Can Large Language Models Transform Computational Social Science?

Caleb Ziems
Stanford University
Computer Science Department
cziems@stanford.edu

William Held
Georgia Institute of Technology
College of Computing
wheld3@gatech.edu

Omar Shaikh
Stanford University
Computer Science Department
oshaikh@stanford.edu

Jiaao Chen
Georgia Institute of Technology
College of Computing
jiaaochen@gatech.edu

Zhehao Zhang
Dartmouth College
Department of Computer Science
zhehao.zhang.gr@dartmouth.edu

Diyi Yang[*]
Stanford University
Computer Science Department
diyiy@stanford.edu

[*] Contribution distributed as follows. Caleb, Will, and Diyi decided the project scope and research questions. Caleb performed the CSS literature review, as well as the 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. Caleb ran the OpenAI and Flan-T5/UL2 prompt perturbation experiments and few-shot experiments. Zhehao was responsible for all baseline experiments. Caleb managed the human evaluations. All authors contributed to discussion, results, error analysis, and paper writing.

Action Editor: Vivek Srikumar. Submission received: 26 April 2023; revised version received: 29 August 2023; accepted for publication: 25 October 2023.

https://doi.org/10.1162/coli_a_00502

Large language models (LLMs) are capable of successfully performing many language processing tasks zero-shot (without training data). If zero-shot LLMs can also reliably classify and explain social phenomena like persuasiveness and political ideology, then LLMs could augment the computational social science (CSS) pipeline in important ways. 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 25 representative English 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 the performance of today's LLMs can 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 underlying attributes of a text). In summary, LLMs are posed to meaningfully participate in 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).

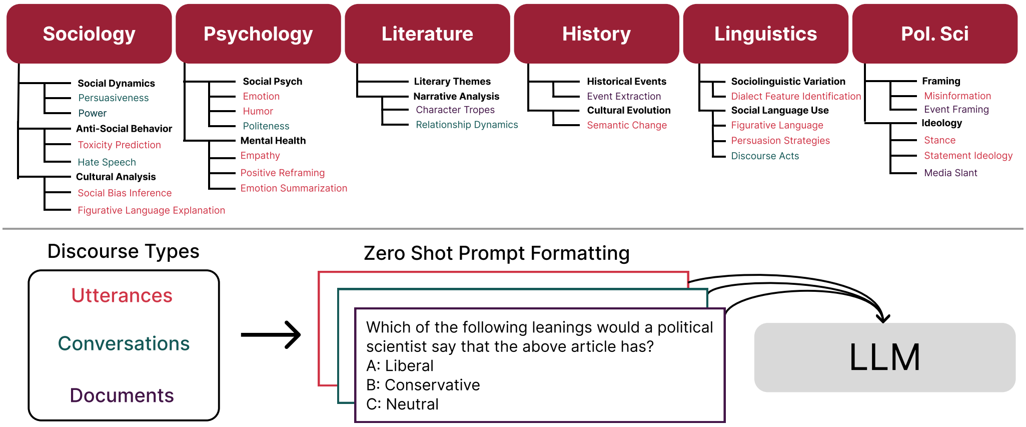
Computational social science (CSS) (Lazer et al. 2020) was born from the immense growth of human data traces on the Web and the rapid acceleration of computational resources for processing this data. These developments allowed researchers to study language and behavior at an unprecedented scale (Lazer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

The goal of this work is to assess the degree to which LLMs can transform 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 labeling efforts to unseen texts (also called *coding* in the social sciences). 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. In the status quo, data resources constrain the theories and subjects CSS can be applied to.

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; Zhuo et al. 2023; Goyal, Li, and Durrett 2022). If LLMs can similarly provide reliable labels and summary codes through zero-shot prompting, CSS research can be broadened to 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 a roadmap with answers to the following research questions:

- **(RQ1) Viability:** Are LLMs able to augment the human annotation pipeline? Can they match or exceed the reliability of human annotation?
- **(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 (Section 2). Then we operationalize each use-case with a set of representative tasks (Section 3). Specifically, classification and parsing methods can help researchers code for linguistic, psychological, and cultural categories (Section 3.1–Section 3.3) while generative models can explain underlying constructs (e.g., figurative language, emotional reactions, hate speech, and misinformation), and restructure text according to established theories like cognitive behavioral therapy (Section 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 (Section 5, Section 6). This allows us to suggest actionable steps for social scientists interested in co-opting LLMs



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

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 and closed-source LLMs across this benchmark using zero-shot prompting (bottom).

for research (Section 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 performance is often too low to entirely replace human annotation. However, LLMs *can* achieve fair levels of agreement with humans on labeling tasks (RQ1). 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 (RQ2). Furthermore, the benefits of LLMs are compounded as models scale up (RQ3). 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. Leading models can achieve parity with the quality of dataset references, and can even exceed them in terms of relevance, coherence, faithfulness, and fluency. 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 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 sociobehavioral phenomena across both scientific and humanistic disciplines. We focus primarily on the most tractable text classification, structured parsing, summarization, and natural language generation tasks for CSS. Some other techniques like aggregate mining of massive datasets or 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*; Section 2.5), communicate (*linguistics*; Section 2.3), establish governance and value-systems (*political science, economics*; Section 2.4), collectively operate (*sociology*; Section 2.6), and create culture (*literature, anthropology*; Section 2.2) across time (*history*; Section 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 Section 3.3.1 and Section 3.3.2, respectively. Historians also work to understand the influence of events on historical shifts in *discourse* (DiMaggio, Nag, and Blei 2013) and *meaning* (Hamilton, Leskovec, and Jurafsky 2016a). We further discuss NLP for discourse and semantic change in Section 2.4 and Section 2.3.

2.2 Literature

Literary studies are closely tied to the analysis of *themes* (Jockers and Mimno 2013), *settings* (Piper, So, and Bamman 2021), and *narratives* (Sap et al. 2022; Saldias and Roy 2020; Boyd, Blackburn, and Pennebaker 2020). Settings can be identified using named entity recognition (Brooke, Hammond, and Baldwin 2016) and toponym resolution (DeLozier et al. 2016), which are already demonstrably solved by prompted models like GPT 3.5 Turbo (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, Chaturvedi, and Mitchell 2016), and the *events* (Sims, Park, and Bamman 2019) they participate in. We cover social role labeling and event extraction methods in Section 3.3.4 and Section 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 Section 3.2.4 and Section 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 Section 3.1.3 and Section 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, Lampos, and Aletras 2015; Del Tredici and Fernández 2017) to personal attributes like age and gender (Bamman, Eisenstein, and Schnoebelen 2014). In Section 3.1.1 and Section 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, Rowe, and Stacey 2016; Eisenstein et al. 2014; Ryskina et al. 2020; Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 2016b; Carlo, Bianchi, and Palmonari 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, belief, and stance (Ahmed and Xing 2010; Baly et al. 2020; Bamman and Smith 2015; Iyyer et al. 2014; Johnson, Lee, and Goldwasser 2017; Preoţiuc-Pietro et al. 2017; Luo, Card, and Jurafsky

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, Budak, and Jurgens 2021; Demszky et al. 2019; Field et al. 2018), with applications to public opinion (Bhatia 2017; Garg et al. 2018; Kozlowski, Taddy, and Evans 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, Kovaleva, and Rumshisky 2019; Tufekci and Wilson 2012), and international relations (King and Lowe 2003). We benchmark ideology detection in Section 3.1.6 and Section 3.3.3, stance detection in Section 3.1.9, and entity framing in Section 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 Section 3.1.8 and Section 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, Stillwell, and Graepel 2013), is no longer available, but in this work, we evaluate on a representative set of psychological factors downstream of personality. For example, differences in personality and cognitive processing can impact what people find funny (Martin and Ford 2018) or persuasive (Hirsh, Kang, and Bodenhausen 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 Section 3.1.5, Section 3.1.8, and Section 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 Section 3.2.2, *emotional aspect-based summarization* in Section 3.4.1, and a *positive reframing* style-transfer task (Ziems et al. 2022) based on cognitive behavioral therapy in Section 3.4.5.

