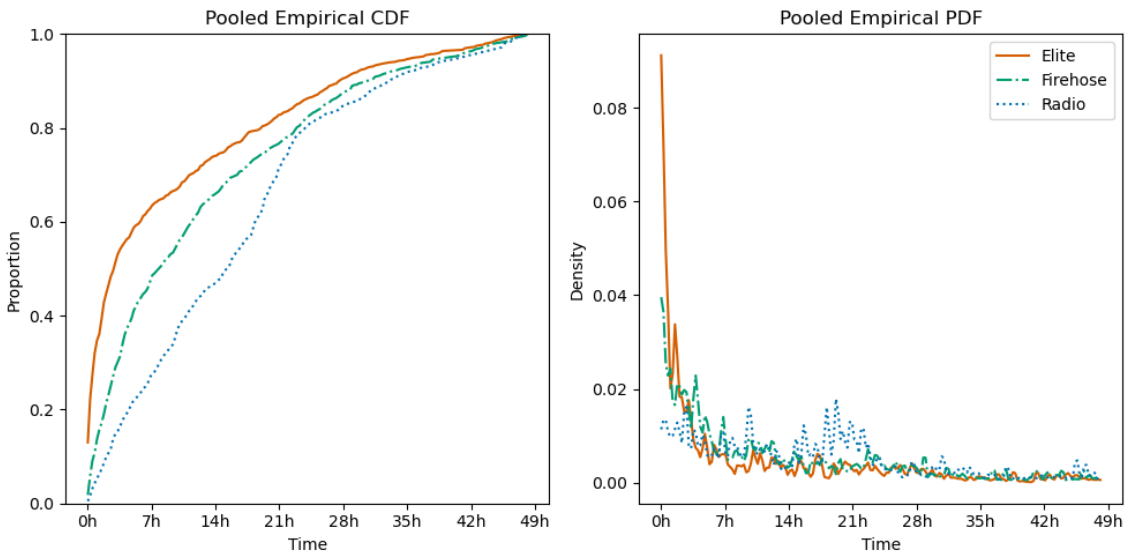


Figure 3.4: The average empirical CDF and PDF of keyword mentions of the 10 manually detected events. For each event in each corpus, we find the time after event occurrence of each mention, compute the empirical CDF at 15-minute resolution and average the CDFs over the 10 events. The PDF is obtained by a central difference estimate of the derivative of the CDF. Only mentions occurring within 48 hours of the event are shown, and we count each mention rather than each piece of content. Note the clear pattern of a spike of early elite-Twitter discussion, with the firehose lagging a bit behind, rapidly decaying into a period of more discussion on the radio than in either segment of Twitter.

The empirical CDFs, presented in Section A.3 and Figure A.3 in the appendix, similarly show elite Twitter discussion rising and falling faster than radio, with firehose discussion in between.

3.5.3 Manually Detected Events

Results for manually detected events are quite similar to those found with the newsLens-detected events. Figure 3.4 summarizes the time courses of the events, and shows that as in Figure 3.2, elite Twitter discussion both rises and falls fastest, followed by the firehose, with radio responding slowest. We also examined relative-time means and SDs. To avoid drift in the real-world events associated with our keywords, only mentions within 4 days of event start are considered. In 9 of 10 cases, except for Purdue Pharma's bankruptcy, Twitter discussion has a lower average relative time than radio discussion. Over all 10 events, the average within-event time of an elite Twitter mention is 65,133 seconds, vs 100,065 for radio, a difference of 53.6%. The SDs show a similar pattern, with 8 of 10 larger for radio than for Twitter. The average of the event-level SDs of Twitter events' relative times is 83,412 seconds, while on radio the analogous figure is 92,063 seconds, a 10.4% difference.

Finally, we looked by hand for newsLens events which matched manually detected events. Because news stories can be large and complicated, with many degrees of freedom in dividing them up, not all manually detected events could be matched to newsLens events. We were able to identify three events,

however, in both Twitter and radio. John Bolton's firing as National Security Advisor, the announcement of the 2019 Trump impeachment inquiry, and Elizabeth Warren's withdrawal from the 2020 presidential race all showed similar patterns through keyword counts to what we found by automatic detection: discussion on Twitter rising sooner and falling faster than on the radio.

3.5.4 Ideological Differences

We also examined whether event lifecycles differ between the liberal and conservative sides of Twitter and radio. Ideology is one of the most important organizing principles of both radio[200] and political Twitter,[252] and differences in event dynamics would point to differences between the ideological sides in how the two media are organized. As discussed in Section 3.4.5, we perform this analysis only for elite Twitter and radio because of the difficulty of drawing ideological inferences for firehose users.

The results are summarized (at the item level) in the rightmost two panes of Figure 3.3. For both liberals and conservatives, Twitter discussion tends to be shorter and finish faster than on radio. Both effects for both ideological groups are statistically significant at the 5σ level ($p < 2.86 \cdot 10^{-7}$), but are slightly stronger for liberals than conservatives, perhaps reflecting a greater role for Twitter in liberal discourse. The empirical CDFs (shown in Section A.4 in the appendix) tell a similar story to the elite and radio curves in Figure 3.2, with Twitter discussion rising and falling faster than on the radio for both liberals and conservatives, and for all three years.

Results are also similar if the manually detected events are broken down by ideology. Of the 20 liberal or conservative sub-events, 18 have a higher average within-event time for radio than for Twitter, and most also have a higher SD of within-event times on the radio side than on Twitter.

3.5.5 Affect Levels

Turning from events' lifecycles to their affective content, we next aim to test whether affect differs between Twitter and radio. We examine three affect metrics: negativity, emotionality and outrage. Our analysis includes the 2020 Covid-related firehose events; results are similar if they are excluded.

Results are shown in Figure 3.5, averaging item-level scores up to the event level. Results at the item level are similar, as shown in the appendix Section A.5 and Figure A.6. We find large differences on all three affect measures, always in the same direction: Firehose Twitter is most negative, emotional and outraged, while radio is least, and elite Twitter is in between. All inter-corpus differences are statistically significant at the 5σ level ($p < 2.86 \cdot 10^{-7}$) by two-sided Mann-Whitney U-tests.

3.5.6 Affect Dynamics

We also find strong, consistent and statistically significant differences between media in the evolution of affect over an event's lifecycle. This analysis again includes the 2020 Covid-related firehose events, and the results are again similar if they are excluded. (For an illustration of this point, see the lower left pane of Figure 3.6.) We test for linear time trends, and especially inter-corpus differences in these trends, for all

three affect metrics, but focus our discussion especially on negativity and outrage, which are of greater theoretical interest.[206] "Emotional," which can also involve positive emotions, is not closely correlated with the other two scores: Item-level emotionality has a 0.364 correlation with outrage and 0.260 with negativity, while the correlation between outrage and negativity is 0.729.

Figure 3.6 shows average negativity and outrage levels over time, revealing substantial differences between our corpora, and between radio and Twitter. Notably, there is one clear commonality across the three datasets: A few very long-lasting events are especially outraged. Out of 1,694 events, 51 persist past 24 hours on elite Twitter or 48 hours in the other two corpora, and content after this point is 17% more negative and 27% more outraged than earlier content.

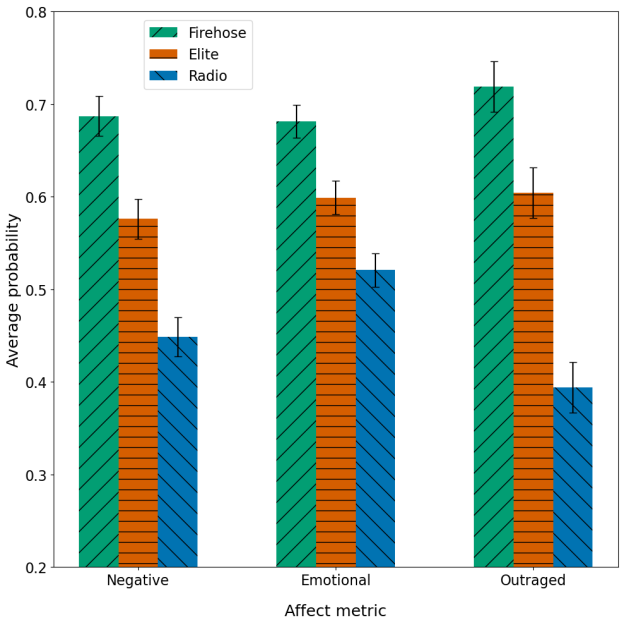


Figure 3.5: Event-level average scores for our three affect metrics. The underlying item-level scores are probabilities, and range from 0 to 1. 95% confidence intervals are shown. All inter-corpus differences are significant at the 5σ level ($p < 2.86 \times 10^{-7}$) by two-sided Mann-Whitney U-tests.

Results of the logistic regression models described in Section 3.4.7 are shown in Table 3.3. All six of the single-corpus models for negativity and outrage have statistically significant linear time trends at the Bonferroni-corrected 5σ level ($p < 3.18 \cdot 10^{-8}$), using both here and below the usual two-sided t-tests of the parameters. The signs, however, differ between radio and Twitter. Both negative and outraged affect show declining trends (negative coefficients) for elite and firehose Twitter, and an increasing trend (positive coefficients) for radio. Emotionality is similar in finding significant declining trends for elite and firehose Twitter, but no significant trend — i.e., a stable level over time — on the radio.

To test our main question of whether these effects are significantly different from each other, rather than only from 0, we fit additional models. We took pairs of corpora (i.e., firehose/elite, elite/radio, radio/firehose) and modeled item-level affect scores as a function of relative time, corpus and the interaction. All of the interaction terms are significant at the Bonferroni-corrected 5σ level, indicating significant differences in event lifecycles between media.

But these patterns leave open an important question: Are the differences because of time trends within individual events, or because of differences between media in how long different kinds of events last? We tested this question with further regression analysis, fitting the same models as above but with event fixed effects. Results are again shown in Table 3.3. If most variability is at the event level, with time trends coming from differences correlated with outrage and negativity in how long events persist, we should expect now not to find significant time trends or differences in time trends. And indeed we do not; neither

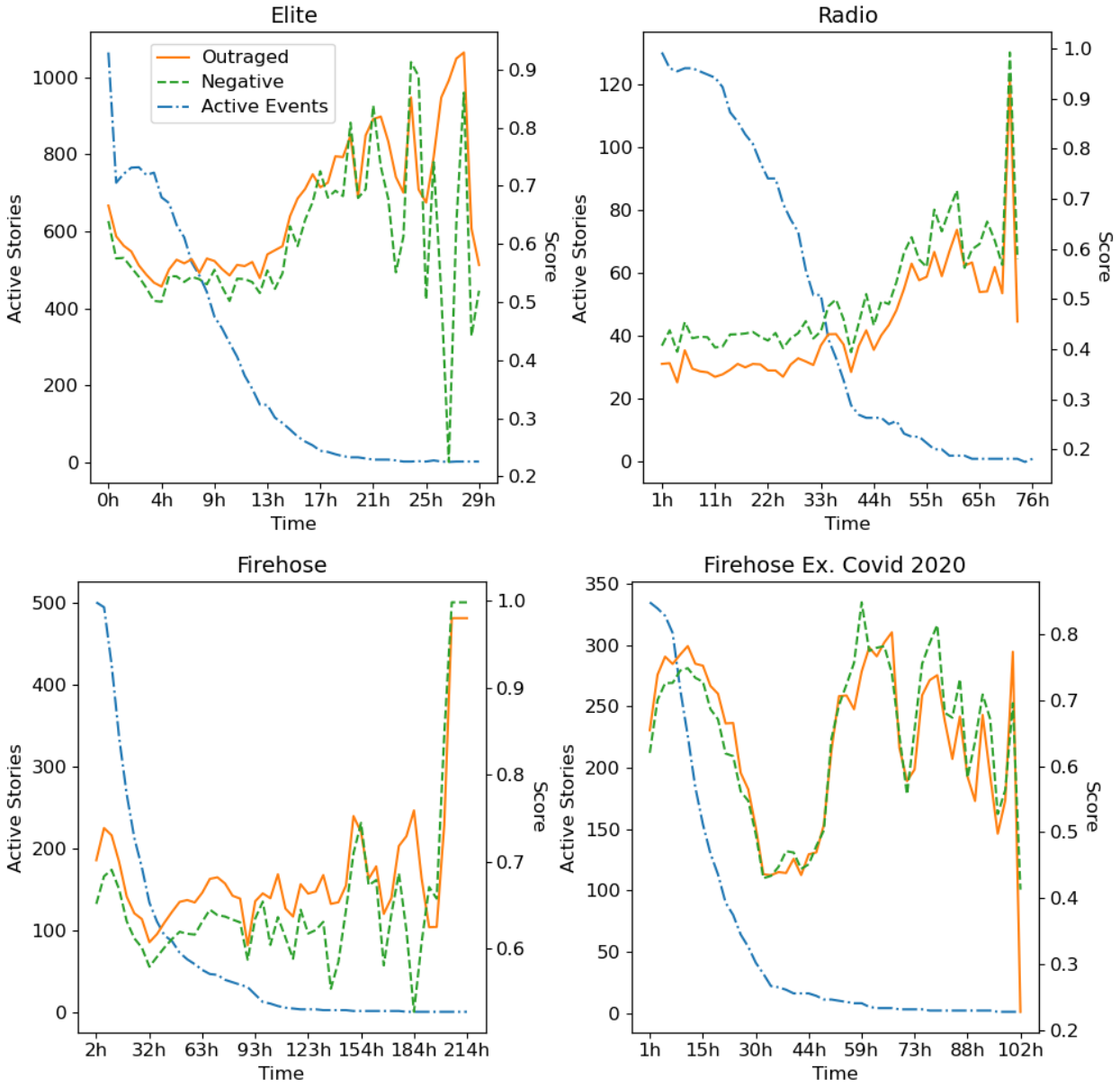


Figure 3.6: Average levels of negativity and outrage over a news event's lifecycle. We also show the number of detected events which are active (i.e., contain tweets or radio speaker turns) at each point in the lifecycle. These figure are calculated by dividing relative time into 200 equal-sized bins (i.e., half-percentiles), within which we average negativity and outrage scores and count the number of active stories. There are notable differences between all three corpora, but especially between radio and the two Twitter datasets. Note that, for the firehose, the distribution of outrage and negativity is similar with and without 2020 Covid-related events but the number of active stories falls off much more quickly without them.

Affect	F.E.	Firehose	Elite	Radio
Emo.	N	**-0.153** (0.003, $p <$ $p_{5\sigma}$)	**-0.160** (0.006, $p <$ $p_{5\sigma}$)	+0.009 (0.004, $p =$ 0.04)
	Y	-0.023 (0.026, $p =$ 0.37)	+0.010 (0.015, $p =$ 0.51)	+0.000 (0.015, $p =$ 0.99)
Neg.	N	**-0.071** (0.007, $p <$ $p_{5\sigma}$)	**-0.113** (0.014, $p <$ $p_{5\sigma}$)	**+0.074** (0.007, $p <$ $p_{5\sigma}$)
	Y	+0.076 (0.054, $p =$ 0.16)	-0.034 (0.025, $p =$ 0.16)	-0.018 (0.030, $p =$ 0.55)
Out.	N	**-0.074** (0.006, $p <$ $p_{5\sigma}$)	**-0.106** (0.012, $p <$ $p_{5\sigma}$)	**+0.104** (0.006, $p <$ $p_{5\sigma}$)
	Y	+0.051 (0.044, $p =$ 0.24)	-0.073 (0.022, $p =$ 0.00)	+0.006 (0.025, $p =$ 0.79)

(a) Linear main effects of relative time, with and without event fixed effects.

Affect	F.E.	Elite vs. Firehose	Elite vs. Radio	Radio vs. Firehose
Emo.	N	**-1.098** (0.055, $p <$ $p_{5\sigma}$)	**+0.425** (0.020, $p <$ $p_{5\sigma}$)	**+0.162** (0.012, $p <$ $p_{5\sigma}$)
	Y	+0.104 (0.123, $p =$ 0.40)	-0.026 (0.044, $p =$ 0.55)	+0.021 (0.049, $p =$ 0.67)
Neg.	N	**-0.815** (0.112, $p <$ $p_{5\sigma}$)	**+0.376** (0.035, $p <$ $p_{5\sigma}$)	**+0.280** (0.023, $p <$ $p_{5\sigma}$)
	Y	-0.350 (0.202, $p =$ 0.08)	+0.069 (0.067, $p =$ 0.30)	-0.121 (0.101, $p =$ 0.23)
Out.	N	**-0.761** (0.097, $p <$ $p_{5\sigma}$)	**+0.393** (0.029, $p <$ $p_{5\sigma}$)	**+0.369** (0.020, $p <$ $p_{5\sigma}$)
	Y	-0.628 (0.180, $p =$ 0.00)	+0.197 (0.065, $p =$ 0.00)	-0.027 (0.082, $p =$ 0.74)

(b) Interaction effects of relative time and corpus from models fit on pooled pairs of corpora with relative time main effects, a corpus dummy and the interaction effect. Dummies are for elite in the first column and radio in the other two (i.e., the estimates are for the difference in time trend going from firehose to elite, from elite to radio, and from firehose to radio).

Table 3.3: Results from regression models predicting affect (indicated in the first column of the tables with "Emo" = emotional, "Neg" = negative, and "Out" = outraged). Coefficients are given first, with standard errors and p-values after in parentheses. Models with event fixed effects are marked with 'F.E.' = Y and use standard errors clustered at the event level. Effects significant at the Bonferroni-corrected 5σ level ($p < 3.18 \cdot 10^{-8}$), under the usual two-sided t-tests of the parameters, are in **bold** and show $p < p_{5\sigma}$. Relative time variables were centered and scaled prior to fitting. Coefficients and SEs are rounded to three decimal places and p-values to two decimal places.

any within-event time trends nor any inter-corpus differences are significant at the Bonferroni-corrected 5σ level after including event fixed effects, and usually have much smaller magnitudes. The observed patterns in the distribution of affect, including the stability over time of emotionality on the radio, are almost all due to differences (or the lack thereof) at the event level.

3.6 Discussion

We examined how news event lifecycles differ between two prominent communications media: Twitter and talk radio. Leveraging large-scale, comprehensive data from both media, we find that Twitter has a systematically shorter attention span than radio, with discussion of events not only beginning sooner

but also decaying faster. These findings are robust, with similar results in all three years, among liberals and conservatives, within pairs of events matched across media, and using an additional set of manually identified events. Twitter is also much more negative, emotional and outraged than radio, with differences both in the average levels of these affect variables and in their dynamics over time. Outrage and negativity in particular are more fleeting on Twitter, with the differences in dynamics driven by differences at the event level in how long more and less outraged events persist.

The scale and representativeness of the datasets leading to these conclusions are an important contribution of our analysis: Far from covering only one station, show, or Twitter account, we can take a systematic look at a major broadcast medium and compare it to both elite and mass audiences on social media. Indeed, including both sides of Twitter allows for a particularly stark demonstration of medium-specific effects: We find that even among the general Twitter user base (i.e., the public rather than the press), news discussion rises and falls faster than among professionals on the radio.

These differences are likely to be rooted in each medium's affordances for discussion,[42] and especially in how social media's affordances differ from those of a broadcast environment.[12] The clearest example we find of a tie to specific medium affordances is in the event-level origin of time trends in affect. Our results show that radio and Twitter differ in how hospitable they are to different kinds of stories, leading emotionality, negativity and outrage to be less persistent (but more intense) on Twitter and more persistent (but less intense) on the radio.

The average levels of affect variables also suggest connections to platform design and affordances. We find that the Twitter firehose is the most outraged, followed by elite Twitter and then by radio. In light of recent findings[253] that elite incivility on Twitter may spread by a reinforcement process, driven by positive feedback from the audience, our results support the idea that audience demand may be an important factor in the higher level of outrage in elite Twitter than on the radio. The underlying affordance is then that on Twitter, journalists and media personalities can get audience feedback easily and very rapidly, much more so than for radio. We cannot, however, say for certain that these audience dynamics are responsible, because of the lack of audience engagement data. It is, rather, a plausible hypothesis for future work.

Another valuable subject for future work is which (if any) useful intervening variables exist between platform affordances and high-level discourse phenomena like sentiment. Differences in news cycle speed, for example, may lead to more opinionated content in one medium than another (because opinion is easier to produce rapidly than reporting or interviews), leading in turn to more outrage. We have not explored such effects, including the difference between news and opinion, in part because of the difficulty of operationalizing and measuring opinion in media which lack a clear editorial separation between it and news. Doing so would, however, be a good subject for future work.

Despite connections to affordances, our findings also rebut a simplistic view of content as entirely determined by properties of media. Firehose discussion of Covid in 2020 deviated sharply from the usual patterns, reflecting the power of real-world events to make themselves felt. The medium may be the message,[42] but not the only message. Determining more precisely how affordances influence both the speed and sentiment of the news cycle is an important subject for future research, including in media other than those we consider here.

The recent sale of Twitter also makes this work timely. There are an increasing number of similar platforms, based on real-time sharing of text posts, which gives us good reason to think the conclusions may apply more broadly. And yet the reduced availability of Twitter data for research has also made it harder to observe journalists' interactions with each other on social media. Using data from when this activity was last concentrated and readily observable, we can still provide a comprehensive investigation.

Overall, our results contribute substantial empirical evidence to the literature on news cycles and their determinants, including the first systematic comparisons of Twitter and traditional media on their a) levels of outrage and b) rates of decay of attention to news. The results also illustrate an important channel of influence Twitter and its propensity for outrage may have on other media: framing and agenda-setting by the simple expedient of being the first to speak.

Media's ongoing shift toward digital formats provides a final reason these results matter. If journalism and media continue to shift to online venues, the social media biases we find here may become more influential. The fast forgetting and proclivity for outrage we find characterize Twitter may thus be not only features of the present, but a preview of the future.

Chapter 4

Treatment Effect Estimation via LLM: Predicting Opinion Change

4.1 Introduction

We are interested in predicting response to persuasive messages and understanding the underlying persuasion process. This work builds on extensive theoretical development of the bases of opinion in political science[10,39] and psychology,[9] and some empirical evaluation of the ease of predicting persuasion.[50] We pursue this goal in the setting of randomized experiments because of the powerful identification guarantees provided by the randomization. Concretely, this question cashes out in the problem of *uplift modeling*,[254] also called *treatment effect estimation* or *heterogeneous treatment effect* modeling: We want to determine how much each experimental subject's measured outcome changed as a result of the treatment.

Both determining these individual-level scores and evaluating the models that assign them is difficult, because of what's known as the *fundamental problem of causal inference*: Counterfactual outcomes, by definition, are not observable. Except in a simulation study, we cannot simply both treat and not treat every experimental subject and then calculate the treatment effect as the difference. Instead, we need to take advantage of the treatment and control groups, which are comparable in expectation by virtue of random assignment.

A number of methods have been developed for doing so (see Section 4.3 and Section 4.5.2), but these methods, diverse though they are, share a common shortcoming: They estimate the individual-level effects de novo for each experiment, rather than take advantage of pretraining. The rise of large pretrained language models, or LLMs, in the last few years suggests the intriguing possibility of improving on these methods. These LLMs have acquired substantial knowledge of the world and people from their pre-(and post-)training, which we may be able to use for more statistically efficient estimation of individual treatment effects.

Methods able to do so would have obvious practical applications, allowing more efficient targeting of interventions in fields from marketing to political communications to medicine. Pilot studies used to target these interventions could become smaller, or even be skipped entirely. But practical applications are

not the only reason to be interested in such methods: They also have substantial theoretical interest for social science. Persuasion and opinion change are complicated, contextual, and difficult to study among people,[50] with the field held back both by the inherent complexity of the phenomenon but also by the difficulty of iterating on new theories. Demonstrating that LLMs can be a useful proxy for humans in this domain would be helpful, allowing hypothesis develoment and refinement through an analog of the animal studies used in medicine. It is exactly methods able to do this that we aim to develop in this chapter.

4.2 Related Work

4.2.1 Public Opinion

Philosophical views of public opinion or related concepts, like Rousseau's "general will,"[255] have a long history. Scientific study of public opinion in the modern sense, however, began with the pioneering work of George Gallup in the 1930s,[109] building on the theoretical work of Lippmann[86] and Dewey.[87] Over the years, the field's conceptions of what opinion consists of have evolved from the early view of crystallized "true attitudes"[9] to theories like those of Converse that stress most people's lack of a coherent ideology.[10] Zaller [39] elaborated on these theories in a particularly influential way, developing a "receive-accept-sample" model of opinion that posits people receive external opinions, accept some according to their prior opinions and latent predispositions, and then randomly sample from those they've accepted when an opinion is needed.

Zaller's view of opinion places a particular emphasis on media effects, as a source of both opinions ("considerations" in his terminology) and cues about the salience of various considerations. He was not the first to emphasize media effects, however, with an appreciation of the role of the media going all the way back to Lippmann[86] and demonstrated in much empirical research over the years.[2,3] This view of the influence of the media motivates earlier chapters' interest in news cycle dynamics and how they may be changing, and provides an answer to the question of why media matters.

Another line of work on public opinion has considered the nature of persuasion and opinion change. From the beginning of modern public opinion research there have been critiques that its paradigm fails to consider the complex, socially situated, contextual nature of opinion,[8] and especially the process of inter-personal influence in forming and changing opinions.[5] In addition to research in disciplines like psychology,[9] political science has considered opinion change in the context of polarization[16,24,191] and in practical survey research.[256,257] The rubber meets the road in a large literature on message testing experiments,[258] which attempt to explore how attitudes shift in response to new information or frames. Our work on language models as models of persuasion response fits most directly into this latter literature, helping predict responses to these experiments and elucidate why and how opinion change occurs.

4.2.2 Language Models

Language models have recently demonstrated enormous improvements in their capabilities at a diverse set of tasks,[32,33] including especially the advent of consumer-facing general-purpose chatbots like ChatGPT. These advances have been powered mainly by increased scale of both models[27] and training datasets.[28,29] In what Sutton [259] called the "bitter lesson," a simple architecture at large scale will tend to outperform a more complex but less powerful approach.

The new large language models (LLMs), based on a deep learning approach,[120] typically use the transformer architecture.[135] They are trained with a masked[138] or more commonly autoregressive[260] objective, with autoregressive models usually preferred for the ease of sampling from them. In keeping with the "bitter lesson," though, the transformer architecture is not essential to making such models work: Large RNN language models[261] have also recently become available.

In addition to the importance of scale, these models are based on certain other fundamental ideas. An essential one is the *distributional hypothesis*, epitomized in J.R. Firth's saying that one "shall know a word by the company it keeps."[262] Structure, in other words, is inferred from co-occurrence. (Indeed, the same general idea occurs even beyond language models, as in vision-language architectures like CLIP[263] which learn from the co-occurrence of images and captions.) The representations that are inferred distributionally are themselves *distributed*,[129] spread across the model weights rather than having concepts localized clearly as in a knowledge graph or expert system. Finally, the pretraining and fine-tuning paradigm[264] based on the notion of *transfer learning* is the dominant way LLMs are trained and used. A "foundation" model[30] learns from a large corpus of content in its pretraining phase, and then is adapted by fine-tuning (further training) to a specific task, or is simply used without further tuning, often by representing other tasks as a text-to-text or sequence completion problem.[31]

Language models have also been used in public opinion-related settings, such as in attempts to extract measures of opinion from social media,[22,175,176] though this problem is challenging.[177,178] Finally, an exciting new direction is using LLMs to simulate respondents to public opinion surveys. Several recent studies have taken this approach,[34,71] part of a broader push to explore generative AI's potential in social science.[37] Our work in this dissertation, while not developing generative agents, falls into a similar paradigm of representing survey-related behavior with LMs.

4.2.3 Experiments and Treatment Effect Estimation

Causal inference has a long history as a research area in statistics. There are two common paradigms: the causal-graph approach pioneered by Pearl [265] and the potential-outcomes framework of Neyman and Rubin,[266,267] building on work by Fisher.[268] Randomized experiments typically employ the latter framework, as we do in this dissertation. The essential idea in an experimental setting is that using exogenous randomization to assign treatment, so that assignment is independent of pre-treatment covariates, *identifies* the average treatment effect (ATE).[269] For our purposes, identification, a fundamental notion in statistical inference, means that the effect could in theory be determined exactly with an infinite amount

of data. (More technically, the statistical model's map between parameter values and observed probability distributions should be one-to-one.) A more technical condition about non-interference between experimental subjects, known as SUTVA for the "stable unit treatment value assumption,"[100] is also commonly employed, though there are methods for dealing with violations of it.[65,102]

We are interested, however, in methods for estimating not average effects but heterogeneity in effects. This question is of obvious practical interest (e.g., figuring out which patients will benefit most from a drug), and has been approached in many ways over the years. Classically, researchers have estimated conditional ATEs within subgroups[266] defined by variables of interest like demographics, and blocking by such variables can improve power.[268] More recently, there has been work on estimating individual-level treatment effects.[270–272] These are never directly observable, except in simulation studies, but the methods can be evaluated over entire experiments with multiple treatment arms.[254,273] The task of estimating these individual-level treatment effects, as mentioned above, has a variety of names in different fields: treatment-effect estimation (TEE), or heterogeneous treatment effect (HTE) modeling, or uplift modeling. Our work in this dissertation fits into this latter literature, exploring the value of scale and pretraining in estimating such effects.

4.3 Problem Formulation

We can formally characterize the problem of interest with the Neyman-Rubin potential outcomes framework.[266,267] We will go through only the basics here; for more information on experimental design, an enormous and rich subject, consult a good textbook like that of Hinkelmann & Kempthorne. [274]

Consider a randomized experiment or randomized controlled trial (RCT) with N subjects and a binary treatment, so that each subject is in exactly one of the treatment and control groups. The experimental design can in general be more complicated than this: There can be more than two treatment levels, they can be assigned factorially or in other blocking designs even as far as a conjoint experiment, treatment levels can be ordinal or continuous (that is, we can ask "how much" treatment each subject got), and so on. To keep the analysis here tractable, however, we consider only simple binary treatments. Treatment is then assigned randomly according to some *assignment mechanism* which is independent of the outcomes; in the simplest case, treatment is i.i.d. Bernoulli with equal-sized groups in expectation: $T_i \overset{\text{iid}}{\sim} \text{Ber}(\frac{1}{2})$. The assignment mechanism can also be more complicated, as with unequal group sizes, complete randomization, stratified assignment and so on. Unlike with the assumption of binary treatment, however, we do not need to specify a particular mechanism; anything satisfying the "stable unit treatment value assumption" (SUTVA)[275,276] of non-interference between units will do.

In the Neyman-Rubin model, each of the N subjects in this case is has two *potential outcomes*, denoted $Y_i(1)$ and $Y_i(0)$, for $i \in \{1, ..., N\}$, which specify the results if subject i respectively is or is not treated. The observed outcome, which is the potential outcome under the actual treatment assignment, is denoted $Y_i \triangleq Y_i(T_i)$. Except in simulation studies, of course, we cannot simply observe both potential outcomes for all subjects. More sophisticated statistical methods are thus required to estimate the effects of the

treatment.

Nevertheless various quantities of interest are identified: Even if difficult to estimate, they admit estimators which, in the limit of infinite data, converge in probability to the true values. In particular, this setup identifies the quantity typically of interest in an RCT, the average treatment effect (ATE) $\tau = \mathbb{E}[Y(1) - Y(0)]$, and lets us estimate it with the usual difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{|S_1|} \sum_{i \in S_1} Y_i(1) - \frac{1}{|S_0|} \sum_{i \in S_0} Y_i(0), \tag{4.1}$$

where $S_j = \{i : T_i = j\}$ for $i = 0, 1$.

There are some subtleties here worth bearing in mind: Different formulations of the Rubin causal model differ in the details of what is fixed and what is random. Potential outcomes are typically taken to be fixed but (partially) unknown, with randomness coming from the treatment assignment. Unless one is conditioning on the observed experimental population, the set of subjects is also often taken to be randomly sampled from a larger "superpopulation" of potential subjects, with this latter source of randomness allowing us to discuss such quantities as the expected value of the ATE.

