Can Large Language Models Transform Computational Social Science?

Caleb Ziems[⋆, GT] **William Held**[⋆, GT] **Omar Shaikh**[⋆, Stanford] **Jiaao Chen**[⋆, GT]
Zhehao Zhang[⋆, SJTU] **Diyi Yang**[⋆, Stanford]

Georgia Institute of Technology, Shanghai Jiao Tong University, Stanford University

{cziems, wheld3, jiaaochen}@gatech.edu, zzh12138@sjtu.edu.cn, {oshaikh, diyiy}@stanford.edu

Abstract

Large Language Models (LLMs) like ChatGPT are capable of successfully performing many language processing tasks zero-shot (without the need for training data). If this capacity also applies to the coding of social phenomena like persuasiveness and political ideology, then LLMs could effectively transform Computational Social Science (CSS). This work provides a road map for using LLMs as CSS tools. Towards this end, we contribute a set of prompting best practices and an extensive evaluation pipeline to measure the zero-shot performance of 13 language models on 24 representative CSS benchmarks. On taxonomic labeling tasks (classification), LLMs fail to outperform the best fine-tuned models but still achieve fair levels of agreement with humans. On free-form coding tasks (generation), LLMs produce explanations that often *exceed* the quality of crowdworkers' gold references. We conclude that today's LLMs can radically augment the CSS research pipeline in two ways: (1) serving as zero-shot data annotators on human annotation teams, and (2) bootstrapping challenging creative generation tasks (e.g., explaining the hidden meaning behind text). In summary, LLMs can significantly reduce costs and increase efficiency of social science analysis *in partnership with humans*.

1 Introduction

> The most surprising scientific changes tend to arrive, not from accumulated facts and discoveries, but from the invention of new tools and methodologies that trigger "paradigm shifts" (Kuhn, 1962).

Computational Social Science (CSS) (Lazer et al., 2020) was born from the immense growth of human data traces on the web and the rapid acceleration of computational resources for processing this data. These developments allowed researchers to study language and behavior at an unprecedented scale (Lazer et al., 2009), with both global and fine-grained observations (Golder and Macy, 2014). From the early days of content dictionaries (Stone et al., 1966), statistical text analysis has facilitated CSS research by providing structure to non-numeric data. Now, Large Language Models (LLMs) may be poised to change the computational social science landscape by providing such capabilities without custom training data.

The goal of this work is to assess the degree to which *LLMs can transform Computational Social Science (CSS)*. Solid computational approaches are needed to help analyze textual data and to understand a variety of social phenomena across academic disciplines. Current CSS methodologies typically use *supervised* text classification and generation in order to scale up manual coding efforts to unseen texts (Nelson et al., 2021). Reliable supervised methods typically demand an extensive amount of human-annotated training data. Alternatively, *unsupervised* methods can run "for free," but the resulting output can be uninterpretable (Lee and Martin, 2015). In the status quo, data resources constrain the theories and subjects CSS can be applied to, especially as studies are largely concentrated on Western, Educated, Industrial, Rich, and Democratic populations (WEIRD; Ignatow and Mihalcea, 2016; Muthukrishna et al., 2020).

LLMs have the potential to remove these constraints. Recent LLMs have demonstrated the striking ability to reliably classify text, summarize documents, answer questions, and generate interpretable explanations in a variety of domains, even exceeding human performance *without the need for supervision* (Bang et al., 2023; Qin et al., 2023;

[⋆] Contribution distributed as follows. Caleb, Will, and Diyi decided the project scope and research questions. Caleb performed the CSS literature review, subject and task selection. Will built the evaluation pipeline and prompting guidelines. Will, Caleb, and Omar all contributed data pre-processing, loading, and evaluation scripts. Will and Omar ran the Flan-T5/UL2 experiments, while Jiaao and Caleb ran the OpenAI prompting experiments. Zhehao was responsible for all baseline experiments. Caleb managed the human evaluations. All authors contributed to discussion, results, error analysis, and paper writing.

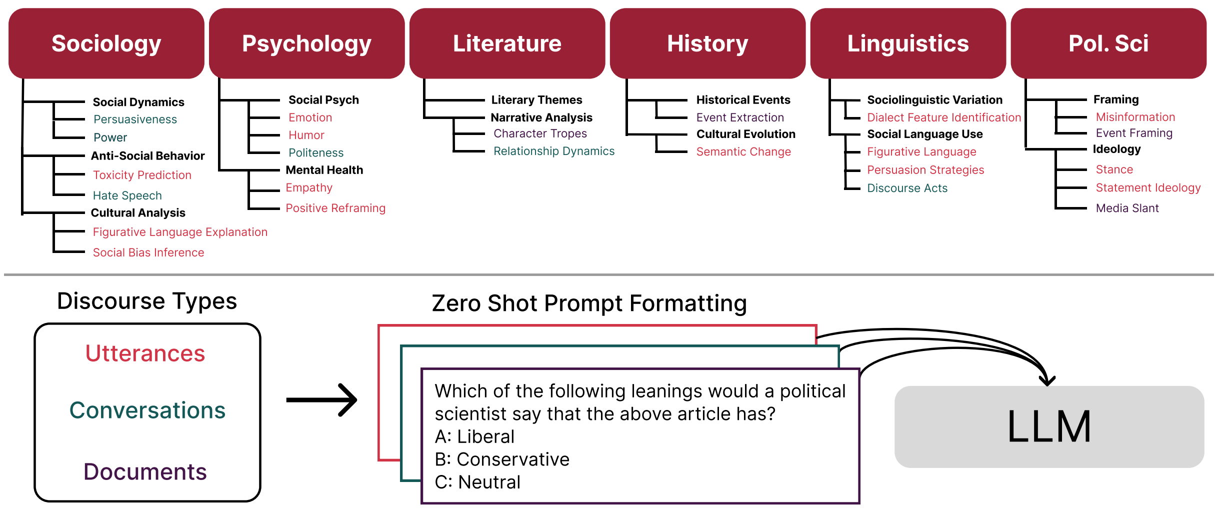


Figure 1: We assess the potential of LLMs as multi-purpose tools for CSS. We identify core subject areas in prior CSS work and select 24 diverse and representative tasks from across these fields (top). Then, we segment tasks into distinct discourse types and evaluate both open-source and industrial LLMs across this benchmark using zero-shot prompting (bottom).

Zhuo et al., 2023; Goyal et al., 2022). If LLMs can similarly provide reliable labels and summary codes through zero-shot prompting, CSS research is broadened a wider range of hypotheses than current tools and data resources support. Zero-shot viability in this space is our primary research question. To effectively harness the power of LLMs, behavioral researchers should understand the pros and cons of different modeling decisions (model-selection), as well as how these decisions intersect with their fields of specialization (domain-utility) and downstream use-cases (functionality). By evaluating LLMs on an extensive suite of CSS tasks, this work provides researchers with a road map with answers to the following research questions:

- **(RQ1) Viability:** Are LLMs able to augment the human annotation pipeline? Can they replace annotation entirely?
- **(RQ2) Model-Selection:** How do different aspects of LLMs (e.g., model size, pretraining) affect their performances on CSS tasks?
- **(RQ3) Domain-Utility:** Are zero-shot LLMs specially adapted for better results in some fields of science rather than others?
- **(RQ4) Functionality:** Are zero-shot LLMs equipped to assist with labeling tasks (classification) or summary-explanatory tasks (generation) or both?

The research pipeline in Figure 1 allows us to answer these questions. First, we survey the social science literature to understand where LLMs could serve as analytical tools (§2). Then we operationalize each use-case with a set of representative tasks (§3). Specifically, classification and parsing methods can help researchers code for linguistic, psychological, and cultural categories (§3.1-3.3) while generative models can explain underlying constructs (e.g., figurative language, hate speech, and misinformation), and restructure text according to established theories like cognitive behavioral therapy (§3.4). With a final evaluation suite of 24 tasks, we test the zero-shot performance of 13 language models with differing architectures, sizes, pre-training, and fine-tuning paradigms (§5, 6). This allows us to suggest actionable steps for social scientists interested in co-opting LLMs for research (§7). Specifically, we suggest a blended supervised-unsupervised scheme for human-AI partnered labeling and content analysis.

Concretely, our analysis reveals that, except in minority cases, prompted LLMs do not match or exceed the performance of carefully fine-tuned classifiers, and the best LLM performances are often too low to entirely replace human annotation. However, LLMs *can* achieve fair levels of agreement with humans on labeling tasks. These results are not limited to a subset of academic fields, but rather span the social sciences across a range of conversation, utterance, and document-level classification tasks. Furthermore, the benefits of LLMs are compounded as models scale up. This suggests that LLMs can augment the annotation process through iterative joint-labeling, significantly speeding up and improving text analysis in the social sciences.

Importantly, some LLMs can also generate informative explanations for social science constructs. In the best case, leading models can achieve parity with quality of dataset references. Humans prefer model outputs 50% of the time, suggesting that human-AI collaboration will extend beyond labeling tasks to the joint coding of new constructs, analyses, and summaries.

2 An Overview of CSS

Following Lazer et al. (2020), we define Computational Social Science as the development and application of computational methods to the scientific analysis of behavioral and linguistic data. Critically, CSS centers around the scientific method, forming and testing broad and objective hypotheses, while similar efforts in the Digital Humanities (DH) focus more on the subjectivity and particularity of events, dialogues, cultures, laws, value-systems, and human activities (Dobson, 2019).

This section surveys the current needs of researchers in both the computational social sciences and digital humanities. We choose to merge our discussion under the banner of CSS, since solid computational approaches are needed to help analyze textual data and to understand a variety of socio-behavioral phenomena across both scientific and humanistic disciplines. We focus primarily on the most tractable text classification, structured parsing, and natural language generation tasks for CSS. Some other techniques like aggregate mining of massive datasets or multi-document summarization and topic modeling may be largely outside the scope of transformer-based language models, which have a fixed processing window size and quadratic space complexity.

The following subsections outline how computational methods can support specific fields of inquiry regarding how people think (psychology; §2.5), communicate (linguistics; §2.3), establish governance and value-systems (political science, economics; §2.4), collectively operate (sociology; §2.6), and create culture (literature, anthropology; §2.2) across time (history; §2.1).

2.1 History

Historians study *events*, or transitions between states (Box-Steffensmeier and Jones, 2004; Abbott, 1990), like the onset of a war. Event extraction is a parsing task from unstructured text to more regular data structures which capture the location, time, cause, and participants in the event (Xiang and Wang, 2019). This task, which is central to a growing number of computational studies on history (Lai et al., 2021; Sprugnoli and Tonelli, 2019), can be broken into (1) event detection, and (2) event argument extraction, which we benchmark in §3.3.1 and 3.3.2 respectively. Historians also work to understand the influence of events on historical shifts in *discourse* (DiMaggio et al., 2013) and *meaning* (Hamilton et al., 2016a). We further discuss NLP for discourse and semantic change in §2.4 and §2.3.

2.2 Literature

Literary studies are closely tied to the analysis of *themes* (Jockers and Mimno, 2013), *settings* (Piper et al., 2021), and *narratives* (Sap et al., 2022; Saldias and Roy, 2020; Boyd et al., 2020). Settings can be identified using named entity recognition (Brooke et al., 2016) and toponym resolution (DeLozier et al., 2016), which are already demostrably solved by prompted models like ChatGPT (Qin et al., 2023). Themes are typically the subject of topic modeling, which is outside the scope of LLMs. Instead we focus on NLP for narrative analysis. NLP systems can be used to parse narratives into chains (Chambers and Jurafsky, 2008) with *agents* (Coll Ardanuy et al., 2020; Vala et al., 2015) their *relationships* (Labatut and Bost, 2019; Iyyer et al., 2016; Srivastava et al., 2016), and the *events* (Sims et al., 2019) they participate in. We cover social role labeling and event extraction methods in Sections 3.3.4 and 3.3.2 respectively. Researchers can also study agents in terms of their *power* dynamics (Sap et al., 2017) and *emotions* (Brahman and Chaturvedi, 2020), which we benchmark in §3.2.4 and 3.1.2. *Figurative language* (Kesarwani et al., 2017) and *humor* classification (Davies, 2017) are two other relevant tasks for the study of literary devices, and we evaluate these tasks in §3.1.3 and §3.1.5.

2.3 Linguistics

Computational sociolinguists use computational tools to measure the interactions between society and language, including the stylistic and structural features that distinguish speakers (Nguyen et al., 2016). Language variation is closely related to social identity (Bucholtz and Hall, 2005), from group membership (Del Tredici and Fernández, 2017), geographical region (Purschke and Hovy, 2019), and social class (Preoţiuc-Pietro et al., 2015) (Del Tredici and Fernández, 2017) to personal attributes like age and gender (Bamman et al., 2014). In Section 3.1.1 and 3.1.10, we use LLMs to identify the structural features of English dialects, which linguists can use to classify and systematically study dialects, measure different feature densities in different population strata, and study the onset and diffusion of language change (Kershaw et al., 2016; Eisenstein et al., 2014; Ryskina et al., 2020; Kulkarni et al., 2015; Hamilton et al., 2016b; Carlo et al., 2019; Zhu and Jurgens, 2021b; Schlechtweg et al., 2020).

2.4 Political Science

Political scientists study how political actors move *agendas* (Grimmer, 2010) by persuasively *framing* their discourse "to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (Entman, 1993). These agendas cohere within *ideologies*. Computational social scientists have advanced political science through the detection of political leaning, ideology, and stance (Ahmed and Xing, 2010; Baly et al., 2020; Bamman and Smith, 2015; Iyyer et al., 2014; Johnson et al., 2017; Preoţiuc-Pietro et al., 2017; Luo et al., 2020a; Stefanov et al., 2020), as well as *issue* (Iyengar, 1990) and *entity* framing (van den Berg et al., 2020). Applications for persuasion, framing, ideology, and stance detection in the social sciences are numerous. Analysts can uncover fringe issue topics (Bail, 2014) and frames (Ziems and Yang, 2021; Mendelsohn et al., 2021; Demszky et al., 2019; Field et al., 2018), with applications to public opinion (Bhatia, 2017; Garg et al., 2018; Kozlowski et al., 2019; Abul-Fottouh and Fetner, 2018), voting behavior (Black et al., 2011), policy change (Flores, 2017), social movements (Nelson, 2021; Sech et al., 2020; Rogers et al., 2019; Tufekci and Wilson, 2012), and international relations (King and Lowe, 2003). We benchmark ideology detection in §3.1.6 and §3.3.3, stance detection in §3.1.9, and entity framing in §3.3.4. Furthermore, understanding the discourse structure and persuasive elements of political speech can help social scientists measure political impact (Altikriti, 2016; Hashim and Safwat, 2015). We benchmark persuasion strategy and discourse acts classification in §3.1.8 and 3.2.1.

2.5 Psychology

As the science of mind and behavior, psychology intersects all other adjacent social sciences in this section. For example, an individual's personality, or their stable patterns of thought and behavior across time, will correlate with their political leaning (Gerber et al., 2010), social status (Anderson et al., 2001), and linguistic expression (Pennebaker and King, 1999). The most influential personality modeling benchmark, MyPersonality (Kosinski et al., 2013), is no longer available, but in this work, we evaluate on a representative set of psychological factors down-stream of personality. For example, differences in personality and cognitive processing have significant impact on what individuals find funny (Martin and Ford, 2018) or persuasive (Hirsh et al., 2012). These psychological factors then exert influence over a range of social interactions. Humor and politeness (Brown and Levinson, 1987) are correlated with subjective impressions of psychological distance between speakers (Trope and Liberman, 2010), while persuasive techniques bind agents in social commitments, with applications in the science of management and organizations. We evaluate on humor, persuasion, and politeness classification in §3.1.5, 3.1.8, and 3.2.6 respectively.

We also consider LLMs as tools for counseling, mental health and positive psychology in text-based interactions. Specifically, we evaluate on *empathy detection* in online mental health platforms (Sharma et al., 2020) in §3.2.2, and on a *positive reframing* style-transfer task (Ziems et al., 2022b) based on cognitive behavioral therapy in §3.4.4.

2.6 Sociology

Sociologists want to understand the structure of society and how people live collectively in social groups (Wardhaugh and Fuller, 2021; Keuschnigg et al., 2018). By tracing the diffusion and recombination of linguistic, political, and psychological content between actors in a community across time, sociologists can begin to understand social processes at both the micro and macro scale. At the micro scale, there is the computational sociology of power (Danescu-Niculescu-Mizil et al., 2012; Bramsen et al., 2011; Prabhakaran et al., 2014, 2012) and social roles (Welser et al., 2011; Fazeen et al., 2011; Zhang et al., 2007; Yang et al., 2015; Maki et al., 2017). LLMs can assist sociological research by predicting power relations (§3.2.4) and unhealthy conversations (§3.2.5). At the macro-scale, there are computational analyses of social norms and conventions (Centola et al., 2018; Bicchieri, 2005), information diffusion (Leskovec et al., 2009; Tan et al., 2014; Vosoughi et al., 2018; Cheng et al., 2016), emotional contagion (Bail, 2016), collective behaviors (Barberá et al., 2015), and social movements (Nelson, 2021, 2015). Again, LLMs can detect constructs like emotion (§3.1.2) and the speech of hateful social groups (§3.1.4). Furthermore, social movements rely on the diffusion of norms and idiomatic slogans, which carry meaning through figurative language that LLMs can decode (§3.1.3).