2.6 Sociology

Sociologists want to understand the structure of society and how people live collectively in social groups (Wardhaugh and Fuller 2021; Keuschnigg, Lovsjö, and Hedström 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, Reid, and Rambow 2014; Prabhakaran, Rambow, and Diab 2012) and social roles (Welser et al. 2011; Fazeen, Dantu, and Guturu 2011; Zhang, Tang, and Li 2007; Yang, Wen, and Rosé 2015; Maki et al. 2017). LLMs can assist sociological research by predicting power relations (Section 3.2.4) and unhealthy conversations

(Section 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, Backstrom, and Kleinberg 2009; Tan, Lee, and Pang 2014; Vosoughi, Roy, and Aral 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 (Section 3.1.2) and the speech of hateful social groups (Section 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 (Section 3.1.3).

3. Representative CSS Task Selection

While not exhaustive, our task selection is designed to provide a representative survey of the CSS needs in Section 2. This will allow us to answer Research Questions 1–4, which are pertinent to social science researchers. Thus our work is distinct and complementary to BIG-Bench (Srivastava et al. 2023) and other efforts to benchmark the logical, physical, and social reasoning capabilities of LLMs. Our attention is more carefully focused on the affordances of LLMs for social science.[1] Our tasks are field-specific and come with the field-specific challenges of expert taxonomies, large label spaces, temporal grounding, and domain-specific parsing schemes (see Section 7.6).

To help answer RQ3 and RQ4, 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 (Section 3.1), 6 conversation-level tasks (Section 3.2), and 4 document-level tasks for the analysis of media (Section 3.3). In addition to these 20 classification tasks, we evaluate 5 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, Smith, and Xing 2011) and ideolects (Zhu and Jurgens 2021a), with numerous applications in sociolinguistics, education, and the sociology of class and community membership (see Section 2.3). These features can be used to study the sociolinguistics of language change (Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 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

[1] To elaborate, BIG-Bench, among its 200 tasks, has some overlap in *figurative language, humor, emotion, empathy*, and *toxicity detection*, but it does not cover *dialect, discourse relations, character tropes, event detection, ideology, misinformation, persuasion, politeness, power relations, semantic change*, or *stance*. BIG-Bench covers few document-level analyses and no conversation-level analysis. We are the first to run extensive experiments to understand patterns of LLM performance on tasks critical to social scientists.

here because syntactic and morphological features are all defined on subtrees of the sentence node (Ziems et al. 2023; Eisenstein et al. 2023).

We evaluate on the Indian English dialect feature detection task of Demszky et al. (2021) 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—namely, 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. For example, the sentence *"Two years I stayed alone"* exemplifies *Preposition Omission*.

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, Mao, and Zeng 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), who use Plutchik's 8 emotional categories: *anger, anticipation, disgust, fear, joy, sadness, surprise,* and *trust*. For example, the following sentence would express *fear*:

> I started the steroids on Saturday and I had some really bad side effects, like my eyes started feeling weird.

Plutchik's model is one of the three most recognized discrete emotion models, and it is also used in our later Emotion Summarization Task (Section 3.4.1).

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 content (Jacobs and Kinder 2018) and political texts (Huguet Cabot et al. 2020), detecting hate speech (Lemmens, Markov, and Daelemans 2021), and identifying mental health self-disclosure (Iyer et al. 2019).

We use the FLUTE (Chakrabarty et al. 2022) benchmark because it is, at this time, the most comprehensive, with examples from many prior datasets (Chakrabarty et al. 2021; Srivastava et al. 2023; Stowe, Utama, and Gurevych 2022). FLUTE contains 9k premise sentences, each paired to a hypothesis with figurative language:

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

The classification task is to recognize whether the figurative sentence contains (1) *sarcasm* (Joshi, Bhattacharyya, and Carman 2017), (2) *simile* (Niculae and Danescu-Niculescu-Mizil 2014), (3) *metaphor* (Gao et al. 2018), or (4) an *idiom* (Jochim et al. 2018). In the above example, the hypothesis contains the idiom "put your heads together."

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*. For example, the following sentence contains *white grievance*: "jewish harvard professor noel ignatiev wants to abolish the white race."

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, Lee, and Wong 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* Web site, and a set of short jokes from Kaggle, summing to ∼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, Kogut, and Naidu 2014). While political scientists care most about identifying the underlying ideologies and partisan organizations behind these actors (Section 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. For example, one sentence associated with a strongly *conservative* ideology is: "*the feminist movement, with its mockery of marriage and demands for absolute sexual freedom. . . was a frontal assault on the meritocracy and the traditional family.*"

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. For example, an article with the headline "*White House Ousts Top Climate Change Official*" is marked as likely to contain misinformation.

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. Section 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, Danescu-Niculescu-Mizil, and Jurafsky 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. An example of an *Evidence* sentence is *"There is a Pizza Hut and a Dominos near me,"* since it provides concrete facts relevant to the request. An example of Scarcity is *"I haven't had a meal in two days."*

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 six topics: *atheism, climate change, the feminist movement, Hillary Clinton, Donald Trump*, and the *legalization of abortion.* Stance is given as *favor*, *against*, or *none*. For zero-shot experiments, we use the test set where the target is Donald Trump for all evaluations. An example tweet *against* Donald Trump is as follows:

> @realDonaldTrump needs to learn when to stop talking. You are making it worse Donald…so much worse.

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, Dowman, and Griffiths 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).[2]

We evaluate LLMs as binary word-sense discriminators using the popular Temporal Word-in-Context benchmark (TempoWiC; 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. For example, consider the different senses of the word "impostor" in the following texts.

[2] Additional works in this area can be found under the 3rd Workshop on Computational Approaches to Historical Language Change. `https://www.aclweb.org/portal/content/3rd-workshop-computational-approaches-historical-language-change`.

> 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

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) can have as many as 40 categories, tailored to spoken communication. We use Zhang, Culbertson, and Paritosh's (2017) simpler and more focused 9-class taxonomy 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, Culbertson, and Paritosh 2017). The model input is a comment from a Reddit thread, along with the utterance to which the comment is responding. For example,

> [userABC]: So i went thinking to myself this fine day ḧey lets check out Levetihanänd then i found out that this DLC does not appear in my Origin store...
>
> [userXYZ]: Did you go into ME3 game and access d̈ownloadable content?̈

The expected output is the category from the above 9-class taxonomy which best describes the comment's speech act: *Question* in the above example. Since *Announcements* and *Negative reactions* have fewer than 10 examples total in the dataset, 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 resources, and this motivates our evaluation on the TalkLife dataset of Sharma et al. (2020), a clinically motivated empathy detection dataset. The paper's 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.* Consider the example conversation:

> 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.

The above response is an example of *No exploration.*

3.2.3 Persuasion. In Section 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, Smith, and Denton Jr 2012).

We evaluate our persuasion prediction task on the Winning Arguments Corpus (Tan et al. 2016), which contains 3,051 conversations from `r/ChangeMyView` in which the persuader tries to convince the persuadee to change their mind. Models receive as input the reply thread (starting from a top-level comment) and perform binary prediction on whether the persuadee awarded the persuader a "delta" for a successful argument: *If you were the original poster, would this reply convince you?* Consider this example of an unsuccessful argument by UserA below:

> [UserA]: "Right on red", when it's allowed, is primarily because you're going from the innermost lane TO the intersecting innermost lane... this part of the state is infamous for a**hat drivers blocking people from turning, changing lanes, etc... could you imagine the chaos? :D
>
> [UserB]: Er, the image I posted, I found on Google images. It may triple the number of lanes to be concerned about, but one of them shouldn't usually be a problem if people stay in their lanes. And the other two, you still have to look in the same direction."

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, Reid, and Rambow 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. In the following example, EditorA uses a high degree of politeness and hedging langauge, which indicates that he is not in a position of power:

> [EditorA]: 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.
>
> [EditorA]: That could be the case. I've seen a few of those tonight.