We are interested, however, not in the well-explored average treatment effect but in conditional treatment effects that depend on properties of each subject. Let each subject i have a set of covariates $X_i^{(j)}$ indexed by $j = 1, ..., D$ where there are D covariates in total, arranged in the same way for each subject into vectors X_i. A simple and common way to approach this question of conditional treatment effects is to condition on certain values of the covariates, letting us ask questions about *subgroup effects* or *conditional ATEs* of the form

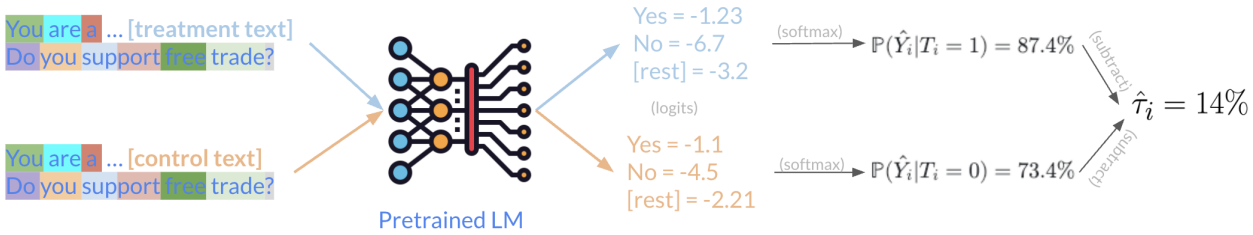
$$\hat{\tau}_{j:c} = \mathbb{E}[Y(1) - Y(0) | X^{(j)} = c], \tag{4.2}$$

where the expectation is conditioned on a categorical variable taking a particular level (or in general, the intersection of multiple such events). Statistical analysis of experiments often asks about quantities of this form, which include questions like how well a drug works for women specifically, rather than all people, and whether voter contact is more effective at driving turnout among those who voted in the previous election. There is a natural and well-behaved estimator as well: Simply restrict the sample to those subjects with the appropriate values for the variables, and estimate the ATE among them.

But this formulation is less general than we'd like. It works only for categorical covariates, and the variables and subgroups must be specified in advance. A more general quantity, which we'd like to estimate, is the individual-level treatment effect $\tau_i = \mathbb{E}[Y_i(1) - Y_i(0)]$. Regrettably, this quantity is identified only under strong assumptions[270] which don't hold in general. We can, however, identify a closely related function, a generalization of the subgroup effects above that Künzel *et al.* call the *CATE function*:[270]

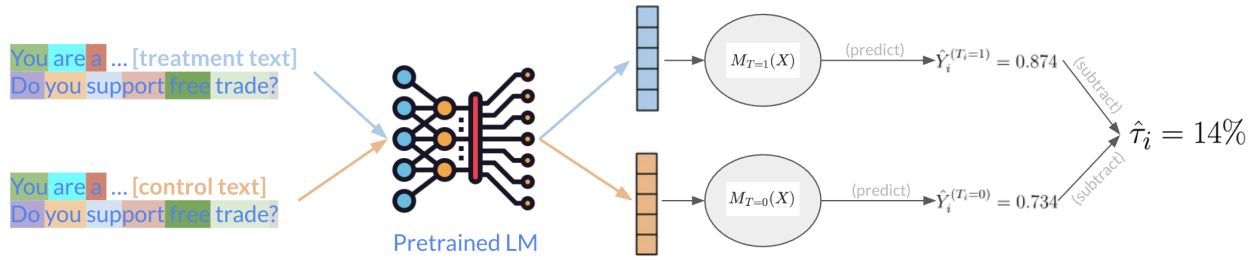
$$\hat{\tau}(x) = \mathbb{E}[Y(1) - Y(0) | X = x] \tag{4.3}$$

which conditions on the vector of covariates in general, not just a small number of discrete subgroups. The



(a) Text2Text.

(b) Persona / Inference-Time Compute.



(c) Representational Regression (RepReg).

Figure 4.1: Explanatory diagrams for our LLM-based treatment effect estimation methods.

best estimator of the CATE function (in the sense of minimizing mean squared error) can also, happily, be shown to be the best estimator by MSE of the individual treatment effect. Only mild assumptions are needed for identification here, namely the usual observational causal-inference assumptions of conditional ignorability and overlap. See the full derivation in Künzel *et al.* [270] for more details.

There are a number of methods in the literature to estimate $\hat{\tau}(x)$, some of which we discuss as baselines below. We'll next outline our methods for estimating this quantity, present the baselines, evaluation metrics and datasets, and discuss results.

4.4 Methods

This section presents several methods we have developed and evaluated (see below) for this task. The methods are summarized visually in Figure 4.1.

4.4.1 Approach 1: Text2Text

Our first approach (diagrammed in Figure 4.1a) has a language model perform entirely zero-shot inference about the probabilities of experimental subjects (persons) having particular outcomes. We represent each subject as a prompt, which

1. Describes the situation (a survey is being taken);
2. States the characteristics, from input data, of the person taking the survey (demographics, etc);
3. Gives the survey text, with different versions for treatment and control conditions;
4. Finally, asks the opinion question the real human respondents faced.

For each subject, we feed the treatment and control versions of this prompt into a language model and calculate the probability of each answer choice (as the next token) for each condition. For answers on a Likert scale,[*] as our evaluation datasets generally are, the predicted treatment effect is simply the expected change in the answer from treatment to control. That is, if the answer choices are integers from m to n, we estimate $\tau_i = \mathbb{E}[Y_i(1) - Y_i(0)]$ as

$$\hat{\tau}_i^{(\text{exp})} = \sum_{j=m}^{n} j \cdot (\mathbb{P}_{LM}(Y_i(1) = j) - \mathbb{P}_{LM}(Y_i(0) = j)) \tag{4.4}$$

where $\mathbb{P}_{LM}$ is the distribution over answers from a language model prompted as described above. The distribution depends on the subject i and the treatment condition, but we've omitted both from the notation for simplicity. Note that we do not fine-tune the input LM in any way; evaluation is entirely zero-shot. This approach can also be cast as a variant or relative of the T-learner meta-algorithm from Künzel *et al.*.[270]

4.4.2 Approach 2: Persona / Inference-Time Compute Scaling

Our second treatment-effect estimation approach builds on the text2text method by applying more compute at inference time. This approach fits into the same paradigm as much recent work on inference-time compute and the performance improvements available from scaling it. Techniques vary – from chain-of-thought prompting[277] to RL-driven reasoning methods[278,279] and others – but it often helps.[280]

In our case, rather than simply asking for the LM's probability distribution over possible answers, we prompt[†] the model to first a) work up a paragraph-length persona for the person who took the survey, based on the provided demographic characteristics, and then b) reason through how such a person would think about the question. The prompt specifies that the model should not answer the question before

[*] The outcome scales used in survey questions normally are, or for our purposes are equivalent to, Likert scales. See Section 4.4.4.

[†] Generation in both this chapter and the persona experiments of Chapter 5 uses beam search multinomial sampling with 5 beams, sampling temperature of 0.5, and a maximum of 512 new tokens.

doing this, so that more compute is applied, and then to output an answer marker "=>" before the answer. (In the occasional case where this marker is not output before the end-of-text token, we manually add it at the end of the generation.) The final step is to run the generation output again through the forward pass and obtain the distribution over answers, as in the text2text approach, but conditioned now on both the survey text and the generated persona and reasoning. We discuss the changes to prompts with examples in Section B.1.

4.4.3 Approach 3: Representational Regression (RepReg)

Our third approach to estimating treatment effects (depicted in Figure 4.1c) uses the LM as a feature preprocessor or "subject encoder." We take the same zero-shot prompt as used for the text2text approach above and obtain from the LM some representation of it, with our experiments evaluating both

3(a) The mean-pooled last hidden states, and

3(b) The vector of probabilities assigned by the LM over the answers.

These vector representations, one per subject or (treatment, subject) pair for counterfactual prediction tasks, are then fed into the same baseline prediction methods discussed in Section 4.5.2. As this method, unlike the others, requires training data, we employ three-fold cross-validation to generate an out-of-fold prediction for each subject, training three times on 2/3 of the data and predicting to the other third. (The splits used are the same ones as for the baselines.)

At first glance, it may seem odd to think that this method could improve performance: We're simply taking the same predictor variables being given to the baselines, combining them with a script that's constant for all subjects, and generating vector representations. The hope, however, is that it can achieve greater *statistical efficiency*: By relying on the LM's pretrained world knowledge, we may be able to get better representations than a simple vector of predictors, and achieve better performance for the same amount of experimental training data (or, equivalently, the same level of performance with less data). Note also that 3(b) is effectively a bet on miscalibration: If it works better than the text2text approach that simply computes the expectation, it implies the probabilities contain useful signal but are systematically skewed or miscalibrated in some easily correctable way.[281]

4.4.4 Adaptability to Various Outcome Scales

The outcome measures in our evaluation datasets are on Likert scales, which figures into the design of the approaches discussed above. In general, of course, there will be other kinds of outcomes we might wish to model. Our approaches are adaptable to a number of other such settings.

We'll consider the text2text and persona-based cases first, where our choices are constrained by the fact that language models generate (distributions over) discrete tokens:

Binary outcomes Choices between two outcomes, especially yes/no questions, are common in public opinion research and for our purposes are a special case of the Likert scale. The same expected-value

method of generating a prediction applies. If there is not a natural ordering of the two answer choices, we can pick one arbitrarily and (up to a change of sign) obtain a scalar prediction of the treatment effect.

Categorical outcomes Handling a categorical set of answer choices, without an ordering over them, is more difficult. In general, however, we can always reduce it in a one-vs-rest fashion to binary outcomes and apply the same techniques outlined above. The structure of any particular set of answer choices may also allow other options.

Continuous outcomes People do not generally come up with and recite arbitrary real-valued numbers in response to a survey question. Certain question formats do come close to eliciting continuous responses, such as feeling thermometers[44] or questions about the respondent's income. These can be handled in different ways; answers to income questions in surveys are commonly binned, while feeling thermometer questions have few enough answer choices to be able to handle them using the same techniques as for Likert scales.

Notably, because of LMs' tokenized outputs, none of these cases have a role for common regression or classification methods: The language model itself fills that role by predicting probabilities of outcome classes, from which we need to derive a scalar score.

For what we've called representational regression, however, there is more freedom in how to generate predicted treatment effects, including not only regression but also classification models where appropriate. For Likert-scale data, a rigorous and statistically motivated, but complicated, approach is to use an ordered logit model[282] to predict the response as a function of covariates. We adopt instead a commonly used approximation[283] which treats the Likert outcomes as ordinal data, and uses a regression model (linear regression or a random forest regressor, depending on the experiment). For our purposes, this approach has the advantage of simplicity: It directly predicts the sort of single scalar outcome we need for an estimate of treatment response.

4.5 Experiment Setup

4.5.1 Evaluation Metrics

Evaluating HTE models suffers from the same fundamental problem as building them: We can't observe counterfactual outcomes. A number of evaluation metrics have been used or proposed to get around this problem.[273] These metrics generally rely on the same basic insight: While we can't observe counterfactual potential outcomes, for any subset of the experiment containing both treatment groups, we *can* observe and compare both a) the means of the predicted individual-level effects and b) the conditional ATE. We'll discuss a few of these metrics here and highlight the one used in results below.

The simplest, oldest and most interpretable metric finds n-tiles of the predicted treatment effect, and then computes the conditional ATE (CATE) within each n-tile. A model producing accurate predictions

of individual treatment effect will feature CATEs increasing in n, so that (for example) the lowest tercile has a lower CATE than the middle tercile which in turn features less treatment effect than the top tercile. While this metric is easy to implement and understand, it has a number of drawbacks: It loses information, the within-bin CATEs have high variance, the choice of n is arbitrary, and perhaps most of all it does not provide an optimizable scalar score.

A more modern and useful methodology is the so-called uplift curve, computed by 1) sorting the experimental subjects in descending order according to their predicted treatment effect, and 2) at each subject index i, computing the conditional ATE among subjects $j \leq i$ (i.e., i and those predicted to be more responsive than i). Each such conditional ATE is called an "uplift," especially in the marketing literature. The 0-th uplift, for subject index 0, is defined to be 0, and the uplift value for the last subject is equal to the overall ATE. The area between this uplift curve and the theoretical random baseline is called the area under the uplift curve, or AUUC. It provides a scalar score useful for comparing models.

The uplift curve, and thus also the area under it, are also frequently normalized in various ways. A common variant is the "gain curve," which multiplies each CATE by i to get the cumulative 'gain' at each subject, and divides by the total gain to make the curve independent of the size of the treatment arms. Our evaluation employs a closely related and commonly used variant of the uplift curve called the Qini curve,[254] which computes the cumulative gain slightly differently. Instead of computing CATEs at each subject i, we compute the sum of the treatment outcome up to i, and subtract the sum of the control outcome up to i. The control outcome is weighted by the treatment:control size ratio up to subject i, so that we estimate the gain as the treatment outcome minus the control-based estimate of what the treatment outcome would have been. This curve is also normalized by dividing it by the cumulative gain, measured as the gain at the last subject. Examples of Qini curves are shown in Figure 4.2.

The Qini curve, which is normalized to start at 0 and end at ± 1, can be thought of as an analogy to the ROC curve for binary classifiers. (The curve ends at $+1$ if the ATE has a positive sign and -1 if the ATE has a negative sign.) The corresponding summary statistic is the area between this curve and the random baseline, known as the Qini coefficient; if the ATE is negative, the random baseline will also slope downward to -1. As the name might suggest, this statistic also has connections to the better-known Gini coefficient.[254]. As with the ROC curve, if the predictions are worse than random, the Qini curve will lie below the random baseline; if they're bad enough, the coefficient (as the integral of the difference between the curves) can be negative.

The analogy to the ROC curve and AUC statistic carries through in another way: For a given dataset, we can't tell upfront what the best possible performance is, just as the Bayes error rate on a classification problem is generally unobservable. Doing so would require having a perfect uplift model already, which assumes the can opener we want to find.

A notable difference from classification metrics, however, is that while it has an unobservable maximum for any given experiment, the Qini coefficient over experiments is not even theoretically bounded, either above or below. With an unbounded, continuous outcome variable, a sufficiently pathological distribution of individual-level treatment effects can make it arbitrarily large (or, if the model performs very poorly, arbitrarily negative). Large Qini coefficients require negative true individual-level treatment

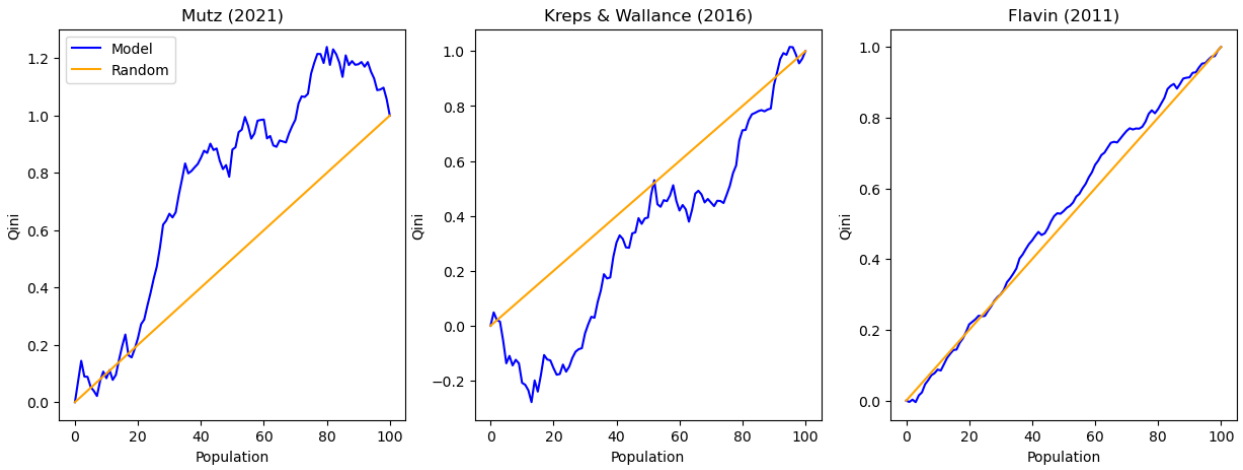


Figure 4.2: Examples of Qini curves demonstrating better-than-random, worse-than-random, and approximately random performance. The curves shown here are generated from the out-of-fold predictions of 3-fold cross-validation on the ensemble baseline discussed in Section 4.5.2.

effects (backlash, in a persuasion setting): a polarized experiment, with highly responsive and highly anti-responsive groups.[254]

Finally, when our results require examining average HTE prediction performance across multiple experiments, we will often need to average Qini curves together, whether for display or to calculate the Qini coefficient of the average curve. This is not straightforward for two reasons: The curves will have different numbers of sample points, with the different numbers of subjects in each experiment, and they may end at different values. In particular, despite the normalization described just above, the curve will end at $+1$ if the ATE has a positive sign and -1 if the ATE has a negative sign. When averaging Qini curves, we deal with these problems in the following way: First, if the ATE has a negative sign, we swap the treatment labels (0 => 1, 1 => 0) and multiply the HTE predictions by -1. This ensures that the ATE has a positive sign and the Qini curve ends at a y-value of $+1$, without changing the area between the curve and the random baseline (which, if the ATE is negative, would also slope downward toward -1). Second, to complete the averaging process, we linearly interpolate each curve to a common set of query points (in our case, the integers from 0 to 100 to get percentiles). The curves can then be averaged pointwise to get a single curve representing average performance.

4.5.2 Baselines

We evaluate against a suite of strong baselines from the existing literature:

- The T-, S-, and X-learners from the meta-learner framework of Künzel *et al.*,[270] each fit twice with a random forest[284] and a linear model as the base function approximator.
- Causal random forests.[272]
- The recent neural model of Shi *et al.*,[271] known as DragonNet or CEVAE.

To provide the strongest possible baseline (and to simplify evaluation) we primarily compare our results to the ensemble average of predictions from these methods. We generate a prediction for each subject in each experiment by using three-fold cross-validation: Each baseline is trained three times, on 2/3 of the experiment's data, and the out-of-fold predictions for the other third are stored. The splits used for each experimental dataset are the same for all baselines (and for the representational-regression methods of Section 4.4.3). Performance of both the baselines and our models are discussed beginning in Section 4.6 below.

4.5.3 Models

We evaluate all of these approaches with a number of specific language models. Most of our results come from three local LMs via the HuggingFace transformers library,[285] namely Meta's Llama3-8B and Llama3-8B-instruct models[286,287] and the OpenRLHF project's RLHF finetune of the instruct variant.[288] Having all three of these models both ensures robustness of the results and lets us evaluate the effect of instruction and preference tuning as compared to a base model. For budget reasons, we do not evaluate the persona method with all of the local models, but rather only with the Llama3-8B-instruct model. Some of our experiments also use two other sets of models, especially to evaluate scaling behavior:

OpenAI We use OpenAI's `gpt-4o-mini-2024-07-18`[‡] to run many of the same experiments as with the local models. This model is part of the GPT-4o family of models that build on and update GPT-4[289] and GPT-3[27]. For representational regression with hidden states, however, because the GPT-4o series does not offer embeddings, we use OpenAI's `text-embedding-3-large` model instead.[§]

Pythia The scaling experiments discussed in Section 4.10.1 use the Pythia "scaling suite,"[290] a line of models from 14M parameters up to 12B, to examine how ATE and HTE prediction performance vary with model scale.

4.5.4 Datasets

We evaluate on a variety of datasets representing survey experiments. Many of our experiments are drawn from the Time-Sharing Experiments for the Social Sciences project,[¶] an NSF-funded collaborative survey program that allows academics to run randomized survey experiments. The experiments cover a range of subject matter areas (from international trade to racism to public health), field times (from the early 2000s to recent years) and sizes (from several hundred to tens of thousands of subjects). In total, we have 28 experiments from 18 studies, comprising 58,144 participants in total.[‖] We summarize these datasets in Table 4.1, and provide more details about them and the original source publications in Section B.1.

[‡] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[§] https://platform.openai.com/docs/guides/embeddings
[¶] https://www.tessexperiments.org/
[‖] This count of participants is not deduplicated across studies, because participants were anonymized.

Dataset	# Treat	# Ctrl	Y Range	ATE	Field Start	Field Days
B. & T. (2024)	405	387	1 — 7	−0.44	2022-07-08	25
B. (2022)	973	993	0 — 1	+0.12	2021-07-29	15
C. & D. (2010)	365	384	1 — 7	+1.39	2009-12-01	1,946
D. et al. (2022) - Dem.	731	701	0 — 100	−4.69	2022-01-14	18
D. et al. (2022) - P.H.	757	701	0 — 100	−2.11	2022-01-14	18
D. et al. (2022) - Rep.	728	701	0 — 100	+2.16	2022-01-14	18
F. & U. (2023)	998	1,006	1 — 7	+0.11	2021-05-28	24
F. (2011)	405	385	1 — 5	−0.83	2011-11-30	7
G. & M. (2010)	764	244	1 — 4	+0.23	2010-01-07	14
G. (2024)	413	381	1 — 5	−0.20	2021-05-28	24
G. & B. (2024) - Q1	1,010	1,004	0 — 1	−0.02	2022-01-10	14
G. & B. (2024) - Q2	1,009	1,002	0 — 1	+0.01	2022-01-10	14
G. & B. (2024) - Q3	1,009	1,000	0 — 1	−0.09	2022-01-10	14
G. & B. (2024) - Q4	1,007	1,007	0 — 1	−0.12	2022-01-10	14
G. & B. (2024) - Q5	1,011	1,000	0 — 1	−0.02	2022-01-10	14
G. & B. (2024) - Q6	1,004	994	0 — 1	−0.00	2022-01-10	14
G. & B. (2024) - Q7	1,010	998	0 — 1	−0.06	2022-01-10	14
G. & B. (2024) - Q8	1,008	999	0 — 1	−0.05	2022-01-10	14
H. (2006) - Orig.	628	950	0 — 1	+0.11	2003-06-01	60
H. (2006) - GfK	1,046	1,038	0 — 1	+0.09	–	–
H. (2006) - MTurk	1,480	1,492	0 — 1	+0.12	–	–
J. & B. (2016) - Orig	1,017	1,028	0 — 1	+0.12	2013-08-01	30
J. & B. (2016) - MTurk	7,422	7,503	0 — 1	+0.20	2013-08-01	30
K. & W. (2016)	481	932	1 — 5	−0.18	2013-09-06	17
L. (2018)	523	522	0 — 100	+1.66	2016-04-15	10
M. (2021)	326	342	1 — 4	+0.16	2017-10-05	17
S. (2023)	516	505	1 — 7	−0.59	2021-02-16	14
W. (2023)	1,279	620	1 — 5	−0.00	2021-08-12	13

Table 4.1: Summary statistics of the experimental datasets used for analysis in Chapter 4 and Chapter 5. The ATEs shown are simple unweighted differences in means. Y values are all discrete and consist of the integers between the minimum and maximum values shown. The very long field period shown for the Chong & Druckman dataset[291] reflects our pooling together of an original and replication study conducted in separate years. The GfK and MTurk replications of Hiscox[292] have no field dates or durations shown because the field dates are not given in Coppock.[48] References and repository IDs are in Table B.1 in the appendix.

These datasets directly test persuasion and opinion change in response to experimentally manipulated messages, and cover a range of time periods with different political events on the agenda. They also pose a difficult test for our approach: Because of the need to preserve respondent anonymity, each has only a few demographic data points available about the respondents. Few of our datasets, in particular, have any respondent-written text from open-ended questions.

4.6 Results: Heterogeneity Prediction

Our presentation of results starts with the main task of predicting individual-level treatment response, where we observe substantial improvement on average over the ensemble baseline. Subsequent sections also examine the closely related task of reproducing *average* treatment effects, scaling behavior, and some fairness concerns.

Overall, our results in this section show that for treatment effect estimation, language models and methods that incorporate them are at least competitive with, and often outperform, even an ensemble of strong traditional baselines. The performance ceiling on any given dataset or group of datasets is unknowable (see Section 4.5.1), but our results are at least quite good and would be practically useful in treatment-targeting applications.

4.6.1 Average Performance

Broadly speaking, our results demonstrate high performance: Both zero-shot methods outperform an ensemble of strong baselines that have access to experiment-specific training data, as shown in the top panes of Figure 4.3. In one case (the text2text method), we get a 133% improvement in the model's performance over a random baseline ($0.426/0.183 \sim 2.33$). The three representational-regression variants we test also put up better-than-random performance, and one (using text2text answer probabilities as inputs) outperforms the ensemble baseline as well.

Notably, however, these methods, which also have experiment-specific training data, do not do as well as either zero-shot approach; this is also true when comparing zero-shot text2text to RepReg on text2text probabilities and zero-shot persona to RepReg on persona probabilities. This strong performance from the simple expected value under the LM distribution over answers, even when compared to a data-driven recalibration with a second-stage model, implies that the language models we study are well-calibrated under both treatment and control conditions. Another conclusion we can derive from Figure 4.3 relates to the language modeling head: Note that the hidden-states RepReg method underperforms both zero-shot methods, as well as the very same RepReg method with text2text probabilities used as inputs to the second stage. Swapping out hidden states for LM-head probabilities and observing a substantial performance increase suggests that the LM head itself contains a large amount of information relevant to the treatment-effect estimation task.

We also observe a striking lack of improvement from inference-time compute scaling, indeed apparently the opposite: The persona method, which in our setup spends up to 512 tokens thinking about the answer

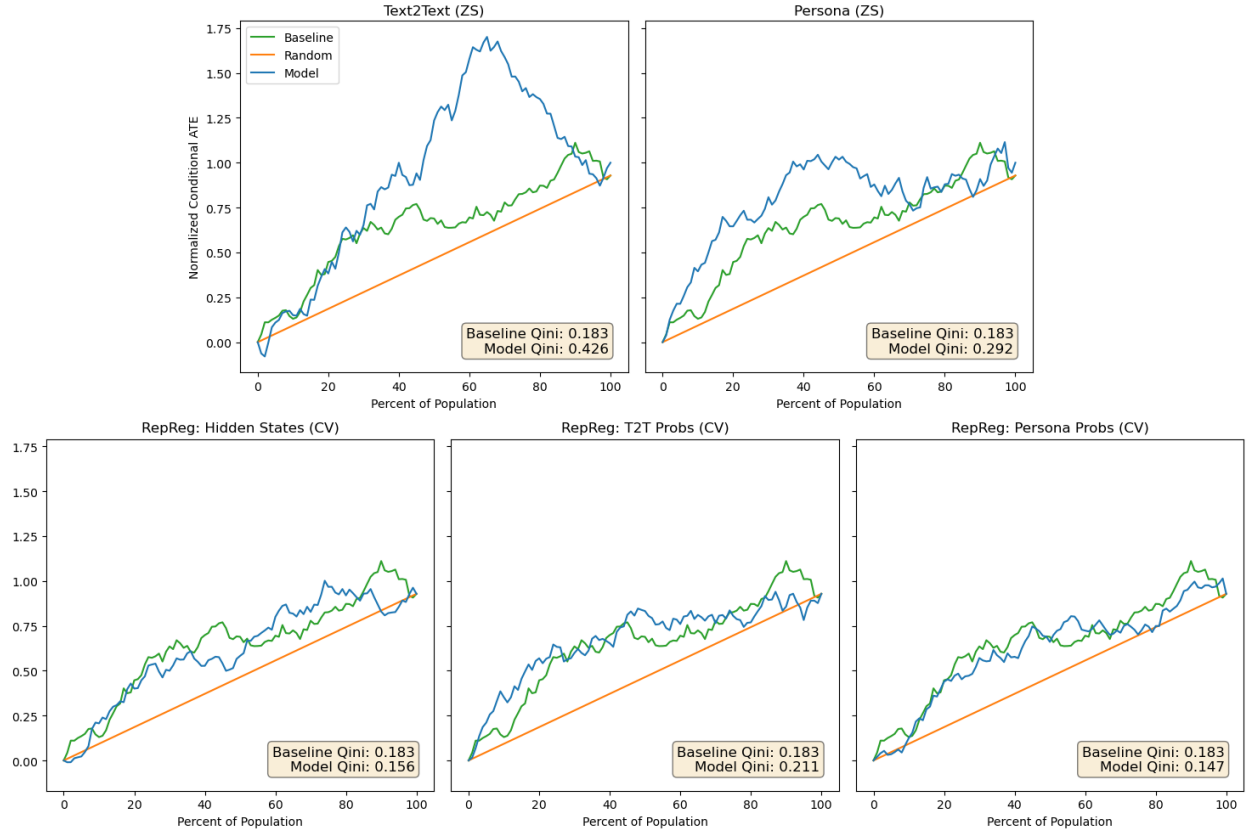


Figure 4.3: The overall average performance observed from models in our experiments, as compared to the ensemble baseline and random scores. The top panes show both zero-shot ("ZS") methods outperforming the baseline on average, while the cross-validated ("CV") RepReg methods most often underperform despite their access to training data from each experiment. Very good performance from some RepReg configurations (see Figure 4.4) suggests that this method is sensitive to the model and choice of second-stage prediction algorithm. For each subfigure, the model Qini curves and corresponding coefficients shown are computed as follows: for each dataset we find the ensemble-average prediction of all included models (for RepReg, language models x second-stage prediction methods); we then standardize the Qini curves to end at $+1$ (rather than -1 if the ATE is negative), interpolate the Qini curves for the different datasets (which have different numbers of points) to a common set of X points, and average them pointwise, computing and displaying the Qini coefficient of this doubly averaged curve. (See Section 4.5.1 for more information on Qini curve averaging.)

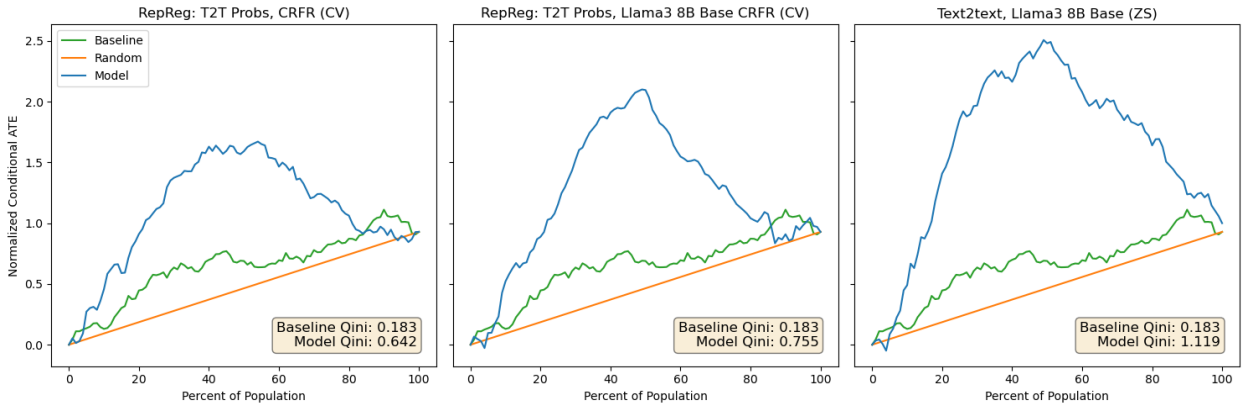


Figure 4.4: The best-performing configurations in our experiments, as compared to the ensemble baseline and random scores. We observe high performance from both (left pane) use of the models' text2text probabilities over the answers as inputs to cross-validated causal random forests (CRFR), and from (right pane) the Llama3 8B base model, zero-shot. The combination of these two (center pane), with cross-validated CRFR on the base model's probabilities, performs in between. Better performance from RepReg on text2text probabilities than on hidden states, even with the same second-stage CRFR algorithm, indicates that much relevant information is contained in the LM head. The curves shown here are calculated in the same way as those shown in Figure 4.3 and as described in Section 4.5.1. ("CV" = cross-validated; "ZS" = zero-shot.)

before providing it, underperforms the text2text method which generates answer probabilities immediately. This same pattern is apparent in the RepReg variants, with persona probabilities performing worse as inputs to a second stage model than text2text probabilities. (See Section 4.6.4 below for a likely explanation.)

4.6.2 Significance Tests

The unobservability of counterfactual outcomes and the complexity of the ranking-based uplift metrics makes it difficult to develop closed-form statistical tests for whether an uplift model has detected a statistically significant amount of heterogeneity. We instead approach this question with a Monte Carlo (simulation-based) test, testing the null hypothesis that the model's HTE predictions find no more heterogeneity (conditioning on the actual treatment assignment) than randomly generated predictions would, against the alternative hypothesis that they find more. The test procedure is laid out in full in Section B.2 in the appendix.