3 Representative CSS Task Selection

Now we operationalize the core CSS needs from §2 with concrete tasks. While not exhaustive, our task selection is designed to provide a representative benchmark of the required capabilities of a general-purpose CSS tool. We organize this section according to our division of tasks into functional categories based on the unit of text analysis: 10 utterance-level classification tasks (§3.1), 6 conversation-level tasks (§3.2), and 4 document-level tasks for the analysis of media (§3.3). In addition to these 20 classification tasks, we evaluate 4 generation tasks in Section 3.4 for explaining social science constructs and applying psychological theories to restructure text.

3.1 Utterance-Level Classification

An utterance is a unit of communication produced by a single speaker to convey a single subject, which may span multiple sentences (Bakhtin, 2010). CSS researchers can use utterance data to study linguistic phenomena like the syntax of dialect, the semantics of figurative language, or the pragmatics of humor. Utterance-level analysis also reflects human states like emotion and communicative intent, or stable traits like stance and ideology (Evans and Aceves, 2016). We evaluate LLMs on utterance classification tasks for dialect, hate speech, figurative language, emotion, humor, misinformation, ideology, persuasion, semantic change, and stance classification.

3.1.1 Dialect Features

Linguistic feature detection is critical to the study of dialects (Eisenstein et al., 2011) and ideolects (Zhu and Jurgens, 2021a), with numerous applications in sociolinguistics, education, and the sociology of class and community membership (see §2.3). These features can be used to study the sociolinguistics of language change (Kulkarni et al., 2015; Hamilton et al., 2016b) or the linguistic biases in educational assessments (Craig and Washington, 2002) and online moderation (Sap et al., 2019). The utterance is an appropriate level of analysis here because syntactic and morphological features are all defined on subtrees of the sentence node (Ziems et al., 2022a).

We evaluate on the Indian English dialect feature detection task of Demszky et al. (2019) because this is one of the only available datasets to be hand-labeled by a domain expert. Additionally, Indian English is the most widely-spoken low-resource variety of English, so the domain is representative. The task is to map utterances to a set of 22 grammatical features: i.e., a lack of inversion in wh-questions, the omission of copula *be*, or features related to tense and aspect like the *habitual progressive*, found in Indian varieties of English.

3.1.2 Emotions

Emotion detection, the cornerstone of affective computing (Picard, 2000), is highly relevant to psychology and political science, among other disciplines, since stable emotional patterns in-part define an individual's personality, and targeted emotions outline the political stances she has. Additional application domains for the task include emotional contagion (Bail, 2016) and human factors behind economic markets (Bollen et al., 2011; Nguyen and Shirai, 2015).

Expert-labeled emotion detection datasets are not common. We evaluate emotion detection with weakly labeled Twitter data from Saravia et al. (2018), which uses Plutchik's 8 emotional categories: *anger, anticipation, disgust, fear, joy, sadness, surprise,* and *trust*. This is one of the three most recognized discrete emotion models, besides Paul Ekman et al.'s 6-category model and the 22-category model of Ortony et al. (2022).

3.1.3 Figurative Language

Figurative expressions are where the speaker meaning differs from the utterance's literal meaning. Recognizing figurative language is a first step in understanding literary (Jacobs and Kinder, 2018) and political texts (Huguet Cabot et al., 2020), detecting hate speech (Lemmens et al., 2021) and identifying mental health self-disclosure (Iyer et al., 2019).

We use the FLUTE (Chakrabarty et al., 2022) benchmark because, presently, FLUTE is the most comprehensive resource with examples from wide range of prior datasets (Chakrabarty et al., 2021; Srivastava et al., 2022; Stowe et al., 2022). FLUTE contains 9k figurative sentences. The classification task is to recognize whether the figurative sentence contains (1) *sarcasm* (Joshi et al., 2017), (2) *simile* (Niculae and Danescu-Niculescu-Mizil, 2014), (3) *metaphor* (Gao et al., 2018), or (4) an *idiom* (Jochim et al., 2018).

3.1.4 Hate Speech

Hate speech is language that disparages a person or group on the basis of protected characteristics like race. Beyond the societal importance of detecting and mitigating hate speech, this is a category of language that is salient to many social scientists. By not only detecting, but also systematically understanding hate speech, political scientists can track the rise of hateful ideologies, and sociologists can understand how these hateful ideas diffuse through a network and influence social movements.

Thus we evaluate on the more nuanced task of fine-grained hate speech taxonomy classification from Latent Hatred (ElSherief et al., 2021). This task requires models to infer a subtle social taxonomy from the coded or indirect speech of U.S. hate groups. Utterances should be classified with one of six domain-specific categories: *incitement to violence, inferiority language, irony, stereotypes and misinformation, threatening and intimidation language*, and *white grievance*.

3.1.5 Humor

Humor is a rhetorical (Markiewicz, 1974) and literary device (Kuipers, 2009) that modulates social distance and trust (Sherman, 1988; Graham, 1995; Kim et al., 2016). However, different audiences may perceive the same joke differently. In the study of sociocultural variation, communication, and bonding, humor detection will be of prime interest to sociologists and social psychologists, as well as to literary theorists and historians. Computational social scientists have effectively detected punchlines (Mihalcea and Strapparava, 2005; Ofer and Shahaf, 2022) and predicted audience laughter (Chen and Soo, 2018), demonstrating the computational tractability of the domain.

Our evaluation uses a popular dataset from Weller and Seppi (2019) to focus on binary humor detection across a wide range of joke sources, from Reddit's r/Jokes, a *Pun of The Day* website, and a set of short jokes from Kaggle, summing to around 16K jokes.

3.1.6 Ideology

A speaker's subtle decisions in word choice and diction can betray their beliefs and the political environment to which they belong (Jelveh et al., 2014). While political scientists care most about identifying the underlying ideologies and partisan organizations behind these actors (§2.4), sociolinguists can study the correlation between language and social factors.

We evaluate ideology detection on the Ideological Books Corpus (Gross et al., 2013) from Iyyer et al. (2014), which contains 2,025 liberal sentences, 1,701 conservative sentences, and 600 neutral sentences. The corpus was designed to disentangle a speaker's overall partisanship from the particular ideological beliefs that are reflected in an individual utterance. Thus labels reflect *perceived* ideology according to annotators and not the speaker's ground truth partisan affiliation.

3.1.7 Misinformation

Misinformation is both a political and social concern as it can jeopardize democratic elections, public health, and economic markets. The effort to combat misinformation is multi-disciplinary (Lazer et al., 2018), and it depends on reliable misinformation detection tools.

We evaluate on the Misinfo Reaction Frames corpus (Gabriel et al., 2022), a dataset of 25k news headlines with fact checked labels for the accuracy of the related news articles about COVID-19, climate change, or cancer. Models perform binary misinformation classification on news article headlines alone, which the authors found was a tractable task for fine-tuned models.

3.1.8 Persuasion

Persuasion is the art of changing or reinforcing the beliefs of others. Understanding persuasive strategies is central to behavioral economics and the psychology of advertising and propaganda (Martino et al., 2020). Utterances are a natural unit for the analysis of individual persuasive strategies, which may be combined in dialogue for an overall persuasive effect (c.f. §3.2.3).

While multi-modal persuasion detection tasks exist, we focus on the popular text-based persuasion dataset, Random Acts of Pizza (RAOP) (Althoff et al., 2014), where Reddit users attempt to convince community members to give them free food. This dataset was labeled by Yang et al. (2019a) with a fine-grained persuasive strategy taxonomy based on Cialdini (2003) that includes *Evidence, Impact, Politeness, Reciprocity, Scarcity*, and *Emotion*. The task objective is to classify utterance-level RAOP requests according to this 6-class taxonomy.

3.1.9 Stance

Although stance detection can be formalized in different ways, the most common task design is for models to determine whether a text's author is in favor of a target view, against the target, or neither. With this formulation, sociologists can understand consensus and disagreement in social groups, psychologists can measure interpersonal attachments, network scientists can build signed social graphs, political scientists can track the views of a voter base or the policies of candidates, historians can plot shifting opinions, and digital humanities researchers can quickly summarize narratives via character intentions and goals.

We evaluate stance detection on the earliest and most established SemEval-2016 Stance Dataset (Mohammad et al., 2016), which contains 1,250 tweets and their associated stance towards five topics: *atheism, climate change, the feminist movement, Hillary Clinton,* and the *legalization of abortion.* Stance is given as *favor*, *against*, or *none*.

3.1.10 Semantic Change

In addition to its more stable features, researchers can plot the change of language over time for a fixed community. Semantic change detection can serve as a proxy measure for the spread and change of culture (Kirby et al., 2007), both on the internet (Eisenstein, 2012; Eisenstein et al., 2014) and in historical archives (Mihalcea and Nastase, 2012; Kim et al., 2014; Kulkarni et al., 2015; Rudolph and Blei, 2018)[1].

We evaluate LLMs as binary word-sense discriminators using the popular Temporal Word-in-Context (TempoWiC) benchmark (Pilehvar and Camacho-Collados, 2019). TempoWiC measures the core capability of drawing discrete boundaries between word-level semantics. Given two sentences with the same lexeme, the task is binary classification with positive indicating both sentences use the same sense of the word and negative indicating different senses of the word. A perfect classifier for this task can be used to cluster all usage of a surface-form into sense groups using pairwise comparison.

3.2 Conversation-Level Classification

Conversations are multi-party exchanges of utterances. They are critical units for analysis in the social sciences (Hutchby and Wooffitt, 2008; Silverman, 1998; Sacks, 1992), since they richly reflect social *relationships* (Evans and Aceves, 2016) — a key factor that was missing in utterance-level analysis. Sociological frameworks like ethnomethodology (Garfinkel, 2016) focus particularly on conversations. The tasks in this section are drawn largely from the ConvoKit toolkit of Chang et al. (2020).

3.2.1 Discourse Acts

Discourse acts are the building blocks of conversations and are thus relevant to conversation analysis in sociology, genre analysis in literature, pragmatics, and ethnographic studies of speech communities (see Paltridge and Burton for example). Some popular discourse act taxonomies like DAMSL (Stolcke et al., 2000) and DiAML (Bunt et al., 2010) are tailored to spoken communication and can have as many as 40 categories. We use the simpler and more focused 9-class taxonomy of Zhang et al. (2017) since it was designed to cover *online* text conversations—the focus of CSS research. The taxonomy includes *questions, answers, elaborations, announcements, appreciation, agreements, disagreements, negative reactions*, and *humor*.

We evaluate on the Coarse Discourse Sequence Corpus (Zhang et al., 2017). The model input is a comment from a Reddit thread, along with the utterance to which the comment is responding. The expected output is the category from the above 9-class taxonomy which best describes the comment's speech act. However, *announcements* and *negative reactions* have fewer than 10 examples total in the dataset, so they are omitted from our evaluation along with the catch-all *other* category.

3.2.2 Empathy

Since the early days of internet access, users have looked to internet communities for support (Preece, 1998). Thus web communities can provide CSS researchers with empathetic communication data in naturalistic settings (Pfeil and Zaphiris, 2007; Sharma et al., 2020). By better understanding community-specific affordances (Zhou and Jurgens, 2020) and the most common triggers for empathetic responses (Buechel et al., 2018; Omitaomu et al., 2022), CSS can reciprocally inform the design of empathetic communities (Coulton et al., 2014; Taylor et al., 2019), as well as community-specific tools like counseling dialogue systems (Sharma et al., 2021; Ma et al., 2020).

Understanding is the first step towards building more effective online mental health re-

[1]Additional works in this area can be found under the Workshop on Computational Approaches to Historical Language Change

sources, and this motivates our evaluation on EPITOME (Sharma et al., 2020), a clinically-motivated empathy detection dataset. EPITOME measures empathy using a multi-stage labeling scheme. First, a listener communicates an *Emotional Reaction* to describe how the seeker's disclosure makes the listener feel. Then the listener offers an *Interpretation* of the emotions the seeker is experiencing. Finally, the listener moves into *Exploration*, or the pursuit of further information to better understand the seeker's situation. Clinical psychologists labeled the listener's effectiveness at each stage of a listener's top-level reply. Here we focus on *Exploration*, as prior work has shown open-questions to be especially effective for peer-support (Shah et al., 2022). Given a seeker's post and a top-level listener's reply, we classify whether the listener offered: *Strong Exploration* (specific questions about the seekers situation), *Weak Exploration* (general questions), or *No Exploration.*

3.2.3 Persuasion

In §3.2.3, we considered utterance-level analysis of fine-grained persuasive strategies. However, social scientists are also interested in the overall persuasive effect that one speaker has on another through sequences of rhetorical strategies in dialogue (Shaikh et al., 2020). Persuasive outcomes are particularly important for the political science of successful campaigns (Murphy and Shleifer, 2004) and the sociology of idea propagation and social movements (Stewart et al., 2012).

We evaluate our persuasion prediction task on the Persuasion for Good Corpus (Wang et al., 2019), which contains 1,017 conversations where the persuader tries to convince the persuadee to donate to a target charity. Models receive as input the truncated conversation thread and perform binary prediction on whether the persuasion was successful: *did the persuadee donated a non-zero amount to the charity after the conversation?*

3.2.4 Power and Status

Sociologists, political scientists, and online communities researchers are interested in understanding hierarchical organizations, social roles, and power relationships. Power is related to control of the conversation (Prabhakaran et al., 2014) and power dynamics shape both behavior and communication. Specifically, text analysis can uncover power relationships in the degree to which one speaker accommodates to the linguistic style of another (Danescu-Niculescu-Mizil et al., 2012). We anticipate that this task is tractable for LLMs.

We evaluate on the Wikipedia Talk Pages dataset from Danescu-Niculescu-Mizil et al. (2012). Conversations are drawn from the debate forums regarding Wikipedia edit histories, and power is a binary label describing whether or not the Wikipedia editor is an administrator. All models are given an editor's entire comment history from the Talk Pages, and the objective is binary classification.

3.2.5 Toxicity Prediction

Toxicity is a major area of social research in online communities, as online disinhibition (Suler, 2005) makes antisocial behaviour especially prevalent (Cheng et al., 2015). Predictive models can be used to understand the early signs of later toxicity (Cheng et al., 2017) for downstream causal analysis on the evolution of toxicity (Mathew et al., 2020) and the effectiveness of intervention methods (Kwak et al., 2015). Even without clearly-interpretable features, a predictive system can serve causal methods as a propensity score.

Using the Conversations Gone Awry corpus (Zhang et al., 2018), we investigate whether LLMs can predict future toxicity from early cues. As context, the model takes the first two messages in a conversation between Wikipedia users. The model should make a binary prediction whether or not the Wikipedia conversation will contain toxic language at any later stage.

3.2.6 Politeness

Before overt toxicity is evident in a community, researchers can measure its health and stability according to members' adherence to politeness norms. Polite members can help communities grow and retain other valuable members (Burke and Kraut, 2008), while rampant impoliteness in a community can foreshadow impending toxicity (Andersson and Pearson, 1999). Text-based politeness measures also reflect other societal factors that we explore in this work, like gender bias (Herring, 1994; Ortu et al., 2016, §3.1.4), power inequality (Danescu-Niculescu-Mizil et al., 2013, §3.2.4), and persuasion (Shaikh et al., 2020, §3.1.8).

We evaluate on the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013). The dataset is foundational in the computational study of politeness and its relation to other social dynamics. The corpus contains requests made by one Wikipedia contributor to another. Each request is classified into one of three categories, *Polite, Neutral*, or *Impolite*, according to Mechanical Turk annotators' interpretation of workplace norms. High zero-shot performance on this task will strongly indicate a model's broader ability to recognize conversational social norms.

3.3 Document-Level Classification

Documents provide a complementary view for the social scientist's research. Like conversations, documents can encode sequences of ideas or temporal events, as well as interpersonal relationships; these were not present in isolated utterances. Unlike the dyadic communication of a conversation, a document can be analyzed under a unified narrative (Piper et al., 2021). Thus for our purposes, a document is a collection of utterances that form a single *narrative*. Our document-level classification tasks cluster around *media*, which has been the subject of content analysis in the social sciences since the time of Max Weber in 1910. In this section, we focus on computational tools for content analysis (Berelson, 1952) to code media documents for their underlying *ideological* content (§3.3.3), the *events* they portray (§3.3.1,3.3.2), as well as the *agents* involved and the specific *roles* or character tropes they exhibit (§3.3.4).