3.2.5 Toxicity Prediction. Toxicity is a major area of social research in online communities, as online disinhibition (Suler 2005) makes antisocial behavior especially prevalent (Cheng, Danescu-Niculescu-Mizil, and Leskovec 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, Blackburn, and Han 2015). Even without 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. For example:

> [UserA]: 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.
>
> [UserB]: In fact, Pappe is already mentioned twice in the proper place.

The conversation above contains overt confrontation that will later devolve into toxicity.

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, Section 3.1.4), power inequality (Danescu-Niculescu-Mizil et al. 2013, Section 3.2.4), and persuasion (Shaikh et al. 2020, Section 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. For example,

> I am looking for help improving the dermatology content on Wikipedia. Would you be willing to help, or do you have any friends interested...

Each request is classified into one of three categories, *Polite, Neutral,* or *Impolite*, according to Amazon Mechanical Turk annotators' interpretation of workplace norms (the example above is *Polite*). 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 social science. Like conversations, documents can encode sequences of ideas or temporal events, as well as interpersonal relationships not present in isolated utterances. Unlike the dyadic communication of a conversation, a document can be analyzed under a unified narrative (Piper, So, and Bamman 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 (Section 3.3.3), the *events* they portray (Section 3.3.1, Section 3.3.2), as well as the *agents* involved and the specific *roles* or character tropes they exhibit (Section 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. For example:

> 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 looked great as ever.

The above lines A and D denote new events.

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, Park, and Bamman 2019), as well as in the wider domains of legal studies (Shen et al. 2020), public health (Jenhani, Gouider, and Said 2016), and policy.

WikiEvents (Li, Ji, and Han 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. An example of this task is to take the following document

> 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...

and produce the following structured output

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

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, Card, and Jurafsky 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 (Section 3.1.6), this task provides an entire news article as context. This tests the ability of the model to understand the relationship that a *sequence* of stances taken across an entire article might have with political leaning. For example, one Left-leaning article in this dataset contains the strongly indicative phrase: "*it was hard not to think about the insularity and cossetting the super-wealthy enjoy,*" and then goes on to talk at length about former LA Clippers owner Donald Sterling. In other articles, political views are more diffuse and less starkly concentrated into particular phrases or slogans. Still, each article must be classified into exactly one of the three ideological categories above.

3.3.4 Roles and Tropes. 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, So, and Bamman 2021).

Others have considered character role labeling for narratives (Jahan, Mittal, and Finlayson 2021) and news media (Gomez-Zara, Boon, and Birnbaum 2018). We evaluate this task with the CMU Movie Corpus dataset from Bamman, O'Connor, and Smith (2013) as it was extended and modified by Chu, Vijayaraghavan, and Roy (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 character's identity best fits (e.g., the *coward* or the *casanova*). The following example quotations are from an *absent-minded professor*:

> Now, THIS makes any fabric instantly impervious. Dirt proof, stain proof. . . Ouch! And bullet proof! It's still not perfected yet! It's hell on the dry-cleaning bill.
>
> This baby is the ultimate corrosive. I call it - DON'T TOUCH IT! - I call it "hydrochloricdioxynucleocarbonium". Well, the name needs work. But it'll eat through a Buick! OR -

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 and summaries. This section describes our natural language generation tasks, where LLMs might be used to summarize relevant aspects (Section 3.4.1), elucidate the hidden social meaning behind a text (Section 3.4.2–Section 3.4.4), or implement social theory by stylistically restructuring an utterance (Section 3.4.5).

3.4.1 Emotion-Specific Summarization. Prior work has already demonstrated LLMs' skill at generic summarization tasks (Goyal, Li, and Durrett 2022; Qin et al. 2023). Here, we consider a more domain-specific task, **aspect-based summarization**. The key idea of aspect-based summarization is that different elements of a document will be relevant to different users. This is especially true for social scientists and other domain specialists. For example, scientists who study population-level emotional responses to crisis events will need to know not only which emotions are represented in text (i.e., emotion detection; Section 3.1.2), but also, in brief, what specific experiences triggered such emotions. The scientist may not have highly focused queries like in QA tasks, but this use still demands summaries about broad subtopics or themes (Ahuja et al. 2022).

We use the CovidET dataset of Zhan et al. (2022), which contains 1,883 Reddit posts from the COVID-19 pandemic. Given a post and one of Plutchik's target emotions, the task is to summarize from the post what triggered the author to feel the target emotion.

3.4.2 Figurative Language Explanation. Our interests in figurative language are covered in Section 3.1.3, where we introduce the FLUTE dataset. 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."

3.4.3 Implied Misinformation Explanation. Both scientific understanding and real-world intervention strategies depend on more than black-box classification. This motivates our implied statement generation task, which is specified in the Misinfo Reaction Frames corpus (Gabriel et al. 2022) as covered in Section 3.1.7. 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."*

3.4.4 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.

3.4.5 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 using CBT strategies without contradicting the original speaker meaning. Take the example source sentence:

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

Using the *growth* and *neutralizing* strategies, the author can reframe this thought more positively as follows:

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

4. Evaluation Methods

4.1 Model Selection and Baselines

Our goal is to evaluate LLMs in **zero-shot settings through prompt engineering** (Section 4.2) and to identify suitable model architectures, sizes, and pre-training/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 `gpt-3.5-turbo` (Qin et al. 2023; Gilardi, Alizadeh, and Kubli 2023), which is the conversation-based LLM trained through RLHF (Christiano et al. 2017). Finally we include **GPT-4** (OpenAI 2023), which is a multimodal model that, at 1.7 trillion parameters, scales up the GPT-3 architecture by 1000×.

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 F1 score 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. These baselines will prove competitive in Table 3, matching or exceeding the best reported performances from the original publications in *Event Surprisal, Event Argument Extraction*, and the classification of *Emotions, Empathy, Figurative Language, Implicit Hate, Persuasion, Persuasion Strategies, Political Ideology*, and *Semantic Change*. Still, it is important to note that greater performances might be achievable by fine-tuning alternative architectures.

4.2 Prompt Engineering

One strength of 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 these instructions (Sanh et al. 2021; 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.

The author who is most familiar with the task writes an initial prompt for it based on the task description and the *design guidelines* below. Then we generate four semantically equivalent perturbations of that prompt using `gpt-3.5-turbo` as a zero-shot paraphrase model. All results are averaged across **prompt perturbations** to remove instruction-based variance (Perez, Kiela, and Cho 2021; Zhao et al. 2021).

In order to receive consistent, reproducible results we utilize a temperature of zero for all LLMs. For models that provide probabilities directly, we constrain decoding to the valid output classes.[3] For other models, such as `gpt-3.5-turbo`, we use logit bias to encourage valid outputs during decoding.[4] All other generation parameters are left at the default settings for each model.

CSS Prompt Design Guidelines. 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 guidelines in Table 1, drawn from both the literature and our own experience with non-CSS tasks as NLP researchers. We explicitly share these guidelines to help CSS practitioners control LLMs for their purposes. There is no claim

[3] Probability outputs for HuggingFace and GPT-3. `https://platform.openai.com/docs/api-reference/completions/create#completions/create-logprobs`.

[4] Logit Bias reference for `gpt-3.5-turbo`. `https://platform.openai.com/docs/api-reference/chat/create#chat/create-logit_bias`.

Table 1
LLM Prompting Guidelines 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.

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 new line (⏎) 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:

that the resulting prompts are optimally-engineered for each task; they instead provide reasonable approximations to the kinds of prompts a non-AI expert could design after considering established guidelines. By averaging our results over five prompt pertubations, we reduce the variance in this approximation of standard CSS tool-use.

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 2. All datasets, prompts, and model outputs are released for future comparison and analysis.[5]

4.4 Evaluation Metrics

Automatic Evaluation. Each dataset has a different structure with a different number of labels (see Table 2), so the use of accuracy is not the most informative metric. Instead, we compute the macro F1 score for all classification and structured parsing tasks and average over the 5 prompt perturbations. 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 (Section 4.2), evaluation scripts use straightforward string-matching.