Both text2text ($p < 0.001$) and persona ($p = 0.002$) identify a statistically significant amount of excess heterogeneity under this test, at the Bonferroni-corrected $p = 0.05$ level. Neither the ensemble baseline nor any of the RepReg predictions are significantly different from random scores.

4.6.3 Best Performance

A few particular configurations also demonstrate substantially better performance than the average, with the three best-performing shown in Figure 4.4. (Note, though, that we found these results after the fact,

and did not pre-specify these hypotheses in particular.) Both the causal random forest regressor (for RepReg) and the base model without post-training do particularly well (in the text2text setting), with the combination of these two also posting good performance.

4.6.4 Effect of Post-Training

As stated in Section 4.5.3, we used Llama3 because of the availability of base, instruct and RLHF checkpoints that would allow testing the effect of post-training on HTE prediction performance. Our results, shown in Figure 4.5, are unequivocal: Post-training severely degrades the model's HTE estimation abilities. In all three method evaluations where we compare them head-to-head (persona methods were evaluated only on the instruct variant, both to reduce costs and because of the base model's lack of instruction tuning and resulting prompt-following ability), the base model does much better than the other two variants. Given the importance of calibration visible in Figure 4.3 and discussed above, a likely culprit is the miscalibration known to be induced by instruction and preference tuning.[293]

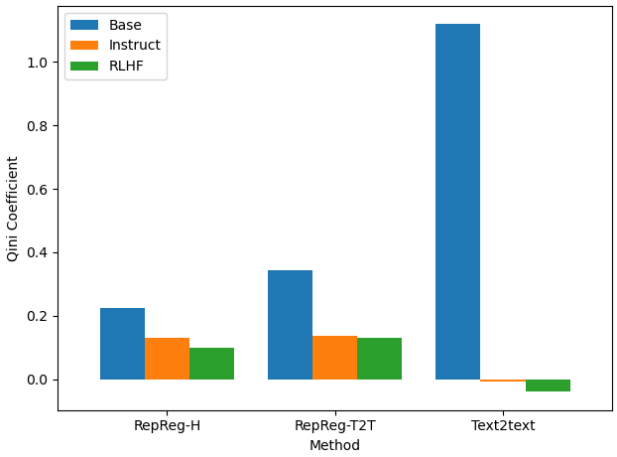


Figure 4.5: Qini coefficients for all head-to-head comparisons of the performance of base, instruction-tuned and preference-tuned models on HTE prediction tasks, demonstrating a large and consistent outperformance by the base model compared to those with post-training. Where more than one set of predictions exists for a dataset (such as from using CEVAE, causal random forests, etc, for second-stage predictions in RepReg), these predictions are ensembled for each dataset and the Qini curves are averaged across datasets. We then compute the Qini coefficient for this final curve.

These results also provide an explanation for what at first blush appears to be inverse scaling from text2text to persona. The instruct model, which we used for persona methods, does improve with test-time compute scaling. As shown in Figure 4.5 and Figure 4.3, the model's performance goes from a Qini of -0.008 to 0.292. This effect is masked by the different, larger set of models used for the text2text experiments, with the base model performing much better.

4.7 Results: Reproducing ATEs

We also examine the closely related task of ATE prediction. Given a set of individual-level treatment effect predictions, it's easy to get a prediction for the average treatment effect: Just average the individual-level predictions. We can also observe the true ATE easily, with the usual Neyman difference-in-means estimator, and compare them. If language models can accurately predict ATEs, we should see the predictions and the observed ATEs correlate well.

As shown in Figure 4.6, this is exactly what we do see. (As in the figure, our analysis here standardizes both ATEs and predicted ATEs as effect sizes for cross-dataset comparability.) RepReg and baseline

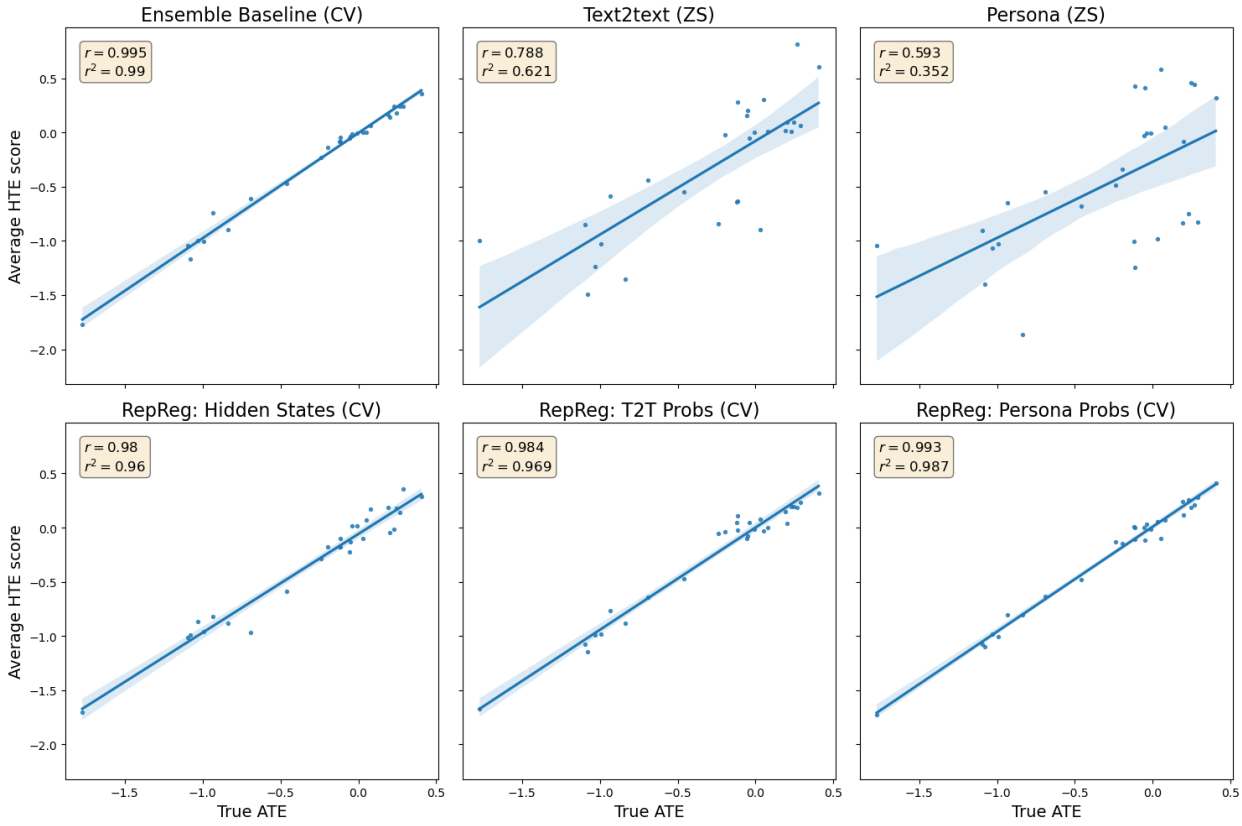


Figure 4.6: ATE prediction performance of various zero-shot ("ZS") and cross-validated ("CV") models from our experiments. The zero-shot approaches have a much harder task: With training data, estimating the difference in means between the treatment arms is straightforward, while doing well without it requires far more world knowledge. The y-coordinate of each dot represents the ensemble prediction on a single dataset of all implementations (models for zero-shot methods, models x second-stage predictors for RepReg) on that dataset, with the correlation taken at the dataset level. The blue line is the line of best fit, and the shaded area around it is a 95% confidence region. True ATEs and average model predictions are first divided by the pooled outcome standard deviation, to make both interpretable across studies as a Cohen's d effect size.

methods excel here, unsurprisingly, because they have training data. If even the simple difference-in-means estimator can learn the average treatment effect from data, a more sophisticated method should be able to as well. More interesting is the performance of the zero-shot methods, which face a much more difficult task. Without any data to learn from, the model needs to very accurately represent how persuasive messages are to predict ATEs well. Our experiments indicate quite good performance at this task, with both the text2text and persona methods putting up correlations far above that expected by chance. The text2text method's predictions in particular are able to explain almost 2/3 of the variance in the observed effect size ($0.788^2 = 0.621$).

These results agree directionally with the $r = 0.91$ correlation that Hewitt *et al.* report a similar method reaching between predicted and true ATEs in a larger sample of effects, though we find a lower correlation. Their results, however, were with a larger model (GPT-4).[**] Capabilities at the level we find would still be quite useful in some experimental design applications, where a different point on the performance for model cost tradeoff curve is best.

If we break down these zero-shot results by model, familiar patterns from the previous section reappear. Individual-level predictions from the Llama3 base model average up to the best ATE predictions, with the text2text method using this model reaching a correlation of 0.831 against only 0.641 for the instruction-tuned checkpoint and 0.707 with RLHF. (The persona method was once again only evaluated with the instruct checkpoint.) We also find a clearer demonstration of inverse compute scaling: The persona method with the instruction-tuned Llama3 model has a true ATE / predicted ATE correlation of 0.593, against the same model's 0.641 with the text2text method. Inverse scaling is unusual[295] and deserves further study.

4.8 Results: Relationship of ATE to HTE Prediction Performance

A natural next question to ask is how these two capabilities relate to each other. That is, are models which better predict ATEs also better at predicting HTEs? The answer, quite surprisingly, is a comprehensive 'no'. The correlations between Qini coefficients and various measures of ATE prediction accuracy are shown in Table 4.2. None of these coefficients are large in magnitude, for any method, and some of them run in the opposite direction of that expected, with a positive correlation between Qini and ATE prediction error. (ATEs are, once again, expressed as effect sizes for comparability between experiments.) Disaggregating the average results given in the table by underlying model, second-stage prediction method where appropriate, and so on (not shown for space) still fails to reveal any strong relationship. Spearman rank correlations also tell a similar story to the Pearson ones we've shown in the table. We might think that ceiling effects in ATE prediction distort the findings for RepReg methods (and the baseline), but the relationship is only slightly stronger for the zero-shot methods.

Now, a possible explanation of this pattern is that it's a coincidence: We are simply seeing an artifactual relationship between which datasets have the easiest-to-predict ATEs and which have the easiest-to-predict

[**] We do find performance at this level in the gpt-4o-mini results of Section 4.10.2, making that model's poor HTE performance all the more striking.

	ATE Err.	Abs. ATE Err.
Text2text (ZS)	0.003	−0.128
Persona (ZS)	0.008	−0.226
RepReg: Hidden States (CV)	0.008	−0.026
RepReg: T2T Probs (CV)	0.004	−0.025
RepReg: Persona Probs (CV)	0.078	−0.049
Ensemble Baseline (CV)	0.010	−0.075

Table 4.2: Average correlations between various ATE prediction performance measures and the corresponding Qini coefficients. Results are Pearson correlation coefficients, computed over pooled results for all datasets and every model evaluated for each approach. ATEs and predictions are first expressed as Cohen's d effect sizes. We find surprisingly low correlations between performance at ATE and HTE prediction performance, suggesting these tasks may draw on quite different model capabilities. (Performance measures: "ATE Err." = signed difference between true ATE and average of individual-level predictions; "Abs. ATE Err." = unsigned difference, absolute value of ATE Err. Methods: "ZS" = zero-shot; "CV" = cross-validated.)

HTEs. We can test this idea by examining, instead of the correlation across datasets within each method, the correlation at the method level within each dataset. If performance at ATE and HTE prediction tasks actually co-vary, we should expect to see much higher correlations. The results, however, decisively reject the coincidence explanation: The average of the within-dataset correlations is -0.017, and only three of 28 datasets have correlations larger in absolute value than 0.2.

ATE and HTE prediction performance are not *completely* unrelated, as Table 4.2 shows; all of the correlations of absolute ATE prediction error with Qini are in the same direction, with lower error correlating with higher Qini coefficients. But they correlate very weakly, and in even the strongest relationship (zero-shot persona experiments) ATE prediction performance at $r = -0.226$ explains only about 5% of the variance in HTE prediction performance.

The likeliest explanation is that these two tasks simply draw on different capabilities of the underlying models. This conclusion finds further support from the ATE/HTE results in Section 5.2, and in the sharply different patterns of demographic disparity observed for ATE and HTE prediction in our very next section.

4.9 Results: Robustness and Demographic Analysis

Having demonstrated that LMs can perform well at both HTE and ATE prediction, and that these are substantially separate capabilities, it's also natural to ask whether there are any important group disparities in these tasks. We are limited in the kinds of groups that can be analyzed by which demographic and other variables are present in all or most of the experimental datasets we have available, and by how they are (or are not) consistently coded. For our experimental datasets, the most useful choices are:

Race This is present in every dataset, and the only consistently usable coding scheme is to dichotomize it to white and nonwhite. We drop from this analysis the Mutz[296] dataset, which included only white respondents.

Sex / gender Every dataset reports this, and does so dichotomously as male or female. No dataset consisted exclusively of one sex or the other, so no datasets are dropped from this portion of the analysis.

Political party Each of our datasets includes the political party identification of the respondents. These are coded in varying ways, from a Democrat/Republican/independent categorical variable to a strong Democrat to strong Republican Likert scale. We dichotomize these variables to Democrats and Republicans, taking leaners on Likert scales as part of the party toward which they lean. Independents and third party identifiers in each dataset were dropped. We also dropped all three datasets from DeMora *et al.*,[297] whose entire experimental sample consisted of Republicans.

We further dropped, from each of these analyses, any subject who refused to answer the relevant demographic question.

To analyze how performance varies along these dimensions, we look at two measures of prediction quality, corresponding respectively to the ATE and HTE analyses of the preceding sections: the conditional ATE within each demographic subgroup and the Qini coefficient computed only among each subgroup. (Note that the CATE is both an analog of the overall ATE for each group and at the level of the entire experiment is also a measure of heterogeneity prediction performance in its own right.) For the Qini coefficient, we do not rerun the entire experiment with the variable of interest removed from the text representations of each subject, which would be both too expensive and represent a subtly different question. Rather we simply take the existing predicted individual-level effects, which were computed with access to the focal variable in each analysis (race, sex or party), and filter down to only the appropriate demographic group. Doing so addresses the more common practical question of how predictions generated with all available data differ between groups, rather than the less relevant question of how the inherent predictability of response varies by group.

The results are shown in Table 4.3, and differ substantially between CATE and Qini measures. For the CATEs, predictions (computed as the average of the individual-level predictions) are off by very similar amounts between groups for all methods. No large or substantively important disparities are apparent.

There are, however, much larger differences across groups in how well individual-level predictions rank subjects *within* groups, which is what the Qini coefficients measure. We do not observe gaps in any consistent direction, however: As Table 4.3 shows, on average we see better prediction performance for white respondents than for non-white ones (average $\Delta = 0.656$), but no similarly large or consistent gaps are evident for party (average $\Delta = 0.219$) or especially gender (average $\Delta = -0.078$). The ensemble baseline, composed of non-neural models trained from scratch, does not always reproduce the same patterns as the other methods, making it possible that biases introduced in language model training or data curation contribute.

			BL	Per.	RR-H	RR-P	RR-T	T2T
Abs. ATE Err.	Race	W	0.049	0.431	0.227	0.122	0.160	0.350
		N	0.051	0.424	0.230	0.138	0.167	0.337
		Δ	−0.002	0.007	−0.003	−0.016	−0.008	0.012
	Party	R	0.078	0.469	0.253	0.163	0.194	0.408
		D	0.053	0.420	0.230	0.140	0.175	0.359
		Δ	−0.025	−0.050	−0.024	−0.023	−0.019	−0.048
	Gender	M	0.056	0.417	0.233	0.142	0.175	0.349
		F	0.062	0.448	0.236	0.144	0.181	0.358
		Δ	−0.005	−0.031	−0.003	−0.001	−0.006	−0.008
	All		0.041	0.425	0.220	0.110	0.149	0.346
Qini Coef.	Race	W	0.059	1.736	0.181	−0.017	0.007	1.736
		N	0.130	0.391	−0.049	0.032	−0.126	−0.607
		Δ	−0.071	1.345	0.231	−0.049	0.133	2.344
	Party	R	−1.020	0.023	0.400	−0.705	0.127	0.164
		D	0.054	0.146	0.033	0.061	0.009	0.001
		Δ	1.074	0.123	−0.367	0.766	−0.118	−0.163
	Gender	M	0.332	0.130	0.268	0.117	0.185	−0.334
		F	0.738	0.003	−0.051	0.102	0.239	0.135
		Δ	−0.405	0.126	0.319	0.015	−0.054	−0.469
	All		0.242	0.264	0.111	0.093	0.104	0.252

Table 4.3: Average values of prediction performance metrics for CATE ("Abs. ATE Err." = absolute value of prediction error for scaled CATE) and HTE (Qini coefficient), by demographic group and method, and intergroup differences. The results demonstrate fairly even performance across demographics in predicting demographic-level conditional ATEs, but much more variability within each demographic (and by method) in HTE prediction. Method names: BL, ensemble baseline; Per, zero-shot persona; RR-H, representational regression with hidden states; RR-P, representational regression with persona probabilities; RR-T, representational regression with text2text probabilities; T2T: zero-shot text2text. Variable levels: Race, W (white) and N (nonwhite); Gender, M (male) and F (female); Party, R (Republican) and D (Democratic).

These results provide further confirmation of the apparent lack of relationship between ATE and HTE prediction performance, and suggest that disparities in HTE prediction performance warrant further research and consideration in practical applications.

4.10 Results: Scaling Behavior

How performance varies with scale is an important question both theoretically and practically. Here we address this question with several new experiments, and discuss and contextualize results from earlier in the chapter.

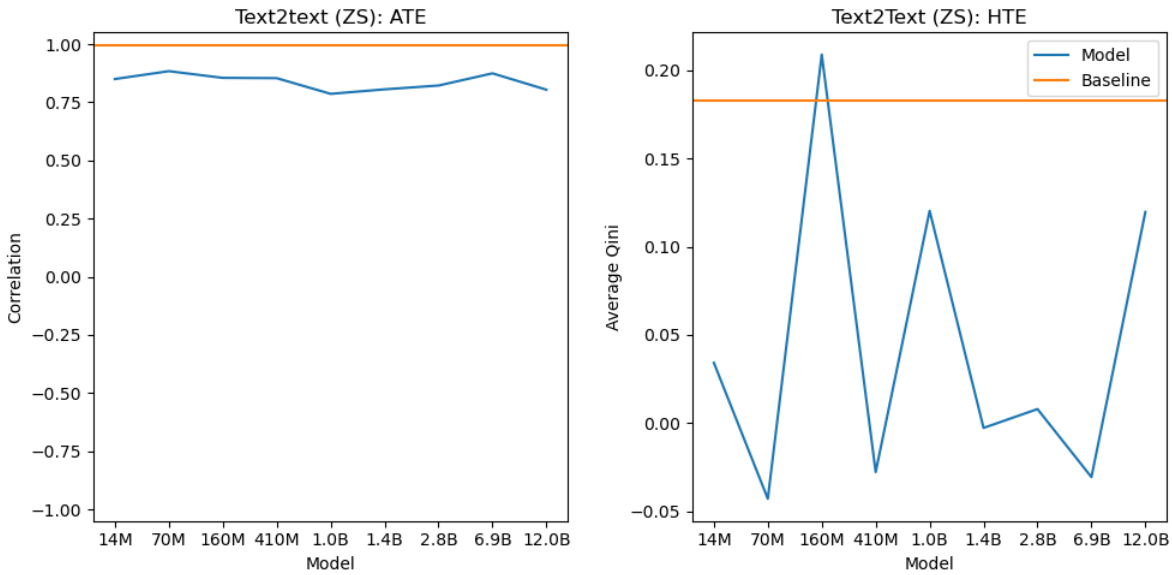
4.10.1 Pythia Scaling Study

As mentioned in Section 4.5.3, we run some experiments with the Pythia scaling suite of models.[290] These range in size from 14M parameters to 12B, and are not instruction- or preference-tuned. We evaluate both ATE and HTE prediction performance of the zero-shot text2text method with these models, as a function of scale. The results are shown in Figure 4.7, and demonstrate a curious lack of returns to scale. ATE performance is approximately constant, even down to quite small models, and HTE performance is inconsistent. It sometimes, but not always, exceeds that of the ensemble baseline, and does not appear to increase with model scale as measured in number of parameters. This is another example of the divergence between ATE and HTE prediction performance: The Pythia models perform comparably to Llama3 variants at ATE prediction (compare Figure 4.6), and notably worse at HTE prediction.

4.10.2 OpenAI Results

As discussed in Section 4.5.3, we repeated the ATE and HTE experiments of earlier in this chapter with models from OpenAI in order to evaluate the impact of scaling on performance. Most experiments use gpt-4o-mini,[††] but the RepReg experiments with hidden-state representations use text-embedding-3-large because OpenAI chat models do not make internal representations available through the API. The chat models also do not make the full vector of probabilities over the vocabulary available, but only logprobs for the most likely completions of the prompt. Our prompts (the same as used for the local models) were quite effective in eliciting the desired completions, and we were able to get logprobs for all possible answers for virtually all examples. (The one exception is a few studies whose final questions are "feeling thermometers" that ask for a number from 0 to 100; we took logprobs for any allowable number which occurred in the API response and set the probabilities of the others to 0. Because the minimum logprob returned was generally very small, the resulting approximation error to the true distribution under the model is also small.)

[††] OpenAI has not publicly stated exactly how large gpt-4o-mini is, nor how it was trained. Some sources say that it is about 8B parameters,[298] the same size as our Llama models, though it may be larger. It is also widely believed to be distilled from larger and more capable models, unlike the trained-from-scratch Llama3,[287] and is thus at least an example of compute scaling if not necessarily parameter scaling.



(a) ATE prediction performance.

(b) HTE prediction performance.

Figure 4.7: ATE and HTE prediction performance of the Pythia scaling suite models. ATE results are the correlation across datasets of the true ATE with the average of individual-level HTE predictions. For ATE results, both the outcome variable and HTE predictions are expressed as a Cohen's *d* effect size, by diving by the pooled standard deviation of the outcome variable. HTE results for each model size are the area under the average Qini curve over datasets. Both subplots show the performance of the ensemble baseline for comparison. ("ZS" = zero-shot.)

The results have both clear similarities to and clear differences from those observed with the Llama models. As with Llama, ATE prediction performance is consistently very high with RepReg methods (unsurprisingly, given their use of training data). The lowest correlation coefficient observed is for RepReg with hidden states ($r = 0.946$) and the other two RepReg correlations do not differ from their Llama analogues by more than 0.001. Zero-shot performance, meanwhile, is notably better than Llama. The average individual-level prediction of the zero-shot text2text method with gpt-4o-mini correlates with true ATEs at $r = 0.865$ (cf. $r = 0.788$ for Llama), and the persona method at $r = 0.894$ (cf. Llama's $r = 0.593$). These improvements are what we might expect from a larger and more capable model.

The HTE prediction performance observed from the OpenAI models, on the other hand, is surprisingly poor. We consider all pairwise comparisons to any Llama model: 3 head-to-head comparisons of the Llama variants with text-embedding-3-large for RepReg with hidden states, and 8 for gpt-4o-mini (1 against Llama3 8B Instruct in zero-shot persona, 1 against Llama3 8B Instruct in RepReg with persona probabilities, and 3 each against the Llama variants in zero-shot text2text and RepReg with text2text probabilities). The text-embedding-3-large model outperforms 0 times out of 3, and gpt-4o-mini only twice out of 8. Neither is the best-performing model in any experiment. In the zero-shot text2text and persona experiments, gpt-4o-mini performs *worse* than the theoretical random baseline, with Qini coefficients of respectively -0.052 and -0.167. Performance is not uniformly terrible, with RepReg using the text2text probabilities from gpt-4o-mini demonstrating the second-best Qini of 0.232 (the best performance in this experiment came with the base Llama3 8B model, at a Qini coefficient of 0.345). But it is certainly underwhelming, and unexpected from what is otherwise a more capable model than the 8B Llama variants. That performance with the text2text probabilities improves substantially when recalibrated by a second-stage RepReg method may point to a partial explanation: the OpenAI models' more aggressive post-training has degraded their calibration, and thus their performance, more thoroughly than the Llama variants (cf the analysis in Section 4.6).

Finally, when we examine the correlation between ATE and HTE performance as in Section 4.8, there are even more notable changes from the Llama pattern. We do observe stronger correlations between ATE prediction error measures and Qini coefficients, with the largest in absolute value being 0.425 (about 18% of the variance). But these usually run in the *other* direction: Both zero-shot methods have positive correlations between absolute ATE error and Qini coefficient (text2text: 0.425, persona: 0.397), implying a tradeoff between performance at the two tasks. The RepReg methods are in line with the Llama models' pattern, having negative but small correlations no larger in magnitude than -0.101.

4.10.3 Sample Efficiency: How Much Pilot Data Are Zero-Shot LMs Worth?

An important practical question for users of zero-shot LM-based methods for predicting treatment response is how much experimental pilot data they can replace. In other words, if one can supplement a pilot study to target an intervention with language-model simulations, how much pilot data is that worth?

We test this question by conducting a scaling study on our ensemble baseline. We refit each model in the ensemble several times, on progressively larger fractions of each dataset, and compute the Qini

coefficient of the average predictions on the rest of the dataset, which is out-of-sample. This gives us a measure, for each dataset and each fraction of the dataset, of how well the baseline performs. We use 16 evenly spaced query points from 0.15 to 0.9 (i.e., every 0.05) to ensure both train and test sets are always large enough. Averaging over the datasets and query points gives us a measure of how much an additional slice of the data is worth, in terms of incremental Qini, to which we can compare the zero-shot methods' performance. Because there is some added variance introduced by the subsampling, we fit a linear model with 0 intercept and estimate the marginal benefit of each unit of data from its slope.

This procedure estimates that the marginal 5% slice of our datasets adds about 0.017 to the Qini coefficient of an ensemble of classical baselines trained on data from them. A simpler approach fitting a line through a) the expected Qini value of 0 with no data and b) the performance of the ensemble baseline shown in Figure 4.3, each prediction of which came from models trained on 2 out of 3 cross-validation folds (66% of the data), provides a slightly lower estimate of 0.014 for each 5%.

In other words, then, the average performance observed from the text2text and persona methods in Figure 4.3 is what we'd expect from models trained on experiments from 125% to 158% the size of the actual sample we have (for text2text) and 86% to 108% (for persona). Even larger values are obtainable from the best-performing models in Figure 4.4. Taking advantage of LMs for HTE prediction, then, can substantially reduce the costs of pilot studies – or, conversely, allow much more accurate targeting for the same budget.

4.11 Conclusion

Our evaluation datasets pose a difficult task for any method: As replication data, they contain only limited and not personally identifying information about respondents. Nevertheless, on average our zero-shot methods demonstrate considerable improvement at HTE prediction over the already nontrivial performance of a strong ensemble baseline. While their performance at predicting ATEs is notably worse than baseline, this is not surprising: Given a significant amount of training data, even a quite simple model ought to be able to estimate the difference in means between the treatment arms. Achieving a respectable correlation, far better than chance, between average individual-level predictions and the true ATE without any experiment-specific data at all is much more difficult.

Because the language models used in our methods have no additional subject-specific data, this improved performance suggests greater statistical efficiency. The LM is able to leverage its world knowledge from pretraining, which includes information about how different kinds of people relate to and react to different kinds of messages, and surface it in better representations or more accurate predictions of their response. None of these methods are, of course, necessary to improve predictive performance: One could simply accumulate a large amount of data and estimate the CATE function with low variance, but its statistical inefficiency makes this idea very expensive.

We have, by contrast, demonstrated entirely zero-shot methods able to beat baselines that have access to substantial amounts of training data. Only each subject's predictor variables and a textual treatment description (and, of course, an internet's worth of pretraining data) are necessary. The results are a powerful

demonstration of transfer learning.

4.11.1 Open Questions

Our results in this chapter answer some questions and raise others, especially around scaling behavior and the variability of performance between models and tasks.

Variability of performance. Performance at HTE prediction tasks is highly variable, along a number of puzzling dimensions. First, we observe a huge impact on HTE prediction performance of post-training. The Llama3 base model greatly outperforms both instruction- and preference-tuned checkpoints derived from it, most likely because of degraded calibration from the post-training process. Model families also perform differently, especially the better performance observed from the Llama models than from gpt-4o-mini or Pythia. It is not at all clear why this happens[‡‡] or which ingredients in a model's training process matter, and further research should take this up.

Moreover, the apparent independence of ATE and HTE prediction as tasks is quite puzzling. Good performance at each seems to draw on very different capabilities from the language models used. Understanding further why this happens and what the key techniques or training data are for each would be valuable.

Scaling. Overall, we find a confusing mix of scaling behaviors.

Model and data scaling Larger models do not reliably have better performance. Larger Pythia models do not consistently do better at either ATE or HTE prediction; the OpenAI models we evaluate do quite poorly at HTE prediction (though they do better than Llama at ATE prediction, further highlighting the independence of these tasks). While we do not directly test variation in the size or composition of pretraining data, we do observe significantly worse performance from models which have undergone instruction or preference tuning than from the Llama3 base model. Applying more data and the compute to train on it, in this case, sharply worsens performance at HTE prediction, by much more than the instruction-tuned model can subsequently recover with test-time compute scaling.

Test-time compute scaling We do not consistently see performance improve with more test-time compute. Llama3 8B Instruct does perform better at HTE prediction with the persona method than the text2text method that spends no additional test-time compute (Qini coefficients of respectively 0.292 and -0.008), but we see the opposite pattern, inverse compute scaling, with gpt-4o-mini. This latter model goes from a worse-than-random Qini of -0.052 with the text2text method to an even

[‡‡] The simplest, superficially plausible, explanation is leakage: that the Llama3 base model's training corpus included our experimental datasets. On further consideration, though, this is not likely: The texts we actually present to the models, rendered versions of the survey scripts with participant demographic traits, are to the best of our knowledge entirely novel and do not appear anywhere else. The scripts alone, which to boot are often contained only in obscure replication data archives, are not sufficient to learn this participant-traits-to-outcome mapping.

worse value of -0.167 in the persona experiments. Llama3 8B Instruct also posts worse performance at ATE prediction ($r = 0.593$) with persona than with text2text ($r = 0.641$), another example of inverse scaling.

In brief, compute scaling helps, except when it hurts, which is also what we observe for model scaling. Inverse scaling of any kind is unusual, and the examples we find here should prompt further research. They should also prompt caution in evaluating how well various models work for practical use cases. Bigger is not necessarily better!

Chapter 5

Transfer Learning for Persuasion: Generalization in Treatment-Effect Estimation

5.1 Introduction

Our results in Chapter 4 demonstrate that language models are capable of good performance, indeed better than baselines, at the task of individual-level treatment effect estimation. Our findings also agree with other recent work[294] that these predictions, averaged over an experiment, are accurate predictors of the empirical average treatment effect.

But simply demonstrating that LMs can do well at this task doesn't explain why or what data is informative. Enormous pretraining datasets (and smaller but still sizable post-training datasets) go into creating them, and there's both practical and theoretical value in knowing which sorts of data are useful for treatment-effect estimation. We are interested, that is, in generalization: What kinds of data allow transfer learning to the setting of persuasion and opinion change?

This is in fact a concrete instance of a central theoretical question in the study of persuasion, especially in politics: what determines persuadability and how it relates to context. A prominent perspective in the study of this question is that persuasion usually occurs fairly uniformly,[48,49] without much heterogeneity along practically measurable dimensions.[*] Where there is meaningful heterogeneity, it appears to be highly contextual[50] and difficult to give a clear theoretical account of. (This complexity makes intuitive sense: A general, simple factor of persuadability without regard to personal traits or message might less charitably be called gullibility.)

There has been particular emphasis on the role of the media,[39] especially traditional media[299] but more recently also social media and the internet.[300] Our first experiments in this chapter take advantage of a

[*]Note, though, that the hypothesis tests of Section 4.6 provide strong evidence this is not true, on average at least, for our corpus: Enough heterogeneity exists for it to be predicted zero-shot by language models, and to be statistically significantly different from what random scores would produce.

unique news dataset, discussed below, to test some of these theories and examine media's relationship to opinion change. These experiments build on pioneering work by Chu *et al.*[34,301], which developed a related notion of 'media-diet models' and applied them to predicting answers in cross-sectional national surveys. The work here differs in a number of important ways (including an updated language-modeling methodology), the most important of which is that we consider causal settings with randomized treatment assignment. Rather than exploring media data's applicability to expressed opinion in surveys, we want to see how it can apply to *changes* in such opinion.