3.3.1 Event Detection

Following a massive effort to digitize critical documents, social scientists depend on event extraction to automatically code and organize these documents into smaller and more manageable units for analysis. Events are the "building blocks" from which historians construct theories about the past (Sprugnoli and Tonelli, 2019); they are the backbone of narrative structure (Chambers and Jurafsky, 2008). Event detection is the first step in the event extraction pipeline. Hippocorpus (Sap et al., 2020b) is a resource of 6,854 stories that were collected from crowdworkers and tagged for sentence-level events (Sap et al., 2022) . Events can be further classified into minor or major events, as well as expected or unexpected. We evaluate on the simplest task: binary event classification at the sentence level.

3.3.2 Event Argument Extraction

Where event detection was concerned with identifying event triggers, event argument extraction is the task of filling out an event template according to a predefined ontology, identifying all related concepts like participants in the event, and parsing their roles. Historians, political scientists, and sociologists can use such tools to extract arguments from sociopolitical events in the news and historical text documents, and to understand social movements (Hürriyetoğlu et al., 2021). Economists can use event argument extraction to measure socioeconomic indicators like the unemployment rate, market volatility, and economic policy uncertainty (Min and Zhao, 2019). Event argument extraction is also a key feature of narrative analysis (Sims et al., 2019), as well as in the wider domains of legal studies (Shen et al., 2020), public health (Jenhani et al., 2016), and policy.

WikiEvents (Li et al., 2021) is a document-level event extraction benchmark for news articles that were linked from English Wikipedia articles. WikiEvents uses DARPA's KAIROS ontology with 67 event types in a three-level hierarchy. For example, the `Movement.Transportation` event has the agentless `Motion` subcategory and an agentive `Bringing` subcategory. Both include a `Passenger`, `Vehicle`, `Origin`, and `Destination` argument, but only the agentive `Bringing` has a `Transporter` agent. KAIROS's event argument ontology is richer and more versatile than the commonly used ACE ontology, which only has 33 types of events.

3.3.3 Ideology

CSS is extremely useful for understanding and quantifying real and perceived political differences. For a variety of specific phenomena (Amber et al., 2013; Baly et al., 2018; Roy and Goldwasser, 2020; Luo et al., 2020b; Ziems and Yang, 2021), this takes the form of gathering articles from across the political spectrum, processing each one further for a phenomenon of interest, and evaluating the relative differences for the articles from different ideological groups. The first step in such studies is to separate articles according to the overarching political ideology they represent.

We evaluate ideology detection on the Article Bias Corpus from Baly et al. (2020), which collects a set of articles from media sources covering the United States of America and labels them according to Left, Right, and Centrist political bias. Unlike the task of utterance-level ideology prediction (§3.1.6), this task provides an entire news article as context. This tests the ability of the model to understand the relationship that the stances taken across an entire article have with political bias. Each article must be classified into exactly one of the three ideological categories above.

3.3.4 Roles

Social roles are defined by expectations for behavior, based on social interaction patterns (Yang et al., 2019b). Similarly, personas are simplified models of personality (Grudin, 2006), like a trope that a character identifies within a movie. The ability to infer social roles and personas from text has immediate applications in the psychology of personality, the sociology of group dynamics, and the study of agents in literature and film. These insights can help us understand stereotypical biases and representational harms in media (Blodgett et al., 2020). Downstream applications also include narrative psychology (Murray et al., 2015), economics, political polarization, and mental health (Piper et al., 2021).

Others have considered character role labeling for narratives (Jahan et al., 2021) and news media (Gomez-Zara et al., 2018). We evaluate this task with the CMU Movie Corpus dataset from Bamman et al. (2013) as it was extended and modified by Chu et al. (2018) to include character trope labels and IMBD character quotes. The *character trope classification* task involves identifying from a character's quotes alone which of 72 movie tropes that characters identity best fits; e.g., the *absent-minded professor* or the *coward* or the *casanova*.

3.4 Generation Tasks

Regarding RQ4 **Functionality**, we want to understand whether LLMs are best suited to classify taxonomic social science constructs from text, or whether these models are equally if not better suited for generative explanations. This section describes our natural language generation tasks, where LLMs might be used to summarize the hidden social meaning behind a text (§3.4.1-3.4.3) or to implement social theory by stylistically restructuring an utterance (§3.4.4).

3.4.1 Figurative Language Explanation

FLUTE contains 9k (literal, figurative) sentence pairs with either entailed or contradictory meanings. The goal of the explanation task is to generate a sentence to explain the entailment or contradiction. For example, the figurative sentence "she absorbed the knowledge" entails the literal sentence "she mentally assimilated the knowledge" under the following explanation: "to absorb something is to take it in and make it part of yourself."

3.4.2 Implied Misinformation Explanation

Both scientific understanding and real-world intervention strategies depend on more than black-box classification. This motivates the implied statement generation task. Models take the headline of a news article and generate the underlying meaning of the headline in plain English. This is called the *writer's intent.* Consider, for example, the misleading headline, "*Wearing a face mask to slow the spread of COVID-19 could cause Legionnaires' disease.*" Here, the annotator wrote that the writer's intent was to say "*wearing masks is dangerous; people shouldn't wear masks.*"

3.4.3 Social Bias Inference

While hate speech detection focuses on the overall harmfulness of an utterance, specific types of hate speech are targeted towards a demographic subgroup. To this end, the Social Bias Inference Corpus (SBIC) (Sap et al., 2020a) consists of 34K inferences, where hate speech is annotated with free-text explanations. Importantly, explanations highlight *why* a specific subgroup is targeted. For example, the sentence *"We shouldn't lower our standards just to hire more women."* implies that *"women are less qualified."* To model these explanations, Sap et al. (2020a) treat the task as a standard conditional generation problem. We mirror this setup to evaluate LLMs.

3.4.4 Positive Reframing

NLP can help scale mental health and psychological counseling services by training volunteer listeners and teaching individuals the techniques of cognitive behavioral therapy (CBT; Rothbaum et al., 2000), which is used to address mental filters and biases that perpetuate anxiety and depression. Positive reframing is a sequence-to-sequence task which translates a distorted negative utterance into a complementary positive viewpoint without contradicting the original speaker meaning.

4 Evaluation Methods

4.1 Model Selection and Baselines

Our goal is to evaluate LLMs in **zero-shot settings through prompt engineering** (§4.2) and to identify suitable model architectures, sizes, and pretraining/fine-tuning paradigms for CSS research (RQ 1,2). We choose **FLAN-T5** (Chung et al., 2022) as an open-source model with strong zero-shot and few-shot performance. Although it follows a standard T5 encoder-decoder architecture, FLAN's zero-shot performance is due to its instruction fine-tuning over a diverse mixture of sequence to sequence tasks. The added benefit is that FLAN-T5 checkpoints exist at six different sizes ranging from small (80M parameters) to XXL (11B) and UL2 (20B), allowing us to investigate scaling laws. Next, we consider OpenAI's **GPT-3** (Brown et al., 2020; Zong and Krishnamachari, 2022) including `text-001`, `text-002` learning with instructions and `text-003`, which is further learned from human preferences (RLHF) (Christiano et al., 2017) series, and **ChatGPT** (Qin et al., 2023; Gilardi et al., 2023) which is the conversation-based LLM trained through RLHF (Christiano et al., 2017).

Traditional supervised fine-tuned models can serve as **baselines** for each task. These baselines are intended to provide a comparison point for the utility of LLMs for CSS, rather than providing a fair methodological comparison between approaches. For classification tasks, we use RoBERTa-large (Liu et al., 2019) as the backbone model and tune hyperparameters based on average accuracy on the validation set. For generation tasks, we use T5-base (Raffel et al., 2020) as the backbone model and tune hyperparameters based on average BLEU score on the validation set. We use a grid search to find the most suitable hyperparameters including learning rate `{5e-6, 1e-5, 2e-5, 5e-5}`, batch size `{4, 8, 16, 32}` and the number of epochs `{1, 2, 3, 4}`. Other hyperparameters are set to the defaults defined by the HuggingFace Trainer. We average results across three different random seeds to reduce variance.

4.2 Prompt Engineering

A key advantage over current LLMs is their ability to be "*programmed*" through natural language instructions (Brown et al., 2020). This capability has been further improved by training models to explicitly follow instructions provided in natural language (Sanh et al.; Wang et al., 2022; Chung et al., 2022; Ouyang et al., 2022). CSS tools can then be developed directly by subject-matter experts using natural language instructions rather than explicit programming language interpretations. In order to evaluate LLMs, each task requires a prompt designed to elicit the desired behavior from the model. However, as discussed in Perez et al. (2021), LLMs can have varied performance in response to prompts which are semantically quite similar. In practice, users prompt iteratively until the model behavior seems locally reasonable to them (Zamfirescu-Pereira et al., 2023). However, the lack of systematic procedures in this process makes it difficult to compare multiple LLMs on a broad suite of tasks, as it is unclear whether performance discrepancies stem intrinsically from the model or from the prompt engineering.

We evaluate the zero-shot performance and do not tailor prompts specifically for each model and task. Instead, when evaluating a task, the author who is familiar with it writes a prompt based on the task description. The same prompt was used across all models. This removes the confounding factor of prompt variation when comparing different LLMs and prevents data snooping via prompt engineering. While the use of a single set of instructions is common in recent broad LLM benchmarks (Liang et al., 2022; Kocoń et al., 2023; Qin et al., 2023), it does not capture instruction-based variance as we discussed further in §7.7 (Zhao et al., 2021).

Best Practices for CSS Prompt Design While no task-specific prompt engineering was done, we produced a set of best practices for CSS prompt design. CSS tasks often require models to make inferences about subtext and offensive language. Additionally, CSS codebooks often project complex phenomena into a reduced set of labels. This raises challenges for the use of LLMs which have been refined for general use. When initially exploring LLM behavior, we found that models would hedge in the case of uncertainty, refuse to engage with offensive language, and attempt to generalize beyond provided labels. While desirable in a general context, these behaviors make it difficult to use LLMs inside a CSS pipeline. Therefore, we built a set of best practices drawn from both the literature and our own experience with non-CSS tasks as NLP researchers. We explicitly share these best practices to help CSS practitioners control LLMs for their purposes. We list our guidelines for retrieving consistently-structured responses from LLMs in Table 1 alongside references to prior work on prompting when available.

Our Table 1 guidelines largely assure structured output for use of an LLM within a larger piece of software. This was necessary in order to score and evaluate many models on structured tasks, however they do not guarantee optimal performance of each model. While it is likely that we could have further improved performance for each LLM with further prompt-engineering, our true zero-shot

Effective Prompt Guideline	Reference	Guideline Example
When the answer is categorical, enumerate options as alphabetical **multiple-choice** so that the output is simply the highest-probability token ('A', 'B').	Hendrycks et al. (2021)	{$CONTEXT} Which of the following describes the above news headline? ⏎ **A:** Misinformation ⏎ **B:** Trustworthy ⏎ {$CONSTRAINT}
Each option should be separated by a newline (⏎) to resemble the natural format of online multiple choice questions. More natural prompts will elicit more regular behavior.	Inverse Scaling Prize	
To promote instruction-following, **give instructions *after* the context** is provided; then **explicitly state any constraints**. Recent and repeated text has a greater effect on LLM generations due to common attention patterns.	Child et al. (2019)	{$CONTEXT} **{$QUESTION}** **Constraint:** Even if you are uncertain, you **must pick either "True" or "False"** without using any other words.
Clarify the expected output in the case of uncertainty. Uncertain models may use default phrases like "*I don't know*," and clarifying constraints force the model to answer.	No Existing Reference	
When the answer should contain multiple pieces of information, **`request responses in JSON format`**. This leverages LLM's familiarity with code to provide an output structure that is more easily parsed.	MiniChain Library	{$CONTEXT} {$QUESTION} **`JSON Output:`**

Table 1: **LLM Prompting Best Practices** to generate consistent, machine-readable outputs for CSS tasks. These techniques can help solve overgeneralization problems on a constrained codebook, and they can force models to answer questions with inherent uncertainty or offensive language. See full example prompts in the Appendix.

process provides a fair comparison across all models. Additionally, it is a reasonable estimate of the performance of a prompt written by a non-AI expert using LLMs to build a CSS tool. However, further work is needed to understand the upper-bound prompted performance for each LLM with task-specific prompt engineering.

In order to receive consistent, reproducible results we utilize a temperature of zero for all LLMs. For models which provide probabilities directly, we constrain decoding to the valid output classes [2]. For other models, such as ChatGPT, we use logit bias to encourage valid outputs during decoding[3]. All other generation parameters are left at the default settings for each model.

4.3 Test Set Construction

For each task, we evaluate a class-stratified sample of at most 500 instances from the dataset's designated test set. If the designation is missing, we take the class-stratified sample from the entire dataset. Our sampled test sizes and class counts are in Table 8. All datasets, prompts, and model outputs are released for future comparison and analysis.[4]

[2]Probability outputs for HuggingFace and GPT-3

[3]Logit Bias reference for ChatGPT

[4]Data Directory of our Github Project

4.4 Evaluation Metrics

Automatic Evaluation Apart from the multi-label classification of Event Detection and the structured parsing task of Event Argument Extraction, all classification tasks are evaluated using accuracy. Since we mapped the label space for each task to an alphabetical list of candidate options and set the logit bias to favor these options (§4.2), evaluation scripts are straightforward string matching procedures. For Event Detection, we use F1 scores.

Human Evaluation For high-variation tasks like dialogue, word-overlap-based machine translation metrics like BLEU and ROUGE have low correlation with human quality judgments (Liu et al., 2016). For open-ended generation tasks in particular, embedding-similarity metrics like BERTScore are insufficient (Novikova et al., 2017) and human evaluation is strongly preferable (Santhanam and Shaikh, 2019). However, even human evaluations can exhibit high variance and instability due to cultural and individual differences (Peng et al., 1997). Pilot rounds revealed a high degree of variance and unpredictability in our evaluation, especially from crowdworkers (see Appendix A), and thus we opted to use expert annotations for generation results in this work. We discuss implications and solutions to CSS evaluation challenges in Section 7.4.

The authors opt to serve as expert annotators. Annotators are blinded to the corresponding models and evaluate only on the targets. Instead of scoring or rating target generations on a standard Likert scale, annotators rank these explanations in terms of their *accuracy* at describing the target construct. The ranking-style evaluation is more reliable and less variable than scoring for generation tasks (Harzing et al., 2009; Belz and Kow, 2010).

All ranking tasks follow the same format. For the Social Bias Frames explanation task, the annotator reviews a *hateful message* and an associated *hate target* (see Figure 8 in the Appendix). Then they review four *Implied Statements* generated by one of the OpenAI models or pulled from the SBIC's gold human annotations. They are asked to rank these statements from 1 (best) to 4 (worst) according to how accurate the *implied statement* is at describing the hidden message from the *hateful message*. In this forced-choice ranking scheme, ties are not allowed, but we use a unanimous vote to determine when a given model outranks human performance. Unanimous vote flattens the variance for explanations of similar quality and reflects only significant differences in quality. See Appendix A for more evaluation details.

5 Classification Results

Table 2 presents all zero-shot results for utterance, conversation, and document-level classification tasks. We use these results to answer Research Questions 1-3. The results suggest that LLMs are a **viable tool for augmenting human efforts in CSS**. For classification tasks specifically, results show that **larger, instruction-tuned open-source LLMs like FLAN-UL2 are preferable.**

5.1 Viability (RQ1)

There are two related questions regarding the viability of LLMs as CSS tools. We first ask whether prompted models perform well enough to directly label text out-of-the-box. The answer is, at best, a contingent *yes,* **LLMs may be ready for research-grade zero-shot classification for some tasks**. Still, carefully fine-tuned models outperform prompted models in 7/10 utterance-level tasks and ∼50% of conversation/document level tasks. Overall, we recommend human-in-the-loop methods to mitigate bias and risk (§7.6, 7.7), and we encourage readers to proceed cautiously.

LLMs achieve the lowest absolute performance on *Event Argument Extraction, Character Tropes, Implicit Hate, and Empathy Classification* with below 40% accuracy (Table 2). These tasks are either structurally complex (event arguments), or have subjective expert taxonomies whose semantics differ from definitions learned in LLM pretraining (tropes, hate, empathy). This may explain our error analysis in Figure 2 where ChatGPT often defaults to the neutral, more colloquially recognizable label *stereotype* (64% of errors) rather than use a more taxonomy-specific label like *white grievance* (for details on the error analysis, see Appendix B). Few-shot prompting might help address the misalignment between model and ground-truth definitions.