[5] Data directory of our Github Project.

Table 2
Dataset size and classes count across all selected CSS benchmarks. Datasets are sorted by class count for each task category.

Dataset	Size	Classes
Generation Tasks	500	–
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

Dataset	Size	Classes
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

For high-variation domains like our generation tasks, on the other hand, word-overlap-based machine translation metrics like BLEU (Post 2018) are expected to have low correlation with human quality judgments (Liu et al. 2016). Here, even embedding-similarity metrics like BERTScore (Zhang et al. 2020) may be insufficient (Novikova et al. 2017). Manual inspection revealed high-quality generation outputs, but the BLEURT (Sellam, Das, and Parikh 2020) score reported zero semantic overlap, and variation in BLEU and BERTScores failed to follow any discernible patterns with regard to model preference or scaling laws that we observed by manual inspection. For generation tasks, human evaluation is strongly preferable (Santhanam and Shaikh 2019), and we describe human evaluation in the following paragraphs.

Human Scoring Evaluation. To get a sense of the generation quality for each task, we recruit a domain expert to blindly evaluate 100–400 model outputs. Evaluations are on 1–5 Likert scales for 4 standard metrics from the NLG literature. Following Fabbri et al. (2021), we define these metrics as follows:

- **Faithfulness:** *The generation is consistent with the source document and with the definition of the task.*
- **Coherence:** *The generation is well-structured and well-organized. The generation is not just a heap of unrelated information, but forms a coherent body of information about a topic.* (Dang 2005)
- **Relevance:** *The generation includes only important information from the source document; no redundancies or excess information.*
- **Fluency:** *The generation has no formatting problems, capitalization errors, or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.*

All experts are recruited and paid through the Upwork platform. Annotator backgrounds and expertise are summarized in the bottom right pane of Table 6. For CovidET summarization, we recruit a former CDC health communication specialist with a B.S. in Public Health and an M.S. in Health Education. For *Misinformation*, we enlist a Public Policy Graduate Student with a B.A. in Political Science. For *Figurative Language*, we hire a former writing expert at Grammarly with an M.F.A. For *Social Bias*, we find a Graduate Student with a B.S. in Journalism. And for *Positive Reframing*, we hire a Nurse in Clinical Behavioral Health with a B.A. in Psychology.

Human Ranking Evaluation. Instead of scoring or rating target generations on a standard Likert scale, annotators can also rank the model explanations in terms of their *faithfulness* at describing the target construct. The ranking-style evaluation can be less variable than scoring for generation tasks (Harzing et al. 2009; Belz and Kow 2010). We will use Social Bias Frames as an example to illustrate the general setup for Human Ranking Evaluations. Here, the annotator reviews a *hateful message* and an associated *hate target*. 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 best to 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.

Pilot annotation proved that crowdworker evaluations can exhibit high variance and instability due to cultural and individual differences, as well as different interpretations of the task. Thus the authors served as blind annotators for this evaluation. Two authors evaluated each task and unanimous voting determined the reported metrics.

5. Classification Results

Table 3 presents all zero-shot results for utterance, conversation, and document-level tasks. We use these results to answer RQ 1–3. The results suggest that LLMs are a *viable tool for augmenting human CSS annotation*. For classification tasks, results show that *larger, instruction-tuned open-source LLMs are preferable.*

5.1 Viability (RQ1)

5.1.1 Zero-Shot Viability. To understand the viability of LLMs as CSS tools, we ask if zero-shot LLMs match or exceed the reliability of human annotation. If the overall performance is high and the expected agreement between humans and prompted models is as high as that between humans alone, then we might expect LLMs to viably augment the human annotation process. According to one paradigm, an LLM can serve as just one of many human and AI labelers, and gold labels would be decided by majority vote across these independent annotations. According to another complementary paradigm, LLM pseudo-labels can be used with a small set of gold-labels to compute unbiased estimators in downstream regressions following methods like Design-based Semi-supervised Learning (DSL; Egami et al. 2023) and others.

Table 3 shows the best zero-shot models achieve as high as 77.4 F1 on misinformation detection. On this task, the best-performing FLAN-UL2 model achieves an agreement of $\kappa = 0.55$ with gold labels (see Table 4), which is higher than the $\kappa = 0.51$ inter-human agreement reported by Gabriel et al. (2022) in their original paper. In fact, for 8 of 17 tasks in Table 4, or 47% of classification tasks, models achieve moderate

Table 3
Zero-shot classification results across our selected CSS benchmark tasks. All tasks are evaluated with macro F-1, which is averaged across 5 prompt permutations for zero-shot models. Supervised baseline results are averaged over 3 random seeds. Best zero-shot models are in green. A dash indicates a model did not follow instructions.

Model	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
Data	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
								Utterance Level Tasks								
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0
								Conversation Level Tasks								
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4
								Document Level Tasks								
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	–	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

Table 4
(Acc.) Best model F1 scores. F1 scores above 70% are bolded. (κ) Agreement scores between zero-shot model classification and human gold labels. Out of ten utterance-level tasks, six have at least moderate M and only two have poor agreement P. Three (50%) of the conversation tasks have at least fair agreement F.

Dataset	Best Model	F1	κ	Agreement
		Utterance-Level		
Dialect	flan-ul2	32.9	0.15	poor
Emotion	flan-ul2	**70.8**	0.65	good
Figurative	flan-ul2	62.3	0.52	moderate
Humor	gpt-4	61.3	0.23	fair
Ideology	davinci-002	60.0	0.40	moderate
Impl. Hate	flan-ul2	32.3	0.20	fair
Misinfo	flan-ul2	**77.4**	0.55	moderate
Persuasion	gpt-4	56.4	0.51	moderate
Semantic Chng.	flan-t5-large	56.9	0.14	poor
Stance	gpt-3.5-turbo	**72.0**	0.58	moderate

Dataset	Best Model	F1	κ	Agreement
		Convo-Level		
Discourse	flan-t5-xxl	50.6	0.45	moderate
Empathy	flan-t5-xxl	35.9	0.04	poor
Persuasion	davinci-003	53.9	0.14	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	gpt-4	59.7	0.26	fair
Toxicity	gpt-4	55.4	0.11	poor
		Document-Level		
Ideology	gpt-4	51.5	0.51	moderate
Event Det.	gpt-4	23.0	n/a	-
Tropes	gpt-4	44.9	n/a	-

to good agreement scores ranging from κ = 0.40 to 0.65. These tasks also correspond to the highest absolute performances on *Stance* (76.0 F1), *Emotion* (70.8 F1), *Figurative Language* (62.3 F1), and utterance-level *Ideology Classification* (60.0 F1). *In these cases of high viability, we recommend that CSS researchers consider the DSL and augmented-annotator*

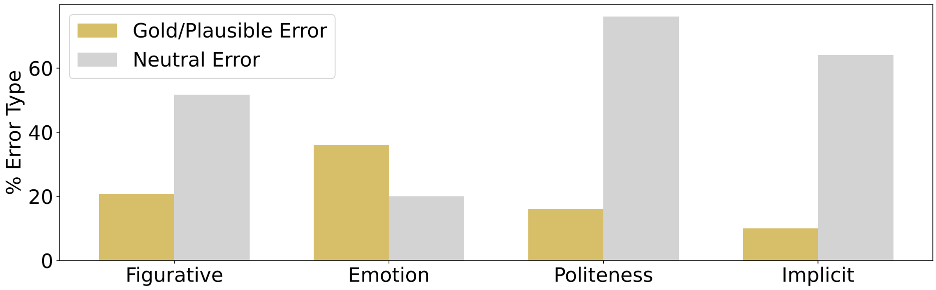


Figure 2
Breakdown of shared error types. For a representative subset of classification tasks, we conduct an analysis of up to 50 shared errors across evaluated models. We focus specifically on the best performing model in a class (e.g. the best variant of FLAN models or the best OpenAI model). Plausible/gold 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).

paradigms. Such strong performances are on tasks that 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). Our qualitative error analysis in Section 5.1.2 will show that here, models are less likely to default to neutral categories, and errors are more likely to come from annotation mistakes in the gold dataset according to the author's own manual error analysis (see lower neutral and higher gold error in Figure 2).