In addition to media, we hope to probe some of the hidden architecture of opinion by examining other kinds of transfer. Our results so far, demonstrating improvements in treatment effect estimation from the use of large pretrained models, already demonstrate a kind of transfer learning (from the model pretraining distribution to our experimental datasets). To allow more focused follow-on inquiry, we'll drill down here on a few kinds of transfer:

Contextual Can we predict persuasion more effectively after pretraining on contextually relevant information? In the political and public-affairs settings of our experimental datasets, this information is most naturally contemporaneous news data, which captures the political agenda and would have been available to (though not necessarily actually accessed by) the experimental subjects.

Cross-Subject Does pretraining on some kinds of subjects, such as those with particular demographics, help generalization to other kinds of subjects? Can some kinds of people be informative about other kinds of people?

Cross-Experiment Finally, most ambitiously, does pretraining on some experiments result in improved HTE prediction performance on others? Has the trained model, in other words, learned anything about persuadability in general?

Positive answers to any of these questions would be both practically and theoretically useful. On the practical side, predicting which messages will be persuasive to which kinds of people has applications in marketing, political communication, and other fields. Knowing which sorts of data are useful for this task allows it to be done better and more efficiently. On the theoretical side, our results (especially if positive) can make a useful entry in long-standing debates[39,48,50] about what factors influence persuasion. Even a positive finding would not prove that (say) exposure to news *causes* opinion change, but it would provide strong evidence in support of that position.

5.2 Contextual Influence: News Data

This section presents our experiments with models fine-tuned on news data. We leverage a novel and large dataset of TV news closed captions to represent news, and fine-tune iteratively on all data up to and including each successive month of this dataset. A new checkpoint is generated each month, re-initializing training from the previous month's checkpoint, to avoid a quadratic increase in training time while

still training on the news an actual human respondent would have been exposed to. We then use the appropriate monthly checkpoint for each experiment conducted during the news-data collection period.

We train three kinds of models for each month: one using data from all four stations (Fox News, MSNBC, CNN and CNBC), one using only MSNBC (to reflect a liberal perspective), and one using only Fox News (to reflect a conservative perspective). These models are then evaluated in a number of different settings: Fox models on only Republican subsets of each experiment, MSNBC models on Democratic subsets, all-data models on all respondents, and finally all three kinds of models ensembled on all respondents to provide the broadest possible test of contextual and media influence.

5.2.1 Dataset

Large-scale datasets of recent news content are hard to come by. Among other reasons, this is because of publishers' intellectual property concerns, which have only intensified in the era of generative AI.[302,303] Accordingly, rather than using a standard preassembled corpus of news, we used one recently developed at the MIT Center for Constructive Communication.

The dataset consists of closed-captioning (CC) transcripts collected from several cable-TV networks via tuner cards: CNN, MSNBC, Fox News and CNBC. These stations cover a range of ideological perspectives, from stridently right-wing (Fox) to comparably left-wing (MSNBC) to left- and right-leaning perspectives in the middle (CNN and CNBC). Our dataset covers the period from May of 2020 to December of 2023.

The raw CC outputs contain timestamps, a field indicating the method of display onscreen (whether the captions are pop-on or roll-up) and the text displayed. No other fields are available, in particular whether the content was part of a program or an advertisement. To more accurately reflect what a viewer of these channels might experience, we elected not to try to detect and remove ads.

We preprocessed the news content to remove certain formatting conventions in the captions, such as the use of a musical-note character to indicate music being played, and split it into speaker turns based on the annotations present in the captions.[†] A small fraction of speaker turns which had been garbled by poor TV signal, which we defined as those with at least a 20% rate of mis-spelled words according to nltk-based spellchecking,[304] were dropped. To help models learn from relatively long-context dynamics, rather than decontextualized short statements, we then sorted the speaker turns in chronological order within each station and batched them into approximately 2048-token chunks. (Longer speaker turns, which occur in host monologues especially, were split up at sentence boundaries; shorter chunks were iteratively combined until consisting of at least 2048 tokens.) Each such chunk became one training example.

All told, after preprocessing we have 761,239 examples comprising 4.84 GB of text. The content is nearly evenly balanced between stations, ranging from a low of 1.11GB for Fox News to a high of 1.29GB for MSNBC.

[†]The raw data, following a common closed-captioning convention, indicates scene changes with ">>>" and speaker changes with ">>".

5.2.2 Training Details

Our finetuning follows the straightforward, month-by-month schedule outlined at the start of this section: Training on each month loads the previous month's checkpoint, discarding its optimizer states, or the stock model for the first month, trains on that month's news data, and then saves the resulting checkpoint. We do this three separate times, accumulating three series of monthly checkpoints, on three separate datasets: Fox News only, MSNBC only, and all four channels pooled together.

To control costs, we avoid using all models from Chapter 4 and instead finetune only the Llama3 8B Instruct variant. We selected the instruction-tuned model in particular because, unlike the base model, it should be usable for persona experiments which need to follow prompts. With results for these experiments available in the last chapter, we can then compare performance of all techniques, not just text2text and RepReg, before and after news tuning.

To avoid ablating the stock model's instruction-following capabilities, the training data mixture is 15% from three standard instruction datasets (Alpaca,[305] Dolly-15k,[306] and OpenAssistant1[307]) and 85% news data. We use all available news data and randomly downsample the very large OpenAssistant dataset to hit the target proportion; the final training data mixture is simply a bit larger than a news-only dataset would be.

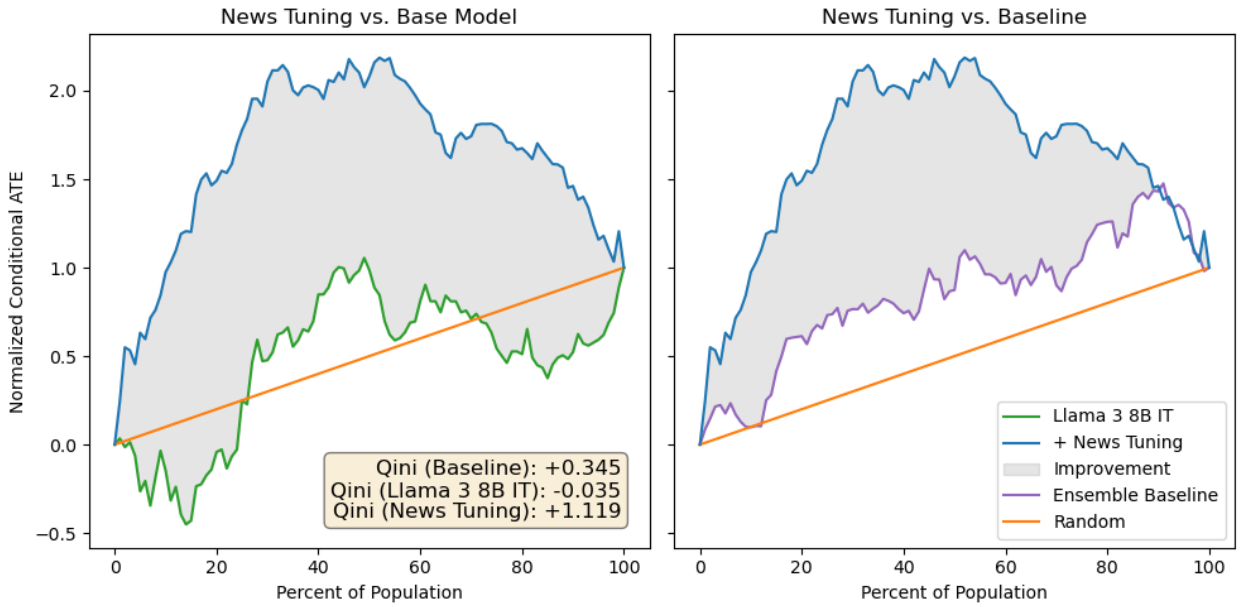
We use full fine-tuning rather than LoRA, because of the large dataset and desire to pick up possibly subtle signals from the news corpus. Training used the 8-bit AdamW optimizer[308,309] from the bitsandbytes package, with the default betas of (0.9, 0.999) and no L2 weight decay. The initial learning rate was `2.0e-5`, and we used a cosine learning-rate scheduler with the default hyperparameters. To stabilize training, the gradients were clipped to a maximum norm of 1.0. The logical batch size was 64, reflecting a physical batch size of 1 and gradient accumulation. (Recall from Section 5.2.1 that each example is already a sequence of several consecutive sentences.) We trained for 3 epochs and took the model at the end of training, without early stopping or model selection based on a validation set. For efficiency, we use gradient checkpointing,[310] mixed-precision training[311] with bfloat16, Flash Attention 2,[312] and Liger Triton kernels.[313] Training used 80GB A100 GPUs from a cloud service and also employed the transformers,[285] trl,[‡] and accelerate[§] packages from Hugging Face.

The datasets were split randomly into 90% train and 10% validation, with the validation set used to confirm that the model was learning and out-of-sample loss was declining. We did not split out a test set, with the models instead evaluated on ATE and HTE prediction tasks on our experimental datasets.

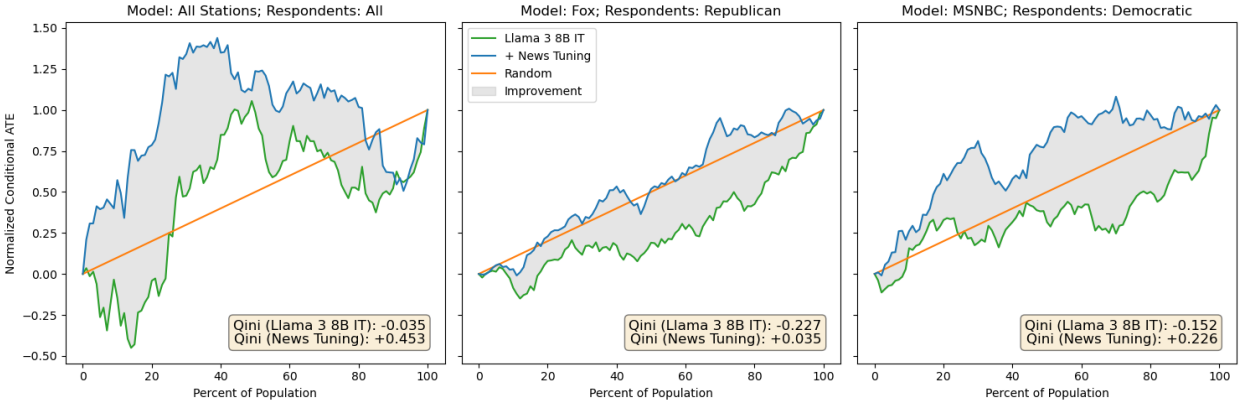
5.2.3 Results

At inference time, we take the appropriate checkpoint for each experiment, given the month it was conducted, and repeat the text2text, persona and RepReg studies from Chapter 4 exactly, using the same parameter settings and code. 17 of our 28 experiments fall within the window for which closed-captioning

[‡] https://github.com/huggingface/trl
[§] https://github.com/huggingface/accelerate



(a) Performance of all news-tuned models (all-data, MSNBC-only and Fox-only) ensembled on all respondents.



(b) Performance of each set of news-tuned models on the matching set of respondents: all-data models on all respondents, Fox-only models on Republican respondents, and MSNBC-only models on Democratic respondents.

Figure 5.1: Performance of the zero-shot text2text approach (see Section 4.4.1) with news-tuned models on various sets of respondents, compared to both the same text2text approach using the Llama3 8B Instruct model without news tuning and the ensemble baseline. The top pane shows the performance of all news tuned models ensembled together, on all respondents, while the bottom pane breaks each set of models out and compares them to the ideologically matching set of respondents. We find substantial improvement in all cases, over both the model without news tuning and the baseline, from contemporaneous news tuning. The curves shown here are calculated in the same way as those shown in Figure 4.3 and as described in Section 4.5.1.

data is available and are used for evaluation. We do this three times, for the all-data, MSNBC-only and Fox-only series of checkpoints. MSNBC-only and Fox-only models are evaluated respectively on Democratic and Republican respondents only, rather than on all respondents, because the greater ideological similarity they have to these populations ought to make them more informative there. We also evaluate the ensemble average of all three models on all respondents, to see whether using the broadest measure of news and news-derived perspectives brings additional benefits. We analyze both ATE prediction and HTE prediction quality. Average HTE prediction performance is evaluated by averaging Qini curves as described in Section 4.5.1.

Broadly speaking, the results demonstrate that we can realize large increases in performance with contextual news tuning. The zero-shot text2text results are shown in Figure 5.1, and show that all tested configurations improve with news tuning. Both the ensemble of all-data, Fox-only and MSNBC-only models and each of these models individually outperforms the stock Llama3 8B Instruct model, usually by a large margin. On the all-respondent evaluations, where we have the ensemble baseline from the last chapter (trained on cross-validation splits of all respondents) to compare to, both the all-data/Fox/MSNBC ensemble and the all-data model alone outperform the baseline while stock Llama3 8B Instruct does not. The ensemble model in particular provides a very large increase in performance, to a level that would be clearly useful in practice for targeting interventions.

Not all of these experiments produce good final performance, with the Fox-only model on Republicans failing to predict much better than chance. For theoretical purposes, though, it does *improve* over the stock model, which performs worse than random. News tuning, in other words, does impart useful information even here, though it also suggests the need for careful evaluation in practical settings.

Our other methods (not shown in the figure) are less improved by the news tuning process. Persona experiments consistently post worse performance than before tuning, and here the cause is made clear by inspecting the generated text: Despite our inclusion of instruction-tuning data to avoid it, the training process ablated some of the instruction-following capabilities. The trained models are much more likely than the stock checkpoint to produce repetitive generations, end output after simply reciting their assigned demographic traits, and otherwise fail to follow the prompt. The RepReg results for both hidden states and text2text probabilities as inputs, meanwhile, show improvement. Over the same four model configurations depicted in Figure 5.1 (All-data/Fox/MSNBC ensemble, all-data only, Fox only, MSNBC only) news tuning improves both sets of RepReg methods' Qini coefficient by an average of 0.103, to 0.212, worse than in the text2text case but still a 94% increase from the 0.109 observed without it.

ATE prediction results are shown in Table 5.1, with the stock Llama3 8B Instruct checkpoint's performance shown for comparison. The zero-shot methods' performance lines up well with what we observe for HTE prediction, with the text2text predicted ATEs becoming substantially more correlated with the true ATEs, while the persona method posts a small improvement but is likely impaired by the same ablation of instruction-following capabilities that hurt HTE performance. The RepReg methods, which have training data to work from and are near the ceiling of ATE prediction performance to begin with, see on average positive but much smaller improvement. Notably, the good results we observe here from RepReg on the persona probabilities combined with the poor performance of zero-shot persona may suggest the

	Llama3 8B IT	News Tuning	Δ
Text2text (ZS)	0.336	0.762	0.426
Persona (ZS)	0.546	0.589	0.043
RepReg: Hidden States (CV)	0.948	0.947	−0.001
RepReg: T2T Probs (CV)	0.963	0.965	0.002
RepReg: Persona Probs (CV)	0.982	0.992	0.010
Ensemble Baseline (CV)	0.992		–

Table 5.1: ATE prediction performance of the Llama3 8B Instruct model, before (left column) and after (right column) news tuning. As in Figure 4.6, the figures shown are Pearson correlation coefficients between (at the dataset level) the true ATE and the average of the individual-level predictions. Performance is evaluated on the 17 experiments conducted during the period for which we have cable-news data, with predictions for each experiment generated by the model trained on news data up to the month of the experiment. These results are for the three-way ensemble of all-data, MSNBC-only, and Fox-only models, and are evaluated on all respondents. ("ZS" = zero-shot; "CV" = cross-validated.)

presence of useful underlying signal and ablation-induced poor calibration. Results are also similar if we break out all-data models on all respondents, MSNBC-only models on Democrats and Fox-only models on Republicans. (Notably, ATE prediction performance is better for Democratic subsets than Republican ones, as in the HTE case.)

Finally, we should note that the differing performance of these methods (while practically important) is secondary from a theoretical perspective. That the text2text method experiences large performance improvement from news tuning demonstrates clearly that contemporaneous, contextual news data contains a great deal of relevant information for HTE and ATE prediction. This fact bears on debates in the theory of persuasion, lending support to views like that of Zaller[39] that stress contextual influence, whatever may be practically most useful for leveraging data about such influence.

5.3 Generalization: In-Distribution

News data, while useful, is not the only thing from which we might want to (and think we can) learn about persuasion. Persuasion experiments themselves are also a natural source of training data, but how useful they are for learning generalizable opinion-change dynamics is disputed.[48,50] In this section, we want to address this question by evaluating transfer learning in models trained or fine-tuned on these datasets. There are two kinds of generalization that are of interest: from some experiments to others and from some kinds of subjects to others.

This is considerably more complicated than working on experiments individually because not only the available predictors but, crucially, the meaning of the treatment indicator will differ between experiments. To be concrete, $T_i = 1$ does not convey the same information when the actual treatment administered was (say) a free-trade message as it does when the treatment was a change to the name of a candidate. We

thus need a generalizable way to tell the model what the treatment was, as well as to encode varying sets of predictors. Simple dummies for the different treatments are not useful because they are inherently unable to generalize to new experiments.

Language models, however, excel at this task, at least for the kind of text-based survey experiments used here. We can, as elsewhere in this dissertation, simply provide the text of the survey and available predictor variables as a prompt. Text provides a common format[31] and allows us to leverage the language model's world knowledge about how treatments ought to inter-relate. Transformer LMs also provide a powerful and high-capacity model architecture to learn potentially complex processes of persuasion.

Before exploring generalization or extrapolation out of the training distribution, to new experiments or new kinds of subjects, we first need to validate that our modeling approach can work at all. This section presents experiments testing whether we can learn to predict well within the training distribution, before moving on to extrapolation in Section 5.4.

5.3.1 Evaluation

For multi-dataset training, we can evaluate models in more than one way. As usual, we can compute the Qini coefficient on each individual dataset, and report a single metric by averaging them together. But we can also evaluate the Qini coefficient on pooled datasets; in this approach we simply combine all datasets together, taking a treatment assignment of 1 (or 0) from a given dataset to be the same as from any other dataset, and compute the overall Qini. Now, of course, because the treatments are not actually all the same, it would not make sense to compute this latter pooled metric in the usual tabular data setting. The detailed textual description of the treatment that we provide to the LM, on the other hand, should allow it to identify what treatment is actually being provided and predict appropriately (especially after training).

The pooled- and individual-dataset Qinis are both worth considering, because they test subtly but importantly different things. The average individual-dataset Qini measures, as usual, the model's ability to predict heterogeneity within a given experiment. The pooled-dataset Qini also measures the ability to predict a subject's relative responsiveness to *different* treatments. A high pooled Qini coefficient implies a model that has learned to predict the overall effect of treatments on different groups of subjects, as well as the heterogeneity in their response to each.

Both measures also have practical value: Individual-dataset Qini indicates how well a model can target a fixed treatment to a group of subjects, and the pooled Qini also captures the ability to assign people to treatments. Each of these problems arises frequently in practice.

5.3.2 Estimation Approach

Because we want to learn directly from these experimental datasets, statistical efficiency requires we leverage their causal structure in training. We can't simply treat them as text datasets on which to do the usual autoregressive, supervised fine-tuning. We need instead to specify how the model will learn causal structure, and fortunately (as discussed in Chapter 4) there are a number of statistical meta-algorithms for doing so,

into which one can plug a black-box function approximator like a neural network.

Our experiments in this section adopt perhaps the simplest approach: the S-learner, which trains one model on the consolidated experimental dataset, together with a treatment indicator variable, to predict the outcome variable. In the classical statistical formulation, the S-learner is trained with a discrete variable T indicating the actual treatment assignment, which is 1 or 0 as appropriate for each subject. At prediction time, the S-learner predicts to new subjects by generating two sets of predictions of the outcome, representing the two potential outcomes and with T set to respectively 1 and 0, and differences them to get an HTE estimate. (For a more formal definition of the S-learner and its relatives the T- and X-learners, see the original paper by Künzel *et al.*.[270])

In this section, however, we represent subjects in the same way as for the text2text approach of Section 4.4.1: a description of demographic variables and a filled-out survey, ending in a final question, *without* the final answer choice.[¶] We can thus avoid using any additional discrete variables, because the text itself encodes the treatment assignment. It also encodes the experiment identity, which allows us to skip experiment-specific dummies or other ways of trying to generalize from some experiments to others; the LM's world knowledge and the text's description of the experiment take care of that. The S-learner has the additional advantage of maximizing the amount of data available to each training process, in contrast to other algorithms like the T-learner (two separate models, trained on the treatment and control groups) or the X-learner (a two-stage, group-specific estimation strategy with a propensity score model).

Our setup, then, is simple: Train to predict the outcome on the pooled experimental datasets, and at test time generate the two sets of predictions and difference them.

5.3.3 Loss Function

The S-learner framework leaves open the question of exactly how to train the model to predict the outcome variable. Likert-scale data like ours can be represented in a number of ways: as classification, regression, or with more complex but more statistically principled ordinal models. As in the last chapter, we opt for the simplest approach of treating the Likert data as continuous and the problem as regression.[283] Our scales are all evenly spaced, and a regression loss will be able to capture the ordinal component of the data. There are several standard regression loss functions, but we employ the canonical and most commonly used mean squared error loss.

Minibatch moment-matching loss. Based on encouraging results in preliminary experiments, we've also evaluated what is to the best of our knowledge a novel training objective for HTE prediction problems. This loss function stems from a desire to employ HTE prediction directly as an optimization target, which unfortunately is difficult to do. No fundamental statistical principle forbids it, but the Qini coefficient and similar ranking-based uplift metrics are not differentiable in the parameters of the model producing the HTE predictions. Our core idea is instead to recognize that while HTE predictions are always evaluated by comparing their means to empirical CATEs, we need not consider only subsets defined by ranking

[¶] The prompts / textual representations used for each subject are exactly the same as in Section 4.4.1.

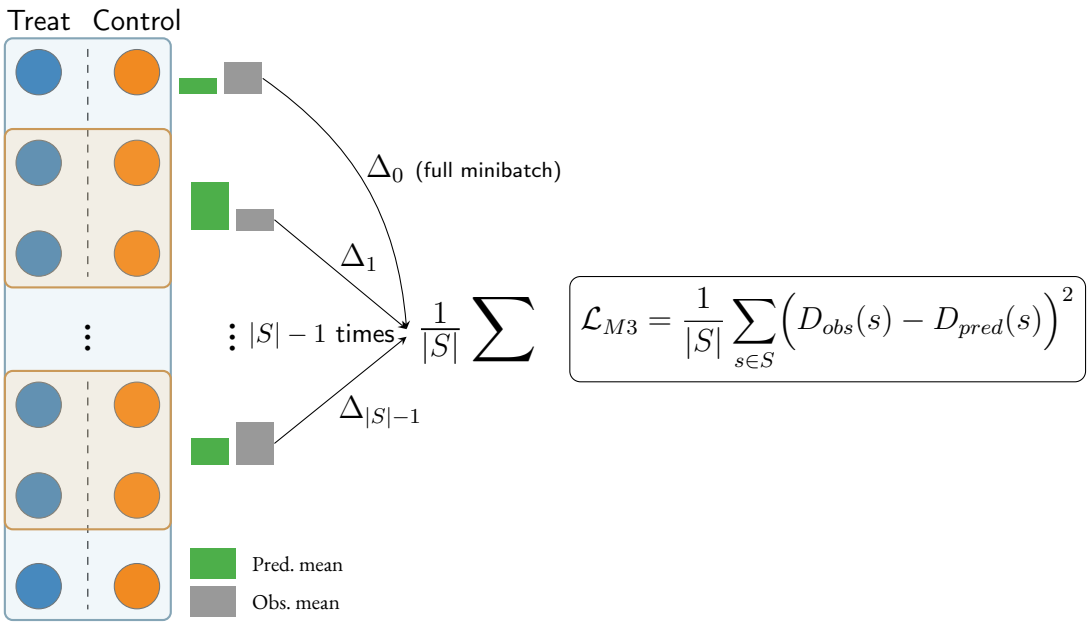


Figure 5.2: A graphical illustration of the M3 loss for neural HTE learning, demonstrating how the loss is computed from the minibatch and its subsamples.

the experimental subjects, as Qini and similar metrics do. We can instead simply sample appropriate (treatment-stratified) subsets from each minibatch, compute their mean HTE predictions and empirical CATEs, and input both to a regression loss like the mean squared error. (For efficiency, we stratify selection into each minibatch by treatment as well.) This approach is simple, easily implemented, differentiable in the model parameters, and should pair well with another loss that teaches the model directly to predict the outcome. It can thus be viewed as a kind of regularization that focuses learning on those predictors of the outcome that are also predictors of response heterogeneity.

Prior work has also incorporated ways to directly constrain ATE or uplift learning to match moments, including some we've covered already. Some of the baselines in Section 4.5.2 can be viewed this way. Tree-based methods like causal random forests[272], for example, can use as a splitting or termination criterion a measure of how accurately a leaf of the tree reproduces the conditional ATE among cases split into that leaf. DragonNet, a neural model, approaches this task in a different way closer to ours, adding learnable parameters and constraints to make the model an unbiased estimator of the overall ATE.[271] Still other work has considered ranking-based losses, despite the general difficulty of optimizing them, such as the recent model of He *et al.* that trains a multi-part architecture to rank subjects correctly in marketing contexts.[314] There are also differentiable surrogate losses, like upper bounds on AUUC,[315] and perhaps closest to our work attempts to enforce calibration guarantees over particular specified subgroups.[316] None of these methods, however, provide the same combination of features: (i) stochastic calibration on every training step, (ii) applied to arbitrary subsets and providing many auxiliary constraints rather than just one on the overall ATE, (iii) imposed on the cardinal values of actual and predicted CATEs rather than ordinal ranks, and finally (iv) usable with one simple pooled model, rather than multiple components.

To give a more formal treatment, take some positive and reasonably large minibatch size B. What follows is relative to the examples in this batch, and we index examples by $i \in \{1, \ldots, B\}$. As in the notation of Section 4.3, let $T_i \in \{0, 1\}$ be the treatment indicator for subject i, Y_i the observed outcome (the potential outcome under treatment T_i) and $\hat{\tau}_i$ the individual-level HTE prediction. We refer to the set of (indices of) all treatment cases by $I_T \triangleq \{i : T_i = 1\}$ and all control cases by $I_C \triangleq \{i : T_i = 0\}$, and refer to their cardinalities as $n_T \triangleq |I_T|$, $n_C \triangleq |I_C|$.

We can then define a collection S of *comparisons*, pairs of subsets of these arm-specific index sets, as follows: Let $S^{(0,0)} = (I_T, I_C)$ be the full-batch comparison. Then for each subset size k from some user-specified minimum size m up to $\min(n_T, n_C)$, we can sample without replacement s_k comparisons $S^{(k,j)}$, indexed by j, of balanced arm size k: $S^{(k,j)} = (I_T^{(k,j)}, I_C^{(k,j)})$, with $|I_T^{(k,j)}| = |I_C^{(k,j)}| = k$. As the notation suggests, $I_T^{(k,j)} \subset I_T$ and $I_C^{(k,j)} \subset I_C$. Each comparison $s = (s_T, s_C) \in S$ gives rise to observed and predicted differences

$$D_{obs}(s) = \frac{1}{|s_T|} \sum_{i \in s_T} Y_i - \frac{1}{|s_C|} \sum_{i \in s_C} Y_i \tag{5.1}$$

$$D_{pred}(s) = \frac{1}{|s_T|} \sum_{i \in s_T} \hat{\tau}_i - \frac{1}{|s_C|} \sum_{i \in s_C} \hat{\tau}_i \tag{5.2}$$

The M3 loss, as the mean squared error of the predicted CATEs on these subsets, is then

$$\mathcal{L}_{M3} = \frac{1}{|S|} \sum_{s \in S} \Big(D_{obs}(s) - D_{pred}(s) \Big)^2. \tag{5.3}$$

This estimator is shown graphically in Figure 5.2, and a numpy-like pseudocode version of it is given in Section B.3.

5.3.4 Model Architecture

We use a simple architecture with an MLP head applied on top of the base model's transformer layer stack. The original LM head is discarded. The MLP head consists of two linear layers, with a final linear layer mapping from the MLP head's hidden dimension down to a 1D regression prediction. Each layer is separated from the next by GELU and dropout. In our experiments, the MLP head's hidden dimension was 256 and the dropout probability was 0.1.

The input to the MLP head is the last hidden states of the transformer, pooled as follows:

1. Use mean-pooling; compute the mean of the hidden states over the sequence indices.
2. Take the hidden state of the last token specifically, the one right before generation would produce an answer to the final survey question.
3. Concatenate these two (for each example in the batch) into a vector of twice the dimension of the transformer's hidden states.

4. Append three additional numbers to the end (and extend the MLP head's input dimensionality appropriately), the same three numbers for each example in the batch: the minimum and maximum value of the answer Likert scale, and the number of points on the scale. The hidden states already capture this information (because it's in the prompt), but we observed benefit in preliminary experiments from making it even more explicit.

We considered and rejected an alternative approach which both a) treats the problem as classification and b) uses the outputs of the LM head at the last token as logits. This approach resembles SFT, but without allowing any position before the last one in the token sequence to contribute to gradient updates. Preliminary experiments did not suggest an advantage in performance, and GPU memory requirements were much higher.

5.3.5 Training Details

For these experiments, we use the Pythia-410m model[290] (of Section 4.10.1) rather than the Llama models. This model provides an advantageous combination of being fairly recent (2023), large enough (bigger than BERT-Large) to have useful world knowledge, and small enough to considerably simplify our implementation. Because we're evaluating a batch-level loss, the usual style of gradient accumulation will not work and the size of the physical batch becomes important. Designing and implementing an out-of-core accumulation method for this scenario would be too involved.

We pool all 28 experimental datasets together and randomly split them at the example level into 60% train, 20% validation and 20% test. We then train three different random initializations for both MSE-only and MSE + M3 (on the same splits), and report their average results in the next section. During training, selection into each batch is stratified by treatment to help stabilize the gradients. For better gradients, the target values are scaled to mean 0 and variance 1 within the model using dataset-wide statistics (and converted back to the original scale when making predictions). To ensure comparable M3 loss values between train and validation, selection into validation batches is stratified similarly. To measure how much the model learns as well as its final performance, we measure performance on the validation and test sets before training (with the untrained models) as well as after (and for validation, during).

Because, unlike in the news-data case, our experimental datasets are fairly small, we also train with LoRA rather than full finetuning. We use LoRA parameters r = 16, alpha = 32 and dropout = 0.1. The physical batch size is 24, which is also the logical batch size in the absence of gradient accumulation. The initial learning rate was `1.0e-4`, with a linear-warmup-decay schedule and a warmup ratio of 0.1. We trained with bitsandbytes' 8-bit AdamW optimizer,[308,309] as in Section 5.2.2, using the default hyperparameters and no weight decay. We again used gradient clipping, to a maximum norm of 1.0, mixed-precision training[311] with bfloat16, and Flash Attention 2.[312] We trained models in this section on 48GB NVIDIA RTX 6000 Ada Generation GPUs and used the Hugging Face transformers package[285] to load and manage the underlying language model.