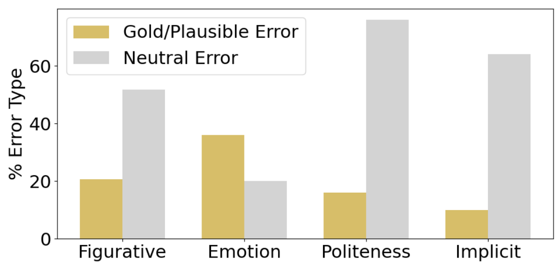


Figure 2: **Breakdown of Error Types in ChatGPT.** Plausible/gold errors occur when gold labels are incorrect or the model identifies a valid secondary label. Neutral errors occur when a model over-predicts a category in a respective task (*metaphor* in Figurative; *surprise* in Emotion; *neutral* in Politeness, and *stereotypical* in Implicit)

On the other hand, LLMs achieve the highest absolute performance on *Misinformation, Stance,* and *Emotion Classification* with above 70% accuracy (bolded in Table 3). These tasks either have objective ground truth (fact checking for misinformation) or have labels with explicit colloquial definitions in the pretraining data (emotional categories like *anger* are part of everyday vernacular; political stances are well-documented and explicit in online forums). Here, models are less likely to default to neutral categories, and errors are more likely to come from annotation mistakes in the gold dataset (see lower neutral and higher gold error in *Emotion* classification in Figure 2).

In the most exceptional best cases, LLMs match or even exceed our reported baselines. In some lower-stakes or aggregate population analyses, 70% may be a sufficient threshold for direct use in downstream analyses.[5] In such scenarios, zero-shot prompted models could replace fine-tuned models and thus remove the need for expensive train-

[5] For example, Prabhakaran et al. (2014) trained a power relations classifier with 70.74% accuracy. The leading submission to the SemEval-2016 shared task (Mohammad et al., 2016), MITRE, achieved 67.82 F1 (Zarrella and Marsh, 2016). MITRE and other models with similar accuracies have been applied to downstream CSS applications like COVID-19 vaccination opinions (Cotfas et al., 2021), political opinions (Siddiqua et al., 2019), and debates (Lai et al., 2020).

Model / Data	Baselines		FLAN-T5					FLAN	Chat	text-001				text-002	text-003
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	ChatGPT	Ada	Babb.	Curie	Dav.	Davinci	Davinci
Utterance Level Tasks															
Dialect	4.5	41.5	1.9	2.3	15.8	16.5	22.6	23.7	15.0	5.3	5.6	6.0	10.9	10.5	16.9
Emotion	16.7	91.7	23.9	65.3	69.1	65.9	66.7	70.3	46.2	44.6	16.1	18.7	19.3	39.8	36.5
Figurative	25.0	94.4	23.6	29.0	25.4	40.2	56.0	64.0	50.2	25.0	24.4	25.0	28.8	52.0	60.6
Humor	50.0	73.1	52.0	51.8	56.2	59.0	50.6	58.8	55.4	55.2	59.0	58.6	50.4	51.4	51.0
Ideology	33.3	61.9	33.1	39.2	48.6	49.2	54.4	48.2	54.8	–	33.3	33.3	34.3	57.6	48.2
Impl. Hate	14.3	69.9	17.7	22.7	17.9	36.3	34.5	35.9	29.7	17.1	18.6	15.7	21.3	22.7	27.1
Misinfo	50.0	82.3	50.0	55.4	69.2	70.2	71.2	77.6	69.0	–	50.4	52.2	52.6	75.6	75.0
Persuasion	12.5	40.4	14.3	19.8	43.9	43.4	†51.6	49.4	40.9	–	16.5	17.0	18.8	26.3	26.3
Sem. Chng.	50.0	65.7	50.3	50.0	†66.9	55.5	51.2	53.7	56.1	50.0	50.5	54.3	39.5	45.9	50.0
Stance	33.3	47.0	34.7	47.8	51.3	52.6	55.9	55.4	†72.0	–	33.1	31.0	48.0	57.4	41.3
Conversation Level Tasks															
Discourse	14.3	47.5	14.7	26.4	37.2	44.3	†52.5	41.9	44.5	13.1	16.5	14.3	17.0	39.8	37.8
Empathy	33.3	33.3	33.3	33.3	35.1	33.7	36.8	†39.8	37.6	–	33.1	35.3	33.3	33.3	33.3
Persuasion	50.0	50.0	48.4	55.3	†57.1	53.0	53.5	53.2	52.9	50.2	50.0	50.0	50.0	50.8	55.9
Politeness	33.3	75.9	33.9	44.2	53.0	59.2	54.2	52.8	50.8	33.1	33.1	32.1	42.2	55.6	47.8
Power	50.0	74.0	47.6	47.2	50.4	56.8	58.8	60.8	61.6	–	52.2	50.6	49.6	50.5	57.0
Toxicity	50.0	64.6	46.8	50.6	49.4	54.2	50.0	56.6	53.0	44.6	50.6	49.0	50.8	52.2	51.2
Document Level Tasks															
Event Arg.*	–	59.4	–	–	–	–	-	–	22.3	–	–	8.6	8.6	21.6	22.9
Event Det.*	–	75.8	9.8	7.0	1.0	10.9	41.8	50.6	51.3	29.8	47.3	47.4	44.4	48.8	52.4
Ideology	33.3	51.0	33.1	34.1	34.1	32.1	49.6	40.3	58.8	32.9	35.1	33.6	25.6	48.7	44.0
Tropes	1.4	0.8	0.9	4.4	8.8	7.9	10.5	16.7	25.4	4.3	7.0	9.6	10.5	18.4	18.4

Table 2: **Zero-shot Classification Results** across our selected CSS benchmark tasks. All tasks are evaluated with accuracy, except for Event Arg. and Event Detection, which use F-1. Models which did not always follow instructions are marked with a dash. Best zero-shot models are in green ; zero-shot models that are not significantly worse ($P > .05$; Paired Bootstrap test (Dror et al., 2018)) are marked blue ; and † denote cases where zero-shot LLMs match or beat finetuned baselines.

ing datasets. Humans could focus their efforts on validating LLM outputs and tuning prompts (§4.2) rather than coding unstructured text. Still, high-risk and sensitive domains like misinformation and hate speech detection will demand higher performances. Practitioners should consider the advantages of using zero-shot prompted LLMs to replace human coding against the risks of producing incorrect labels. Our results can help researchers understand this boundary and guide the decision-making process across a broad range of common CSS tasks.

Next, we consider a less aggressive shift in methodology: *Can LLMs augment the human annotation process?* According to this paradigm, an LLM could serve as just one of many human and AI labelers, and gold labels would be decided by majority vote across these independent labels. The validity of this paradigm depends on the expected agreement between humans and prompted models (Chaganty et al., 2018). We report Fleiss' κ agreement in Table 3 and find that, for a substantial subset of tasks (6/17 = 35.3%), models achieve moderate to good agreement, ranging from $\kappa = 0.42$ to 0.64. For another 6 tasks, we see fair agreement. Only 5/17 = 29.5% of tasks have poor agreement where social scientists might not consider annotation augmentation via LLMs. We conclude that **CSS researchers should strongly consider the augmented annotator paradigm** discussed above for analysis of utterances, conversations, or documents. See §7 for further discussion.

5.2 Model-Selection (RQ2)

CSS researchers should understand how their choice of model can decide the reliability of their method. Our results show that, for structured parsing tasks like event extraction, OpenAI's `text-davinci-003` code-instructed model is ideal, while for most classification tasks, open-source LLMs like FLAN-UL2 are best.

Model Size. LLMs generally follow scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) where performance increases with the size of the model and training data. We investigate scaling laws in the two families of instruction-tuned LLMs: FLAN and OpenAI. Results show larger FLAN models are preferable.

Dataset	Best Model	Acc.	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	23.7	0.15	poor
Emotion	flan-ul2	**70.3**	0.64	good
Figurative	flan-ul2	64.0	0.52	moderate
Humor	flan-t5-xl	59.0	0.16	poor
Ideology	davinci-002	57.6	0.36	fair
Impl. Hate	flan-ul2	36.3	0.23	fair
Misinfo	flan-ul2	**77.6**	0.55	moderate
Persuasion	flan-t5-xxl	51.6	0.42	moderate
Semantic Chng.	flan-t5-large	66.9	0.34	fair
Stance	chatgpt	**72.0**	0.58	moderate
Convo-Level				
Discourse	flan-t5-xxl	52.5	0.44	moderate
Empathy	flan-ul2	39.8	0.04	poor
Persuasion	flan-t5-large	57.1	0.13	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	chatgpt	61.6	0.23	fair
Toxicity	flan-ul2	56.6	0.01	poor
Document-Level				
Ideology	chatgpt	58.8	0.36	fair

Table 3: *(Acc.)* **Best model accuracy.** Accuracies above 70% are bolded as high enough for possible downstream use. *(κ)* **Agreement scores between zero-shot model classification and human gold labels.** Out of ten utterance-level tasks, five have at least moderate M and only two have poor agreement P. Three (50%) of the conversation tasks have at least fair agreement F, as does the document-level task.

FLAN's CSS task performance roughly matches Kaplan et al.'s predicted power-law effects from pure model size. Figure 3 shows FLAN classification performances scaling nearly logarithmically with the parameter count. All FLAN-T5 models use the same stable corpus, pretraining objective, and architecture, which gives us a controlled environment to observe stable scaling laws.

OpenAI's GPT-3 `001` models, on the other hand, do not monotonically benefit from scaling.[6] Although performance improves on the lower end of model scale (from ada to babbage), there is minimal performance improvement from babbage to `davinci`, despite a size increase of two orders of magnitude. Instead, the largest performance improvements come from variations in *pretraining*, *fine-tuning*, and *reinforcement learning*.

[6]This analysis relies on estimates which combine community estimates, the OpenAI research documentation, and the assumption that all models named or "improved" from `davinci` have the same parameter counts. These estimates may be incorrect, as hypothesized by other community estimates. This is a limitation of research on these models as exact model size and training data are a trade secret of OpenAI.

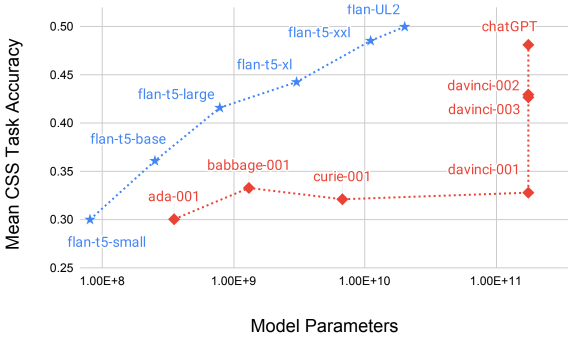


Figure 3: **Effects of Scaling** on the mean performance on our CSS benchmark tasks. FLAN models and `davinci-001/002` are instruction fine-tuned. `davinci-003` and ChatGPT are instruction fine-tuned and refined with Reinforcement Learning from Human Feedback. GPT Parameter counts reported based on approximates[7].

Pretraining & Instruction Fine-tuning. Besides scale, two key factors play a major role in model performance: *pretraining data* and *instruction fine-tuning*. Pretraining data is the raw text upon which an LLM learns to model the general generative process of language. Instruction fine-tuning refines the raw LLM to perform specific tasks based on human-written instructions.

OpenAI's `davinci` models significantly benefit from pretraining and instruction fine-tuning tricks. For classification tasks (Table 2, we see an outsized increase in CSS performance (↑ absolute 10 pct. pts.) moving from `davinci-001` to `davinci-002`, larger than any performance increase from scale alone. Both `davinci-001` and `davinci-002` use the same supervised instruction fine-tuning strategy, but `davinci-002` is based on OpenAI's base-code model, which had access to a larger set of instruction fine-tuning data. Most importantly, `davinci-002` was pre-trained on both text and code. This difference clearly benefits structured tasks like Event Argument Extraction with its JSON-formatted outputs. While `davinci-001` often fails to generate JSON, `davinci-002` succeeds with markedly improved performance (+13.0 F1).

Learning From Human Feedback. We see that *RLHF does not systematically improve LLM performance on CSS classification tasks.* Although RLHF has been lauded as the major catalyst behind the success of instruction-following models (Ouyang et al., 2022), we do not see uniform performance benefits with our selected tasks. Instead, we see equivalent mean task performances from `text-davinci-002` without RLHF and `003` with RLHF. ChatGPT achieves significantly better classification performance than `text-davinci-003`, but the causal factor is not clear. ChatGPT's training details have not been fully disclosed, and this is a limitation of research on OpenAI models.

5.3 Domain-Utility (RQ3)

The survey and taxonomy of social science need in Section 2 allows us to understand whether the utility of LLMs is limited to certain domains or certain data types. To do so, we partition all classification results from Table 2 into bins corresponding to the academic field most impacted by the task.[8] Although we recognize the multi-disciplinary utility of *all* tasks, this type of 1:1 organization is appropriate for understanding the academic scope of our results. We acknowledge that the partitioning and selection of the dataset influence the performance distributions that we observe. We urge readers to interpret the results with caution and focus on broader conclusions rather than the fine numerical details of these distributions.

The box plot in Figure 4 shows that field-specific performances significantly overlap. Thus overall, **we do not observe a strong bias against or proclivity for a particular field of study.** In political science, we see the highest overall performance on misinformation detection (77.6%) and much lower performance on ideology (51%) and implicit hate detection (36.3%). For historical and literary analysis, we observe high performance on event detection (52.4%) and low performance on both event argument extraction (22.9%) and character trope classification (25.4%). High and low peformances span the full range of disciplines. This suggests that performance is not tied to academic discipline.

In terms of data type, Figure 5 suggests that **performance may be more closely determined by the complexity of the input**. In particular, documents encode complex sequences of ideas or temporal events, and overall, the two lowest task performances are on the document-level tasks: character trope classification and event argument extraction. All other document-level accuracies are at or below 50%. The most challenging utterance and conversation-level tasks are also a function of their label space complexity. Implicit hate (36.3%), empathy (39.8%), and dialect feature (41.5%) annotations are expert-labeled on a subtle, theoretical taxonomy.

[8]This partitioning follows Figure 1, with stance and ideology detection in the *political science* bin and dialect feature classification under *linguistics*, for example.

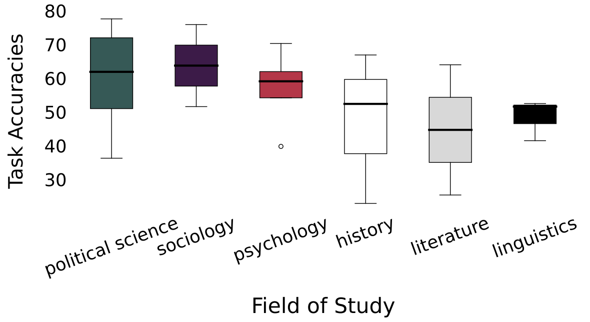


Figure 4: **Task Performance By Field of Study**. Significant overlap in the distributions suggests that neither high nor low performance is exclusive to any particular discipline. Caution: The distributions depend on the particular choices of this study, which datasets to select and how to partition them.

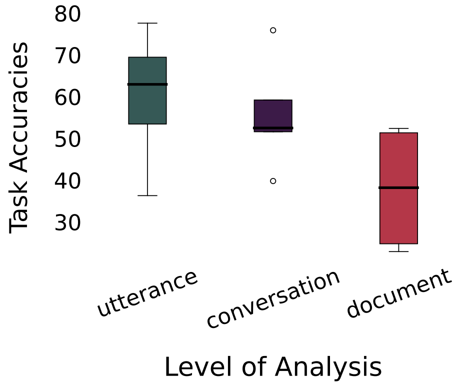


Figure 5: **Task Performance By Level of Analysis**. Document-level tasks are challenging for their input length and complexity, and this is reflected in their accuracies near or below 50%. Utterance and conversation-level task performance varies also with the complexity of the task.

6 Generation Results

In this section, we answer *RQ4: Are prompted LLMs useful for generatively implementing theories and explaining social scientific constructs with text?* Will generative models replace or augment human analysis? To answer this question, we rely on the human evaluation setup described in Section 4.4. Results are in Table 4, with evaluator preferences for misinformation (MRF), figurative language (FLUTE), and hate speech (SBIC) explanations in the middle three columns. The right column displays preferences for the positive psychology reframing task. Note that FLAN models are excluded from this table because FLAN models failed to follow instructions by manual inspection.

We found that **prompted LLMs produce helpful and informative generations in all four evaluation tasks.** Model generations outrank the dataset's gold human reference at least 38% of the time. The best models approach parity with humans where it is a near 50-50 coin toss to decide which is preferred. Furthermore, we see significant performance benefits from both RLHF models, ChatGPT and `text-davinci-003`. Unlike classification (§5.2), our selected generation tasks seem to systematically benefit from human feedback.

Despite strong performances, no model substantially outperforms human annotation. This suggests that current **LLMs cannot replace human analysis**. Still, **LLMs can powerfully augment the analytical pipeline and reduce human coders' cognitive load**. Instead of coding text with summary explanations from scratch, researchers and annotators could apply minor edits to correct model generations. [9] The results in Table 4 suggest that, for every five model generations, 2 to 3 of these outputs will demand no additional annotator effort, thus significantly increasing the efficiency of the social scientist's research pipeline.