Are zero-shot models ready to label text out-of-the-box? Zero-shot results rarely exceed the carefully tuned supervised RoBERTa baselines in Table 3. However, the best observed performances here match that of classifiers used in published studies on stance detection (67.8 F1; Zarrella and Marsh 2016), COVID-19 vaccination opinions (Cotfas et al. 2021), political opinions (Siddiqua, Chy, and Aono 2019), and debates (Lai et al. 2020). In such scenarios, *zero-shot models could offer a data-efficient alternative to fine-tuned models* by removing the need for expensive training sets. Humans could focus all of their efforts on validating LLM outputs and tuning prompts (Section 4.2) rather than coding unstructured text. However, we encourage practitioners to proceed cautiously, especially in sensitive domains, and we recommend human-in-the-loop methods to mitigate bias and risk. See Section 7.7 and Section 7.8 for further discussion.

It is important to consider that LLM performance could be unusably poor for some CSS tasks. For a non-negligible subset of tasks we considered, LLMs have poor agreement (5/17 = 29.5%), and here social scientists might not consider zero-shot annotation augmentation via LLMs. On these poor agreement tasks, LLM absolute performance is not significantly better than random guessing: see 56.9 F1 vs 50 Random on *Semantic Change*; 35.9 F1 vs 33.3 Random on *Empathy*; 53.9 F1 vs 50 Random on conversation-level *Persuasion*; and 55.4 F1 vs 50 Random on *Toxicity*. Some of these low-performance tasks like *Event Argument Extraction* are structurally complex and may require additional engineering efforts. Others like *Empathy* and *Tropes* have challenging and subjective expert taxonomies whose semantics differ from definitions learned in model pretraining. This is confirmed by our error analysis in Figure 2 where GPT3.5 often defaults to the neutral, more colloquially recognizable label *stereotype* (64% of errors) rather than use a more taxonomy-specific label like *white grievance*. In Section 5.1.3, we test if few-shot prompting can reduce misalignments between model and ground-truth definitions.

5.1.2 Zero-Shot 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 break down the error types for `gpt-3.5-turbo`.

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 10 unanimous errors, 3 were idioms mistaken as metaphors, and 7 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.

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 of 50 examples *across all evaluated models* were judged as gold mislabels. Additionally, for FLAN-UL2 and ChatGPT, 17 of 50 and 15 of 50 predictions, respectively, could be considered as valid—even if they differed from the gold label.[6]

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

Politeness Prediction. We first visualize the per-category accuracy of the different best-performing models (FLAN-T5-XL, `Text-davinci-002`, and ChatGPT). We observe that: (1) The XL model tends to predict more polite labels. It is more accurate in terms of the utterances that were polite and neutral with 70.4% and 62.0% accuracy. Most errors come from impolite cases (with a 45.2% accuracy). (2) `davinci-002` performs best in judging neutral utterances. `davinci-002` is the most accurate for neutral utterances

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

(82.9% accuracy) while making significantly more errors for polite and impolite utterances (43.9% and 40.9% accuracy, respectively). (3) ChatGPT performs worst in finding impolite utterances while making more neutral predictions, with only 9.0% accuracy for the impolite category, whereas it achieves 75.9% and 66.8% for neutral and polite cases.

We then consider the 81 of 498 cases where the three models are all making errors. We find that the three models make the same errors in most cases (54/81) and `davinci-002` models make errors more similar to ChatGPT (17/81 cases). Among these common error cases, we observe that 79 of 81 cases are related to the 1st and 2nd person mention strategy (Danescu-Niculescu-Mizil et al. 2013) and all of them are direct or indirect questions, while 38 of 81 are related to counterfactual modal and indicative modal (Danescu-Niculescu-Mizil et al. 2013). This indicates that all three models struggle with direct or indirect questions with 1st and 2nd person mentions.

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).

5.1.3 Few-Shot Viability. Zero-shot models may not be naturally aligned with the non-standard or technical meanings of certain key terms in the social sciences. To address this issue, we consider the viability of open-source FLAN models for few-shot classification. Specifically, we try 3-shot and 5-shot experiments with no further prompt engineering. Table 5 shows that any improvements from these methods are spotty and inconsistent. For some challenging tasks like Empathy and Persuasion that have subjective definitions or non-standard taxonomies, few-shot learning can improve performance in 2 and 5 out of 6 model sizes, respectively. However, these gains are small and not widespread among other tasks. We conclude that *additional engineering efforts may be needed to achieve significant gains on CSS tasks via few-shot learning.*

5.2 Model-Selection (RQ2)

CSS researchers should understand how their choice of model can decide the reliability of zero-shot methods. Our results show that, for structured parsing tasks like event extraction, it is best to use a code-instructed model like OpenAI's `gpt-3.5-turbo`, 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 open-sourced models are preferable.

FLAN's CSS performance roughly matches Kaplan et al.'s predicted power-law effects from pure model size. Figure 3 shows FLAN classification performances scaling

Table 5
Few-shot classification does not uniformly improve performance across our selected CSS benchmark tasks. All tasks are evaluated with macro F-1.

Model	FLAN Small			FLAN Base			FLAN Large			FLAN XL			FLAN XXL			FLAN UL2		
Shot	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5
Dialect	0.2	0.0	**0.4**	**4.5**	0.0	1.4	**23.4**	0.7	14.1	**24.8**	8.0	20.5	**30.3**	0.2	29.9	**32.9**	12.6	27.5
Emotion	**19.8**	10.6	10.1	**63.8**	42.7	42.0	**69.7**	67.6	67.4	**65.7**	62.1	62.5	**66.2**	61.8	57.4	**70.8**	70.0	69.8
Figurative	**16.6**	10.0	9.2	23.2	**29.1**	27.3	18.0	**21.8**	19.6	**32.2**	27.9	28.5	53.2	52.6	**66.2**	**62.3**	52.7	62.0
Humor	51.8	52.8	**53.1**	**37.1**	35.1	34.7	**54.9**	54.0	53.8	56.9	**57.0**	56.7	29.9	34.8	**35.3**	**56.8**	55.5	54.1
Ideology	18.6	16.7	**24.0**	**23.7**	22.6	38.3	43.0	**47.3**	45.5	47.6	**48.8**	50.4	53.1	52.9	**57.7**	46.4	36.9	**51.5**
Impl. Hate	**7.4**	6.8	6.2	14.4	**21.1**	7.4	7.2	**9.3**	4.7	32.3	28.5	**34.6**	29.6	31.6	**35.1**	**32.0**	29.5	25.9
Misinfo	**33.3**	33.3	33.3	53.2	45.3	**59.7**	**64.8**	64.8	64.2	68.7	67.2	**69.7**	69.6	**74.9**	74.4	**77.4**	53.7	76.4
Persuasion	**3.6**	3.6	3.6	10.4	**10.8**	7.3	37.5	**39.0**	37.7	32.1	**44.3**	41.8	45.7	44.6	**48.6**	**43.5**	42.2	40.1
Sem. Chng.	33.5	33.3	**34.0**	41.0	35.7	**41.7**	56.9	48.8	**60.4**	**52.0**	40.8	35.6	**36.3**	34.0	33.3	41.6	**62.5**	34.6
Stance	25.2	16.7	**29.6**	**36.6**	18.1	36.6	**42.2**	41.8	39.8	43.2	**52.1**	46.2	**49.1**	46.0	48.7	48.1	**55.6**	54.7
Discourse	4.2	4.0	**7.5**	**21.5**	18.1	20.7	33.6	3.6	**34.6**	37.8	3.6	**38.0**	**50.6**	3.6	43.4	**39.6**	3.6	39.1
Empathy	**16.7**	16.7	16.7	**16.7**	16.7	16.7	**22.1**	16.7	17.1	21.2	**30.4**	22.8	**35.9**	29.8	28.2	34.7	**41.5**	39.6
Persuasion	9.2	**55.9**	45.0	11.0	**55.0**	48.7	11.3	**54.6**	51.7	8.4	42.8	**43.8**	**41.8**	38.8	35.2	43.1	**44.9**	46.1
Politeness	**22.4**	16.7	20.1	**42.4**	23.9	35.4	44.7	44.5	**51.9**	**57.2**	27.7	50.4	**51.9**	44.2	50.3	53.4	43.6	**53.9**
Power	**46.6**	44.5	33.3	**48.0**	39.8	41.4	40.8	**45.5**	43.5	55.6	58.9	**60.2**	52.6	52.0	**62.6**	56.9	57.2	**57.5**
Toxicity	43.8	**46.7**	33.3	40.4	34.7	**54.4**	**42.5**	34.7	36.7	43.4	38.7	**49.2**	34.0	33.3	**35.1**	48.2	44.7	**52.5**
Ideology	**24.0**	16.7	19.2	19.2	16.6	**21.3**	**28.3**	17.0	17.9	29.0	**31.7**	27.0	42.4	**48.5**	47.9	38.8	**38.9**	39.7
Tropes	1.7	**5.1**	3.4	**8.4**	5.1	3.4	**13.7**	10.0	11.6	**14.6**	8.4	10.0	**19.0**	8.4	6.8	**28.6**	27.3	24.6