The model selection strategy was as follows:

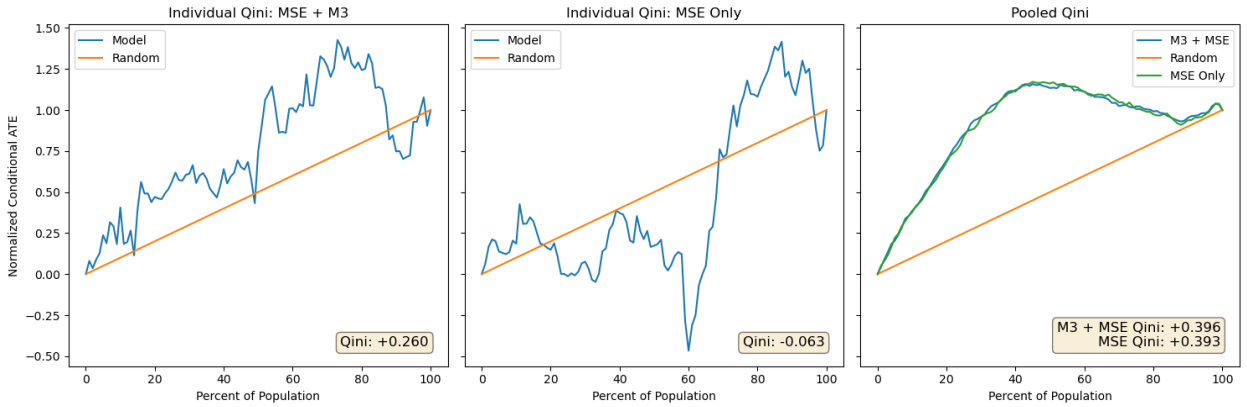


Figure 5.3: Test-set performance of the neural S-learner from the in-distribution training experiments, with and without the minibatch moment-matching loss term. The results show that a batch-level moment-matching loss term yields better generalization and improved prediction performance in the more difficult individual-dataset case. Both models perform almost identically on evaluation with pooled Qini. (Note that pooled-dataset Qini reflects both the ability to predict within-experiment heterogeneity as well as relative responsiveness to different treatments. Individual-dataset Qini addresses only HTE prediction performance. Before computing pooled Qini, we normalize both the outcome and the predictions at the experiment level, by subtracting the minimum outcome value and dividing by the range.)

- We trained each model for 15 epochs, evaluating and saving the adapter weights every half-epoch, without early stopping.
- At each evaluation, we computed the pooled-dataset and average individual-dataset Qini coefficients.[‖]
- At the end of training, we took the best model by validation-set pooled Qini.

We ultimately trained two model configurations: one with MSE loss only, and one with both MSE and M3 (summed, equally weighted). Training was in all other respects identical for both configurations. Both configurations were trained three times on random initializations with the same dataset splits; results reported below are the averages of each set of runs.

5.3.6 Results

The results, shown graphically in Figure 5.3 and in the first column of Table 5.2, are clear. Within the training distribution, our models can and do learn to predict treatment response heterogeneity across experiments. The S-learner with the minibatch moment-matching term in the loss also learns to do so meaningfully better than the MSE-only model does: Individual Qini coefficients are higher for the M3 model, and indeed in the individual-dataset case the MSE-only model performs worse than random.

[‖]An important but subtle implementation detail is that computing the Qini requires evaluating the model also on counterfactual data points, with treatment assignments not actually used in the experiment, while the loss computation requires only factual data points.

Notably, we also find competitive performance with what Chapter 4 reports on average for its six methods. The individual-dataset Qini for our trained M3 model is 0.260, which would be the third-best evaluated, ahead of all RepReg methods and the ensemble baseline (all of which also had data, in fact more data – two of three cross validation folds, or 66%, vs 60% here – to train on). This good performance occurs even though we use a much smaller and less capable model (Pythia-410m) rather than one or more Llama3 variants. Both architectural improvements and data may contribute to these results.

The individual-dataset case can be interpreted just like prior results in this chapter and Chapter 4: Given a treatment, the model displays good performance at predicting how responsive subjects will be to it. Good performance in the pooled case instead demonstrates that the model can predict *which* treatments which people will respond to. Now, of course, how hard this task is depends substantially on how much heterogeneity there is in each experiment. In the sharp-null case of exactly constant treatment effect in all experiments for everyone, achieving the best possible performance at it is trivial: Simply identify which treatment the subject has been assigned to (which the embeddings encode), and predict its ATE. According to the hypothesis tests in Section 4.6.2, however, we have significantly more heterogeneity than this. The pooled case is accordingly at least somewhat more difficult than just predicting the ATE, even if the within-experiment prediction task that does not depend at all on ATE prediction is harder (as lower performance there shows). Pooled-dataset Qini will provide a stricter test in the following sections, where demographics and entire experiments in the test set are held out from the training distribution.

Notably, the M3 model in particular performs well at both pooled- and individual-dataset Qini evaluations. The training objective, which matches moments within the minibatch, does not just teach the model to learn the ATE, because we use *mini*-batch gradient descent and further sample subsets of each batch. It pushes the model, in other words, toward being able to accurately line up the differences in means across arbitrary subsets. This optimization pressure ought to, and from our results apparently does, lead to improvement in multiple components of prediction: experiments' ATEs, prediction of responsiveness within an experiment, and even prediction of relative responsiveness across experiments.

All in all, these experiments fulfill our goal for this section: Demonstrating that our modeling approach is capable of learning generalizable information about treatment heterogeneity across experiments. This finding in the within-distribution setting adds credibility to either positive or negative results in the following section's cross-dataset and cross-demographic experiments. It has two additional benefits:

- Validating the moment-matching loss as a viable option for improving HTE modeling performance with neural models. Future work should compare this objective to other models, such as doubly robust estimation methods.[317]

- Demonstrating a way to learn not just *how* responsive a subject will be to a treatment, but *which* of several treatments the subject will respond best to. The out-of-sample predictions of our models, as evaluated by the pooled-dataset Qini coefficient, demonstrate good performance at this task.

5.4 Generalization: Cross-Experiment and Cross-Demographic

Next we move on to perhaps the most ambitious set of experiments in this dissertation, testing whether there's a useful degree of generalization across experimental contexts and subject populations. An affirmative answer would have obvious practical and theoretical value. Building on the validation of the M3 objective in the previous section, we employ it here as well and also compare to the MSE-only variant.

5.4.1 Data Splits and Training Details

We use exactly the same training setup as for the within-distribution experiments (Section 5.3.5). Only the way train, test and validation splits are created differs:

Cross-Experiment We randomly split the set of 28 experiments into a train set of 18 train experiments and 10 test experiments. The train experiments were further split randomly at the subject level, without regard to experiments, into an 80% train set and a 20% validation set. The test set is thus entirely experiments the model has not seen during training. We generated three independent such splits and report the average results over all of them in the next subsection.

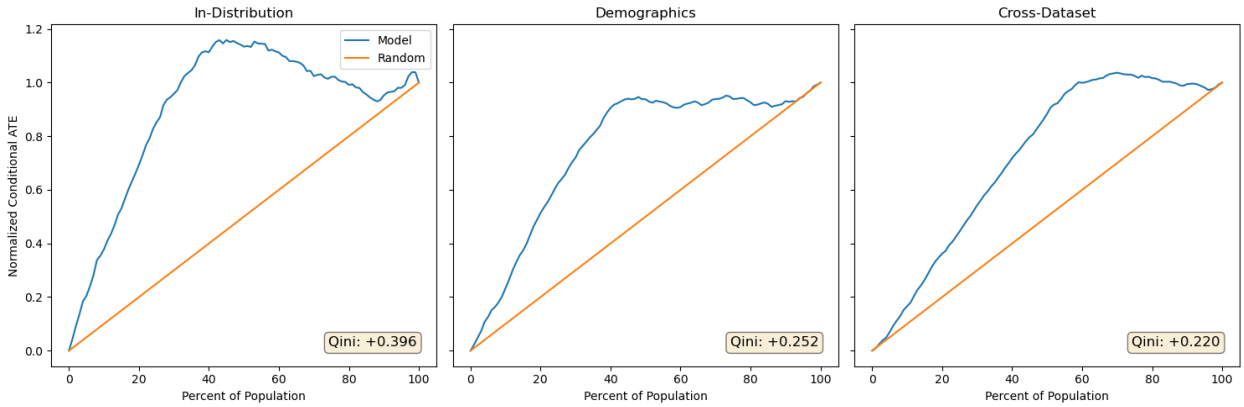
Cross-Demographic We take three demographic variables present in all experiments and dichotomize them: race, into white and non-white; sex or gender, into male and female; and political party, into Democrat and Republican. Any subject without a value for these variables (or who identified as an independent, third-party, etc, for party) is dropped. For each variable, we train twice, with each group providing the training set (out of which we split a 20% validation set as above) and the other group the test set. There are thus six models all told, each evaluated on a subject population it has not seen during training.

5.4.2 Results

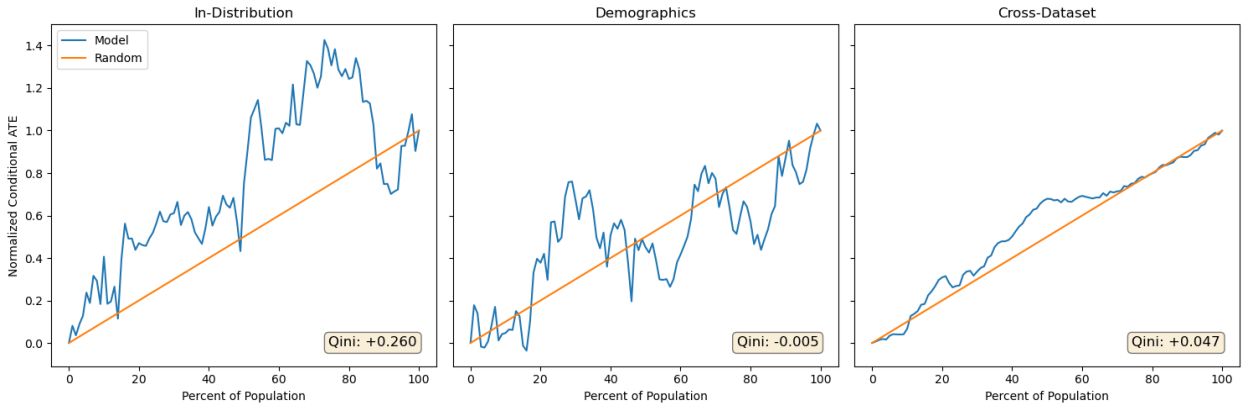
This is a considerably harder prediction task than the in-distribution case of Section 5.3, and performance is correspondingly lower. In fact, as can be seen in Table 5.2, the test-set loss values are systematically higher than in the in-distribution case. The model has acquired less generalizable knowledge, plausibly because there is simply less of it to acquire.

HTE prediction performance in the pooled case, while lower, is still nontrivial, with the cross-dataset pooled Qini (the lowest) reaching a value of 0.220. Combined with much lower individual-dataset Qini coefficients, this implies that the model has acquired some ability to predict the effects of treatments and relative responsiveness to them, but little if any of the finer-grained ability to predict how responsive different people will be to the same treatment. We also observe mostly better performance, in 5 out of 6 cases, from the models trained with the MSE + M3 objective compared to the MSE-only one.

How to interpret this result is not entirely clear: The models have learned something during training, but the absolute level of generalization is worse than in our other experiments. Because the ceiling on



(a) Average Qini curve (across runs) over **pooled datasets**.



(b) Average Qini curve (across runs and datasets) over **individual datasets**.

Figure 5.4: Average test-set performance of the S-learner (with both MSE and minibatch moment-matching objective). We show results on all three sets of generalization experiments, where train/test splits are given by, from left to right, random sampling within each dataset ("in-distribution"), demographic variables, and entire datasets. Qini coefficients over pooled datasets (top) correspond to the pooled training task, and reflect both within-experiment HTE prediction performance and the ability to learn to predict responsiveness to different treatments. Average individual-dataset Qinis reflect only the ability to learn to predict response within each experiment. Generally better performance in the pooled case indicates that average treatment effects and relative responsiveness to treatments can more easily be predicted from training on other similar subjects, demographics or experiments than can individual-level responses within a fixed treatment.

	In-Distribution	Demographic	Cross-Dataset
Pooled Qini (MSE + M3)	0.396	0.252	0.220
Pooled Qini (MSE)	0.393	0.309	0.213
Δ	0.003	−0.057	0.007
Indiv. Qini (MSE + M3)	0.260	−0.005	0.047
Indiv. Qini (MSE)	−0.063	−0.343	0.023
Δ	0.323	0.348	0.024
Loss (MSE + M3)	1.245	1.389	1.397
Loss (MSE)	0.868	0.950	0.976

Table 5.2: Test-set performance of models with and without the moment-matching loss term from the in-distribution, demographic and cross-dataset fine-tuning experiments. These results average over three runs each for in-distribution and cross-dataset, and over all tested demographics for the demographic experiments. The results demonstrate generally better performance from the models with the M3 objective, and a practically useful level of prediction ability in all pooled cases and the in-distribution individual-Qini case. The absolute level of performance, however, is low for all models in the individual-Qini demographic and cross-dataset experiments.

performance in this setting is unobservable (see Section 4.5.1), we can't tell how far away these results are from optimal, and whether better models might achieve better results. While our findings have theoretical value — we find at least some generalizable knowledge about opinion change can be learned across datasets and experiments – they suggest the need for further investigation to clarify how much, if any, additional generalization is possible and in what experimental settings.

5.5 Conclusion

These results provide further demonstration that transfer learning can improve performance at predicting opinion change, and provide several practical ways to do so.

First, we observe improved performance from fine-tuning on contemporaneous news data, both in HTE and ATE prediction tasks. The details can be tricky, as with the impairment of instruction-following performance from our fine-tuning process, but the possibility of large performance gains is clear. This is useful for the practical task (in politics, marketing, and indeed also research) of predicting who will be most persuaded by an intervention, and also for what it says theoretically. Such results lend support to views like those of Zaller[39] (and agree with other pieces of empirical evidence[3]) which stress the role of context and especially media in influencing opinion formation. If, in other words, knowing what's on the news help a model learn how people change their opinions, that same information may be helping people themselves do so. Now, of course, few people in our experimental datasets will be obsessive consumers of multiple news channels, but processes of social influence and diffusion[5] can reach even those who are not directly exposed to opinion-leading parts of society.

Our results on transfer learning between experiments and types of subject are more equivocal. We find training on disjoint experiments and subject types yields little if any improvement in performance in HTE prediction within an experiment, and a more moderate amount in the pooled setting. Combined with the predictability of opinion change from within-experiment training data (Section 5.3) and other kinds of data (like news), these results argue against a thoroughgoing version of the "persuasion in parallel" hypothesis[48] and for more contextually dependent views.[50] Our experimental datasets, on average if not always individually, have a statistically significant amount of heterogeneity to predict (see Section 4.6), but the information that is most predictive is broad real-world context: language model pretraining corpora and bulk news coverage.

This chapter has also made some technical contributions:

M3 loss for HTE prediction The M3 objective (for **m**inibatch **m**oment-**m**atching) we've developed and tested here performs well at both within-experiment and cross-experiment HTE prediction, as shown in Section 5.3. Though there is related work, this objective is to the best of our knowledge novel. We find that adding this term to the usual regression loss markedly enhances the S-learner's performance in predicting heterogeneity. Applying it to larger models or batch sizes, where GPU memory would become constraining, does pose some engineering challenges. Multi-pass out-of-core accumulation techniques, which have been used for other cases of training neural nets on batch-level objectives,[318] may be useful.

Predicting relative responsiveness to different treatments The pooled-Qini evaluation in Section 5.3.6 demonstrates that models trained on experiments can perform well in the widely encountered problem of predicting out-of-sample which subjects will respond best to which treatments. We observe nontrivial performance here in both in-distribution settings and extrapolation to other kinds of subjects and other experiments.

A final practical takeaway from our results both here and in Chapter 4 is the sensitivity of HTE prediction performance. It is possible to do well at this task, indeed to do much better than classical baselines, but it is not simple. Details matter, and apparently similar methods can have sharply different results. More research is needed into the causes and extent of this phenomenon.

Chapter 6

Conclusion and Future Directions

6.1 Summary of Contributions

Our work in this dissertation makes both substantive and methodological contributions, from demonstrating faster and more outraged news cycles on social media to developing LLM-assisted methods to predict opinion change. We summarize these contributions briefly in this chapter. (Note that Section 6.1.1, as with Chapter 3, was adapted from the published paper about these results.[194])

6.1.1 Media and the News Cycle

The rapid evolution of the Internet is reshaping the media landscape, with frequent claims in both academic and popular sources of an accelerated and increasingly outraged news cycle. We test these claims empirically, investigating the dynamics of news spread, decay, and sentiment on Twitter (X) compared to US talk radio. We analyze data from 2019 to 2021, including 518,000 hours of radio content and 26.6 million tweets by elite journalists, politicians, and general users. Our automated event detection methods identify 1,694 news events.

We find that news on Twitter consistently both circulates faster and fades faster. These patterns are consistent across various robustness checks: We find similar results from a set of manually identified keyword-based events, breakdowns of the data by ideological affiliation, and matched triples of events (intended to reflect the same real-world occurrence) across elite journalist/politician Twitter, general firehose Twitter and radio. Our comparison of the rate of decay of attention to news is to our knowledge the first such large-scale study for Twitter and traditional media.

We also find that Twitter news is more negative and outraged compared to radio, with Twitter outrage also more short-lived. To the best of our knowledge, this is also the first large-scale comparison of outrage between Twitter and traditional media, and the finding that Twitter is significantly more outraged and negative has implications for media regulation and platform design. Also relevant is our finding that differences in the lifecycle of outrage between media occur almost entirely at the level of events, rather than within events: Twitter as a platform selects for different kinds of stories than radio, ones which are

more outraged up front.

Our results illustrate an important way social media may influence traditional media: framing and agenda-setting simply by speaking first. As journalism evolves with these media, news audiences may encounter faster shifts in focus, less attention to each news event, and much more negativity and outrage.

6.1.2 Heterogeneity Prediction, Generalization, and Opinion Change

Our work in Chapter 4 and Chapter 5 shows that generalist foundation models, both zero-shot and with further training, offer a powerful paradigm for causal inference in social science settings. Chapter 4 explores the capabilities of off-the-shelf models, while Chapter 5 drills down on which kinds of pretraining data are most helpful to this task and explores training strategies and objectives.

Predicting individual and average treatment effects. Our most important result is that using only a pretrained language model, data about the respondents and a survey instrument, one can obtain excellent zero-shot performance at predicting both heterogeneous treatment effects (HTEs) and average treatment effects (ATEs). The findings about average effects agree with and reinforce recent research on that question,[50] and those about heterogeneous effects break new ground. Our zero-shot HTE predictions substantially and significantly outperform an ensemble of strong but data-hungry baselines that do not leverage world knowledge acquired from a pretraining corpus. The baselines max out ATE prediction performance, of course, because estimating a difference-in-means from a substantial amount of data is not difficult. HTE prediction and zero-shot ATE prediction are much harder tasks, ones our methods excel at.

In other words, we've demonstrated methods with greater statistical efficiency. Both ATEs and HTEs can be learned from enough data, but conducting field experiments and randomized trials is expensive. Methods which can substitute for these experiments, or reduce the size of those that need to be conducted, have considerable value.

We also find some puzzling aspects of LMs' performance at these tasks, which ought to spur further research. In our experiments, post-training sometimes degrades performance, larger models do not always perform better, and there are even cases of inverse compute scaling. We also find, consistently across many analyses, that ATE and HTE prediction performance are almost totally uncorrelated, with the most plausible explanation being that they simply draw on different model capabilities. All of this points to the need for deeper investigation of which techniques and data matter most in causal-inference tasks.

Training on bulk news data. The results above clearly show transfer learning from the language models' pretraining data to ATE and HTE prediction tasks. Further experiments in this dissertation took up the closely related question of *which* datasets, as training data, generalize well to these tasks. We explored two kinds of data with strong theoretical motivation: bulk news coverage and other experiments. These analyses also explore the important technical question of which architectures and training setups are able to leverage these datasets for improved performance. For news, we find that fine-tuning in the usual self-supervised autoregressive way on contemporaneous news coverage is extremely effective at boosting

both the ATE and HTE prediction performance of language models. News tuning takes one of the worst-performing models in chapter 4 (Llama3 8B Instruct) all the way up to the best HTE performance observed in any of those experiments, a very large and practically useful improvement. Large and consistent effects are observed in many configurations: training on all available data and evaluating on all respondents, training on Fox and evaluating on Republicans, training on MSNBC and evaluating on Democrats, and best of all ensembling these three models to evaluate on all respondents. There are still some technical difficulties to sort out, including the ablation of instruction-following performance we observed in Section 5.2. But these are not showstopping problems, and the main takeaway is the both practical and theoretical value of showing that news coverage is a strong predictor of opinion change.

Transfer learning from other experiments and kinds of subjects. Our experiments on transfer and generalization between experiments and kinds of subject find more mixed results. We conducted first a phase-1 study to demonstrate that models could learn to predict heterogeneous effects out-of-sample but in-distribution, with test sets randomly split off from a pooled group of experimental datasets. In the course of doing so, we develop and introduce a novel training objective ('M3') for neural HTE estimation, based on a **m**oment-**m**atching loss within each **m**inibatch.

We evaluate these models in two ways: pooled evaluation with the Qini coefficient over all experiments at once, and with the average of performance on each dataset at a time. Pooled evaluation tests a broader array of abilities: simultaneously ATE prediction, experiment-level responsiveness prediction for subjects, and within-experiment heterogeneity prediction relative to the ATE. But because the within-experiment prediction task is the hardest, pooled evaluation is less challenging than the simpler, stricter task of within-experiment HTE prediction. Both are practically relevant, with the pooled case testing in particular the ability to predict how well different subjects will respond to different treatments.

The first, in-distribution phase was a clear success: We find first that our M3 objective increases performance compared to a baseline model which did not use it, and also that the model trained with it can effectively predict heterogeneous effects. Performance is better when pooling over experiments than when evaluating within single experiments, but is nontrivial for both. The model without the M3 loss term performed well only in pooled evaluation, not in the stricter individual-dataset case.

The second phase tested models both with and without the M3 loss term on out-of-distribution generalization, with test sets composed of entire held-out experiments and held-out demographic groups within experiments. In this case, there is also a nontrivial (though smaller) degree of transfer in the easier, but still practically relevant, pooled evaluation, with our models having acquired some ability to predict which subjects are most responsive to which treatments. The harder case of HTE estimation within experiments works better than chance only in-distribution; in these extrapolation experiments, we do not find any generalizable knowledge transferring across datasets or kinds of subjects.

6.2 Limitations and Future Work

While the work presented in this dissertation makes real contributions (as discussed just above), it also has some real limitations. Addressing them provides several natural jumping-off points for future work.

6.2.1 Media and the News Cycle

The limitations of our work on the news cycle are about scope and causality: We've exhaustively examined and tested differences in the news cycle between media, but only for certain media and time periods, and we can't say for sure why these differences exist. Future work should pick up both threads. In particular:

Scope of media The work in this dissertation uses only one social media site and one broadcast medium. The work would be stronger, and future work should replicate our analysis, with data from more social media services as well as broadcast and cable TV.

Recent changes We have limitations of temporal scope as well. Twitter, now X, is rapidly changing and many new platforms like Mastodon and Bluesky are appearing and growing, but these changes came after the end of our data collection period. How well do our conclusions hold in this new world? In particular, does the existence of several Twitter-like services interacting with each other give rise to interesting new dynamics?

Cross-medium influence A natural hypothesis in light of our results is that Twitter's faster news cycles give it (or gave it) influence over other media, and propagated its biases out to their coverage of events. But we can only speculate about this form of intermedia agenda-setting, and future work should address the question more definitively.

Causes of greater speed We are also unable to say much about the causes of faster news cycles and more rapid forgetting on Twitter. Some features of the platform must make this happen, but we have not explored which ones it is. Some possibilities include the existence of an algorithmic recommendation system, the very short nature of the posts (280 characters during this period), and the extremely low friction of sharing others' messages. Future work should take up these and other possibilities and try to answer this question.

Causes of negativity Similarly, the origin of Twitter's much greater propensity to negativity and outrage than radio is unclear. There are plausible guesses, some of which we've discussed in Chapter 3: in particular, a reinforcement process involving audience interaction with journalists, and shorter story lifecycles selecting for outrage. But we have not carefully examined this question, and it would be a very valuable one for future work to take up.

Overall, this work points toward and suggests the need for a science of online platform design. Systematic, empirically informed but theoretically grounded, interdisciplinary investigation into how platforms'

features influence their cultures would pay large dividends. It is possible that the platforms themselves have already done some of this work, but if so (as with Wall Street research on trading strategies) they have not published much of it. Other researchers should fill in this lacuna in the public literature, which would be useful for both science and policy.

6.2.2 Heterogeneity Prediction, Generalization, and Opinion Change

Our work on persuasion and opinion change also has some natural next steps that would address its limitations:

Scaling behavior and variable performance The unusual and sometimes inverse scaling behavior observed in Chapter 4 is in need of further research and explanation, as is the variability in performance between different models. Pulling on these threads may provide both practical performance improvements and greater insight into the underlying mechanisms of opinion change that models learn.

Interpretability Similarly, we have made only limited efforts to understand what happens within the models as they perform these prediction tasks. Interpretability tools may help illuminate *how* they make predictions, and when or whether they can be used as models of human psychology. How much the internals of the models accurately represent human response to persuasion would be a valuable question to explore.

Binary treatment Our work has considered only dichotomous treatment (a control group with a single treatment group), which are simpler and more tractable than general experimental designs. Future work ought to take up this question and develop methods aimed at more complex designs, including such settings as ordinal treatment (e.g., dosage levels) and conjoint experiments.

Observational settings We have considered experimental settings here, but randomized experiments are difficult and expensive to conduct. Does an LM with general world knowledge from pretraining also make it easier to do observational causal inference? The most obvious way for such world knowledge to be useful is more efficient inference of the assignment mechanism, either with lower bias or lower variance or both. Future work might attempt to use LLMs for propensity score estimation in observational settings, though validating the inferences drawn would be difficult.

From persuasion to persuasiveness: improving generation Our work has not targeted the generation performance of the model or tried to optimize the text it generates for persuasiveness. In practical terms, we have not provided much of an aid to writing persuasive messages, only to testing and targeting them. Future work might attempt to tune LMs so that they speak persuasively in some context, leveraging supervision from experimental datasets. Relatedly, one might try to optimize a particular starting message. There are, of course, potentially serious ethical issues with training language models to be persuasive. This limitation of our work is not necessarily a real drawback, and future

work should think carefully about whether to pursue this line of reasoning and how any harms can be mitigated or prevented.

Our work on opinion change fits into a broad recent literature on the use of language models as simulated participants, subjects or agents, which attempts to study human behavior *in silico*. For opinion change in particular, these results suggest the possibility of simulated message tests and other uses of LMs similar to the way medical research uses animal studies. Hypothesis generation and iteration may become easier and cheaper than they currently are.

6.3 Data and Code Availability

Analysis code and replication data for the news-cycle analysis of Chapter 3 are available in the repositories provided with the original publication. For the code, that's https://doi.org/10.6084/m9.figshare.24546400; for the data, https://doi.org/10.6084/m9.figshare.24454777. See the published paper for more information. For the causal-inference work of Chapter 4 and Chapter 5, analysis code and replication data will be released after publication. Code will be available at https://doi.org/10.6084/m9.figshare.28929302 and data at https://doi.org/10.6084/m9.figshare.28929296.

6.4 Ethical Considerations

No human-subjects issues arise in any of this work, because our datasets either did not involve human subjects (the Twitter, radio and cable-news data of Chapter 3 and Chapter 5), or were anonymized replication data (the experimental datasets of Chapter 5). Given the highly public nature of radio, cable television and Twitter, we do not believe there are important privacy concerns to address in connection with those datasets which were not anonymized upfront. We have addressed copyright and contractual obligations by not releasing the full raw data for Twitter, radio and cable-news datasets. See the published news-cycle paper for more information.[194]

In addition to their research value, we hope that greater insight into news-cycle dynamics and processes of opinion change can have social benefits through more informed policymaking and discourse. The work of Chapter 4 and Chapter 5, however, requires more careful attention to its ethical implications. While any work on tools for enhancing persuasion raises some ethical questions, we considered these questions and ultimately concluded that this work met the bar. First, and most importantly, people cannot be simply mind-controlled: Persuasive effects are typically not large. As shown in Table 4.1, most treatments do not make for large changes of opinion (nor are such changes necessarily lasting). The typical practical application of methods like those we study here is to optimize a nudge of some kind: to vote, to buy a product, even to take a vaccine. The fact that we can accurately predict the effect of a treatment does not imply, necessarily or usually, that the effect is large.

Moreover, HTE estimation methods already exist and are widely used in industry; while our work improves upon these methods, it builds on what came before. Conversely, we believe that the methods

and results in Chapter 4 and Chapter 5 have much greater scientific value: If LLMs can be used as in-silico proxies for humans, we can far more easily study aspects of opinion change that are difficult or expensive to study with humans.

It is, of course, possible that better understanding of how persuasion works may allow various kinds of bad actors to be more persuasive. (For example, in silico trials of persuasive messages may allow more accurate selection of effective ones.) This potential for misuse must be weighed against the benefits to prosocial and morally neutral applications. Indeed, technology never has a straightforward moral valence, but rather can always be used for good or ill. Exactly this potential for misuse provides an additional spur to research: Bad actors will (by definition) not unilaterally avoid misusing technology, and there is social value in understanding and being prepared for whatever is possible.

6.5 Discussion

The results we've shown in this dissertation contribute to a number of areas of research, from political science (the study of public opinion) to computer science (the study of language models) to the interdisciplinary field of causal inference. Each of these areas bears in a different way on the questions addressed here, and we hope that our results can prove useful to a diverse set of research communities. Further research, indeed, is needed; rapid change in media ecosystems and in the capabilities of language models mean that what was known yesterday may need to be updated for tomorrow.

Research on the path we pursue in Chapter 4 and Chapter 5, in particular, can contribute to an ambitious goal: building useful foundation models and inference techniques for causal tasks, especially in social-science settings. These models do not need to learn a hypothetical, and unlikely, general common factor of persuadability, and indeed are more useful if there is no such factor. The more opinion change depends on context, the greater value there is in distilling much of the context into a usable computational artifact. Such models could help considerably with predicting response to various interventions, accelerating and reducing the cost of research, and (for our purposes) allowing more effective in-silico tests of theories about opinion change. Making progress on such an agenda obviously requires exploring useful datasets and data mixtures, and figuring out which ones contribute to good performance on which tasks. Because (unlike with next-token prediction) counterfactual outcomes aren't observable, it also may require exploring different training objectives. Our work in this dissertation contributes to both areas.

Nor can the media upstream of these opinion-formation processes be ignored. The influences on opinion formation come from all throughout society, but the rubber meets the road in the increasingly digitized ways people interact with each other. As these platforms develop further and their social impacts become more apparent, research can help inform regulation and platform design. We hope that this work takes a useful step in that direction.

Appendix A

Speed and Sentiment of News: Further Details and Robustness Checks

Here we provide further details, plots and breakdowns for the robustness checks conducted in Chapter 3, as well as one additional check in Section A.1. See the main discussion of results in Section 3.5 for the context this work fits into. The analysis in Section A.1 was first published in the supplementary information to a 2024 article by Brannon & Roy.[194] The analysis in the subsequent sections, presenting robustness checks for Chapter 3, has not previously appeared elsewhere. Though the results are novel, as in Chapter 3 a small amount of text here may have previously appeared in the MS thesis that first developed a similar dataset.[195]

A.1 Twitter-Radio Structural Similarity

Our first robustness check aims to demonstrate that elite Twitter and radio represent similar underlying social and information-disseminating structures. We would expect them to: Both media are part of the same tightly coupled world of political and journalistic elites, and ought to reflect a common (partially offline) social world. Our analysis finds exactly this, providing reassurance that slower decay on radio is an effect of the medium rather than an artifact of differences between our datasets. Our results are therefore not just about measuring differences between who is active on these media. In order to get a closer match between the collection times of radio speech and the Twitter follow graph, the analysis below uses only 2019 and 2020 data.