As a tradeoff for LLM's efficiency, **researchers will face the burden of manually validating generative outputs.** It is well-known that automatic performance metrics fail to capture human preferences (Goyal et al., 2022; Liang et al., 2022). In fact, we found that BLEU (Post, 2018), BERTScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020) that rely on comparisons to human written groundtruth all produced largely uninterpretable scores for generation tasks (see Table 7 in the Appendix). This highlights a fundamental challenge for evaluation of generation systems in CSS, especially if zero-shot performance continues to improve. As zero-shot models approach or outperform the quality of the gold-reference generations, reference-based scoring becomes an *invalid construct* for measuring models' true utility (Raji et al., 2021), even if we assume the semantic similarity metrics are ideal. This motivates our use of reference-free expert evaluation of generations, that is, asking expert annotators which generation is more accurate with regard to the input or preferable. However, this alternative is limited by both cost and reproducibility concerns (Karpinska et al., 2021). There is a clear need for new metrics and procedures to quantify model utility for CSS.

[9]Note that is that machine generated explanations might be limited in terms of their diversity. Although human validation can help refine these machine outputs, such process may not be able to introduce novel edits or perspectives.

Model	% Preferred Over Human Gold Annotations MRF	FLUTE	SBIC	Reframing
Baseline	31.2%	4.6%	16.5%	45.0%
`text-ada-001`	17.6%	1.7%	11.8%	0.0%
`text-babbage-001`	29.4%	6.7%	29.4%	0.0%
`text-curie-001`	29.4%	1.7%	32.4%	11.5%
`text-davinci-001`	21.4%	6.2%	43.9%	30.4%
`text-davinci-002`	21.4%	25.0%	29.3%	10.0%
`text-davinci-003`	38.9%	47.0%	50.0%	48.5%
ChatGPT	27.8%	37.9%	65.9%	56.1%

Table 4: **Expert Human Evaluations for Zero-shot Generation Tasks** give the proportion of all pairwise rankings where authors unanimously ranked the model's generation as more accurate or preferable to a gold-standard explanation drawn from the dataset. Best models are in green and runner-ups are in blue.

7 Discussion

This work presents a comprehensive evaluation of LLMs on a representative suite of CSS tasks. We contribute a robust evaluation pipeline, which allows us to benchmark performance alongside supervised baselines on a wide range of tasks. Our research questions and empirical results are designed to help CSS researchers make decisions about when LLMs are suitable and which models are best suited for different research needs. In summary, we find that **LLMs can radically augment but not entirely replace the traditional CSS research pipeline.**

More concretely, we make the following **recommendations to CSS researchers**:

1. Integrate LLMs-in-the-loop to transform large-scale data labeling.
2. Prioritize open-source LLMs for classification and OpenAI LLMs for generation.
3. Investigate how LLMs produce new CSS paradigms built on the multipurpose capabilities of LLMs in the long term.

7.1 How Can LLMs Transform Annotation

Our work shows that current LLMs can increase the efficiency of data annotation. The human-AI agreement results in Section 5.1 show that augmenting annotation with an LLM annotator yields moderate or better agreement on 12 out of 17 tasks. However, *LLMs are not a wholesale replacement for human annotators*. Even the best LLMs exhibit unusably low performance on CSS tasks. Ensembling prediction does not mitigate this label corruption as LLMs demonstrate high internal agreement, even when

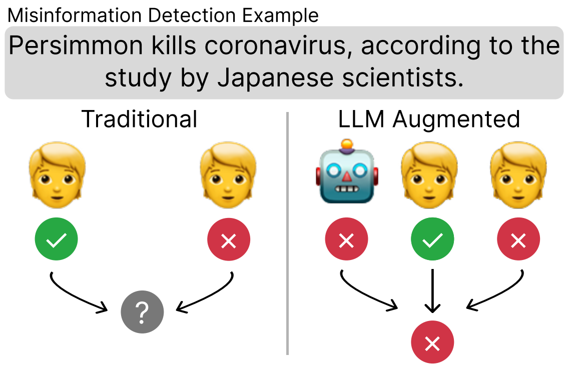


Figure 6: **Human-AI Collaboration** can improve the efficiency and reliability of text analysis. In this misinformation example, the LLM helps scale up annotation while reducing variance in the gold labels. Human annotation serves as validation for model-provided annotations.

inaccurate (Gilardi et al., 2023). Overconfident models, if left unchecked, distort the conclusions of CSS research and subsequently mislead policy and social actions taken in response. Human validation is key to avoiding a replication crisis in CSS caused by LLM hallucinations and inaccuracies.

Instead, *we advocate that CSS researchers integrate LLMs with annotation*, as illustrated in Figure 6. Even in the commonly used Majority Vote annotation scheme (Snow et al., 2008; Potts et al., 2021), an LLM can be used as one of multiple annotators to annotate the same amount of data with between 50% and 33% less human labor for a binary task. For tasks with rich label spaces, researchers can construct a reliable and unbiased gold-labeling system by averaging the differences between human and LLM labels, with prior estimates placing savings at between 7-13% for real-valued scoring (Chaganty et al., 2018).

Moving forward, LLMs can serve as a *flywheel* for dataset collection. Prompted LLMs consistently perform significantly better than chance, providing imperfect labels at low cost. Annotation schemes developed to iteratively improve imperfect data —such as weak supervision (Ratner et al., 2017), targeted data cleaning (Chen et al., 2022), and active learning (Yuan et al., 2020; Li et al., 2022) — avoid LLM pitfalls by allowing human validation to refine the original model. This creates a virtuous cycle which exploits the strengths of LLMs to focus human expertise where it is most needed (Kiela et al., 2021).

Our results show that *LLMs are even more likely to transform annotation for generation tasks*, being rated superior to human gold annotations over 38%

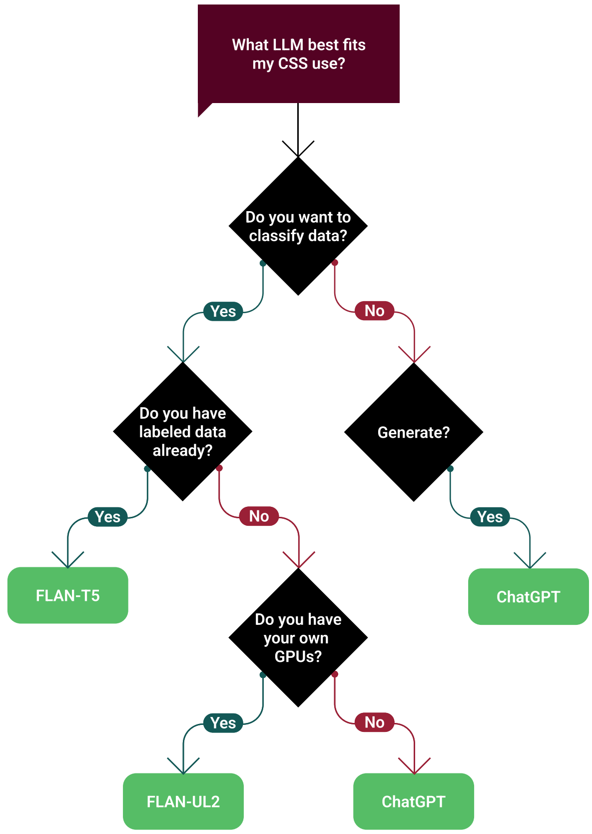


Figure 7: Our high-level model recommendations for CSS researchers looking to utilize LLMs based on our benchmark results, cost of operation, and ease of model adaptability.

of the time in all 4 tasks we evaluate. LLMs can already generate syntactically cohesive and stylistically consistent text. Humans expertise can be used to curate outputs according to accuracy, relevance, and quality. Dataset construction through human curation of LLM generations has already emerged in recent NLP works on decision explanation (Wiegreffe et al., 2022), model error identification (Ribeiro and Lundberg, 2022), and even to build the figurative language benchmark used in this work (Chakrabarty et al., 2022).

We recommend that CSS researchers make use of LLMs as the foundation of such annotation procedures to improve annotation efficiency. CSS researchers should reinvest savings from improved efficiency to train **expert annotators**, reversing the trends of replacing experts with crowdworkers due to cost (Snow et al., 2008). By doing so, LLMs can enable data labeling procedures which more deeply benefit from the non-computational expertise of the social scientists whose theories we build upon.

7.2 When To Use What LLM

We hope these results help CSS researchers to understand LLM alternatives for their use cases. Our general prompt guidelines allow us to quickly design functional prompts for many models. When looking to incorporate LLMs in their work, CSS researchers should consider the advantages and disadvantages of open source and industrial models. As a quick reference, we provide high level recommendations in Figure 7.

For CSS classification, our work shows that open-source models like FLAN are as capable as state-of-the-art industrial LLMs from OpenAI. We recommend researchers who already have access to GPUs capable of running these models prefer FLAN models. For continuous monitoring and enormous-scale analysis, the low marginal cost of these open-source models makes them price-advantageous. For CSS researchers with expertise, open-source LLMs have the added benefit of being able to be fine-tuned on labeled data and constrained programatically for more predictable behaviour. At this time, it is not possible to further fine-tune OpenAI's instruction-tuned models[10].

However, for those without existing hardware infrastructure, OpenAI models are an extremely cost-efficient option. Based on current cloud pricing[11], the hardware necessary to run FLAN-T5-XXL costs 170 dollars per day—the equivalent of processing roughly 50 million words using ChatGPT[12]. In most cases, ChatGPT is more cost-efficient and has a lower operational overhead for hardware-constrained research groups.

For generation tasks, the results are clear-cut. Even the largest open-source models failed to generate meaningful responses for CSS tasks. Even when labeled data is available, the best *OpenAI models outperform fine-tuned baselines consistently* and approach parity with gold human annotations when evaluated by crowdworkers. For CSS experts looking to generate interpretations or explanations of data, ChatGPT is the clear leading LLM by both price and performance. No matter which modeling decision is made, practitioners should keep the limitations of natural language generation in mind, understanding that explanations are not causal and recognizing the risks that come with model errors and hallucinations (see §7.7).

[10] As of March 30th from the OpenAI documentation.
[11] Google Cloud FLAN hosting cost
[12] OpenAI Pricing

Our work shows that **all LLMs struggle to a greater degree with conversational and full document data**. Moreover, **LLMs currently lack clear cross-document reasoning capabilities**, limiting common CSS applications like topic modeling. For CSS subfields that study these discourse types—sociology, literature, and psychology—LLMs have major limitations and are unlikely to have major immediate impact. NLP researchers who aim to improve existing LLMs to empower more CSS tasks should study the unique technical challenges of conversations, long documents, and cross-document reasoning (Beltagy et al., 2020; Caciularu et al., 2021; Yu et al., 2021).

7.3 Blending CSS Paradigms

The few-shot (Brown et al., 2020) and zero-shot capabilities (Ouyang et al., 2022) of LLMs **blur the traditional line between supervised and unsupervised ML methods for the social sciences**. Historically, supervised methods invest in labeled data guided by existing theory to develop a trained model. This model is then used to classify text at scale to gather evidence for the causal effects surrounding the theory. By comparison, unsupervised methods like topic modeling often condense large amounts of information to help researchers discover new relationships, which develop or refine social theories (Evans and Aceves, 2016).

The ability of LLMs to follow instructions and interpret complex tasks is rapidly advancing, with major new models even within the course of this work (OpenAI, 2023). Beyond annotation, LLMs have multi-purpose capabilities to retrieve, label, and condense relevant information at scale. We believe that this can blend the boundaries between supervised and unsupervised paradigms. Rather than using separate paradigms to develop and test theories, a single tool can be used to develop working hypotheses, using generated and summarized data, and test hypotheses, labeling human samples flexibly with low-cost classification capabilities. We believe CSS researchers should use the multi-functionality of LLMs to create new paradigms of research for their fields.

Simulation. Simulation is an early area of example of such innovation in CSS is the use of LLMs as simulated sample populations. Game theorists have used rule-based utility functions to develop hypotheses about the causes of social phenomena (Schelling, 1971; Easley and Kleinberg, 2010) and to predict the effects of policy changes (Shubik, 1982; Kleinberg et al., 2018). However, simulations are limited by the expressiveness of utility functions (Ellsberg, 1961; Machina, 1987). LLMs hold a great potential to provide more powerful simulations, as they replicate human biases without explicit conditioning (Jones and Steinhardt, 2022; Koralus and Wang-Maścianica, 2023). Recently, this capacity of LLMs has been leveraged to simulate social computing systems (Park et al., 2022), community and their members' interactions (Park et al., 2023), public opinion (Argyle et al., 2022; Chu et al., 2023), and subjective experience description (Argyle et al., 2022).

However, there are *dangers and uncertainties* in this area as noted in these works. Since social systems evolve unpredictably (Salganik et al., 2006), simulated samples inherently have limited predictive and explanatory power. While utility-based simulations have similar limitations, their assumptions are explicit unlike the opaque model of human behaviour an LLM provides. Additionally, current models exhibit higher homogeneity of opinions than humans (Argyle et al., 2022; Santurkar et al., 2023). Combining LLMs with true human samples is essential to avoid an algorithmic *monoculture* and could lead to fragile findings covering only the limited perspectives represented (Kleinberg and Raghavan, 2021; Bommasani et al., 2022).

7.4 The Need for A New Evaluation Paradigm

Evaluation will need to adapt if blended methods create a new CSS paradigm. Accuracy-based metrics were ideal for fixed-taxonomy classification tasks in the era of NLP benchmarking. Similarly, word-overlap metrics made sense for natural language generation tasks in which the gold reference was well-defined (e.g., translation). However, open-ended coding and CSS explanation objectives follow neither a pre-defined taxonomy nor a regular output template. For more open-ended data exploration tasks like topic modeling, held-out likelihood helped automatically measure the predictive power of the model (Wallach et al., 2009), but predictiveness does not always correlate with explainability (Shmueli et al., 2010), and these automatic metrics proved to be at odds with human quality evaluations (Chang et al., 2009). In CSS, human evaluations can be unreliable (Karpinska et al., 2021). We observe this directly in our work, as crowd work seems to provide unreliable for FLUTE, a nuanced generative task. New metrics are needed to capture the semantic validity of free-form coding with LLMs as explanation-generators.

7.5 CSS Challenges for LLMs

As shown by our Section 5 results, LLMs face notable challenges that pervade the computational social sciences. The first challenge comes from the subtle and non-conventional language of **expert taxonomies**. Expert taxonomies contain technical terms like the dialect feature *copula omission* (§3.1.1), plus specialized or nonstandard definitions of colloquial terms, like the persuasive *scarcity* strategy (§3.1.8), or *white grievance* in implicit hate (§3.1.4). LLMs may lack sufficient representations for such technical terms, as they may be absent from the pretraining data (Yao et al., 2021). How to *teach* LLMs to understand these social constructs deserves further technical attention. This is especially true for *novel theoretical constructs* that social scientists may wish to define and study in collaboration with LLMs.

Unlike widely used NLP classification tasks, the challenge of expert taxonomies in CSS is compounded by the **size of the target label space**, which, in CSS applications, may contain upwards of 72 classes (see *character tropes*, §3.3.4). This challenges transformer-based LLMs, which have relatively limited memory, finite processing windows, and quadratic space complexity.

Large, complex, and nuanced annotation schemes may also introduce dependencies among labels that are organized into multi-level hierarchies or richly constrained schemas, as in many *event argument extraction* applications. Such complex **structural parsing** tasks pose special challenges to the zero-shot prompting paradigm introduced in this work since prompted models often struggle to generate **consistent outputs** (Mishra et al., 2019). Our prompting best practices in Table 1 all help LLMs generate more consistent machine-readable outputs, but this challenge is not fully solved for all CSS tasks.

Finally, Computational Social Scientists study language, norms, beliefs, and political structures that all *change across time*. To account for these distribution shifts, LLMs will need an extremely high level of **temporal grounding**—knowledge and signals by which to orient a text analysis in a particular place and time (Bommasani et al., 2021).

This is especially challenging wherever researchers are interested in **rapid, synchronous analysis of breaking events**. It may be prohibitively expensive to frequently update LLM's knowledge of current events via continually training (Bender et al., 2021), and this challenge will only be exacerbated as models continue to scale up.

7.6 Issues in Bias and Fairness

Researchers should weigh the benefits of applying prompting methods to CSS, along with the limitations and risks of doing so. Most notably, LLMs are known to amplify social biases and stereotypes (Sheng et al., 2021; Abid et al., 2021; Borchers et al., 2022; Lucy and Bamman, 2021; Shaikh et al., 2022). These biases can emerge in open-ended generation tasks like the explanation and paraphrasing (Dhamala et al., 2021). The performance of LLMs as tools for classification and parsing may vary systematically as a function of demographic variation in the target population (Zhao et al., 2018). With the datasets available, we were unable to perform a systematic analysis of biases and performance discrepancies, but we urge researchers to carefully consider these risks in downstream applications.