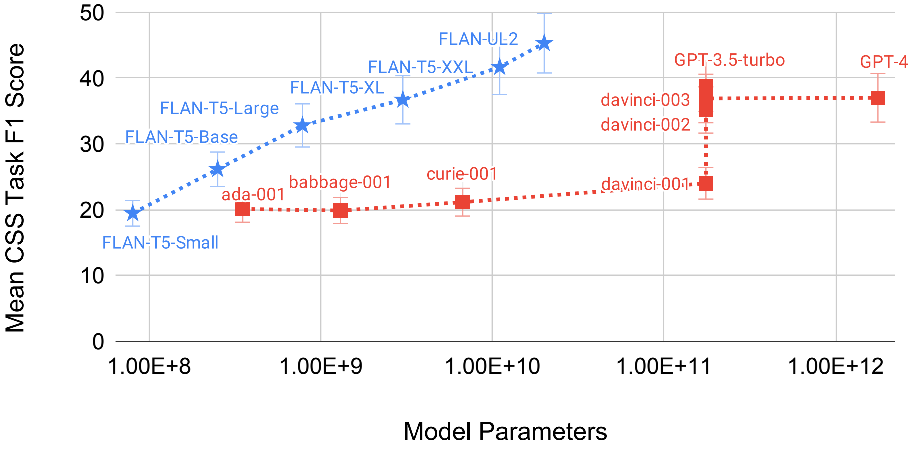


Figure 3
Effects of scaling on the mean zero-shot performance on our CSS benchmark tasks. FLAN models and `davinci-001/002` are instruction fine-tuned. `davinci-003` and `gpt-3.5-turbo` are instruction fine-tuned and refined with Reinforcement Learning from Human Feedback. GPT Parameter counts reported based on approximates.

nearly logarithmically with the parameter count. With each order of magnitude size increase, the median average task improvement in FLAN models is 5.0 absolute percentage points. 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.[7] There is minimal performance improvement from `ada` to `davinci-001`, despite the three orders of magnitude increase in size. Similarly, we observe little benefit from the trillion-parameter GPT-4 over the hundred-billion-parameter GPT-3.5-turbo. Instead, the greatest performance improvements come from variations in *pretraining, fine-tuning*, and *reinforcement learning*.

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 3), we see an outsized increase in CSS performance (↑ absolute 11 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 benefits structured, JSON-formatted tasks like Event Argument Extraction. 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 can improve LLM performance on CSS classification tasks.* RLHF has been lauded as the major catalyst behind the success of instruction-following models (Ouyang et al. 2022), and here we see `text-davinci-003` and `gpt-3.5-turbo` (with RLHF) improves the average F1 of `text-davinci-002` (without RLHF) by 3.5 absolute points.

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 3 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*

[7] 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.

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

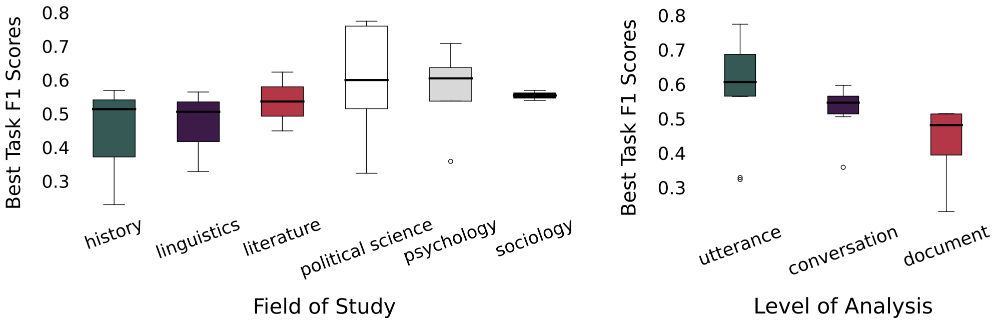


Figure 4
(*Left*) Task performance by field of study. Significant overlap in the distributions suggests that neither high nor low performance is exclusive to any particular discipline. *Caution:* The distributions depend on the particular choices of this study, which datasets to select, and how to partition them.
(*Right*) Task performance by level of analysis. Document-level tasks are challenging for their input length and complexity, and this is reflected in their F1 scores all near or below 50%. Utterance and conversation-level task performance varies also with the complexity of the task.

study. In political science, we see the highest overall performance on misinformation detection (F1 = 77.4) and much lower performance on implicit hate detection (F1 = 32.3). For psychology, we observe high performance on emotion detection (F1 = 70.8) and low performance on empathy detection (F1 = 35.9). Performances span the full range of disciplines. This suggests that performance is not tied to academic discipline.

In terms of data type, Figure 4 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 (F1 = 32.3), empathy (F1 = 35.9), and dialect feature (F1 = 32.9) annotations are expert-labeled on a subtle theoretical taxonomy.

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. Note that FLAN models are excluded from all evaluation tables because FLAN models failed to follow instructions by manual inspection. Instead, we evaluate across the OpenAI suite.

Human Scoring Evaluation. According to the domain experts in Table 6, leading generative models can produce text of a quality that matches or exceeds that of human gold references. For Aspect-Based Emotion Summarization, Misinformation Explanation, and Social Bias Inference, `gpt-3.5-turbo` produce, on average, more domain-faithful, coherent, and fluent text than both gold references and the fine-tuned baseline. For Positive Reframing and Figurative Language explanation, `text-davinci-003` match the

Table 6
Expert scoring evaluations for zero-shot generation tasks show that leading generative models (`davinci-003`, `GPT 3.5`) can match or exceed the faithfulness, relevance, coherence, and fluency of both fine-tuned models (Baseline) and gold references (Human). All scores are average ratings on 1-5 Likert scales. Best models are in green followed by blue. Marks for $^{+}$ and $^{-}$ show performance significantly better or worse than human ($P < .05$) by Paired Bootstrap.