Concretely, we compare the elite Twitter follow and mention graphs to the *coairing graph* for radio, in which two shows are connected if they air on the same radio station (within a week of each other, a restriction intended to minimize the impact of errors in webscraped schedule data). This graph, which does not depend on the content of shows, reflects corporate syndication and programming decisions, and we view it as an observable analog of the Twitter graphs. It embodies the same social and informational aspects[99] as Twitter: Hosts whose shows air on the same station are more likely to have professional connections than random host pairs, and corporate programmers have clearly judged that such shows are

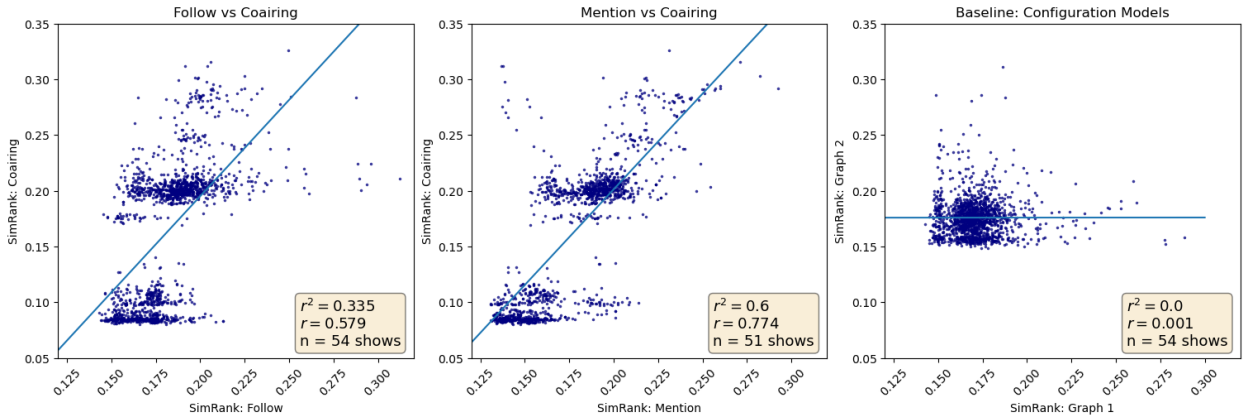


Figure A.1: Relationships between SimRank similarity scores for all pairs of non-isolated nodes in the follow, mention and coairing graphs on shows. The leftmost two plots compare coairing SimRank to follow-graph and mention-graph SimRank; the rightmost plot illustrates a baseline comparison of two configuration-model graphs having the same degree distributions as the follow and coairing graphs. Note that because SimRank is uninformative for nodes with no edges, all comparisons exclude show pairs in which either show was an isolate in either graph being compared. (Isolates are, concretely, shows whose associated Twitter accounts followed or were followed by no other shows' accounts, and analogously for mention and coairing graphs.) The follow-coairing comparison thus includes 54 shows or nodes out of 67, while the mention-coairing comparison includes 51; note that a larger sample of shows would be expected to include more social context and produce fewer isolates. For comparability with the undirected coairing graph, we ignored edge directions in computing follow and mention SimRank scores. One outlier show pair has also been excluded from each of the two left panels (though not the r^2 calculations) for ease of visualization.

similar enough to be of interest to the same audience. If we were comparing to follow or mention graphs involving the shows' audiences, the relationship might plausibly run the other way, with shows similar on Twitter because they air on the same stations and thus acquire the same listeners. Using Twitter data for elites instead mitigates this confounder.

The full coairing graph contains 785 nodes, one for each show in the 2019/2020 portion of the final radio corpus, connected by 21,927 edges. The graph has two connected components: one with 11 shows, from a radio station in Detroit which airs only local content, and another with the other 774 shows from all other stations. To allow a clean comparison with Twitter, we analyze only the Twitter-matched coairing graph: the subgraph over shows which have been matched to Twitter handles, all of which are in the larger component. For brevity, we refer to this graph simply as "the coairing graph" below. There are 67 Twitter-matched shows in total.

The coairing graph has quite similar structure to both the follow and especially the mention graphs. In addition to certain summary statistics presented in Table A.1, we demonstrate this similarity with the SimRank node similarity algorithm.[319] SimRank generates similarity scores for pairs of nodes in a graph, with nodes more similar the more they occur in similar contexts in the graph. Figure A.1 compares SimRank scores for all pairs of nodes in the coairing graph with two other sets of SimRank scores from

Statistic	Coairing	Follow	Mention
Order (# nodes)	56	55	52
Size (# edges)	434	370	343
Average degree	15.50	13.45	13.19
Transitivity	0.67	0.53	0.57
Avg. Clust. Coef.	0.77	0.60	0.65

Table A.1: Summary statistics of the follow, mention, and coairing graphs for the set of Twitter-matched shows, demonstrating a substantial degree of similarity. These statistics cover the large connected component of each graph, omitting a few shows out of the full 67 which were isolates.

the follow and mention graphs over radio shows. (Recall that we aggregated the follow and mention graphs over the 203 radio-linked Twitter accounts, or "radio accounts" for short, up to the level of shows.) Coairing SimRank is strongly correlated with follow and especially mention SimRank, much more so than in a configuration-model baseline, highlighting that shows are in similar social and informational positions on Twitter and on air.

In particular, coairing SimRank is strongly bimodal, with two clear clusters visible above and below a threshold value of about 0.15. Manual inspection suggests these clusters divide along left-right ideological lines: similar pairs are mostly those where both shows are conservative or both are liberal. Follow-graph and especially mention-graph SimRank, while less ideologically polarized, replicate much of this structure.

A.2 Event Lifecycles by Year

Section 3.5.1 describes a consistent pattern in which elite Twitter has the fastest event lifecycles, followed by the firehose, followed by radio (with the single exception of Covid-related events in 2020). We noted that the results are similar if events are broken down by year, and indeed they are: The breakdown by year is shown in Figure A.2 for empirical CDFs and empirical PDFs, and in Table A.2 for statistics of relative times, with consistent results (except for 2020 Covid-related events) in all three years.

A.3 Matched Events

As shown in Figure A.3, the empirical CDFs of mentions of the (elite, radio, firehose) matched triples of events reinforce the results presented in Section 3.5.2 and find that elite Twitter discussion rises and falls faster than radio discussion, with (except for 2020 Covid-related events) firehose discussion intermediate between the two.

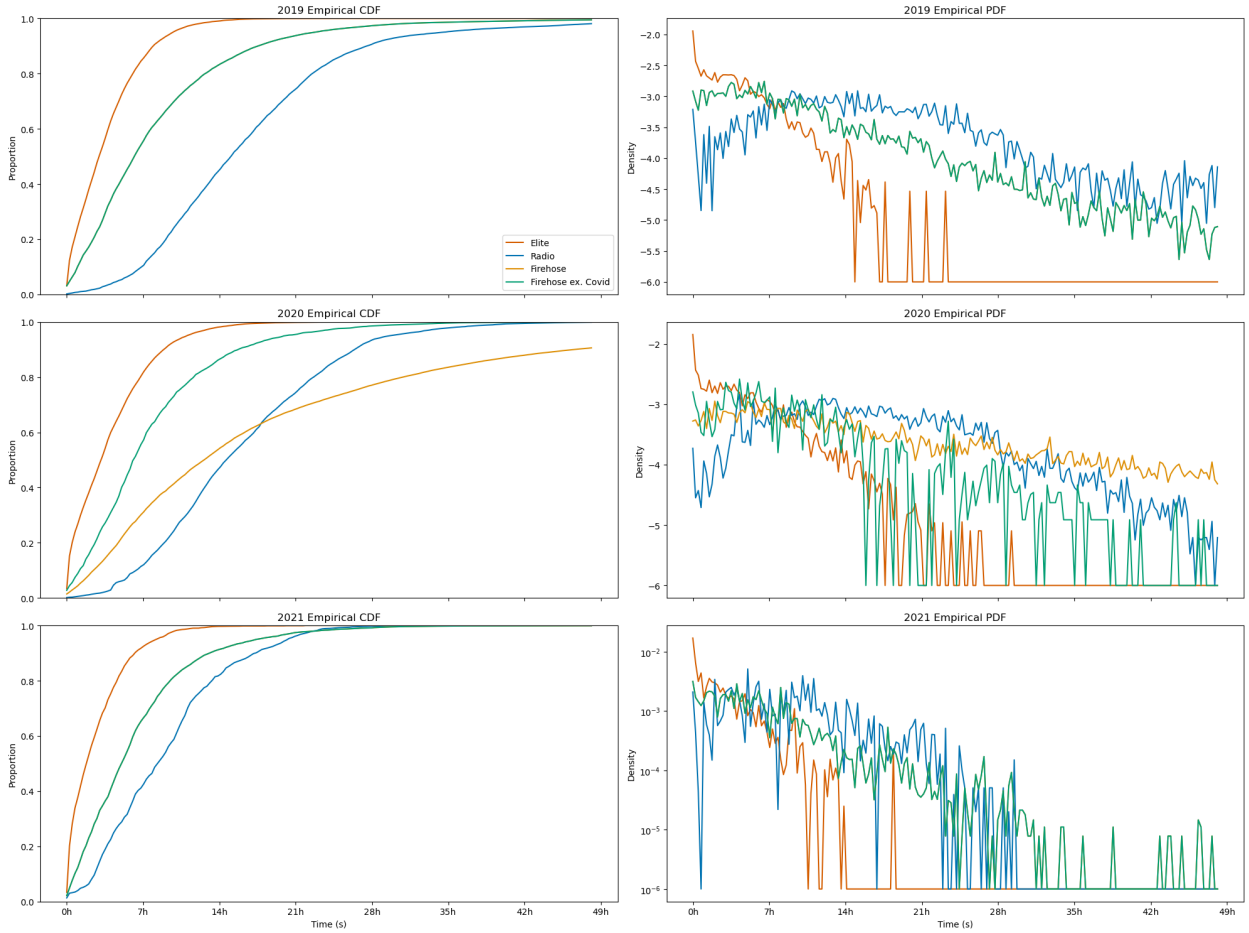


Figure A.2: Average empirical CDFs and PDFs of mentions of the automatically detected events (as in Figure 3.2), broken out by year.

Year	Corpus 1	Corpus 2	Metric	t-stat.	df	p-value
All	Elite	Radio	Avg.	-35.07	1192	$< 10^{-5}$
All	Elite	Radio	Std.	-28.06	1192	$< 10^{-5}$
All	Elite	Firehose-C	Avg.	-17.91	1397	$< 10^{-5}$
All	Elite	Firehose-C	Std.	-18.77	1397	$< 10^{-5}$
All	Elite	Firehose-A	Avg.	-21.33	1562	$< 10^{-5}$
All	Elite	Firehose-A	Std.	-22.46	1562	$< 10^{-5}$
All	Firehose-C	Radio	Avg.	-10.24	463	$< 10^{-5}$
All	Firehose-C	Radio	Std.	-5.26	463	$< 10^{-5}$
All	Firehose-A	Radio	Avg.	-1.46	628	0.14365
All	Firehose-A	Radio	Std.	1.39	628	0.16476
2021	Elite	Radio	Avg.	-11.63	139	$< 10^{-5}$
2021	Elite	Radio	Std.	-9.75	139	$< 10^{-5}$
2021	Elite	Firehose-C	Avg.	-8.16	176	$< 10^{-5}$
2021	Elite	Firehose-C	Std.	-8.87	176	$< 10^{-5}$
2021	Elite	Firehose-A	Avg.	-8.16	176	$< 10^{-5}$
2021	Elite	Firehose-A	Std.	-8.87	176	$< 10^{-5}$
2021	Firehose-C	Radio	Avg.	-3.29	107	0.00132
2021	Firehose-C	Radio	Std.	0.17	107	0.86114
2021	Firehose-A	Radio	Avg.	3.29	107	0.00132
2021	Firehose-A	Radio	Std.	0.17	107	0.86114
2020	Elite	Radio	Avg.	-26.98	617	$< 10^{-5}$
2020	Elite	Radio	Std.	-19.60	617	$< 10^{-5}$
2020	Elite	Firehose-C	Avg.	-7.58	602	$< 10^{-5}$
2020	Elite	Firehose-C	Std.	-7.99	602	$< 10^{-5}$
2020	Elite	Firehose-A	Avg.	-21.16	767	$< 10^{-5}$
2020	Elite	Firehose-A	Std.	-22.53	767	$< 10^{-5}$
2020	Firehose-C	Radio	Avg.	-7.55	89	$< 10^{-5}$
2020	Firehose-C	Radio	Std.	-4.08	89	$< 10^{-5}$
2020	Firehose-A	Radio	Avg.	1.51	254	0.13179
2020	Firehose-A	Radio	Std.	3.58	254	0.00041
2019	Elite	Radio	Avg.	-26.37	432	$< 10^{-5}$
2019	Elite	Radio	Std.	-23.45	432	$< 10^{-5}$
2019	Elite	Firehose-C	Avg.	-12.89	615	$< 10^{-5}$
2019	Elite	Firehose-C	Std.	-12.70	615	$< 10^{-5}$
2019	Elite	Firehose-A	Avg.	-12.89	615	$< 10^{-5}$
2019	Elite	Firehose-A	Std.	-12.70	615	$< 10^{-5}$
2019	Firehose-C	Radio	Avg.	-7.51	263	$< 10^{-5}$
2019	Firehose-C	Radio	Std.	-5.30	263	$< 10^{-5}$
2019	Firehose-A	Radio	Avg.	-7.51	263	$< 10^{-5}$
2019	Firehose-A	Radio	Std.	-5.30	263	$< 10^{-5}$

Table A.2: Results of all hypothesis tests for differences between corpora in means ("Avg.") and standard deviations ("Std.") of within-event relative times, as discussed in Section 3.5.1. Negative t statistics indicate corpus 1 has a lower value than corpus 2, while positive t statistics indicate the reverse. "Firehose-A" indicates the entire firehose corpus, and "Firehose-C" the corpus excluding 2020 Covid-related events.

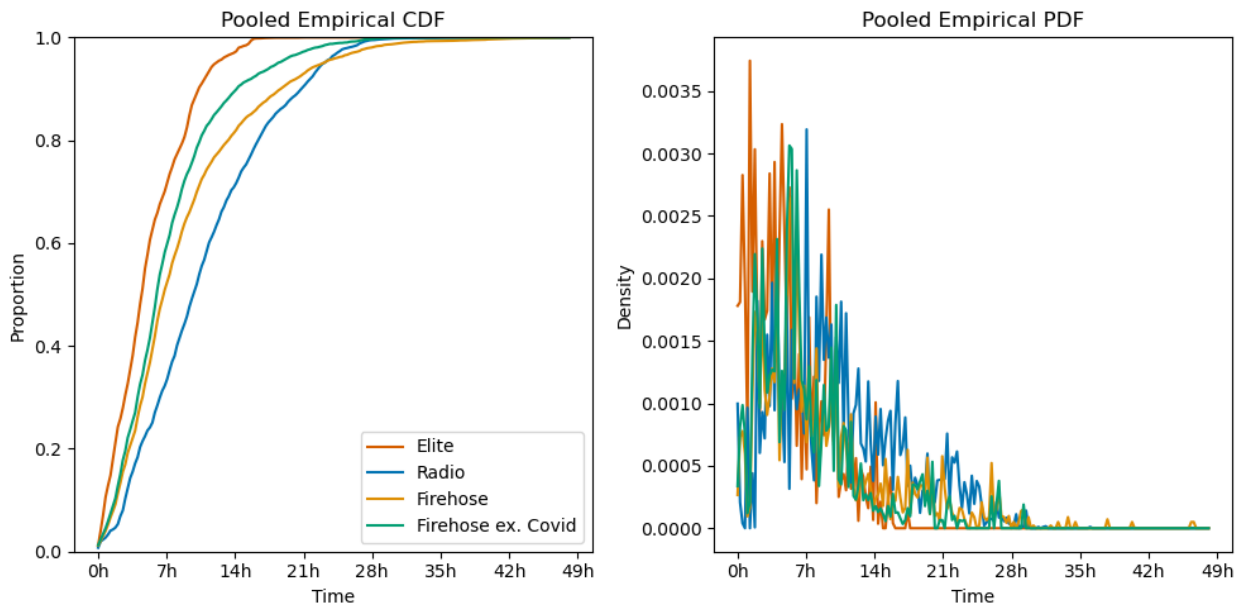


Figure A.3: Average empirical CDFs and PDFs of mentions of the (elite, radio, firehose) matched triples of automatically detected events. Results for all 38 triples and their 114 constituent events are shown.

A.4 Ideological Breakdowns

Figure A.4 shows average empirical CDFs and empirical PDFs of mentions of the automatically detected events, broken down by ideology. Figure A.5 further breaks down by year, as well as by ideology. The estimates shown in both figures reinforce the conclusions drawn in Section 3.5.4: elite Twitter discussion is systematically faster than on radio, among both liberals and conservatives and across all three years of our datasets.

A.5 Affect Levels

We show item-level average scores for our affect metrics in Figure A.6. Results are quite similar to those shown in Figure 3.5, with firehose Twitter consistently the most negative, emotional and outraged, radio the least, and elite Twitter in between.

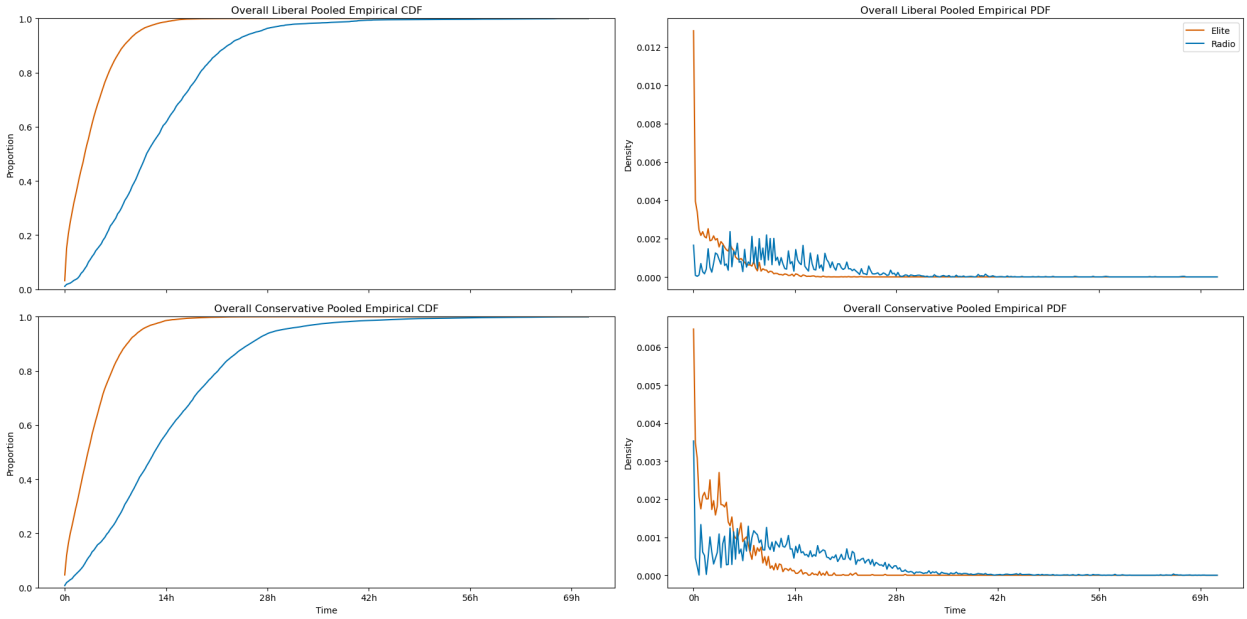


Figure A.4: Average empirical CDFs and PDFs of mentions of the automatically detected events in elite Twitter and radio, broken out by ideology over all three years.

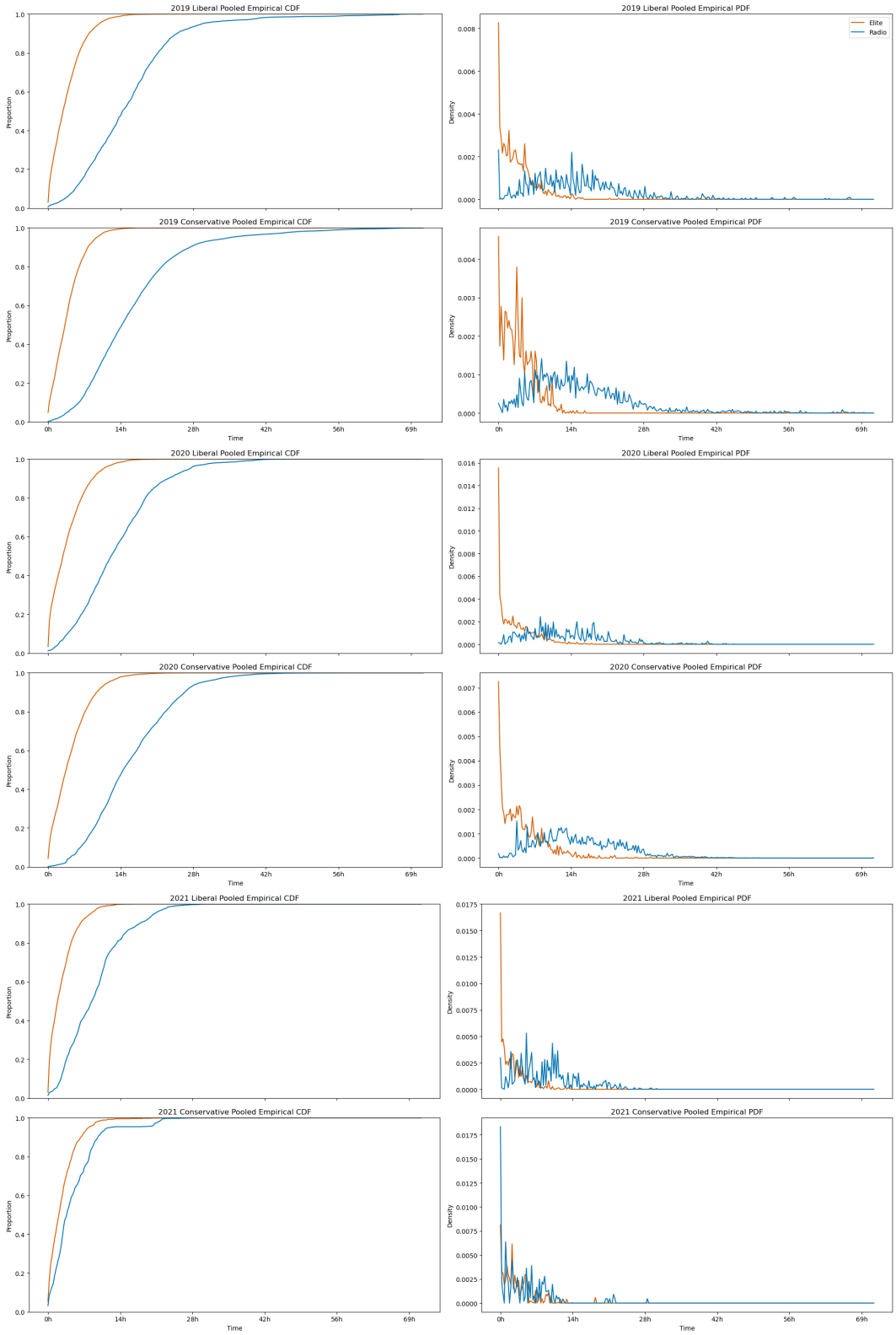


Figure A.5: Average empirical CDFs and PDFs of mentions of the automatically detected events in elite Twitter and radio, broken out by ideology and by year.

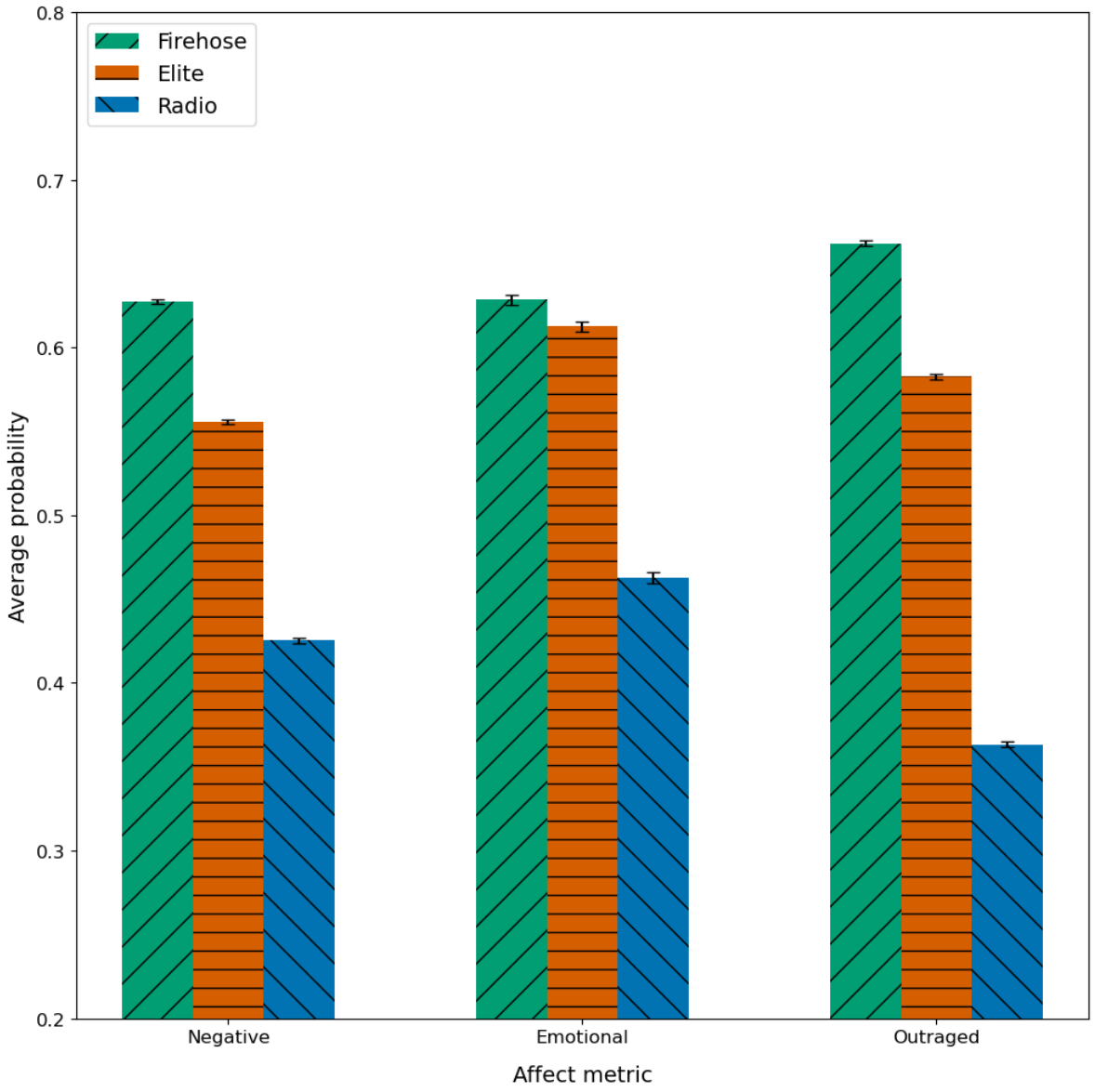


Figure A.6: Item-level average scores for our three affect metrics. The underlying item-level scores are probabilities, and range from 0 to 1. 95% confidence intervals are shown. Note the similarity of the patterns shown here to those at the story level in Figure 3.5.

Appendix B

Predicting Treatment Responsiveness: Further Details

B.1 Experimental Datasets

Here we provide additional details about the datasets used for analysis in Chapter 4 and Chapter 5.

B.1.1 Dataset Repository IDs and Citations

Table B.1 provides repository IDs for our datasets where applicable, and cites relevant papers, books or other sources.

Table B.1: References and repository IDs for the experimental datasets listed in Table 4.1 and used for analysis in Chapter 4 and Chapter 5.

Dataset	TESS ID	OSF ID	Citations
B. & T. (2024)	bolsenM6	yg958	[320, 321]
B. (2022)	brownS80	3cbdj	[322, 323]
C. & D. (2010)	–	–	[48, 291]
D. et al. (2022) - Dem.	newman1436	96rnk	[297]
D. et al. (2022) - P.H.	newman1436	96rnk	[297]
D. et al. (2022) - Rep.	newman1436	96rnk	[297]
F. & U. (2023)	faheyS78	ncs7k	[324, 325]
F. (2011)	flavin235	s2rd5	[48, 326]
G. & M. (2010)	gash-murakami718	82xmd	[48, 327]
G. (2024)	groenendykS79	ar9e3	[328, 329]
G. & B. (2024) - Q1	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q2	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q3	groenendyk1419	fct42	[330, 331]

Continued on next page

Table B.1: References and repository IDs for the experimental datasets.

Dataset	TESS ID	OSF ID	Citations
G. & B. (2024) - Q4	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q5	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q6	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q7	groenendyk1419	fct42	[330, 331]
G. & B. (2024) - Q8	groenendyk1419	fct42	[330, 331]
H. (2006) - Orig.	hiscox136	sxryh	[48, 292, 332]
H. (2006) - GfK	hiscox136	sxryh	[48, 292]
H. (2006) - MTurk	hiscox136	sxryh	[48, 292]
J. & B. (2016) - Orig	johnstonbrief4	5yfgx	[48, 333, 334]
J. & B. (2016) - MTurk	johnstonbrief4	5yfgx	[48, 333, 334]
K. & W. (2016)	wallacebrief13	6mr9e	[48, 335, 336]
L. (2018)	levendusky741	nk9jd	[337, 338]
M. (2021)	Mutz886	ctf7p	[48, 296, 339]
S. (2023)	sahaS66	8khwt	[340, 341]
W. (2023)	wu1435	q9w6s	[342, 343]

B.1.2 Dataset Prompts

All of our datasets, when rendered to text for input to a language model, start by informing the model of the setting and personal traits of the survey respondent. Models which take a system prompt receive this information as the system prompt. The prompt proceeds as follows, with the exact formatting available in the source code for our experiments:

> The year is `{{ field_start_year }}`, and you are a person in the US taking a survey about politics. Here are some facts about you and the kind of person you are:
>
> - `{{ attribute1 }}: {{ value1 }}`
> - `{{ attribute2 }}: {{ value2 }}`
> - `{{ attribute3 }}: {{ value3 }}`
> - ...
>
> As part of the survey, you will be asked questions, which start with 'Q:'; some questions ask for open-ended answers and others are multiple-choice. Multiple-choice questions will list numbered answer options. Please answer these questions by giving ONLY THE NUMBER of your answer choice. For example, if the answer choices are 1: Support, 2: Neutral and 3: Oppose, you might answer '1'.
>
> Survey follows:

and follow the final question with an "A: " on a new line.

B.1.3 Persona Experiment Prompts

The persona-based or inference-time compute experiments discussed in Section 4.4.2 alter the prompts given to models. We replace the instruction paragraph above with:

> As part of the survey, you will be asked questions, which start with 'Q:'; some questions ask for open-ended answers and others are multiple-choice. Multiple-choice questions will list numbered answer options. For multiple-choice questions, you should answer all but the last question in the survey by giving only the number of your answer. For example, if the answer choices are 1: Support, 2: Neutral and 3: Oppose, you might answer "1".
>
> The last question in the survey will be a multiple-choice question, and here you should do something different: Please think step-by-step through your answer and explain why you've chosen the answer you did. When you've decided what your answer is, the absolutely very last thing of all that you write should be ONLY THE NUMBER of your final answer, preceded by "=>". For example, you might write "=> 3" or "=> 4" – but without any periods or quotation marks.

And then instead of "A: ", we follow the final question with the following prompt:

> Please develop a full, paragraph-length persona for yourself, using the demographic characteristics above, such that the paragraph would be informative about your view on the question. You can be creative, making up new details and life experiences. After doing so, please explain your thinking about the question, WITHOUT providing your answer. At the very end, please provide your answer after "=>" (e.g., "=> A" or "=> D"):

B.2 Hypothesis Testing Methodology

The hypothesis tests discussed in Section 4.6 use a Monte Carlo resampling procedure laid out in Algorithm 1. In brief, for each method we resample a number of uniform-random prediction vectors for each dataset, compute the resulting Qini curves, average them across datasets to get the Qini coefficients, and then compare the observed Qini coefficient from the method's predictions to this resampled null-score distribution. The test asks the question, conditional on the actual treatment assignment, of whether our method's scores find more heterogeneity in response to it than would be expected purely by chance. In the terminology of Algorithm 1, we use $N = 1000$ iterations and X a grid of 101 evenly spaced points. We perform six tests for six methods (baseline, text2text, persona, and three variants of RepReg), and apply the corresponding Bonferroni correction to the significance thresholds.