Social science research is often described as overreliant on Western, Educated, Industrial, Rich, and Democratic populations (WEIRD; Muthukrishna et al., 2020), and this is true of CSS as well, where data resources are abundant in English-speaking Western contexts (Ignatow and Mihalcea, 2016). It is again a limitation of current data resources that prevents us from exploring cross-cultural or cross-lingual CSS in this work, and we acknowledge this as a severe limitation in the field.

7.7 Limitations

Task Selection. Our tasks do not represent an exhaustive list of all application domains. Some highly-sensitive domains like mental health (Nguyen et al., 2022), which requires expert annotations, and cultural studies, which requires community-specific knowledge, are rife with additional challenges and ethical concerns. These are largely outside the scope of the current study. More broadly, LLMs should not be used to give legal or medical advice, prescribe or diagnose illness, or interfere with democratic processes (Solaiman and Dennison, 2021).

Evaluation and Prompting. When evaluating LLMs, one notable concern is data leakage. Data from the test set might have been seen by LLMs during the pre-training, and this would artificially inflate test performances. One mitigation strategy is to design explicit prompts that force the model to forget the test set. Another strategy is to design custom test sets from perturbations of existing data to more fairly evaluate models. We leave this for future work. Furthermore, different LLMs might benefit from different types of prompts for different tasks, but in this work, we only utilize a single unified prompt for each task. The performance of LLMs can be highly sensitive to prompt engineering (Zhao et al., 2021), and thus prompt variation can impact downstream tasks. In future work, we will test on prompt perturbations or ensemble prompt variations for more robust results. We can also achieve additional performance gains by prompting models iteratively across multiple rounds of refinement as in Wei et al. (2023).

Finally, we focus on zero-shot learning. Performances might be significantly improved with few-shot in-context learning. In future work, we will identify the number of examples needed to outperform traditional fully-supervised learning.

Causality and Explanations. Explanations are important to social science (Shmueli et al., 2010; Hofman et al., 2017; Yarkoni and Westfall, 2017). In this work, we explored the predictive power of LLMs rather than causal explanations. Predictions serve to expose and elaborate on the underlying social phenomena latent in a text. These explicit phenomena can then be used as structured features for further analysis with causal methods.

However, this may not be sufficient: social scientists often seek causal theories (DiMaggio, 2015), or at least *contrastive* explanations, *why P instead of Q* (Miller, 2019). Because LLMs are not grounded in a causal model of the world (Bender et al., 2021), they are not on their own reliable tools for mining causal relationships in text. We leave it to future work to explore contrastive or causal explanations in LLMs.

Acknowledgements

We are thankful to Tony Wang, Rishi Bommasani, Albert Lu, Myra Cheng, Yanzhe Zhang, Camille Harris, and Minzhi Li for their helpful feedback on the draft. Caleb Ziems is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655.

References

Andrew Abbott. 1990. Conceptions of time and events in social science methods: Causal and narrative approaches. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 23(4):140–150.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Deena Abul-Fottouh and Tina Fetner. 2018. Solidarity or schism: ideological congruence and the twitter networks of egyptian activists. *Mobilization: An International Quarterly*, 23(1):23–44.

Amr Ahmed and Eric Xing. 2010. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1140–1150, Cambridge, MA. Association for Computational Linguistics.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 12–21.

Sahar Altikriti. 2016. Persuasive speech acts in barack obama's inaugural speeches (2009, 2013) and the last state of the union address (2016). *International Journal of Linguistics*, 8(2):47–66.

Boydstun Amber, Gross Justin, Philip Resnik, and Noah Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *New directions in analyzing text as Data workshop*.

Cameron Anderson, Oliver P John, Dacher Keltner, and Ann M Kring. 2001. Who attains social status? effects of personality and physical attractiveness in social groups. *Journal of personality and social psychology*, 81(1):116.

Lynne M Andersson and Christine M Pearson. 1999. Tit for tat? the spiraling effect of incivility in the workplace. *Academy of management review*, 24(3):452–471.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *Political Analysis*, pages 1–15.

Chris Bail. 2014. *Terrified: How anti-Muslim fringe organizations became mainstream*. Princeton University Press.

Christopher A Bail. 2016. Emotional feedback and the viral spread of social media messages about autism spectrum disorders. *American journal of public health*, 106(7):1173–1180.

Mikhail Mikhaĭlovich Bakhtin. 2010. *Speech genres and other late essays*. University of Texas press.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman and Noah A. Smith. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv preprint*, abs/2302.04023.

Pablo Barberá, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. 2015. The critical periphery in the growth of social protests. *PloS one*, 10(11):e0143611.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv preprint*, abs/2004.05150.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bernard Berelson. 1952. Content analysis in communication research.

Sudeep Bhatia. 2017. Associative judgment and vector space semantics. *Psychological review*, 124(1):1.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Ryan C Black, Sarah A Treul, Timothy R Johnson, and Jerry Goldman. 2011. Emotions, oral arguments, and supreme court decision making. *The Journal of Politics*, 73(2):572–581.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter! debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

Janet M Box-Steffensmeier and Bradford S Jones. 2004. *Event history modeling: A guide for social scientists*. Cambridge University Press.

Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, 6(32):eaba2196.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. Bootstrapped text-level named entity recognition for literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–350, Berlin, Germany. Association for Computational Linguistics.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Moira Burke and Robert Kraut. 2008. Mind your ps and qs: the impact of politeness and rudeness in online communities. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 281–284.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm:

Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6326–6334. AAAI Press.

Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. 2018. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Derek Chen, Zhou Yu, and Samuel Bowman. 2022. Clean or annotate: How to spend a limited data collection budget. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168.

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117, New Orleans, Louisiana. Association for Computational Linguistics.

Justin Cheng, Lada A. Adamic, Jon M. Kleinberg, and Jure Leskovec. 2016. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 671–681. ACM.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, volume 9, pages 61–70.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *ArXiv preprint*, abs/1904.10509.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *ArXiv preprint*, abs/2303.16779.

Eric Chu, Prashanth Vijayaraghavan, and Deb Roy. 2018. Learning personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2638–2646, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Robert B Cialdini. 2003. *Influence*. Influence At Work.

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *Ieee Access*, 9:33203–33223.

Paul Coulton, Jonny Huck, Andrew Hudson-Smith, Ralph Barthel, Panagiotis Mavros, Jennifer Roberts, and Philip Powell. 2014. Designing interactive systems to encourage empathy between users. In *Proceedings of the 2014 companion publication on Designing interactive systems*, pages 13–16.

Holly K Craig and Julie A Washington. 2002. Oral language expectations for african american preschoolers and kindergartners.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon M. Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 699–708. ACM.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Catherine E Davies. 2017. Sociolinguistic approaches to humor. In *The Routledge handbook of language and humor*, pages 472–488. Routledge.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Grant DeLozier, Ben Wing, Jason Baldridge, and Scott Nesbit. 2016. Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 188–198, Berlin, Germany. Association for Computational Linguistics.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Paul DiMaggio. 2015. Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2):2053951715602908.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

James E Dobson. 2019. *Critical digital humanities: the search for a methodology*. University of Illinois Press.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Jacob Eisenstein. 2012. Mapping the geographical diffusion of new words. *ArXiv preprint*, abs/1210.5268.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, Oregon, USA. Association for Computational Linguistics.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Daniel Ellsberg. 1961. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, 75(4):643–669.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *ArXiv preprint*, abs/2109.05322.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

James A Evans and Pedro Aceves. 2016. Machine translation: Mining text for social theory. *Annual review of sociology*, 42:21–50.

Mohamed Fazeen, Ram Dantu, and Parthasarathy Guturu. 2011. Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches. *Social Network Analysis and Mining*, 1(3):241–254.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

René D Flores. 2017. Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, 123(2):333–384.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Harold Garfinkel. 2016. Studies in ethnomethodology. In *Social Theory Re-Wired*, pages 85–95. Routledge.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Alan S Gerber, Gregory A Huber, David Doherty, Conor M Dowling, and Shang E Ha. 2010. Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review*, 104(1):111–133.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *ArXiv preprint*, abs/2303.15056.

Scott A Golder and Michael W Macy. 2014. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40:129–152.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the hero, the villain, and the victim? detection of roles in news articles using natural language techniques. In *23rd International Conference on Intelligent User Interfaces*, pages 311–315.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv preprint*, abs/2209.12356.

Elizabeth E Graham. 1995. The involvement of sense of humor in the development of social relationships. *Communication Reports*, 8(2):158–169.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Justin H Gross, Brice Acree, Yanchuan Sim, and Noah A Smith. 2013. Testing the etch-a-sketch hypothesis: a computational analysis of mitt romney's ideological makeover during the 2012 primary vs. general elections. In *APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting*.

Jonathan Grudin. 2006. Why personas work: The psychological evidence. *The persona lifecycle*, 12:642–664.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, et al. 2009. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International business review*, 18(4):417–432.

Safair Safwat Mohammed Hashim and Suhair Safwat. 2015. Speech acts in political speeches. *Journal of Modern Education Review*, 5(7):699–706.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Susan C Herring. 1994. Politeness in computer culture: Why women thank and men flame. In *Cultural performances: Proceedings of the third Berkeley women and language conference*, pages 278–294.

Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. 2012. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *ArXiv preprint*, abs/2203.15556.

Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science*, 355(6324):486–488.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Deniz Yuret, and Aline Villavicencio. 2021. Challenges and applications of automated extraction of socio-political events from text (CASE 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9, Online. Association for Computational Linguistics.

Ian Hutchby and Robin Wooffitt. 2008. *Conversation analysis*. Polity.

Gabe Ignatow and Rada Mihalcea. 2016. *Text mining: A guidebook for the social sciences*. Sage Publications.

Shanto Iyengar. 1990. Framing responsibility for political issues: The case of poverty. *Political behavior*, 12(1):19–40.

Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. 2019. Figurative usage detection of symptom words to improve personal health mention detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1142–1147, Florence, Italy. Association for Computational Linguistics.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Arthur M Jacobs and Annette Kinder. 2018. What makes a metaphor literary? answers from two computational studies. *Metaphor and Symbol*, 33(2):85–100.

Labiba Jahan, Rahul Mittal, and Mark Finlayson. 2021. Inducing stereotypical character roles from plot structure. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 492–497.

Zubin Jelveh, Bruce Kogut, and Suresh Naidu. 2014. Detecting latent ideology in expert text: Evidence from academic papers in economics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1804–1809, Doha, Qatar. Association for Computational Linguistics.

Ferdaous Jenhani, Mohamed Salah Gouider, and Lamjed Ben Said. 2016. A hybrid approach for drug abuse events extraction from twitter. *Procedia computer science*, 96:1032–1040.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew L Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Kristen Johnson, I-Ta Lee, and Dan Goldwasser. 2017. Ideological phrase indicators for classification of political discourse framing on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 90–99, Vancouver, Canada. Association for Computational Linguistics.

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

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv preprint*, abs/2001.08361.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of* *the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 553–562. ACM.

Vaibhav Kesarwani, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. Metaphor detection in a poetry corpus. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–9, Vancouver, Canada. Association for Computational Linguistics.

Marc Keuschnigg, Niclas Lovsjö, and Peter Hedström. 2018. Analytical sociology and computational social science. *Journal of Computational Social Science*, 1(1):3–14.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Tae-Yeol Kim, Deog-Ro Lee, and Noel Yuen Shan Wong. 2016. Supervisor humor and employee outcomes: The role of social distance and affective trust in supervisor. *Journal of Business and Psychology*, 31:125–139.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Gary King and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642.

Simon Kirby, Mike Dowman, and Thomas L Griffiths. 2007. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *ArXiv preprint*, abs/2302.10724.

Philipp Koralus and Vincent Wang-Maścianica. 2023. Humans in humans out: On gpt converging toward common sense in both success and failure. *arXiv preprint arXiv:2303.17276*.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Thomas S. Kuhn. 1962. *The structure of Scientific Revolutions*. The University of Chicago Press.

Giselinde Kuipers. 2009. Humor styles and symbolic boundaries.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 625–635. ACM.

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 3739–3748. ACM.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Computational social science. *Science*, 323(5915):721–723.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Monica Lee and John Levi Martin. 2015. Coding, counting and cultural cartography. *American Journal of Cultural Sociology*, 3:1–33.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. Association for Computational Linguistics.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. 2022. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *ArXiv preprint*, abs/2211.09110.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020a. Desmog: Detecting stance in media on global warming. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3296–3315.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020b. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Mark J Machina. 1987. Choice under uncertainty: Problems solved and unsolved. *Journal of Economic perspectives*, 1(1):121–154.

Keith Maki, Michael Yoder, Yohan Jo, and Carolyn Rosé. 2017. Roles and success in Wikipedia talk pages: Identifying latent patterns of behavior. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1026–1035, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Dorothy Markiewicz. 1974. Effects of humor on persuasion. *Sociometry*, pages 407–422.

Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263, Jeju Island, Korea. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Bonan Min and Xiaoxi Zhao. 2019. Measure country-level socio-economic indicators with streaming news: An empirical study. 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)*, pages 1249–1254, Hong Kong, China. Association for Computational Linguistics.

Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. 2019. A modular architecture for unsupervised sarcasm generation. 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)*, pages 6144–6154, Hong Kong, China. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Kevin M Murphy and Andrei Shleifer. 2004. Persuasion in politics. *American Economic Review*, 94(2):435–439.

Michael Murray et al. 2015. Narrative psychology. *Qualitative psychology: A practical guide to research methods*, pages 85–107.

Michael Muthukrishna, Adrian V Bell, Joseph Henrich, Cameron M Curtin, Alexander Gedranovich, Jason McInerney, and Braden Thue. 2020. Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, 31(6):678–701.

Laura K Nelson. 2015. Political logics as cultural memory: cognitive structures, local continuities, and women's organizations in chicago and new york city. *Work. Pap. Kellogg School Manag., Northwestern Univ*.

Laura K Nelson. 2021. Cycles of conflict, a century of continuity: the impact of persistent place-based political logics on social movement strategy. *American Journal of Sociology*, 127(1):1–59.

Laura K Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, Beijing, China. Association for Computational Linguistics.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha, Qatar. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Dan Ofer and Dafna Shahaf. 2022. Cards against AI: Predicting humor in a fill-in-the-blank party game. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5397–5403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *ArXiv preprint*, abs/2205.12698.

OpenAI. 2023. Gpt-4 technical report.

Andrew Ortony, Gerald L Clore, and Allan Collins. 2022. *The cognitive structure of emotions*. Cambridge university press.

Marco Ortu, Alessandro Murgia, Giuseppe Destefanis, Parastou Tourani, Roberto Tonelli, Michele Marchesi, and Bram Adams. 2016. The emotional side of software developers in jira. In *Proceedings of the 13th international conference on mining software repositories*, pages 480–483.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Brian Paltridge and Jill Burton. 2000. *Making sense of discourse analysis*. Antipodean Educational Enterprises Gold Coast, Queensland.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Kaiping Peng, Richard E Nisbett, and Nancy YC Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological methods*, 2(4):329.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Ulrike Pfeil and Panayiotis Zaphiris. 2007. Patterns of empathy in online communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 919–928.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2012. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.

Jenny Preece. 1998. Empathic communities: Reaching out across the web. *interactions*, 5(2):32–43.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Christoph Purschke and Dirk Hovy. 2019. Lörres, möppes, and the swiss.(re) discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography*, 7(2):113–134.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *ArXiv preprint*, abs/2302.06476.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive testing and debugging of nlp models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3253–3267.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to action on social media: Potential for censorship and social impact. *EMNLP-IJCNLP 2019*, page 36.

Barbara Olasov Rothbaum, Elizabeth A Meadows, Patricia Resick, and David W Foy. 2000. Cognitive-behavioral therapy.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Maja R. Rudolph and David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1003–1011. ACM.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Harvey Sacks. 1992. Lectures on conversation: Volume i. *Malden, Massachusetts: Blackwell*.

Belen Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 78–86, Online. Association for Computational Linguistics.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020a. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020b. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1978, Online. Association for Computational Linguistics.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A Smith, James W Pennebaker, and Eric Horvitz. 2022. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

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

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. 2020. Examining the ordering of rhetorical strategies in persuasive requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306, Online. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *ArXiv preprint*, abs/2212.08061.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Lawrence W Sherman. 1988. Humor and social distance in elementary school children.

Galit Shmueli et al. 2010. To explain or to predict? *Statistical science*, 25(3):289–310.

Martin Shubik. 1982. *Game theory in the social sciences: concepts and solutions*.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using multi-kernel convolution and attentive lstm variants. *IEICE TRANSACTIONS on Information and Systems*, 102(12):2493–2503.