Aspect-Based Summarization (CovidET)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	2.1	2.3	2.1^{-}	2.6^{-}
`ada-001`	1.8^{-}	1.8^{-}	2.4	3.6
`babbage-001`	2.0^{-}	2.0	2.3	3.7
`curie-001`	2.3	2.3	2.6	3.8
`davinci-001`	2.3	2.4	2.5	3.9
`davinci-002`	2.4	2.5	3.2	4.0
`davinci-003`	2.9	2.8	3.0	**4.1**$^{+}$
`GPT 3.5`	**3.9**$^{+}$	**3.5**$^{+}$	**3.8**$^{+}$	**4.5**$^{+}$
`GPT 4`	**3.7**$^{+}$	**3.3**$^{+}$	**3.8**$^{+}$	**4.4**$^{+}$
Human	2.8	2.6	2.8	3.8

Figurative Language Explanation (FLUTE)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	1.4^{-}	1.7^{-}	1.4^{-}	4.2
`ada-001`	1.4^{-}	1.5^{-}	1.5^{-}	3.9
`babbage-001`	1.4^{-}	1.9^{-}	1.5^{-}	3.9^{-}
`curie-001`	1.5^{-}	2.3^{-}	1.7^{-}	4.1
`davinci-001`	1.2^{-}	1.9^{-}	1.5^{-}	4.1
`davinci-002`	2.5	3.4	2.5	4.1
`davinci-003`	3.0	4.0	3.1	**4.1**$^{+}$
`GPT 3.5`	2.1^{-}	3.6	2.5	4.1
`GPT 4`	2.1^{-}	3.3	2.4	4.0
Human	2.8	4.0	2.6	4.2

Positive Reframing

Model	Faithful	Relevant	Coherent	Fluent
Baseline	4.1	4.2	3.9	4.4
`ada-001`	1.8^{-}	1.4^{-}	1.8^{-}	1.6^{-}
`babbage-001`	3.8	2.5^{-}	3.8	3.7
`curie-001`	4.1	3.7^{-}	4.1	3.9
`davinci-001`	3.5^{-}	4.0	3.3^{-}	4.1
`davinci-002`	4.0	3.9^{-}	4.0	4.2
`davinci-003`	4.4	**4.5**$^{+}$	4.2	**4.6**$^{+}$
`GPT 3.5`	4.3	4.3	4.2	4.4
`GPT 4`	4.1	4.3	4.1	4.2
Human	4.2	4.2	4.1	4.2

Implied Misinformation Explanation (MRF)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	3.4	3.5	3.7	4.2
`ada-001`	1.1^{-}	1.1^{-}	2.0^{-}	4.5
`babbage-001`	1.6^{-}	1.7^{-}	2.5^{-}	4.3
`curie-001`	2.6^{-}	2.7^{-}	3.1^{-}	4.4
`davinci-001`	1.7^{-}	1.7^{-}	2.5^{-}	4.5
`davinci-002`	**3.9**$^{+}$	**4.1**$^{+}$	**4.3**$^{+}$	**4.9**$^{+}$
`davinci-003`	3.1^{-}	3.4	3.9	4.5
`GPT 3.5`	**3.7**$^{+}$	3.9	**4.2**$^{+}$	**4.9**$^{+}$
`GPT 4`	3.7	3.9	4.1	4.5
Human	3.5	3.7	3.9	4.4

Social Bias Inference (SBIC)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	1.9^{-}	2.1^{-}	2.1^{-}	1.9^{-}
`ada-001`	2.4	2.2^{-}	2.7	**3.3**$^{+}$
`babbage-001`	3.1	3.1	**3.6**$^{+}$	**3.8**$^{+}$
`curie-001`	3.4	3.3	**3.9**$^{+}$	**4.5**$^{+}$
`davinci-001`	3.4	3.4	**3.8**$^{+}$	**3.9**$^{+}$
`davinci-002`	**3.7**$^{+}$	3.5	**4.1**$^{+}$	**4.2**$^{+}$
`davinci-003`	3.5	3.4	**4.1**$^{+}$	**4.4**$^{+}$
`GPT 3.5`	**4.0**$^{+}$	**3.7**$^{+}$	**4.2**$^{+}$	**4.2**$^{+}$
`GPT 4`	**4.1**$^{+}$	**3.8**$^{+}$	**4.2**$^{+}$	**4.6**$^{+}$
Human	2.9	3.0	3.1	2.6

Annotator Backgrounds

Task	Education	Profession
CovidET	MS, Health Ed.	CDC Health Comm. Specialist
MRF	BA, Poli. Sci.	Grad Student, Public Policy
FLUTE	MFA, Creat. Writing	Writing Expert, Grammarly
SBIC	BS, Journalism	Grad Student, Epidemiology
Reframing	BA, Psychology	Clinical Behavioral Health, Nurse

gold-standard levels of faithfulness, relevance, coherence, and fluency, again outperforming the fine-tuned baseline. For generation performance, scale matters, as smaller models fail to produce explanations and summaries that are faithful to the task specifications, especially in CovidET, FLUTE, and SBIC, where `text-ada-001` earns faithfulness scores of less than 2 out of 5 on average. With scale, however, *LLMs are capable of generating useful, relevant, coherent, and fluent explanations and summaries of underlying social science constructs*.

Human Ranking Evaluation. According to the authors' ranking evaluations in Table 7, *prompted LLMs produce helpful and informative text in all five generation tasks.* Model generations outrank the dataset's gold human reference at least 38% of the time. The

Table 7
Ranking 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 .

	% Model Preferred Over Gold Annotations				
Model	MRF	FLUTE	SBIC	Reframing	CovidET
Baseline	31.2%	4.6%	16.5%	45.0%	37.5%
`ada-001`	17.6%	1.7%	11.8%	0.0%	23.5%
`babbage-001`	29.4%	6.7%	29.4%	0.0%	23.5%
`curie-001`	29.4%	1.7%	32.4%	11.5%	41.2%
`davinci-001`	21.4%	6.2%	39.0%	30.4%	50.0%
`davinci-002`	21.4%	25.0%	29.3%	10.0%	37.5%
`davinci-003`	38.9%	47.0%	50.0%	48.5%	59.1%
`GPT 3.5`	27.8%	37.9%	65.9%	56.1%	68.2%
`GPT 4`	36.4%	51.5%	60.6%	39.4%	36.4%

best models approach parity with humans—a near 50-50 coin toss to decide which is preferred. Furthermore, we see significant performance benefits from both RLHF models, `gpt-3.5-turbo` and `text-davinci-003`. Unlike classification (Section 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 fully replace human analysis*. Still, given LLM's performance parity with humans, the results suggest one avenue for human-AI collaboration: 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 7 suggest that, for every five model generations, 2 to 3 of these outputs would demand no additional annotator effort. If implemented successfully, this partnership could significantly increase the efficiency of the social scientist's research pipeline. However, we leave it to future work in HCI to determine the plausibility of this partnership and the degree to which it might reduce annotators' cognitive load on exploratory analysis and free-coding tasks.

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, Li, and Durrett 2022; Liang et al. 2022). In fact, we found that BLEU, BERTScore, and BLEURT, which rely on comparisons to human written ground truth, all produced uninterpretable scores for generation tasks. 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

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

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. However, this alternative is limited by both cost and reproducibility concerns (Karpinska, Akoury, and Iyyer 2021). There is a clear need for new metrics and procedures to quantify model utility for CSS.

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 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
3. Prioritize faithfulness, relevance, coherence, and fluency in your generations by opting for larger instruction-tuned models that have learned human preferences.
4. Investigate how LLMs produce new CSS paradigms built on the multipurpose capabilities of LLMs in the long term.

In the following subsections, we more specifically detail how annotation fits into the CSS pipeline (Section 7.1), how LLMs can augment annotation (Section 7.2), which LLMs we recommend for this purpose (Section 7.3), and the opportunities LLMs expose for new research paradigms (Section 7.4 and Section 7.5).