Simulation-based tests like this have been used before in the uplift modeling literature,[344] but are not universal and there is not a single consensus test to use. For an overview of Monte Carlo tests in general, see Section 4.2 of Davison & Hinkley.[345]

Algorithm 1 Monte Carlo test for treatment-response heterogeneity

Input: datasets $\mathcal{D} = \{D_1, \ldots, D_K\}$, model scoring function $\hat{s}(\cdot)$, iterations N, common sample point grid for Qini curves X

Output: observed Qini Q_{obs}, two-sided p-value p

procedure AGGREGATEQINICOEF($\mathcal{D}$, $\{\hat{y}_k\}_{k=1}^{K}$)
  initialize $C_{\text{sum}}(x) \leftarrow 0$ for all $x \in X$
  for all $(D_k, \hat{y}_k)$ **do**
    $C_k \leftarrow$ COMPUTEQINICURVE($D_k, \hat{y}_k$)
    $C_k^{(\text{int})} \leftarrow$ INTERPOLATE($C_k,\ X$)
    $C_{\text{sum}} \leftarrow C_{\text{sum}} + C_k^{(\text{int})}$
  end for
  $\bar{C} \leftarrow C_{\text{sum}}/K$
  return COMPUTEQINICOEFFICIENT($\bar{C}$)
end procedure

initialize list $Q_{\text{null}} \leftarrow []$
for $i = 1 \rightarrow N$ **do**
  for $k = 1 \rightarrow K$ **do**
    $u_k \leftarrow \text{UniformRandom}(0, 1)^{|D_k|}$
  end for
  $Q_i \leftarrow$ AGGREGATEQINICOEF($\mathcal{D}$, $\{u_k\}_{k=1}^{K}$)
  append Q_i to Q_{null}
end for
for $k = 1 \rightarrow K$ **do**
  $\hat{y}_k \leftarrow \hat{s}(D_k)$ ▷ model's actual scores
end for
$Q_{\text{obs}} \leftarrow$ AGGREGATEQINICOEF($\mathcal{D}$, $\{\hat{y}_k\}_{k=1}^{K}$)
$\widehat{F} \leftarrow$ EMPIRICALCDF(Q_{null})
$p_{\text{raw}} \leftarrow \widehat{F}\big(Q_{\text{obs}}\big)$
$p \leftarrow 2\ \min\big(p_{\text{raw}},\ 1 - p_{\text{raw}}\big)$
return $Q_{\text{obs}},\ p$

B.3 Sketch of M3 Loss Implementation

The mathematical underpinnings of the M3 minibatch moment-matching loss developed in this thesis are given in Section 5.3.3. Listing B.1 sketches the implementation of the core idea in numpy-like pseudocode.

Listing B.1: Numpy-like pseudocode for the core idea of our M3 loss

```
def compute_m3_loss(
    preds,               # Predicted HTEs of shape (B,)
    labels,              # True outcomes of shape (B,)
    treatment,           # Treatment indicator in {0,1} of shape (B,)
    min_subset_size=2,   # Minimum size of a subset to sample
    subsets_per_size=1,  # How many subsets to sample for each size
):
    t_indices = (treatment == 1).nonzero()
    c_indices = (treatment == 0).nonzero()
    n_t, n_c = len(t_indices), len(c_indices)

    # Start with full-batch CATE; we can always do at least this
    pred_diffs = [preds[t_indices].mean() - preds[c_indices].mean()]
    obs_diffs = [labels[t_indices].mean() - labels[c_indices].mean()]

    # Can't do anything else if not enough samples
    if min(n_t, n_c) < min_subset_size:
        return MSE(pred_diffs, obs_diffs)

    # Sample subsets of appropriate sizes
    for size in range(min_subset_size, min(n_t, n_c) + 1):
        n_subsets = min(subsets_per_size, comb(n_t, size), comb(n_c, size))

        for _ in range(n_subsets):
            t_subset = random_sample(t_indices, size, replace=False)
            c_subset = random_sample(c_indices, size, replace=False)

            # Calculate predicted and observed CATEs
            pred_diffs.append(preds[t_subset].mean() - preds[c_subset].mean())
            obs_diffs.append(labels[t_subset].mean() - labels[c_subset].mean())

    # Return MSE over all differences
    return MSE(pred_diffs, obs_diffs)
```

References

1. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E. & Fowler, J. H. A 61-million-person experiment in social influence and political mobilization. *Nature* **489,** 295–298. doi:10/f3689v (2012).
2. Iyengar, S. & Kinder, D. R. *News that matters: television and American opinion*. Updated edition. ISBN: 978-0-226-38858-8 (University of Chicago Press, Chicago, 2010).
3. King, G., Schneer, B. & White, A. How the news media activate public expression and influence national agendas. *Science* **358,** 776–780. doi:10/gcmbb5 (2017).
4. Bursztyn, L., Rao, A., Roth, C. & Yanagizawa-Drott, D. *Misinformation During a Pandemic*. w27417 (National Bureau of Economic Research, Cambridge, MA, 2020). doi:10/jr3k.
5. Katz, E. & Lazarsfeld, P. F. *Personal influence: the part played by people in the flow of mass communications*. ISBN: 978-1-4128-0507-0 (Free Press, Detroit, 1955).
6. Key, V. O. *The Responsible Electorate: Rationality in Presidential Voting 1936–1960*. ISBN: 978-0-674-49775-7. doi:10/m334 (Harvard University Press, 1966).
7. Hibbs, D. A. *The American Political Economy: Macroeconomics and Electoral Politics*. ISBN: 978-0-674-02736-7 (Harvard University Press, 1989).
8. Blumer, H. Public Opinion and Public Opinion Polling. *American Sociological Review* **13,** 542–549. doi:10/dzqm6x (1948).
9. Hovland, C. I., Janis, I. L. & Kelley, H. H. *Communication and persuasion; psychological studies of opinion change*. ISBN: 978-0-300-00573-8 (Yale University Press, New Haven, CT, US, 1953).
10. Converse, P. E. The nature of belief systems in mass publics. *Critical Review* **18,** 1–74. doi:10/d5pwvk (1964).
11. Zaller, J. R. Monica Lewinsky's Contribution to Political Science. *PS: Political Science and Politics* **31,** 182–189. doi:10/bnk6js (1998).
12. danah boyd. *Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications*. In *A networked self: identity, community and culture on social network sites* (ed Papacharissi, Z.) (Routledge, New York, 2011). ISBN: 978-0-415-80180-5.
13. Bucy, E., Gantz, W. & Wang, Z. *Media Technology and the 24-Hour News Cycle*. In *Communication Technology and Social Change* (eds Lin, C. A. & Atkin, D. J.) (Routledge, 2014). ISBN: 978-1-135-25125-3. doi:10/gfgxdq.
14. Kumar, S., Hamilton, W. L., Leskovec, J. & Jurafsky, D. *Community Interaction and Conflict on the Web*. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18* (ACM Press, Lyon, France, 2018), 933–943. ISBN: 978-1-4503-5639-8. doi:10/ggsk38.

15. Tufekci, Z. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. ISBN: 978-0-300-22817-5. doi:10/hfjk (Yale University Press, New Haven, 2017).
16. Druckman, J., Peterson, E. & Slothuus, R. How Elite Partisan Polarization Affects Public Opinion Formation. *American Political Science Review* **107,** 57–79. doi:10/gfs2dm (2013).
17. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science* **22,** 129–146. doi:10/gfv79s (2019).
18. Sunstein, C. R. *#Republic: divided democracy in the age of social media*. ISBN: 978-1-4008-8471-1 (Princeton University Press, Princeton, 2017).
19. Kennedy, C. & Hartig, H. *Response rates in telephone surveys have resumed their decline*. Pew Research Center. https://www.pewresearch.org/short-reads/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.
20. De Leeuw, E. D., Suzer-Gurtekin, Z. T. & Hox, J. J. *The Design and Implementation of Mixed-mode Surveys*. In *Advances in Comparative Survey Methods* (eds Johnson, T. P., Pennell, B.-E., Stoop, I. A. & Dorer, B.) 1st ed., 385–408 (Wiley, 2018). ISBN: 978-1-118-88498-0. doi:10/mgr3.
21. Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C. & Conrad, F. G. Research synthesis: Social media analyses for social measurement. *Public Opinion Quarterly* **80,** 180–211. doi:10/f8sdgx (2016).
22. Bovet, A., Morone, F. & Makse, H. A. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Scientific Reports* **8,** 8673. doi:10/gdq27h (2018).
23. Smyth, J. D. *Internet Survey Methods: A Review of Strengths, Weaknesses, and Innovations*. In *Social and Behavioral Research and the Internet* (Routledge, 2010). ISBN: 978-0-203-84492-2.
24. Sunstein, C. R. The Law of Group Polarization. *Journal of Political Philosophy* **10,** 175–195. doi:10/bsfm68 (2002).
25. Pariser, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. ISBN: 978-1-101-51512-9 (Penguin, 2011).
26. Turkle, S. *Alone together: why we expect more from technology and less from each other*. ISBN: 978-0-465-03146-7 (Basic Books, New York, 2011).
27. Brown, T. *et al. Language Models are Few-Shot Learners*. In *Advances in Neural Information Processing Systems (NeurIPS)* **33** (2020), 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
28. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. & Amodei, D. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361.
29. Hoffmann, J. *et al. Training compute-optimal large language models*. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2024), 30016–30030. ISBN: 978-1-71387-108-8. arXiv: 2203.15556.
30. Bommasani, R. *et al. On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258.
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21,** 140:5485–140:5551. arXiv: 1910.10683 (2020).
32. Wei, J. *et al.* Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research.* arXiv: 2206.07682 (2022).

33. Bubeck, S. *et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712.
34. Chu, E., Andreas, J., Ansolabehere, S. & Roy, D. *Language Models Trained on Media Diets Can Predict Public Opinion*. 2023. arXiv: 2303.16779.
35. Wang, L. *et al.* A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science* **18,** 186345. doi:10/gtvsz9 (2024).
36. Andreas, J. *Language Models as Agent Models*. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds Goldberg, Y., Kozareva, Z. & Zhang, Y.) (ACL, Abu Dhabi, United Arab Emirates, 2022), 5769–5779. doi:10/gs83vt.
37. Bail, C. A. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences* **121,** e2314021121. doi:10/gtwpxs (2024).
38. Manning, B. S., Zhu, K. & Horton, J. J. *Automated Social Science: Language Models as Scientist and Subjects*. 2024. arXiv: 2404.11794.
39. Zaller, J. *The nature and origins of mass opinion*. ISBN: 978-0-511-81869-1. doi:10/dgwr (Cambridge University Press, 1992).
40. Milgram, S., Bickman, L. & Berkowitz, L. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology* **13,** 79–82. doi:10/cqxbtk (1969).
41. Lippmann, W. *The phantom public*. ISBN: 978-1-56000-677-0 (Transaction Publishers, New Brunswick, NJ, 1925).
42. McLuhan, M. *Understanding media: the extensions of man*. 1st edition. ISBN: 81-14-67535-7 (McGraw-Hill, New York, 1964).
43. Postman, N. *Amusing ourselves to death: public discourse in the age of show business*. ISBN: 978-0-670-80454-2 (Viking, New York, 1985).
44. Campbell, A., Converse, P., Miller, W. & Stokes, D. *The American Voter*. ISBN: 978-0-226-09254-6 (University of Chicago Press, Chicago, 1960).
45. Petty, R. E. & Cacioppo, J. T. *Communication and Persuasion*. ISBN: 978-1-4612-9378-1 (Springer, New York, NY, 1986).
46. Kahneman, D. & Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47,** 263. doi:10/g98 (1979).
47. Bourdieu, P. *Distinction: a social critique of the judgement of taste*. 11. print. Trans. by Nice, R. ISBN: 978-0-674-21277-0 (Harvard Univ. Press, Cambridge, Mass, 2002).
48. Coppock, A. *Persuasion in parallel: how information changes minds about politics*. ISBN: 978-0-226-82184-9. doi:10/m52z (The University of Chicago Press, Chicago London, 2022).
49. Coppock, A., Leeper, T. J. & Mullinix, K. J. Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences* **115,** 12441–12446. doi:10/ggf5zm (2018).
50. Hewitt, L., Broockman, D., Coppock, A., Tappin, B. M., Slezak, J., Coffman, V., Lubin, N. & Hamidian, M. How Experiments Help Campaigns Persuade Voters: Evidence from a Large Archive of Campaigns' Own Experiments. *American Political Science Review,* 1–19. doi:10/gthwk6 (2024).
51. Castells, M. *The rise of the network society*. 2nd. ed. ISBN: 978-1-4051-9686-4 (Wiley-Blackwell, 2011).

52. Dedrick, J., Gurbaxani, V. & Kraemer, K. L. Information technology and economic performance: A critical review of the empirical evidence. *ACM Computing Surveys* **35,** 1–28. doi:10/d8v84k (2003).
53. Ozimek, A. The Future of Remote Work. *SSRN Electronic Journal.* doi:10/ghf92g (2020).
54. Haddock, A., Ward, N., Yu, R. & O'Dea, N. Positive Effects of Digital Technology Use by Adolescents: A Scoping Review of the Literature. *International Journal of Environmental Research and Public Health* **19,** 14009. doi:10/gtg45h (2022).
55. Haidt, J., Rausch, Z. & Twenge, J. *Social Media and Mental Health: A Collaborative Review*. 2023. https://tinyurl.com/SocialMediaMentalHealthReview.
56. Sohn, S. Y., Rees, P., Wildridge, B., Kalk, N. J. & Carter, B. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and GRADE of the evidence. *BMC Psychiatry* **19,** 356. doi:10/df7z (2019).
57. Haidt, J. & Bail, C. *Social Media and Political Dysfunction: A Collaborative Review*. 2023. https://tinyurl.com/PoliticalDysfunctionReview.
58. Finkel, E. J. *et al.* Political sectarianism in America. *Science* **370,** 533–536. doi:10/ghs347 (2020).
59. Kubin, E. & Von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association* **45,** 188–206. doi:10/gprd5d (2021).
60. Gibson, J. J. *The ecological approach to visual perception: classic edition*. ISBN: 978-1-315-74021-8 (Psychology Press, Hove, East Sussex, 2015).
61. Norman, D. A. *The design of everyday things*. ISBN: 978-0-262-64037-4 (MIT Press, London, 2001).
62. danah boyd & Ellison, N. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* **13,** 210–230. doi:10/gzn (2007).
63. Marwick, A. E. & danah boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* **13,** 114–133. doi:10/ffn6sf (2011).
64. Wang Jicheng, Huang Yuan, Wu Gangshan & Zhang Fuyan. *Web mining: knowledge discovery on the Web*. In *Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics (SMC '99)* **2** (IEEE, Tokyo, Japan, 1999), 137–141. ISBN: 978-0-7803-5731-0. doi:10/dskb38.
65. Aral, S. *Networked Experiments*. In. *The Oxford Handbook of the Economics of Networks* (eds Bramoullé, Y., Galeotti, A. & Rogers, B. W.) 375–411 (Oxford University Press, 2016). ISBN: 978-0-19-994827-7. doi:10/hfjj.
66. Eckles, D., Karrer, B. & Ugander, J. Design and Analysis of Experiments in Networks: Reducing Bias from Interference. *Journal of Causal Inference* **5,** 20150021. doi:10/gpbghx (2017).
67. Tu, H., Doupe, A., Zhao, Z. & Ahn, G.-J. *SoK: Everyone Hates Robocalls: A Survey of Techniques Against Telephone Spam*. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP '16)* (IEEE, San Jose, CA, 2016), 320–338. ISBN: 978-1-5090-0824-7. doi:10/gf25sb.
68. Bail, C. A. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences* **113,** 11823–11828. doi:10/f87mw2 (2016).

69. Leskovec, J., Backstrom, L. & Kleinberg, J. *Meme-tracking and the dynamics of the news cycle*. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09* (ACM Press, New York, New York, USA, 2009), 497. ISBN: 978-1-60558-495-9. doi:10/dp59g8.

70. Li, C. & Goldwasser, D. *Encoding Social Information with Graph Convolutional Networks for Political Perspective Detection in News Media*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL, Florence, Italy, 2019), 2594–2604. doi:10/ghndks.

71. Kim, J. & Lee, B. *AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction*. 2023. arXiv: 2305.09620.

72. Habermas, J. *The Structural transformation of the public sphere: an inquiry into a category of bourgeois society*. 10. print. ISBN: 978-0-262-58108-0 (MIT Press, Cambridge, Mass, 1999).

73. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359,** 1146–1151. doi:10/gc3jt6 (2018).

74. Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A. & Hardt, M. Algorithmic Amplification of Politics on Twitter. *Proceedings of the National Academy of Sciences* **119,** e2025334119. doi:10/gn248s (2021).

75. Bail, C. *Breaking the social media prism: how to make our platforms less polarizing*. ISBN: 978-0-691-20342-3 (Princeton University Press, Princeton, 2021).

76. Keeter, S., Hatley, N., Lau, A. & Kennedy, C. *What 2020's Election Poll Errors Tell Us About the Accuracy of Issue Polling*. (Pew Research Center, 2021). https://www.pewresearch.org/methods/2021/03/02/what-2020s-election-poll-errors-tell-us-about-the-accuracy-of-issue-polling/.

77. Kennedy, C. *et al.* An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly* **82,** 1–33. doi:10/gpbzn5 (2018).

78. Bruner, J. The Narrative Construction of Reality. *Critical Inquiry* **18,** 1–21. doi:10/chsms7 (1991).

79. Ganz, M. *Public Narrative, Collective Action, and Power*. In *Accountability through Public Opinion: From Inertia to Public Action* (eds Odugbemi, S. & Lee, T.) 273–289 (The World Bank, 2011). ISBN: 978-0-8213-8505-0. doi:10/hfjm.

80. Shiller, R. J. Narrative Economics. *American Economic Review* **107,** 967–1004. doi:10/f94zr4 (2017).

81. Abbott, H. P. Narrative and Emergent Behavior. *Poetics Today* **29,** 227–244. doi:10/c5bpv8 (2008).

82. Brewer, M. B. The Social Self: On Being the Same and Different at the Same Time. *Personality and Social Psychology Bulletin* **17,** 475–482. doi:10/d6gfsp (1991).

83. Berelson, B., Lazarsfeld, P. F. & McPhee, W. N. *Voting: a study of opinion formation in a presidential campaign*. ISBN: 978-0-226-04350-0 (University of Chicago Press, Chicago, 1954).

84. Backstrom, L., Huttenlocher, D., Kleinberg, J. & Lan, X. *Group formation in large social networks: membership, growth, and evolution*. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)* (ACM Press, Philadelphia, PA, USA, 2006), 44. ISBN: 978-1-59593-339-3. doi:10/dtdzkz.

85. Achen, C. H. Mass Political Attitudes and the Survey Response. *American Political Science Review* **69,** 1218–1231. doi:10/bhkdnd (1975).

86. Lippmann, W. *Public Opinion*. (Harcourt, Brace, New York, 1922).

87. Dewey, J. *The Public and Its Problems*. (Holt, New York, 1927).

88. Meyrowitz, J. *No sense of place: the impact of electronic media on social behavior*. 1st Edition. ISBN: 978-0-19-504231-3 (Oxford Univ. Press, New York, NY., 1986).

89. Ong, W. J. & Hartley, J. *Orality and literacy: the technologizing of the word*. ISBN: 978-1-280-32779-7 (Routledge, New York, 1982).

90. Brandtzaeg, P. B. & Lüders, M. Time Collapse in Social Media: Extending the Context Collapse. *Social Media + Society* **4,** 2056305118763344. doi:10/gj2kwc (2018).

91. Granovetter, M. S. The Strength of Weak Ties. *American Journal of Sociology* **78,** 1360–1380. doi:10/bxwqdn (1973).

92. Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T. & Sprecher, S. Online Dating: A Critical Analysis From the Perspective of Psychological Science. *Psychological Science in the Public Interest* **13,** 3–66. doi:10/b75x (2012).

93. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. *The Anatomy of the Facebook Social Graph*. 2011. arXiv: 1111.4503.

94. Rosen, J. *1. The People Formerly Known as the Audience*. In *The Social Media Reader* (ed Mandiberg, M.) 13–16 (New York University Press, 2020). ISBN: 978-0-8147-6302-5. doi:10/pm3g.

95. Seamans, R. & Zhu, F. Responses to Entry in Multi-Sided Markets: The Impact of Craigslist on Local Newspapers. *Management Science* **60,** 476–493. doi:10/ghc77k (2014).

96. Negroponte, N. *Being Digital*. ISBN: 978-0-679-43919-6 (A. A. Knopf, New York, 1996).

97. Abernathy, P. M. *News Deserts and Ghost Newspapers: Will Local News Survive?* ISBN: 978-1-4696-6130-8 (University of North Carolina Press, 2020).

98. Sullivan, M. What happens to society — and our democracy — when community and regional journalism dries up. *Washington Post.* https://www.washingtonpost.com/magazine/2021/11/30/margaret-sullivan-the-local-news-crisis/ (2021).

99. Kwak, H., Lee, C., Park, H. & Moon, S. *What is Twitter, a social network or a news media?* In *Proceedings of the 19th international conference on World wide web - WWW '10* (ACM Press, New York, New York, USA, 2010), 591. ISBN: 978-1-60558-799-8. doi:10/c2k8cj.

100. Gerber, A. S. & Green, D. P. *Field experiments: design, analysis, and interpretation*. 1st ed. ISBN: 978-0-393-97995-4 (W. W. Norton, New York, 2012).

101. Ugander, J., Karrer, B., Backstrom, L. & Kleinberg, J. *Graph cluster randomization: network exposure to multiple universes*. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)* (ACM, Chicago Illinois USA, 2013), 329–337. ISBN: 978-1-4503-2174-7. doi:10/gg7tcn.

102. Aronow, P. M. & Samii, C. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* **11,** 1912–1947. doi:10/gkkzm6 (2017).

103. Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F. & Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115,** 9216–9221. doi:10/gd4t96 (2018).

104. Nyhan, B. *et al.* Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620,** 137–144. doi:10/gspqt7 (2023).

105. Guess, A. M. *et al.* How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381,** 398–404. doi:10/gsjbzq (2023).

106. Guess, A. M. *et al.* Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381,** 404–408. doi:10/gspqvc (2023).

107. González-Bailón, S. *et al.* Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381,** 392–398. doi:10/gspqvg (2023).

108. Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* **111,** 8788–8790. doi:10/tcg (2014).

109. Gallup, G. Testing Public Opinion. *The Public Opinion Quarterly* **2,** 8–14. doi:10/fpjgmr (1938).

110. Pew Research Center. *The Cell Phone Challenge to Survey Research*. (Pew Research Center, 2006). https://www.pewresearch.org/politics/2006/05/15/the-cell-phone-challenge-to-survey-research/.

111. Van Dam, A. Analysis | Barely a quarter of Americans still have landlines. Who are they? *Washington Post.* https://www.washingtonpost.com/business/2023/06/23/landline-telephone-holdouts/ (2023).

112. Blumberg, S. & Luke, J. *Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, July-December 2022*. (U.S. National Center for Health Statistics, 2023). doi:10/m336.

113. Silver, N. The Real Story Of 2016: What reporters — and lots of data geeks, too — missed about the election, and what they're still getting wrong. *FiveThirtyEight.* https://fivethirtyeight.com/features/the-real-story-of-2016/ (2017).

114. Kennedy, C., Popky, D. & Keeter, S. *How Public Polling Has Changed in the 21st Century*. (Pew Research Center, 2023). https://www.pewresearch.org/methods/2023/04/19/how-public-polling-has-changed-in-the-21st-century/.

115. Evans, J. R. & Mathur, A. The value of online surveys. *Internet Research* **15,** 195–219. doi:10/d8tcnt (2005).

116. Atkeson, L. R. & Adams, A. N. *Mixing Survey Modes and Its Implications*. (eds Atkeson, L. R. & Alvarez, R. M.) doi:10/mgr4 (Oxford University Press, 2017).

117. Schoeni, R. F., Stafford, F., Mcgonagle, K. A. & Andreski, P. Response Rates in National Panel Surveys. *The Annals of the American Academy of Political and Social Science* **645,** 60–87. doi:10/gmw5nk (2013).

118. Igielnik, R., Keeter, S., Kennedy, C. & Spahn, B. *Commercial voter files and the study of U.S. politics*. (Pew Research Center, 2018). https://www.pewresearch.org/methods/2018/02/15/voter-files-in-action/.

119. Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J. & Tucker, J. A. *Measuring Public Opinion with Social Media Data*. In *The Oxford Handbook of Polling and Survey Methods* (eds Atkeson, L. R. & Alvarez, R. M.) (Oxford University Press, 2017). doi:10/gfsczc.

120. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444. doi:10/bmqp (2015).

121. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95,** 245–258. doi:10/gbp987 (2017).

122. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* **2,** 303–314. doi:10/dp3968 (1989).

123. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359–366. doi:10/c4wch5 (1989).

124. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536. doi:10/cvjdpk (1986).

125. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. doi:10/gk7nfp (2021).

126. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* **550,** 354–359. doi:10/gcsmk9 (2017).

127. Mikolov, T., Chen, K., Corrado, G. & Dean, J. *Distributed Representations of Words and Phrases and their Compositionality*. In *Advances in Neural Information Processing Systems (NeurIPS)* (2013). arXiv: 1310.4546.

128. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* **41,** 391–407. doi:10/db4ft5 (1990).

129. Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. *Distributed representations*. In *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations* 77–109 (MIT Press, Cambridge, MA, 1986). ISBN: 978-0-262-68053-0.

130. Minsky, M. & Papert, S. A. *Perceptrons: An Introduction to Computational Geometry*. ISBN: 978-0-262-34393-0. doi:10/gkc8fr (The MIT Press, 2017).

131. Elman, J. L. Finding Structure in Time. *Cognitive Science* **14,** 179–211. doi:10/fwxk7w (1990).

132. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780. doi:10/bxd65w (1997).

133. Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (ACL, Doha, Qatar, 2014), 1724–1734. doi:10/gddmvq.

134. Bahdanau, D., Cho, K. & Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR '15)* (2015). arXiv: 1409.0473.

135. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. *Attention Is All You Need*. In *Advances in Neural Information Processing Systems (NeurIPS)* **30** (2017). arXiv: 1706.03762.

136. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. *Deep Contextualized Word Representations*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (ACL, New Orleans, LA, 2018), 2227–2237. doi:10/gft5gf.

137. Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. *Word Translation Without Parallel Data*. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)* (2018). arXiv: 1710.04087.

138. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) (ACL, Minneapolis, MN, 2019), 4171–4186. doi:10/ggbwf6.

139. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. *Understanding deep learning requires rethinking generalization*. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)* (2017). arXiv: 1611.03530.

140. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. & Sutskever, I. *Deep Double Descent: Where Bigger Models and More Data Hurt*. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '19)* (2019). arXiv: 1912.02292.

141. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* **116,** 15849–15854. doi:10/gf5dmw (2019).

142. Frankle, J. & Carbin, M. *The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks*. In *Proceedings of the 7th International Conference on Learning Representations (ICLR '19)* (2019). arXiv: 1803.03635.

143. Longpre, S. *et al. Bridging the Data Provenance Gap Across Text, Speech, and Video*. In *Proceedings of the 13th International Conference on Learning Representations (ICLR '25)* (2025). arXiv: 2412.17847.

144. Longpre, S. *et al.* A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* **6,** 975–987. doi:10/gt8f5p (2024).

145. Longpre, S., Mahari, R., Muennighoff, N., Chen, A., Perisetla, K., Brannon, W., Kabbara, J., Villa, L. & Hooker, S. *The Data Provenance Project*. In *Proceedings of the 1st Workshop on Generative AI and Law (GenLaw '23) at ICML '24* (Honolulu, Hawaii, USA, 2023). https://blog.genlaw.org/CameraReady/20.pdf.

146. Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. 2023. arXiv: 2302.06590.

147. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. *Language Models are Unsupervised Multitask Learners*. 2019. https://openai.com/blog/better-language-models/.

148. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T. & Chen, E. A survey on multimodal large language models. *National Science Review* **11,** nwae403. doi:10/g8whw6 (2024).

149. Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X. & Liu, Q. *Aligning Large Language Models with Human: A Survey*. 2023. arXiv: 2307.12966.

150. Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. & Du, M. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology* **15,** 1–38. doi:10/gtkmj5 (2024).

151. Hamilton, W. L. Graph Representation Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **14,** 1–159. doi:10/ghfgfg (2020).

152. Ruiz, L., Gama, F. & Ribeiro, A. *Gated Graph Convolutional Recurrent Neural Networks*. In *Proceedings of the 27th European Signal Processing Conference (EUSIPCO '19)* (IEEE, A Coruna, Spain, 2019), 1–5. ISBN: 978-90-827970-3-9. doi:10/ggm5h2.

153. Kipf, T. N. & Welling, M. *Semi-Supervised Classification with Graph Convolutional Networks*. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)* (2017). arXiv: 1609.02907.

154. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P. & Bengio, Y. *Graph Attention Networks*. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)* (2018). arXiv: 1710.10903.

155. Joshi, C. *Transformers are Graph Neural Networks*. NTU Graph Deep Learning Lab. https://graphdeeplearning.github.io/post/transformers-are-gnns/.

156. Grover, A. & Leskovec, J. *node2vec: Scalable Feature Learning for Networks*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (ACM, San Francisco, CA, USA, 2016), 855–864. doi:10/gftdzj.

157. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. & Leskovec, J. *Strategies for Pre-training Graph Neural Networks*. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)* (2020). arXiv: 1905.12265.

158. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. *How Powerful are Graph Neural Networks?* In *Proceedings of the 7th International Conference on Learning Representations (ICLR '19)* (2019). arXiv: 1810.00826.

159. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine* **34,** 18–42. doi:10/gcsgh7 (2017).

160. Morris, C., Frasca, F., Dym, N., Maron, H., Ceylan, I. I., Levie, R., Lim, D., Bronstein, M. M., Grohe, M. & Jegelka, S. *Position: Future Directions in the Theory of Graph Machine Learning*. In *Proceedings of the 41st International Conference on Machine Learning* (PMLR, 2024), 36294–36307. https://proceedings.mlr.press/v235/morris24a.html.

161. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559,** 547–555. doi:10/gdvnpg (2018).

162. Brannon, W., Virkar, Y. & Thompson, B. Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing. *Transactions of the Association for Computational Linguistics* **11,** 419–435. doi:10/gr9cbz (2023).

163. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L. & Zdeborová, L. Machine learning and the physical sciences. *Reviews of Modern Physics* **91,** 045002. doi:10/ggd5qv (2019).

164. T.K., B., Annavarapu, C. S. R. & Bablani, A. Machine learning algorithms for social media analysis: A survey. *Computer Science Review* **40,** 100395. doi:10/gnbp44 (2021).

165. Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H. & Alfakeeh, A. S. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing* **13,** 3433–3456. doi:10/gj4p87 (2022).

166. Monti, F., Frasca, F., Eynard, D., Mannion, D. & Bronstein, M. M. *Fake News Detection on Social Media using Geometric Deep Learning*. 2019. arXiv: 1902.06673.

167. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* **19,** 22–36. doi:10/gf6d5n (2017).

168. Neethu, M. S. & Rajasree, R. *Sentiment analysis in twitter using machine learning techniques*. In *Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT '13)* (IEEE, Tiruchengode, 2013), 1–5. ISBN: 978-1-4799-3926-8. doi:10/ggb6mk.

169. Sims, M., Park, J. H. & Bamman, D. *Literary Event Detection*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (ACL, Florence, Italy, 2019), 3623–3634. doi:10/ggnw5f.

170. Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T. & Rothman, R. E. Influenza Forecasting with Google Flu Trends. *PLoS ONE* **8** (ed Viboud, C.) e56176. doi:10/gmf23d (2013).

171. Jiang, H., Beeferman, D., Brannon, W., Heyward, A. & Roy, D. *Bridging Dictionary: AI-Generated Dictionary of Partisan Language Use*. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '24)* (ACM, San Jose Costa Rica, 2024), 79–82. ISBN: 9798400711145. doi:10/g8wgvz.

172. Brannon, W., Beeferman, D., Jiang, H., Heyward, A. & Roy, D. *AudienceView: AI-Assisted Interpretation of Audience Feedback in Journalism*. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '24)* (ACM, San Jose Costa Rica, 2024), 65–68. ISBN: 9798400711145. doi:10/g8wgvx.