David Silverman. 1998. *Harvey Sacks: Social science and conversation analysis*. Oxford University Press on Demand.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Shashank Srivastava, Snigdha Chaturvedi, and Tom M. Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2807–2813. AAAI Press.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Charles J Stewart, Craig Allen Smith, and Robert E Denton Jr. 2012. *Persuasion and social movements*. Waveland Press.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

John Suler. 2005. Contemporary media forum: The online disinhibition effect. *International Journal of Applied Psychoanalytic Studies*.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological review*, 117(2):440.

Zeynep Tufekci and Christopher Wilson. 2012. Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of communication*, 62(2):363–379.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Esther van den Berg, Katharina Korfhage, Josef Ruppenhofer, Michael Wiegand, and Katja Markert. 2020. Doctor who? framing through names and titles in German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4924–4932, Marseille, France. European Language Resources Association.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Ronald Wardhaugh and Janet M Fuller. 2021. *An introduction to sociolinguistics*. John Wiley & Sons.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *ArXiv preprint*, abs/2302.10205.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. 2011. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA. ACM.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019a. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019b. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 344. ACM.

Diyi Yang, Miaomiao Wen, and Carolyn Rosé. 2015. Weakly supervised role identification in teamwork interactions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1671–1680, Beijing, China. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 460–470.

Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2021. Score: Pretraining for context representation in conversational semantic parsing. In *International Conference on Learning Representations*.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

J Zamfirescu-Pereira, Richmond Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems (CHI'23)*.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 357–366.

Jing Zhang, Jie Tang, and Juanzi Li. 2007. Expert finding in a social network. In *Advances in Databases: Concepts, Systems and Applications*, pages 1066–1069. Springer.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021a. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. *ArXiv preprint*, abs/2109.03158.

Jian Zhu and David Jurgens. 2021b. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218, Online. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *ArXiv preprint*, abs/2301.12867.

Caleb Ziems, William Held, Jingfeng Yang, and Diyi Yang. 2022a. Multi-value: A framework for cross-dialectal english nlp. *ArXiv preprint*, abs/2212.08011.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022b. Inducing positive perspectives with text reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976.

Mingyu Zong and Bhaskar Krishnamachari. 2022. a survey on gpt-3.

Approach	Average Spearman Correlation MRF	FLUTE	SBIC	Reframing
MTurk + Staging	0.074	0.160	0.101	0.029
MTurk + Qual	-	0.283	-	-
Expert	0.367	0.383	0.125	0.300

Table 5: **Reliability of Evaluation Approaches** as given by the average Spearman correlation between annotators' rankings. The near-zero MTurk + Staging agreement on two tasks indicates a need for better quality measures. The qualifying exam (MTurk + Qual) improved FLUTE agreement. The most reliable data comes from Expert evaluations as defined in the body text.

Model	% Preferred Over Human Gold Annotations MRF	FLUTE	+ Qual	SBIC	Reframing
ada-001	4.5%	51.5%	0.0%	15.2%	6.2%
babbage-001	21.2%	45.5%	0.0%	10.6%	-
curie-001	12.1%	50.0%	0.0%	22.7%	13.6%
davinci-001	10.6%	25.8%	0.0%	18.2%	14.8%
davinci-002	15.2%	16.7%	1.5%	21.2%	15.2%
davinci-003	13.6%	21.2%	3.0%	31.8%	23.0%
ChatGPT	24.2%	22.7%	0.0%	39.4%	19.7%

Table 6: **Majority Vote Crowdworker Evaluations for Zero-shot Generation Tasks** give the proportion of all pairwise rankings where, by majority vote, crowdworkers ranked the model's generation as more accurate or preferable to a gold-standard explanation drawn from the dataset. Best models are in green and runnner-ups are in blue. **+Qual** indicates evaluations after annotators completed a qualifying exam to ensure quality. Results are

A Challenges of Evaluating CSS Generation Tasks

In this appendix section, we discuss the inherent challenge of running human evaluation on generation tasks.

Automatic evaluation. Immediately, we found that automatic evaluation metrics in Table 7 were uninformative proxies for generation quality. Manual inspection revealed high-quality generation outputs (see also the following paragraph). However, BLEURT effectively reported zero semantic overlap (see near-zero negative scores). Furthermore, variation in BLEU and BERTScores failed to follow any discernible patters with regards to model preference or scaling laws that we observed by manual inspection. This lead us to develop a more systematic human evaluation harness.

Human evaluation. In all human evaluations, the annotators produce ranked lists of model generations. We iteratively improved upon the reliability of the human evaluation in three stages displayed in Table 5. The first stage (MTurk + Staging) led to low agreement and misunderstandings of the task. The second stage (MTurk + Qual) improved the agreement, but additional misunderstandings persisted. This led us to adopt an Expert evaluation setup.

For our first round of annotation **(MTurk + Staging)**, we recruited crowdworkers from Amazon Mechanical Turk, paying a fair wage based on the federal minimum. We minimized cultural and individual variance in two ways: (1) by recruiting only workers from the United States, and (2) by filtering workers through a staging round. The staging round contained a smaller pool of tasks. Only workers who demonstrated 1-3 examples of agreement with a verified worker was then verified and given access to the full set of tasks. Despite these efforts, the quality of the resulting annotations was not high, with near-zero average pairwise Spearman correlation (ρ) between annotator judgments. Furthermore, the highly counterintuitive inverse scaling on FLUTE proved to be the result of annotators misunderstanding the task.

To address the FLUTE misunderstandings above, we opted for better instructions and a qualifying exam in the second round **(MTurk + Qual)**. Specifically, we updated task instructions to include many observed issues in the generations (i.e., failure to explain the underlying construct). The instructions outlined a desired template, which required explanations to (1) identify the figurative language phrase; (2) translate that phrase; (3) describe how this results in the entailment or contradiction. Most importantly, annotators needed to pass a qualifier by answering at least 4 out of 5 task questions in the same way as our hidden expert judgments. The qualifier and improved instructions induced higher overall agreement ($\rho = 0.283$, Table 5),[13] but as a result, annotators became too fixated on the expected output template, which favored human gold references. The template excluded many accurate model explanations, resulting in the reported 0% model preference over gold in the +Qual column of Table 5—an overly conservative signal.

Finally, the authors decided to serve as **Expert** annotators. OS and JC annotated MRF and SBIC, while CZ and WH annotated FLUTE and Positive Reframing. Expert annotators bypassed misunderstandings, increased coder reliability across the

[13] This is a reasonable level of agreement. We used forced-choice ranking to tease out subtle quality differences and unanimous vote to flatten the variance for explanations of similar quality. Thus our full evaluation setup reflects only real, observable, yet subtle differences in quality.

Model / Eval	Baselines	FLAN-T5					FLAN	Chat	text-001				text-002	text-003
	Finetune	Small	Base	Large	XL	XXL	UL2	ChatGPT	Ada	Babb.	Curie	Dav.	Davinci	Davinci
Social Bias Inference Corpus														
BLEU	29.2	8.9	7.0	5.8	8.3	17.5	7.7	6.1	3.6	4.7	5.1	10.2	7.3	8.4
BLEURT	-0.7	-0.9	-0.9	-0.9	-0.9	-0.8	-0.9	-0.5	-0.8	-0.7	-0.6	-0.6	-0.5	-0.4
BERTScore	90.8	86.3	85.7	85.4	86.0	87.6	86.2	87.2	85.7	86.1	86.3	87.5	87.1	87.8
FLUTE: Figurative Language														
BLEU	14.6	-	-	-	-	-	-	4.7	2.7	6.7	6.5	6.2	5.5	6.0
BLEURT	-0.4	-	-	-	-	-	-	-0.8	-1.1	-0.8	-0.8	-0.9	-0.8	-0.6
BERTScore	89.5	-	-	-	-	-	-	86.0	85.0	87.2	86.9	86.8	86.4	86.9
Misinformation Reaction Frames														
BLEU	7.4	4.4	7.6	7.1	9.6	5.1	5.5	8.3	2.6	3.4	3.8	3.1	7.7	6.0
BLEURT	-0.6	-0.9	-0.8	-1.0	-0.7	-0.9	-1.1	-0.6	-1.1	-0.8	-0.7	-1.1	-0.5	-0.7
BERTScore	86.8	85.1	87.1	86.8	87.7	86.5	85.9	86.9	85.1	85.9	86.3	85.4	87.7	86.3
Positive Reframing														
BLEU	7.1	0.4	9.4	10.5	11.1	9.4	9.1	6.2	0.8	1.6	4.5	6.9	5.7	5.2
BLEURT	-0.8	-1.1	-0.7	-0.6	-0.6	-0.6	-0.7	-0.5	-1.1	-0.9	-0.7	-0.7	-0.6	-0.4
BERTScore	88.6	81.3	86.9	87.9	88.0	88.7	88.3	88.1	82.9	83.3	87.2	87.7	87.9	87.8

Table 7: **Automatic Evaluation of Zero-shot Generation** across our selected CSS benchmark tasks. All three metrics appear to be uninformative, and the results lack patterns or discernable structure. We opt instead for human evaluation on these generation tasks.

board (higher Spearman correlations in Table 5), and produced more sensible results, which demonstrate expected scaling behaviors (see Table 4 in the body text). However, this solution is expensive and infeasible in many application domains. We discuss these points as limitations in Section 7.4.

B Error Analysis

For a representative subset of classification tasks, we conduct an analysis of shared errors across evaluated models. We focus specifically on the best performing model in a class (e.g. the best variant of FLAN models or the best OpenAI model). Finally, in Figure 2, we highlight a breakdown of error types for ChatGPT.

B.1 Figurative Language

We sample all 29 cases in which every model was incorrect. In just under half of these cases (14/29), all models agreed on an incorrect answer, which we call a *unanimous error*. Out of fourteen unanimous errors, the models were at least partially correct four times, which we call a *plausible/gold error* (see Figure 2). There was one mistaken gold label and three cases of correctly-labeled similes nested inside the predicted sarcasm. Of the remaining ten unanimous errors, three were idioms mistaken as metaphors, and seven were similes classified with the more general metaphor label. For humans, this is a common error, but for models, this is surprising, since similes should have easy keyword signals "as" and "like." The baseline method was likely able to exploit these signals to achieve a higher accuracy.

In 5 errors, all models disagreed and missed the intended sarcasm label. In another 5 error cases, only UL2 and `text-davinci-003` agreed on the correct label, but the dataset was mislabeled, with four idioms marked wrongly as metaphors and one simile marked as an idiom. In the remaining 5 errors, ChatGPT showed a preference for the most generic label and predicted metaphor.

B.2 Emotion Recognition

We sample 50 cases where all models differed from the gold labels. Unlike Figurative Language, a minority of examples had the same mismatch across models (9/50). However, a closer analysis of individual errors yields a surprising result: at least 18/50 examples *across all evaluated models* were judged as gold mislabels. Additionally, for FLAN-UL2 and ChatGPT, 17/50 and 15/50 predictions respectively could be considered as valid—even if they differed from the gold label.[14]

Moving to true negatives, we observe that DV2 makes the most errors (28/50) that cannot be cate-

[14]For example, "*i feel that the sweet team really accomplished that*" can be considered both *love - gold* or *joy - predicted*

gorized as a gold mislabel, while UL2 (17/20) and ChatGPT (19/20) make significantly fewer. The distribution of errors differ across each model type: ChatGPT, for example, over-labels with *surprise*: especially instances with a true gold label of *Joy* (8) or *Love* (5). On the other hand, UL2 mislabels *Love* as *Joy* frequently (9); and fear as *Sadness* (4) or *Surprise* (4). Finally, `davinci` mislabels Sadness most frequently as Joy (9) or anger/love (3 each).

B.3 Politeness Prediction

We first visualized the per-category accuracy of the different best-performing models (FLAN-T5-XL, `Text-davinci-002`, and ChatGPT). We observed that: (1) The XL model tended to predict more polite labels. It was more accurate in terms of the utterances that were polite and neutral with 70.4% and 62.0% accuracy. And most of the errors came from impolite cases (with a 45.2% accuracy). (2) `davinci-002` performed the best in judging neutral utterances. `davinci-002` model was the most accurate for neutral utterances (82.9% accuracy) while making significantly more errors for polite and impolite utterances (43.9% and 40.9% accuracy respectively). (3) ChatGPT performed the worst in finding impolite utterances while making more neutral predictions, with only a 9.0% accuracy for the impolite category, whereas it achieved 75.9% and 66.8% for neutral and polite cases.

We then went through the 81/498 cases where the three models are all making errors. We found that the three models are making the same errors in most of the cases (54/81) and `davinci-002` models are making more similar errors with ChatGPT (17/81 cases). Among these common error cases, we observed that 79/81 cases were related to the 1st and 2nd person mention strategy (Danescu-Niculescu-Mizil et al., 2013) and all of them were direct or indirect questions where 38/81 of them were related to counterfactual modal and indicative modal (Danescu-Niculescu-Mizil et al., 2013), which indicated that all three models suffered from making accurate judgments towards direct or indirect questions with 1st and 2nd person mentions.

B.4 Implicit Hate Classification

We first consider the confusion matrix and find that OpenAI models are particularly oversensitive to the "stereotypical" class (71% and and 65% false-positive rates from `davinci-003` and ChatGPT respectively). Our error analysis of 50 samples shows that models fail to apply the definition: stereotypical text must associate the target with particular characteristics. Instead, models are more likely to mark as stereotype any text that contains an identity term (86% of false-positives contain identity terms). All models also fail to recognize strong phrasal signals, like "rip" or "kill white people" for the *white grievance* (all 3/50 cases are errors), or violent terms associated with threats. More subtle false-negatives require sociopolitical knowledge (2/50) or understanding of humor (6/50). Other errors are examples where the model identified a valid secondary hate category (5/50).

C Additional Tables and Prompts

Dataset	Size	Classes
Utterance Level		
Dialect	266	23
Persuasion	399	7
Impl. Hate	498	6
Emotion	498	6
Figurative	500	4
Ideology	498	3
Stance	435	3
Humor	500	2
Misinfo	500	2
Semantic Chng	344	2
Conversation Level		
Discourse	497	7
Politeness	498	3
Empathy	498	3
Toxicity	500	2
Power	500	2
Persuasion	434	2
Document Level		
Event Arg.	283	–
Evt. Surprisal	240	–
Tropes	114	114
Ideology	498	3
Generation Tasks		
MRF	500	–
FLUTE	500	–
SBIC	500	–
Reframing	500	–

Table 8: Dataset size and classes count across all selected CSS benchmarks. Datasets are sorted by class count for each task category.

Discourse Acts

Context #example input

danger67: I'm on my second playthrough and have been trying to practice PvP(which I am terrible at) and last night I invaded the opening section of da...

Prompt #query ChatGPT

Which of the following best characterizes the discourse type of the previous statement?
A: Question
B: Answer
C: Agreement
D: Disagreement
E: Appreciation
F: Elaboration
G: Humor

Constraint: Answer with only the option above that is most accurate and nothing else.

Expected answer

B: Answer

Figurative Language Classification

Context #example input

premise: I said, work independently and come up with some plans.

hypothesis: I said, put your heads together and come up with some plans.

Prompt #query ChatGPT

Which of the following best describes the type of figurative language used in the hypothesis?
A: Idiom
B: Metaphor
C: Sarcasm
D: Simile

Constraint: Answer with only the option above that is most accurate and nothing else.