7.1 How Annotation Fits Into CSS

Social scientists are not often interested in classification labels or generative codes merely for their own sake. Labeled text is almost always used to explain a wider phenomenon using downstream inferential statistics such as regression. To ensure a valid downstream inference, any estimators of the underlying construct need to be asymptotically unbiased and allow for the computation of valid confidence intervals. In a study that follows up our own, Egami et al. (2023) demonstrate that even the most accurate of LLM annotations will produce biased estimates and invalid confidence intervals when they are merely averaged. However, Egami et al. (2023) propose a method called DSL which they use for computing unbiased estimators for prevalence estimation on each of our 17 classification tasks. Using FLAN-UL2 pseudo-labels and only a handful of gold-labeled instances, they find that it is possible to compute unbiased regressions, even for tasks with low accuracy from UL2. With only 25 gold annotations, DSL can compute confidence intervals that have >86% correct coverage for all of our classification tasks *including those with the lowest performance from LLM pseudo-labels.*

7.2 How LLMs Can Augment Annotation

Our work shows that current LLMs perform well enough to augment the CSS annotation pipeline and thus reduce the need for human labor on some tasks. Section 5.1 shows that an LLM annotator yields fair to strong agreement with humans on 12 out of 17 classification tasks. However, *LLMs are not a wholesale replacement for human annotators*. Even the best LLMs exhibit unusably low performance on some CSS tasks. Ensembling prediction does not mitigate this label corruption as LLMs demonstrate high internal agreement, even when inaccurate (Gilardi, Alizadeh, and Kubli 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 human annotation*, as illustrated in Figure 5. When a LLM matches human levels of agreement, it can be used as one of multiple labelers. It is possible from such methods to produce unbiased estimators. The estimation methods of Chaganty, Mussmann, and Liang (2018) are provably optimal, and provide these unbiased estimators with 7–13% monetary savings by leveraging LLM pseudo-labels.

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, Yu, and Bowman 2022), and active learning (Yuan, Lin, and Boyd-Graber 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 stronger at generation tasks*, being rated superior to human gold annotations over 38% of the time in all 5 tasks we evaluate. LLMs can already generate syntactically cohesive and stylistically consistent text, and as we have shown, the best generations are also highly relevant and faithful to core CSS tasks. Human expertise can be used to further curate outputs according to domain-specific

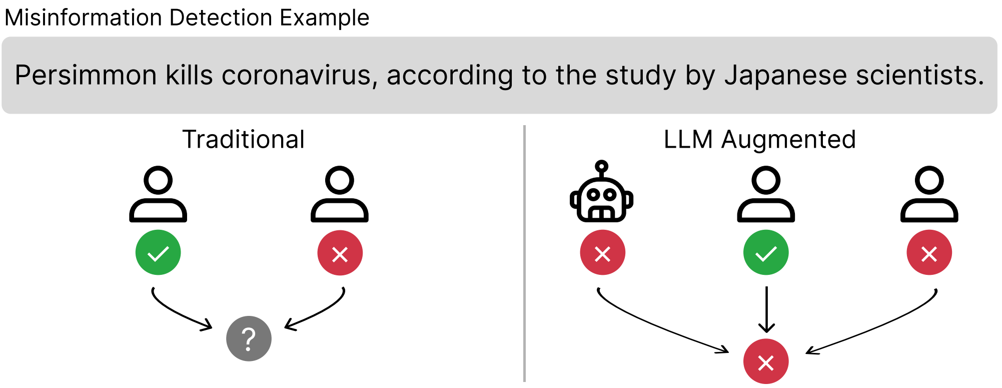

Misinformation Detection Example
Persimmon kills coronavirus, according to the study by Japanese scientists.
Traditional
LLM Augmented
?

Figure 5
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.

accuracy and quality metrics. 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 consider the use of LLMs as the foundation of such annotation procedures, and that future studies measure the degree to which this strategy improves annotation efficiency as we suspect. If the anticipated efficiency is achieved, CSS researchers should reinvest savings to train **expert annotators**, reversing the trends of replacing experts with crowdworkers due to cost (Snow et al. 2008). By doing so, LLMs could enable data labeling procedures that more deeply benefit from the non-computational expertise of the social scientists whose theories we build upon.

7.3 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 and closed-source models.

For CSS classification, our work shows that open-source models like FLAN are as capable as state-of-the-art closed-source LLMs from OpenAI. We recommend that researchers who already have access to GPUs capable of running these models prefer open-source models. For continuous monitoring and enormous-scale analysis, the low marginal cost of these models could make 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 behavior. At this time, it is not possible to further fine-tune all of OpenAI's instruction-tuned models.

For those without existing hardware infrastructure, proprietary APIs appear to be a cost-efficient option. Based on current cloud pricing,[10] the hardware necessary to run FLAN-T5-XXL costs $170 per day—the equivalent of processing ∼50 million words with `gpt-3.5-turbo`.[11] In most cases, `gpt-3.5-turbo` 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 *proprietary 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, `gpt-3.5-turbo` 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 Section 7.8).

Our work shows that *all LLMs struggle most with conversational and full document data*. Also, *LLMs currently lack clear cross-document reasoning capabilities*, limiting extremely common CSS applications like topic modeling. For CSS subfields that often 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

[10] Google Cloud FLAN hosting cost. `https://medium.com/google-cloud/deploy-flan-t5-xxl-on-vertex-ai-prediction-579953afdc88`.

[11] OpenAI pricing. `https://openai.com/pricing`.

technical challenges of conversations, long documents, and cross-document reasoning (Beltagy, Peters, and Cohan 2020; Caciularu et al. 2021; Yu et al. 2021).

7.4 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. An emerging 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 for CSS (Bail 2023), as they replicate human biases without explicit conditioning (Jones and Steinhardt 2022; Koralus and Wang-Maścianica 2023). Recent work leverages this capacity 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, Dodds, and Watts 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 behavior 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).

7.5 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 predefined 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, Akoury, and Iyyer 2021). We observe this directly in our work, as crowd work seems to provide unreliable quality metrics for FLUTE, a nuanced generative task. New metrics are needed to capture the semantic validity of free-form coding with LLMs as explanation-generators.

7.6 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 nonconventional language of **expert taxonomies**. Expert taxonomies contain technical terms like the dialect feature *copula omission* (Section 3.1.1), plus specialized or nonstandard definitions of colloquial terms, like the persuasive *scarcity* strategy (Section 3.1.8), or *white grievance* in implicit hate (Section 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*, Section 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, Tater, and Sankaranarayanan 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.

7.7 Issues in Bias and Fairness

Bias. 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, Farooqi, and Zou 2021;

Borchers et al. 2022; Lucy and Bamman 2021; Shaikh et al. 2022), as well as viewpoint biases in subjective domains (Santurkar et al. 2023). These biases can emerge in open-ended generation tasks like 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.

Risks Inherent to Proprietary LLMs. Researchers will often lack details on the selection, filtering, and formatting of online corpora for training proprietary models like OpenAI's GPT-4. There are unique risks and privacy implications for those who employ these models for research. One risk inherent in an obscured pre-training corpus is that we can't decompose the above issues of bias and social harm into their respective sources for targeted mitigation. Especially with open-ended generative coding (Section 3.4), unattributed biases could jeopardize the reliability and prosocial impact of downstream scientific analyses. Such issues may also escape human review, as humans are prone to falsely attribute *factuality* to texts that bear a more authoritative or expert *style* (Wu and Aji 2023)—a style that modern proprietary LLMs have largely mastered. Furthermore, black-box industrial APIs may not allow for targeted mitigation via fine-tuning.

Researchers who operate with closed-source industrial APIs may also be more prone to privacy issues and legal disputes over intellectual property. Proprietary models are known to replicate copyrighted materials and sensitive personally identifiable information (Carlini et al. 2021). Researchers who use these models may be unknowingly accountable for knowledge based on false or missing attributions as well as private personal information.

7.8 Limitations

Task Selection and Data Leakage. 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). Finally, our task selection was limited by the available data resources in the field, which is largely dominated by text in standard dialects (Ziems et al. 2023) representing members of Western, Educated, Industrial, Rich, and Democratic populations (WEIRD; Ignatow and Mihalcea 2016; Muthukrishna et al. 2020). Future studies should separately consider LLMs' utility for cross-cultural CSS.

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. This problem is especially concerning for closed-source or proprietary models with undisclosed training sets. 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.

Causality and Explanations. Explanations are important to social science (Shmueli et al. 2010; Hofman, Sharma, and Watts 2017; Yarkoni and Westfall 2017). In this work, we