173. Brannon, W., Kang, W., Fulay, S., Jiang, H., Roy, B., Roy, D. & Kabbara, J. *ConGraT: Self-Supervised Contrastive Pretraining for Joint Graph and Text Embeddings*. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing* (eds Ustalov, D., Gao, Y., Panchenko, A., Tutubalina, E., Nikishina, I., Ramesh, A., Sakhovskiy, A., Usbeck, R., Penn, G. & Valentino, M.) (ACL, Bangkok, Thailand, 2024), 19–39. https://aclanthology.org/2024.textgraphs-1.2.

174. Yasunaga, M., Leskovec, J. & Liang, P. *LinkBERT: Pretraining Language Models with Document Links*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACL, Dublin, Ireland, 2022), 8003–8016. doi:10/grhhv2.

175. Gayo-Avello, D. *"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" – A Balanced Survey on Election Prediction using Twitter Data*. 2012. arXiv: 1204.6441.

176. Dwi Prasetyo, N. & Hauff, C. *Twitter-based Election Prediction in the Developing World*. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15* (ACM Press, Guzelyurt, Northern Cyprus, 2015), 149–158. ISBN: 978-1-4503-3395-5. doi:10/gthk3c.

177. Gayo-Avello, D., Metaxas, P. & Mustafaraj, E. Limits of Electoral Predictions Using Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* **5,** 490–493. doi:10/gthk3d (2021).

178. Joseph, K., Shugars, S., Gallagher, R., Green, J., Quintana Mathé, A., An, Z. & Lazer, D. *(Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)* (eds Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-t.) (ACL, Punta Cana, Dominican Republic, 2021), 312–324. doi:10/gs9rq3.

179. Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J. & Botvinick, M. *Fine-tuning language models to find agreement among humans with diverse preferences*. In *Advances in Neural Information Processing Systems (NeurIPS)* **35** (2022), 38176–38189. https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

180. Small, C. T., Vendrov, I., Durmus, E., Homaei, H., Barry, E., Cornebise, J., Suzman, T., Ganguli, D. & Megill, C. *Opportunities and Risks of LLMs for Scalable Deliberation with Polis*. 2023. arXiv: 2306.11932.

181. Fulay, S., Brannon, W., Mohanty, S., Overney, C., Poole-Dayan, E., Roy, D. & Kabbara, J. *On the Relationship between Truth and Political Bias in Language Models*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (ACL, Miami, Florida, USA, 2024), 9004–9018. doi:10/g8wgvw.

182. Bang, Y., Chen, D., Lee, N. & Fung, P. *Measuring Political Bias in Large Language Models: What Is Said and How It Is Said*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Ku, L.-W., Martins, A. & Srikumar, V.) (ACL, Bangkok, Thailand, 2024), 11142–11159. doi:10/g8whxf.

183. Pickard, V. *Democracy without Journalism?: Confronting the Misinformation Society*. 1st ed. ISBN: 978-0-19-094675-3. doi:10/m332 (Oxford University Press, 2020).

184. Harder, R. A., Sevenans, J. & Van Aelst, P. Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times. *The International Journal of Press/Politics* **22,** 275–293. doi:10/gbk7cx (2017).

185. Castillo, C., El-Haddad, M., Pfeffer, J. & Stempeck, M. *Characterizing the life cycle of online news stories using social media reactions*. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)* (ACM, Baltimore, Maryland, USA, 2014), 211–223. ISBN: 978-1-4503-2540-0. doi:10/ggvr2j.

186. Anderson, K. E. Getting acquainted with social networks and apps: it is time to talk about TikTok. *Library Hi Tech News* **37,** 7–12. doi:10/ghjcsz (2020).

187. Hayes, D. & Lawless, J. L. The Decline of Local News and Its Effects: New Evidence from Longitudinal Data. *The Journal of Politics* **80,** 332–336. doi:10/gcvdnc (2018).

188. Mason, L. "I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science* **59,** 128–145. doi:10/f6w534 (2014).

189. Cramer, K. J. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. ISBN: 978-0-226-34908-4 (University of Chicago Press, Chicago, 2016).

190. Iyengar, S., Sood, G. & Lelkes, Y. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly* **76,** 405–431. doi:10/bsck (2012).

191. Iyengar, S. & Westwood, S. J. Fear and Loathing across Party Lines: New Evidence on Group Polarization. *American Journal of Political Science* **59,** 690–707. doi:10/f7jhfs (2015).

192. Yang, J. *et al.* Why Are "Others" So Polarized? Perceived Political Polarization and Media Use in 10 Countries. *Journal of Computer-Mediated Communication* **21,** 349–367. doi:10/f9fjwj (2016).

193. Putnam, R. D. *Bowling alone: the collapse and revival of American community*. ISBN: 978-0-7432-0304-3 (Simon & Schuster, London, 2001).

194. Brannon, W. & Roy, D. The speed of news in Twitter (X) versus radio. *Scientific Reports* **14,** 11939. doi:10/gtwggj (2024).

195. Brannon, W. *Mapping U.S. Talk Radio: A Textual Survey at Scale*. M.S. Thesis (MIT, Cambridge, MA, 2020). https://dspace.mit.edu/handle/1721.1/129270.

196. Lewis, J. *Democratic or Disposable? 24-Hour News, Consumer Culture and Built-In Obsolescence*. In *The rise of 24-hour news television: global perspectives* (eds Cushion, S. & Lewis, J.) 81–98 (Peter Lang, New York, 2010). ISBN: 978-1-4331-0776-4.

197. Barkemeyer, R., Faugère, C., Gergaud, O. & Preuss, L. Media attention to large-scale corporate scandals: Hype and boredom in the age of social media. *Journal of Business Research* **109,** 385–398. doi:10/gspmvm (2020).

198. Robinson, P. The CNN Effect Revisited. *Critical Studies in Media Communication* **22,** 344–349. doi:10/cm5bn3 (2005).

199. The Nielsen Company. *Nielsen Total Audience Report*. 2021. https://www.nielsen.com/insights/2021/total-audience-advertising-across-todays-media/.

200. Rosenwald, B. *Talk Radio's America: How an industry took over a political party that took over the United States*. ISBN: 978-0-674-18501-2 (Harvard University Press, Cambridge, MA, 2019).

201. Rosenberg, H. & Feldman, C. S. *No time to think: the menace of media speed and the 24-hour news cycle*. ISBN: 978-0-8264-2931-5 (Continuum, New York, 2008).

202. Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J. & Ma, K.-L. *Breaking news on Twitter*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (ACM, Austin, TX, 2012), 2751–2754. ISBN: 978-1-4503-1015-4. doi:10/gfzzxt.

203. Lawrence, R. G. *Campaign News in the Time of Twitter*. In *Controlling the Message* (eds Farrar-Myers, V. A. & Vaughn, J. S.) 93–112 (NYU Press, 2020). ISBN: 978-1-4798-6550-5. doi:10/jr3n.

204. Hermida, A. From TV to Twitter: How Ambient News Became Ambient Journalism. *Media/-Culture Journal* **13.** doi:10/gsqft4 (2010).

205. Sobieraj, S. & Berry, J. M. From Incivility to Outrage: Political Discourse in Blogs, Talk Radio, and Cable News. *Political Communication* **28,** 19–41. doi:10/c65wmh (2011).

206. Berry, J. M. & Sobieraj, S. *The outrage industry: political opinion media and the new incivility*. ISBN: 978-0-19-049846-7 (Oxford University Press, Oxford, 2016).

207. Myers, S. A., Sharma, A., Gupta, P. & Lin, J. *Information network or social network? The structure of the Twitter follow graph*. In *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web* (ACM, New York, New York, USA, 2014), 493–498. ISBN: 978-1-4503-2745-9. doi:10/dw9n.

208. Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. J. *Who says what to whom on Twitter*. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)* (ACM Press, New York, New York, USA, 2011), 705. ISBN: 978-1-4503-0632-4. doi:10/cxcjbn.

209. Willnat, L. & Weaver, D. H. Social Media and U.S. Journalists: Uses and perceived effects on perceived norms and values. *Digital Journalism* **6,** 889–909. doi:10/gf8mfp (2018).

210. Linker, D. Twitter is destroying America. *The Week.* https://theweek.com/articles/702389/twitter-destroying-america (2017).

211. Meyer, R. Why Twitter May Be Ruinous for the Left. *The Atlantic*. https://www.theatlantic.com/technology/archive/2020/01/how-twitter-harms-left/605098/ (2020).

212. Barker, D. C. Rush to Action: Political Talk Radio and Health Care (un)Reform. *Political Communication* **15,** 83–97. doi:10/d6kjpx (1998).

213. Barker, D. C. Rushed Decisions: Political Talk Radio and Vote Choice, 1994-1996. *The Journal of Politics* **61,** 527–539. doi:10/fsxxtb (1999).

214. Benkler, Y., Farris, R. & Roberts, H. *Network Propaganda*. ISBN: 978-0-19-092362-4. doi:10/dgx3 (Oxford University Press, 2018).

215. Wang, T. Media, Pulpit, and Populist Persuasion: Evidence from Father Coughlin. *American Economic Review* **111,** 3064–3092. doi:10/gmn2qw (2021).

216. Crockett, M. J. Moral outrage in the digital age. *Nature Human Behaviour* **1,** 769–771. doi:10/gf8chb (2017).

217. Lamba, H., Malik, M. M. & Pfeffer, J. *A Tempest in a Teacup? Analyzing Firestorms on Twitter*. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (ACM, New York, NY, USA, 2015), 17–24. ISBN: 978-1-4503-3854-7. doi:10/gfzztk.

218. Garimella, K., Mathioudakis, M., Morales, G. D. F. & Gionis, A. *Exploring Controversy in Twitter*. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (ACM, New York, NY, USA, 2016), 33–36. ISBN: 978-1-4503-3950-6. doi:10/gft8tz.

219. Gregory, K. & Singh, S. S. Anger in Academic Twitter: Sharing, Caring, and Getting Mad Online. *tripleC: Communication, Capitalism & Critique* **16,** 176–193. doi:10/gm3bzh (2018).

220. Keuchenius, A., Törnberg, P. & Uitermark, J. Why it is important to consider negative ties when studying polarized debates: A signed network analysis of a Dutch cultural controversy on Twitter. *PLOS ONE* **16,** e0256696. doi:10/gqw3hn (2021).

221. Jamieson, K. H. & Cappella, J. N. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. ISBN: 978-0-19-974086-4 (Oxford University Press, 2008).

222. Lorenz-Spreen, P., Mønsted, B. M., Hövel, P. & Lehmann, S. Accelerating dynamics of collective attention. *Nature Communications* **10,** 1759. doi:10/gfzw3q (2019).

223. Berry, J. M. & Sobieraj, S. Understanding the Rise of Talk Radio. *PS: Political Science & Politics* **44,** 762–767. doi:10/d3c6kt (2011).

224. Hermida, A. *Social Journalism: Exploring how Social Media is Shaping Journalism*. In *The Handbook of Global Online Journalism* (eds Siapera, E. & Veglis, A.) 1st ed., 309–328 (Wiley, Chichester, UK, 2012). ISBN: 978-1-4443-3855-3. doi:10.1002/9781118313978.ch17.

225. Wolf, F., Lorenz-Spreen, P. & Lehmann, S. Successive Cohorts of Twitter Users Show Increasing Activity and Shrinking Content Horizons. *Journal of Quantitative Description: Digital Media* **2.** doi:10/gtjtpm (2022).

226. Kim, Y., Kim, Y. & Zhou, S. Theoretical and methodological trends of agenda-setting theory: A thematic analysis of the last four decades of research. *The Agenda Setting Journal* **1,** 5–22. doi:10/ds4f (2017).

227. Su, Y. & Xiao, X. Mapping the intermedia agenda setting (IAS) literature: Current trajectories and future directions. *The Agenda Setting Journal* **5,** 56–83. doi:10/gn99g3 (2021).

228. Atkinson, M. L., Lovett, J. & Baumgartner, F. R. Measuring the Media Agenda. *Political Communication* **31,** 355–380. doi:10/gq9mfd (2014).

229. Sweetser, K. D., Golan, G. J. & Wanta, W. Intermedia Agenda Setting in Television, Advertising, and Blogs During the 2004 Election. *Mass Communication and Society* **11,** 197–216. doi:10/b298r5 (2008).

230. Dezsö, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I. & Barabási, A.-L. Dynamics of information access on the web. *Physical Review E* **73.** doi:10/cwxmw9 (2006).

231. Wu, F. & Huberman, B. A. Novelty and collective attention. *Proceedings of the National Academy of Sciences* **104,** 17599–17601. doi:10/cw6x6p (2007).

232. Lin, Y.-R., Keegan, B., Margolin, D. & Lazer, D. Rising Tides or Rising Stars?: Dynamics of Shared Attention on Twitter during Media Events. *PLOS ONE* **9.** doi:10/f56mgp (2014).

233. Pfeffer, J., Matter, D. & Sargsyan, A. *The Half-Life of a Tweet*. In *Proceedings of the International AAAI Conference on Web and Social Media* **17** (2023), 1163–1167. doi:10/g8whxj.

234. Liu, X. *et al. Reuters Tracer: A Large Scale System of Detecting & Verifying Real-Time News Events from Twitter*. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management* (ACM, New York, NY, USA, 2016), 207–216. ISBN: 978-1-4503-4073-1. doi:10/gq82rf.

235. Beeferman, D., Brannon, W. & Roy, D. *RadioTalk: A Large-Scale Corpus of Talk Radio Transcripts*. In *Proceedings of Interspeech* (ISCA, Graz, Austria, 2019), 564–568. doi:10/gpcff2.

236. Povey, D. *et al. The Kaldi Speech Recognition Toolkit*. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (IEEE Signal Processing Society, Waikoloa Village, HI, 2011). ISBN: 978-1-4673-0366-8. https://infoscience.epfl.ch/record/192584?v=pdf.

237. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. *Robust Speech Recognition via Large-Scale Weak Supervision*. In *Proceedings of the 40th International Conference on Machine Learning - ICML '23* (2023). arXiv: 2212.04356.

238. Lazer, D. Studying human attention on the Internet. *Proceedings of the National Academy of Sciences* **117,** 21–22. doi:10/ggh6fp (2020).

239. Laban, P. & Hearst, M. *newsLens: building and visualizing long-ranging news stories*. In *Proceedings of the Events and Stories in the News Workshop* (ACL, Vancouver, Canada, 2017), 1–9. doi:10/ghwpkd.

240. Zhang, Y., Guo, F., Shen, J. & Han, J. *Unsupervised Key Event Detection from Massive Text Corpora*. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), 2535–2544. doi:10/grhskd.

241. Staykovski, T. & Barron-Cedeno, A. *Dense vs. Sparse Representations for News Stream Clustering*. In *Proceedings of the Text2StoryIR Workshop @ ECIR 2019* (2019). https://ceur-ws.org/Vol-2342/paper6.pdf.

242. Reimers, N. & Gurevych, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (ACL, Hong Kong, China, 2019), 3980–3990. doi:10/ggfrnw.

243. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008,** P10008. doi:10/bxz74q (2008).

244. Galil, Z. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys* **18,** 23–38. doi:10/d42twn (1986).

245. Hagberg, A., Schult, D. & Swart, P. *Exploring network structure, dynamics and function using NetworkX*. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (eds Varoquax, G., Vaught, T. & Millman, J.) (Pasadena, CA USA, 2008), 11–15. https://aric.hagberg.org/papers/hagberg-2008-exploring.pdf.

246. Vijayaraghavan, P., Vosoughi, S. & Roy, D. *Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (ACL, Stroudsburg, PA, USA, 2017), 478–483. doi:10/gpcfgz.

247. Barberá, P. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* **23,** 76–91. doi:10/f63r45 (2015).

248. Green, J. *Identifying and Estimating the Ideologies of Twitter Pundits*. Data For Progress. https://www.dataforprogress.org/blog/2018/11/19/identifying-and-estimating-the-ideologies-of-twitter-pundits.

249. Timm, J. *Twitter, political ideology & the 115th US Senate*. R-bloggers. https://www.r-bloggers.com/2018/11/twitter-political-ideology-the-115th-us-senate/.

250. Yin, W., Hay, J. & Roth, D. *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (ACL, Hong Kong, China, 2019), 3912–3921. doi:10/ghhzw9.

251. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL, 2020), 7871–7880. doi:10/gg8hcp.

252. Halberstam, Y. & Knight, B. Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of Public Economics* **143,** 73–88. doi:10/f88stf (2016).

253. Frimer, J. A., Aujla, H., Feinberg, M., Skitka, L. J., Aquino, K., Eichstaedt, J. C. & Willer, R. Incivility Is Rising Among American Politicians on Twitter. *Social Psychological and Personality Science* **14,** 259–269. doi:10/gspvv5 (2023).

254. Radcliffe, N. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal,* 14–21. https://www.research.ed.ac.uk/en/publications/using-control-groups-to-target-on-predicted-lift-building-and-ass (2007).

255. Rousseau, J.-J. *On The Social Contract*. Trans. by Cole, G. ISBN: 978-1-945644-99-3 (Dent, New York, 1913).

256. Hovland, C. I. & Lumsdaine, A. A. *Experiments on Mass Communication*. ISBN: 978-1-4008-8680-7 (Princeton University Press, 2017).

257. Lazarsfeld, P. F., Berelson, B. & Gaudet, H. *The people's choice: how the voter makes up his mind in a presidential campaign*. (Duell, Sloan and Pearce, New York, 1944).

258. Mutz, D. C. *Population-based survey experiments*. ISBN: 978-0-691-14451-1 (Princeton University Press, Princeton, 2011).

259. Sutton, R. *The Bitter Lesson*. Incomplete Ideas. http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

260. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* **3,** 1137–1155. https://www.jmlr.org/papers/v3/bengio03a.html (Feb 2003).

261. Peng, B. *et al. RWKV: Reinventing RNNs for the Transformer Era*. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (ACL, Singapore, 2023), 14048–14077. doi:10/gtzvjh.

262. Firth, J. R. *A Synopsis of Linguistic Theory, 1930-1955*. In *Studies in linguistic analysis* 1–31 (1957).

263. Radford, A. *et al. Learning Transferable Visual Models From Natural Language Supervision*. In *Advances in Neural Information Processing Systems (NeurIPS)* **139** (2021), 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

264. Howard, J. & Ruder, S. *Universal Language Model Fine-tuning for Text Classification*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACL, Melbourne, Australia, 2018), 328–339. doi:10/ggsc9c.

265. Pearl, J. *Causality: Models, Reasoning, and Inference*. 2nd ed. ISBN: 978-0-511-80316-1. doi:10/ggd72q (Cambridge University Press, 2009).

266. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701. doi:10/cqq7sc (1974).

267. Splawa-Neyman, J., Dabrowska, D. M. & Speed, T. P. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* **5.** doi:10/gcjqm9 (1990).

268. Fisher, R. *The Design of Experiments*. (Oliver and Boyd, 1935).

269. Rubin, D. B. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association* **100,** 322–331. doi:10/cqsb7s (2005).

270. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116,** 4156–4165. doi:10/gfwcnc (2019).

271. Shi, C., Blei, D. M. & Veitch, V. *Adapting Neural Networks for the Estimation of Treatment Effects*. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS '19)* (2019). arXiv: 1906.02120.

272. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* **113,** 1228–1242. doi:10/gfdc56 (2018).

273. Gutierrez, P. & Gérardy, J.-Y. *Causal Inference and Uplift Modelling: A Review of the Literature*. In *Proceedings of the 3rd International Conference on Predictive Applications and APIs* (PMLR, 2017), 1–13. https://proceedings.mlr.press/v67/gutierrez17a.html.

274. Hinkelmann, K. & Kempthorne, O. *Design and analysis of experiments*. 2nd ed. ISBN: 978-0-471-72756-9 (Wiley-Interscience, Hoboken, N.J, 2008).

275. Rubin, D. B. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. *Journal of the American Statistical Association* **75,** 591. doi:10/bh6d6x (1980).

276. Rubin, D. B. Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association* **81,** 961. doi:10/d6v8dn (1986).

277. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. & Zhou, D. *Chain-of-thought prompting elicits reasoning in large language models*. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS '22)* (Curran Associates Inc., Red Hook, NY, USA, 2022), 24824–24837. ISBN: 978-1-71387-108-8. arXiv: 2201.11903.

278. OpenAI *et al. OpenAI o1 System Card*. 2024. arXiv: 2412.16720.

279. Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E. & Hashimoto, T. *s1: Simple test-time scaling*. 2025. arXiv: 2501.19393.

280. Snell, C. V., Lee, J., Xu, K. & Kumar, A. *Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning*. In *Proceedings of the 13th International Conference on Learning Representations (ICLR '25)* (2025). arXiv: 2408.03314.

281. Jiang, Z., Araki, J., Ding, H. & Neubig, G. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* **9,** 962–977. doi:10/gpmkrh (2021).

282. McCullagh, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **42,** 109–127. doi:10/ggntw8 (1980).

283. Norman, G. Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education* **15,** 625–632. doi:10/bzds22 (2010).

284. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. doi:10/d8zjwq (2001).

285. Wolf, T. *et al. Transformers: State-of-the-Art Natural Language Processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (ACL, 2020), 38–45. doi:10/ghs3bd.

286. Touvron, H. *et al. LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971.

287. Grattafiori, A. *et al. The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783.

288. Hu, J., Wu, X., Wang, W., Xianyu, Zhang, D. & Cao, Y. *OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework*. 2024. doi:10.48550/arXiv.2405.11143.

289. OpenAI *et al. GPT-4 Technical Report*. 2024. arXiv: 2303.08774.

290. Biderman, S. *et al. Pythia: a suite for analyzing large language models across training and scaling*. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)* **202** (JMLR.org, Honolulu, Hawaii, USA, 2023), 2397–2430.

291. Chong, D. & Druckman, J. N. Dynamic Public Opinion: Communication Effects over Time. *American Political Science Review* **104,** 663–680. doi:10/c6b472 (2010).

292. Hiscox, M. J. Through a Glass and Darkly: Attitudes toward International Trade and the Curious Effects of Issue Framing. *International Organization* **60,** 755–780. doi:10/bp25j9 (2006).

293. Zhu, C., Xu, B., Wang, Q., Zhang, Y. & Mao, Z. *On the Calibration of Large Language Models and Alignment*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)* (Singapore, 2023). https://openreview.net/forum?id=Of2xc2GVid.

294. Hewitt, L., Ashokkumar, A., Ghezae, I. & Willer, R. *Predicting Results of Social Science Experiments Using Large Language Models*. 2024. https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf.

295. McKenzie, I. R. *et al.* Inverse Scaling: When Bigger Isn't Better. *Transactions on Machine Learning Research.* arXiv: 2306.09479 (2023).

296. Mutz, D. C. (Mis)Attributing the Causes of American Job Loss. *Public Opinion Quarterly* **85,** 101–122. doi:10/g8w478 (2021).

297. DeMora, S., Merolla, J., Newman, B. & Zechmeister, E. *Values-consistent messaging to a vaccine skeptical population*. 2022. https://www.tessexperiments.org/study/newman1436.

298. Abacha, A. B., Yim, W.-w., Fu, Y., Sun, Z., Yetisgen, M., Xia, F. & Lin, T. *MEDEC: A Benchmark for Medical Error Detection and Correction in Clinical Notes*. 2025. arXiv: 2412.19260.

299. McCombs, M. E. & Shaw, D. L. The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly* **36,** 176. doi:10/dnd7p4 (1972).

300. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. *The role of social networks in information diffusion*. In *Proceedings of the 21st international conference on World Wide Web* WWW 2012: 21st World Wide Web Conference 2012 (ACM, Lyon France, 2012), 519–528. ISBN: 978-1-4503-1229-5. doi:10/gf3gv2.

301. Chu, E. *Learning human beliefs with language models*. PhD thesis (Massachusetts Institute of Technology, Cambridge, MA, 2022). https://hdl.handle.net/1721.1/143376.

302. Longpre, S. *et al. Consent in Crisis: The Rapid Decline of the AI Data Commons*. In *Proceedings of the 38th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS '24)* (2024). arXiv: 2407.14933.

303. Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K. I., Pentland, A. & Kabbara, J. *Position: Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?* In *Proceedings of the 41st International Conference on Machine Learning* (PMLR, 2024), 32711–32725. arXiv: 2404.12691.

304. Bird, S., Klein, E. & Loper, E. *Natural language processing with Python*. ISBN: 978-0-596-51649-9. https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ (O'Reilly, Cambridge, 2009).

305. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. & Hashimoto, T. B. *Stanford Alpaca: An instruction-following LLaMA model*. 2023. https://github.com/tatsu-lab/stanford_alpaca.

306. Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M. & Xin, R. *Free Dolly: Introducing the world's first truly open instruction-tuned LLM*. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

307. Köpf, A. *et al. OpenAssistant Conversations - Democratizing Large Language Model Alignment*. In *Proceedings of the 37th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS '23)* (2023). arXiv: 2304.07327.

308. Loshchilov, I. & Hutter, F. *Decoupled Weight Decay Regularization*. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)* (2018). arXiv: 1711.05101.

309. Dettmers, T., Lewis, M., Shleifer, S. & Zettlemoyer, L. *8-bit Optimizers via Block-wise Quantization*. In *Proceedings of the 9th International Conference on Learning Representations (ICLR '21)* (2021). arXiv: 2110.02861.

310. Chen, T., Xu, B., Zhang, C. & Guestrin, C. *Training Deep Nets with Sublinear Memory Cost*. 2016. arXiv: 1604.06174.

311. Micikevicius, P. *et al. Mixed Precision Training*. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)* (2018). arXiv: 1710.03740.

312. Dao, T. *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*. In *Proceedings of the 12th International Conference on Learning Representations (ICLR '23)* (2023). arXiv: 2307.08691.

313. Hsu, P.-L., Dai, Y., Kothapalli, V., Song, Q., Tang, S., Zhu, S., Shimizu, S., Sahni, S., Ning, H. & Chen, Y. *Liger Kernel: Efficient Triton Kernels for LLM Training*. 2025. arXiv: 2410.10989.

314. He, B., Weng, Y., Tang, X., Cui, Z., Sun, Z., Chen, L., He, X. & Ma, C. *Rankability-enhanced Revenue Uplift Modeling Framework for Online Marketing*. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2024), 5093–5104. ISBN: 9798400704901. doi:10/g9hhjb.

315. Betlei, A., Diemert, E. & Amini, M.-R. *Uplift Modeling with Generalization Guarantees*. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, New York, NY, USA, 2021), 55–65. ISBN: 978-1-4503-8332-5. doi:10/gn3r6n.

316. Yuan, Y. *Multi-Calibrated Heterogeneous Treatment Effect Estimation*. 2023. doi:10/g9hg4b.

317. Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* **17,** 3008–3049. doi:10/g83tth (2023).

318. Cheng, Z., Zhang, H., Li, K., Leng, S., Hu, Z., Wu, F., Zhao, D., Li, X. & Bing, L. *Breaking the Memory Barrier: Near Infinite Batch Size Scaling for Contrastive Loss*. 2024. arXiv: 2410.17243.

319. Jeh, G. & Widom, J. *SimRank: a measure of structural-context similarity*. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)* (ACM Press, Edmonton, Alberta, Canada, 2002), 538. ISBN: 978-1-58113-567-1. doi:10/c4h8xv.

320. Bolsen, T. & Thornton, J. R. Do norm-based appeals affect the acceptance of the singular use of they/them pronouns? *Research & Politics* **11.** Publisher: SAGE Publications Ltd, 20531680241264078. doi:10/g8w2sg (2024).

321. Thornton, J. R. & Bolsen, T. *Do Messages That Encourage the Use of They/Them Pronouns Influence Public Opinion?* 2022. https://www.tessexperiments.org/study/bolsenM6.

322. Brown, N. D., Jacoby-Senghor, D. S. & Raymundo, I. If you rise, I fall: Equality is prevented by the misperception that it harms advantaged groups. *Science Advances* **8.** Publisher: American Association for the Advancement of Science, eabm2385. doi:10/htcg (2022).

323. Brown, N. D. *Equality is Misperceived as Harmful to Advantaged Groups: An Intervention to Increase Vaccine Equity Support*. 2021. https://www.tessexperiments.org/study/brownS80.

324. Fahey, J. J., Roberts, D. C. & Utych, S. M. Principled or Partisan? The Effect of Cancel Culture Framings on Support for Free Speech. *American Politics Research* **51,** 69–75. doi:10/grpcr7 (2023).

325. Fahey, J. & Utych, S. *Cancel Culture for Friends, Consequence Culture for Enemies: The Effects of Ideological Congruence on Perceptions of Free Speech*. 2021. https://www.tessexperiments.org/study/faheyS78.

326. Flavin, P. *Public Attitudes about Political Equality*. 2011. https://www.tessexperiments.org/study/flavin235.

327. Gash, A. & Murakami, M. *Understanding How Policy Venue Influences Public Opinion*. 2010. https://www.tessexperiments.org/study/gash-murakami718.

328. Groenendyk, E., Krupnikov, Y., Ryan, J. B. & Connors, E. C. Selecting Out of "Politics": The Self-Fulfilling Role of Conflict Expectation. *American Political Science Review,* 1–16. doi:10/g8w425 (2024).

329. Groenendyk, E. *What Motivates Interest in Politics?* 2021. https://www.tessexperiments.org/study/groenendykS79.

330. Pickup, M., Kimbrough, E. O., Groenendyk, E. W. & Banks, A. J. *What Counts as Racism? American Norms, Partisan Norms, and Personal Beliefs About Racism*. Rochester, NY, 2024. doi:10.2139/ssrn.5006267.

331. Groenendyk, E. & Banks, A. *What Counts as Racism? How Conceptual Disagreement Fuels Social and Political Division*. 2022. https://www.tessexperiments.org/study/groenendyk1419.

332. Hiscox, M. J. *The Political Backlash Against Globalization? Framing Effects, Information, and Attitudes Toward Internation Trade*. 2003. https://tessexperiments.org/study/hiscox136.

333. Johnston, C. D. & Ballard, A. O. Economists and Public Opinion: Expert Consensus and Economic Policy Judgments. *The Journal of Politics* **78,** 443–456. doi:10/f8qpcf (2016).

334. Johnston, C. D. *Economists and Public Opinion: Expert Consensus and Economic Policy Judgments*. 2013. https://tessexperiments.org/study/johnstonbrief4.

335. Kreps, S. E. & Wallace, G. P. International law, military effectiveness, and public support for drone strikes. *Journal of Peace Research* **53,** 830–844. doi:10/f9bmfd (2016).

336. Wallace, G. & Kreps, S. E. *Elites, International Law, and Public Support for Warfare*. 2013. https://tessexperiments.org/study/wallacebrief13.

337. Levendusky, M. S. Americans, Not Partisans: Can Priming American National Identity Reduce Affective Polarization? *The Journal of Politics* **80,** 59–70. doi:10/gcvdqv (2018).

338. Levendusky, M. *Can We Reduce Affective Polarization in the Mass Public?* 2016. https://www.tessexperiments.org/study/levendusky741.

339. Mutz, D. *The Political Impact of Others' Job Loss: Personifying the Enemy*. 2017. https://tessexperiments.org/study/Mutz886.

340. Saha, S. Why don't politicians talk about meat? The political psychology of human-animal relations in elections. *Frontiers in Psychology* **14,** 1021013. doi:10/g8w2sj (2023).

341. Saha, S. *Why Don't Politicians Talk About Meat? Voter Backlash and the Limits of Innovation Credit*. 2021. https://www.tessexperiments.org/study/sahaS66.

342. Wu, J. D. National Origin Identity and Descriptive Representativeness: Understanding Preferences for Asian Candidates and Representation. *Public Opinion Quarterly* **87,** 978–999. doi:10/g8w2sh (2023).

343. Wu, J. *Candidate Asian-ness and Descriptive Representativeness*. 2021. https://www.tessexperiments.org/study/wu1435.

344. Verbeken, B., Guerry, M.-A. & Verboven, S. *Uplift model evaluation with ordinal dominance graphs*. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases-ECML/PKDD'22* (ECML/PKDD'22 Uplift Modeling Tutorial & Workshop., 2022). https://upliftworkshop.ipipan.waw.pl/pdf/Presentation_Workshop_Uplift.pdf.

345. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and their Application*. 1st ed. ISBN: 978-0-521-57391-7. doi:10.1017/cbo9780511802843 (Cambridge University Press, 1997).