Expected answer

A: Idiom

Figurative Language Explanation

Context #example input

```
premise: not impressive or
remarkable in any way, but I
thought it looked nice.

hypothesis: Nothing to write home
about, but I thought it looked
nice.
```

Prompt #query ChatGPT

```
Which of the following describes
the relationship between the
premise and the hypothesis?
A: Entailment
B: Contradiction

Constraint: Explain the
figurative language in the
hypothesis in one sentence and
then answer with which option is
the most accurate.
```

Expected answer

```
A: Entailment

To nothing to write home about
means that the thing is not
impressive or remarkable, which
is in line with how speaker felt
even though they felt it was nice.
```

Latent Hatred

Context #example input

```
" : jewish harvard professor
noel ignatiev wants to abolish
the white race via #wr " " "
```

Prompt #query ChatGPT

```
Which of the following categories
of hate speech best describes the
sentence above?
A: White Grievance (frustration
over a minority group's perceived
privilege and casting majority
groups as the real victims of
racism)
B: Incitement to Violence
(flaunting in-group unity and
power or elevating known hate
groups and ideologies)
C: Inferiority Language (implies
one group or individual is
inferior to another, including
dehumanization and toxification)
D: Irony (the use of sarcasm,
humor, and satire to attack or
demean a protected class or
individual)
E: Stereotypes and Misinformation
(associating a protected class
with negative attributes)
F: Threatening and Intimidation
(conveys a speaker commitment to
a target's pain, injury, damage,
loss, or violation of rights)

Constraint: Answer with one or
more of the options above that is
most accurate and nothing else.
Always choose at least one of the
options.
```

Expected answer

```
A: White Grievance
```

Event Surprisal

Context `#example input`

```
A: Four months ago, I had a big
family reunion.
B: We haven't had one in over 20
years.
C: This was a very exciting event.
D: I saw my Grandma who said I
liked great as ever.
```

Prompt `#query ChatGPT`

```
This is an Event Extraction task.
Which sentences above indicate
new events?
```

Expected answer

```
A, D
```

Utterance Ideology

Context `#example input`

```
Union shop proponents point out
that the `` free rider '' option
weakens labor unions because
fewer people are likely to join a
labor union and pay me...
```

Prompt `#query ChatGPT`

```
Which of the following leanings
would a political scientist say
that the above article has?
A: Liberal
B: Conservative
C: Neutral

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
Conservative
```

Humor Classification

Context `#example input`

```
If a mass of beef fat is
'tallow', and mass of pig fat is
'lard', what is a mass of human
fat called?_____'American'.Just
kidding, it's actually calle...
```

Prompt `#query ChatGPT`

```
Is the above joke humorous to
most of the people? Constraint:
You must pick between "True" or
"False" You cannot use any other
words except for "True" or "False"
```

Expected answer

```
True
```

Dialect Features

Context #example input

```
two years I stayed alone
```

Prompt #query ChatGPT

```
Which of the following features
would a linguist say that the
above sentence has?
A: Article Omission (e.g.,
'Person I like most is here.')
B: Copula Omission (e.g.,
'Everything busy in our life.')
C: Direct Object Pronoun Drop
(e.g., 'He didn't give me.')
D: Extraneous Article (e.g,
'Educated people get a good
money.')
E: Focus Itself (e.g, 'I did it
in the month of June itself.')
F: Focus Only (e.g, 'I was there
yesterday only'.)
G: General Extender "and all"
(e.g, 'My parents and siblings
and all really enjoy it'.)
H: Habitual Progressive (e.g.,
'They are getting H1B visas.')
I: Invariant Tag "isn't it, no,
na" (e.g., 'It's come from me,
no?')
J: Inversion In Embedded Clause
(e.g., 'The school called to ask
when are you going back.')
K: Lack Of Agreement (e.g., 'He
talk to them.')
L: Lack Of Inversion In
Wh-questions (e.g., 'What are you
doing?')
M: Left Dislocation (e.g., 'My
parents, they really enjoy
playing board games.')
N: Mass Nouns As Count Nouns
(e.g., 'They use proper grammars
there.')
O: Non-initial Existential "is /
are there" (e.g., 'Every year
inflation is there.')
P: Object Fronting (e.g., 'In
fifteen years, lot of changes we
have seen.')
Q: Prepositional Phrase Fronting
With Reduction (e.g., 'First of
all, right side we can see a
plate.')
R: Preposition Omission (e.g., 'I
stayed alone two years.')
...
```

Expected answer

```
R: Preposition Omission
```

Article Ideology

Context #example input

```
Has the media gotten worse ? Or
am I just grouchier ?
Every day I see things that are
wrong or that so miss the point I
want to scream .
As this week ...
```

Prompt #query ChatGPT

```
Which of the following leanings
would a political scientist say
that the above article has?
A: Left
B: Right
C: Center

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
B: Right
```

Misinformation Detection

Context #example input

```
White House Ousts Top Climate
Change Official
```

Prompt #query ChatGPT

```
Which of the following describes
the above news headline?
A: Misinformation
B: Trustworthy

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
A: Misinformation
```

Implied Misinformation Explanation

Context #example input

```
White House Ousts Top Climate
Change Official
```

Prompt #query ChatGPT

```
What is the implied message of
the above news headline?

Constraint: Answer with a short
phrase like "some masks are
better than others."
```

Expected answer

```
The white house lost confidence
in their top climate change
official.
```

Persuasion

Context #example input

```
Amablue: At some point, the sum
of all your actions becomes nil.
No one remembers and no one cares
that everyone forgot.

Why does this mean your life is
pointless? It had a point *to
you*. That's all the meaning you
can hope for. No matter what else
happens, even in the heat death
of the universe when every
particle has decayed and there's
nothing left, nothing can change
that you existed for a period of
time. Your existence and your
actions still happened even if
there's no record of them. Do
your actions need permanence to
have meaning?

Senecatwo: I am simply a
biological machine looking to
make more biological machines.
The meaning I find in my actions
is there thanks to biological
imperatives to survive and
reproduce. Once I'm dead, the
meaning leaves with me.
```

Prompt #query ChatGPT

```
If you were the original poster,
would this reply convince you?
True
False

Constraint: Even if you are
uncertain, you must pick either
True or False with without using
any other words.
```

Expected answer

```
False
```

Politeness

Context #example input

```
user: I am looking for help
improving the dermatology content
on wikipedia. Would you be
willing to help, or do you have
any friends interested in der...
```

Prompt #query ChatGPT

```
Based on formal workplace social
norms, which of the following
best describes the above
conversation?
A: Polite
B: Neutral
C: Impolite

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
A: Polite
```

Positive Reframing

Context #example input

```
Always stressing and thinking
about loads of things at once
need I take it one at a time
overload stressed need to rant
```

Prompt #query ChatGPT

```
Rephrase the above sentence to be
more positive using the following
strategies: ['growth',
'neutralizing']

Strategies are defined below

growth: viewing a challenges as
an opportunities for the author
to grow or improve themselves
impermanence: saying bad things
don't last forever, will get
better soon, and/or that others
have experienced similar struggles
neutralizing: replacing a
negative word with a neutral word
optimism: focusing on things
about the situation itself, in
that moment, that are good
self-affirmation: talking about
what strengths the author already
has, or the values they admire
thankfulness: expressing gratitude
```

Expected answer

```
Loads of things on my mind, I
need to make a list, prioritise
and work through it all calmly
and I will feel much better.
```

Power

Context #example input

```
Mike 7: That's odd. Somehow, I
came across one of that user's
edits, though I believe it was on
recent changes. As you can see,
most of the older edits are
vandalism, but I guess due to the
time that wouldn't warrant much
of a block. I don't know how I
happened to come across that
since it's so old.

Mike 7: That could be the case.
I've seen a few of those tonight.
```

Prompt #query ChatGPT

```
Based on the conversation is Mike
7 in a position of power?

Constraint: Even if you are
uncertain, you must pick either
True or False with without using
any other words.
```

Expected answer

```
False
```

Persuasion Strategies

Context #example input

```
Hi there, I was supposed to be
paid today.
```

Prompt #query ChatGPT

```
If a user in a community for
giving away free pizza saw the
above sentence, they would
consider it to use which of the
following persuasion strategies?
A: Evidence (Providing concrete
facts or evidence for the
narrative or request, like "There
is a Pizza Hut and a Dominos near
me.")
B: Politeness (The usage of
polite language in requests, like
"Thank you so much!")
C: Reciprocity (Responding to a
positive action with another
positive action. People are more
likely to help if they have
received help themselves. Example
messages are like "I'll pay it
forward with my first check")
D: Impact (Emphasizing the
importance or impact of the
request, like "I'll use this
pizza to celebrate")
E: Emotion (Making requests full
of emotional valence and arousal
affect to influence others, like
"I've been in the lowest
depressive state of my life")
F: Scarcity (People emphasizing
on the urgency, rare of their
needs, like "I haven't ate a meal
in two days")
G: Other

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
A: Evidence
```

Social Bias Inference

Context #example input

```
"""Boys convince themselves they
are a man by drinking lots of
beer and having a vagina nearby
to stick themselves into.""
FUCKING NAILED IT"
```

Prompt #query ChatGPT

```
Explain the stereotype that this
sentence communicates? Please use
a single sentence like "[GROUP]
are ______"
```

Expected Answers

```
["men need to prove themselves as
manly", "men only drink.", "men
only want a woman to have sex
with.", "men will have sex with
anyone.]
```

Stance

Context #example input

```
@2014voteblue @ChrisJZullo
blindly supporting political
parties is willfull ignorance,
repub or dem. #SemST
```

Prompt #query ChatGPT

```
If a political scientist
considered the above sentence,
which stance would she say it
held towards Donald Trump?
A: Against
B: Favor
C: None

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
C: None
```

Empathy

Context #example input

```
Seeker: I spent today either
staring blankly at a computer
screen or my phone. Was too hurt
to do anything today, really.

Response: I wish I even had the
will to play games. For me it's
excessive daydreaming.
```

Prompt #query ChatGPT

```
Explorations are when a mental
health counselor shows active
interest in a seeker by asking
about unstated experiences. What
level of exploration is expressed
in the counselor's message above?
A: Strong exploration
(specifically labels the seeker's
experiences and feelings, like
"Are you feeling alone right
now?")
B: Weak exploration (a generic
question, like "What happened?")
C: No exploration

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
C: No exploration
```

Temporal Semantic Change

Context `#example input`

```
text1: Having a rough start to my
doctorate program in both the
student and teacher roles and
feel down and ashamed. I spoke to
faculty and know how to move
forward, but while they believe
in me I find it hard to believe
in myself. How do you fight
impostor syndrome @AcademicChatter

text2: laughed so hard running
from impostor friend around the
lab table that I gave myself an
headache lmao what a good day

word: impostor
```

Prompt `#query ChatGPT`

```
If a linguist considered the word
above in text1 and text2, would
she consider the meaning of this
word to be the
A: Same
B: Different

Constraint: Answer with only the
option above that is most
accurate and nothing else.
```

Expected answer

```
B: Different
```

Toxicity Prediction

Context `#example input`

```
Shrike: I have removed recent
edition  of pappe to the lead
though Pappe view might notable
currently without attribution and
proper context of other views it
WP:NPOV  violation.

MelissaLond: In fact, Pappe is
already mentioned twice in the
proper place.
```

Prompt `#query ChatGPT`

```
Will the previouse conversation
eventually derail into a personal
attack?

Constraint: Even if you are
uncertain, you must pick either
"True" or "False" with without
using any other words.
```

Expected answer

```
True
```

Character Tropes

Context #example input

```
You don't know how hard it is
being a woman looking the way I
do.
You don't know how hard it is
being a man looking at a woman
looking the way you do....
```

Prompt #query ChatGPT

```
Given quotes from the character
above, which of the following
tropes would you say this
character represents?

A: Absent Minded Professor
B: Adventurer Archaeologist
C: Arrogant Kungfu Guy
D: Big Man On Campus
E: Bounty Hunter
F: Brainless Beauty
G: Broken Bird
H: Bromantic Foil
I: Bruiser With A Soft Center
J: Bully
K: Byronic Hero
L: Casanova
M: Chanteuse
N: Charmer
O: Child Prodigy
P: Classy Cat Burglar...
```

Expected answer

```
M
```

Event Extraction

Context #example input

```
The Taliban <tgr>killed <tgr>more
than 100 members of the Afghan
security forces inside a military
compound in central Maidan Wardak
province on Monda...
```

Prompt #query ChatGPT

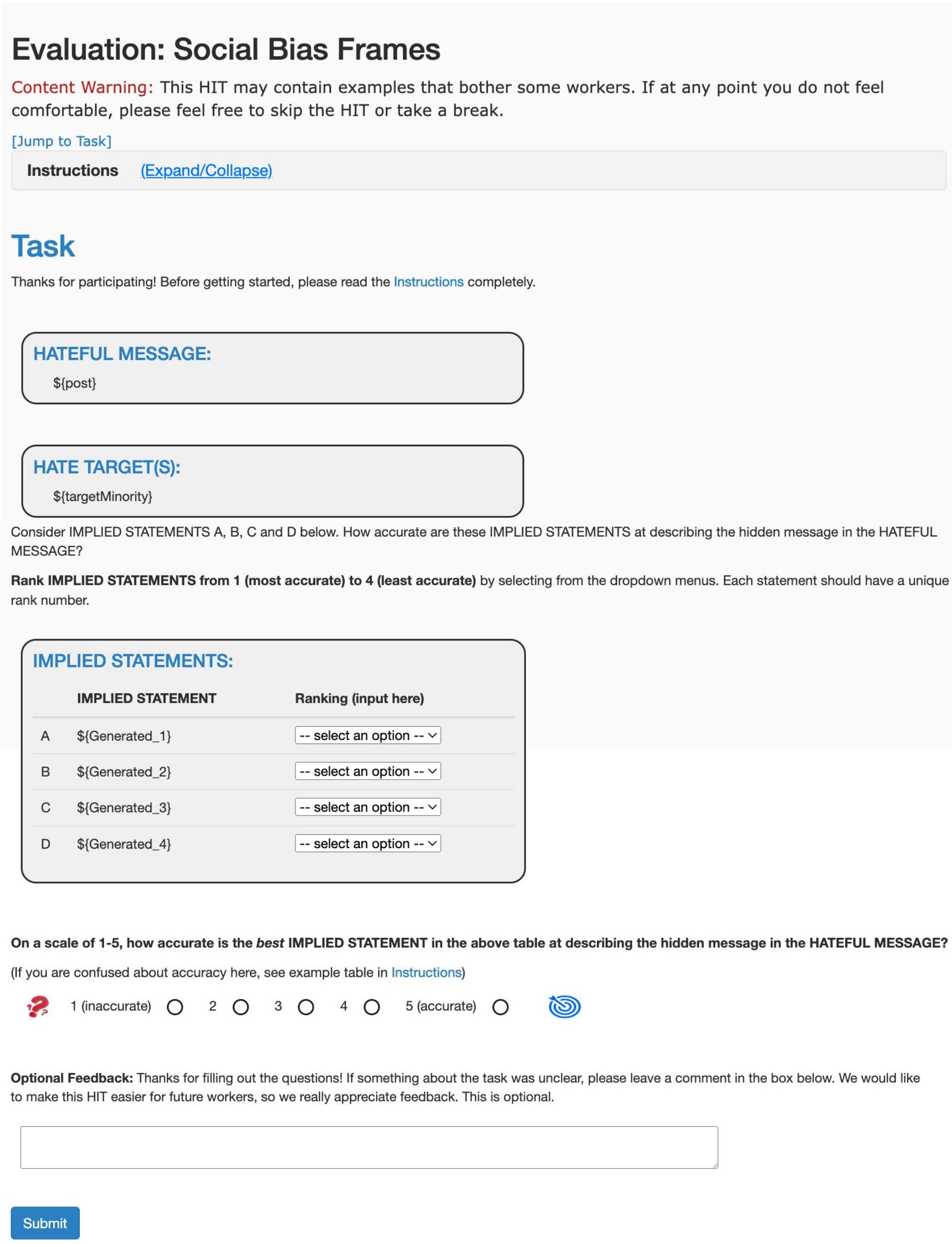
```
{'Victim': 'BLANK', 'Place':
'BLANK', 'Killer': 'BLANK',
'MedicalIssue': 'BLANK'}

Replace the BLANKs with the
extracted information about the
event in <tgr> . Leave the keys
of the JSON unchanged.

JSON Output:
```

Expected answer

```
{'Victim': 'members', 'Place':
'undefined', 'Killer': 'The
Taliban', 'MedicalIssue':
'undefined'}
```



Evaluation: Social Bias Frames

Content Warning: This HIT may contain examples that bother some workers. If at any point you do not feel comfortable, please feel free to skip the HIT or take a break.

[Jump to Task]

Instructions (Expand/Collapse)

Task

Thanks for participating! Before getting started, please read the Instructions completely.

HATEFUL MESSAGE:

${post}

HATE TARGET(S):

${targetMinority}

Consider IMPLIED STATEMENTS A, B, C and D below. How accurate are these IMPLIED STATEMENTS at describing the hidden message in the HATEFUL MESSAGE?

Rank IMPLIED STATEMENTS from 1 (most accurate) to 4 (least accurate) by selecting from the dropdown menus. Each statement should have a unique rank number.

IMPLIED STATEMENTS:

	IMPLIED STATEMENT	Ranking (input here)
A	${Generated_1}	-- select an option --
B	${Generated_2}	-- select an option --
C	${Generated_3}	-- select an option --
D	${Generated_4}	-- select an option --

On a scale of 1-5, how accurate is the *best* IMPLIED STATEMENT in the above table at describing the hidden message in the HATEFUL MESSAGE?

(If you are confused about accuracy here, see example table in Instructions)

1 (inaccurate) 2 3 4 5 (accurate)

Optional Feedback: Thanks for filling out the questions! If something about the task was unclear, please leave a comment in the box below. We would like to make this HIT easier for future workers, so we really appreciate feedback. This is optional.

Submit

Figure 8: **MTurk Human Evaluation for Social Bias Inference Corpus.** Workers review a *hateful message* and an associated *hate target*. Then they review four *Implied Statements* generated by models or pulled from the SBIC's gold human annotations. They are asked to rank these statements from 1 (most accurate) to 4 (least accurate) according to how accurate the *implied statement* is at describing the hidden message from the *hateful message*.