What's in a Prompt?: A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, Libby Hemphill

University of Michigan School of Information
Ann Arbor, Michigan, USA
satreja@umich.edu

Abstract

Manually annotating data for computational social science tasks can be costly, time-consuming, and emotionally draining. While recent work suggests that LLMs can perform such annotation tasks in zero-shot settings, little is known about how prompt design impacts LLMs' *compliance* and *accuracy*. We conduct a large-scale multi-prompt experiment to test how model selection (GPT-4o, GPT-3.5, PaLM2, and Falcon7b) and prompt design features (definition inclusion, output type, explanation, and prompt length) impact the compliance and accuracy of LLM-generated annotations on four highly relevant and diverse CSS tasks (toxicity, sentiment, rumor stance, and news frames). Our results show that LLM compliance and accuracy are prompt-dependent. For instance, prompting for numerical scores instead of labels reduces all LLMs' compliance and accuracy. Concise prompts can significantly reduce prompting costs but also lead to lower accuracy on tasks like toxicity. Furthermore, minor prompt changes like asking for an explanation can cause large changes in the distribution of LLM-generated labels. By assessing the impact of prompt design on the quality and distribution of LLM-generated annotations, this work serves as both a practical guide and a warning for using LLMs in CSS research.

Introduction

NLP systems for computational social science tasks have traditionally relied on manually annotating large datasets, which can yield high-quality labels but at the expense of time, money, and emotional labor. Many studies are thus turning to prompting LLMs for text annotations for many tasks such as toxicity (Li et al. 2024) and news frame detection (Gilardi, Alizadeh, and Kubli 2023). Results (Li et al. 2024; Qin et al. 2023; Gilardi, Alizadeh, and Kubli 2023) show that LLMs like ChatGPT and PaLM can perform these text annotation tasks in *zero-shot setting*, i.e., through prompts containing instructions on how to annotate the data. However, there is little large-scale, systematic, empirical evidence about what prompt designs are most effective across computational social science tasks.

Most research on benchmarking LLMs' performance report results using just *one prompt design* (Wang et al. 2023; Qin et al. 2023; Gilardi, Alizadeh, and Kubli 2023). While numerous guides for LLM prompting exist (DAIR 2023; Giray 2023; Akin 2023; Bach et al. 2022), they do not all offer the same guidance, and leave many empirical questions unanswered. For example, most guides suggest making the prompts as "descriptive and detailed" as possible (DAIR 2023). However, longer prompts make tasks more expensive as LLM costs depend on the number of input tokens. Can re-writing prompts for concision still maintain accuracy?

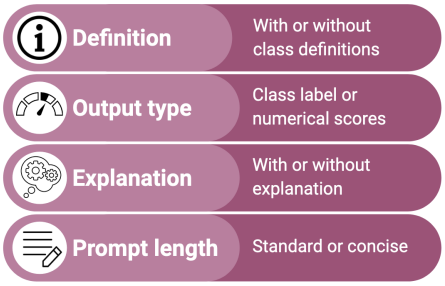

Figure 1: Prompt variations used in our experiments

Separate from prompt designs that lead to accurate outputs, there is little systematic evidence on the extent of LLMs' *compliance* with input prompt instructions. It is important that LLMs generate valid output that conforms to the instructions provided in the prompt since non-compliance wastes both time and money. Qin et al. (2023) report some examples where GPT-3.5 does not comply with the input prompt – despite explicit instructions to generate "positive" or "negative" sentiment labels only, GPT-3.5 generates "neutral" or "mixed" as the label. In the absence of any systemic evidence, however, it remains unclear whether certain prompt designs are more or less likely to generate compliant outputs.

To understand the relationship between prompt design and LLM compliance and accuracy, we conducted a large-scale multi-prompt experiment to annotate four datasets including toxicity, sentiment, rumor stance, and news frames, using four LLMs (GPT-4o, GPT-3.5, PaLM2, Falcon7b). Inspired by a combination of popular prompting practices (DAIR 2023) and practical constraints (e.g., prompting costs), we vary prompts along four dimensions (see Fig-

ure 1): i) definition inclusion (yes/no), ii) output type (label or numerical score), iii) explanation (yes/no), and iv) prompt length (standard/concise). We follow a complete factorial design to generate 16 different prompts (2*2*2*2) for each task and produce a large multi-task, multi-model, multi-prompt design experiment with a combined 483,904 annotations.

Our results show that LLM compliance and accuracy are highly prompt-dependent, especially for multi-class tasks. Below, we report our key findings for individual prompt designs:

- Prompting for **numerical scores** instead of labels reduces both compliance and accuracy for most LLMs and tasks.
- Prompting with **definitions** improves the accuracy of GPT-3.5 and GPT-4o without reducing their compliance. Prompting with definitions reduces PaLM2's and Falcon7b's compliance.
- The impact of **concise prompts** on accuracy and compliance is highly task and model-dependent. For example, using concise prompts for sentiment reduces the cost of annotations without decreasing accuracy. In most cases, however, concise prompts adversely impact either accuracy or compliance.
- Prompting LLMs to **explain their input** improves their compliance with prompt instructions. However, this also changes the distribution of generated labels. For example, GPT-3.5 annotates 34% more content as neutral when prompted to explain its output.

Taken together, we point to several best practices for researchers and practitioners on prompting LLMs for CSS tasks. Additionally, we also highlight inconsistent effects of prompt design (e.g., changing the distribution of generated labels) that can affect social science research results.

Related Work

LLMs for CSS

Computational systems for social science tasks often require manual annotations to train classifiers and evaluate the effectiveness of unsupervised models (Gilardi, Alizadeh, and Kubli 2023). In many instances, these applications demand the support of crowd-workers sourced from platforms such as MTurk to annotate data samples (Huynh, Bigham, and Eskenazi 2021; Gilardi, Alizadeh, and Kubli 2023). However, the financial cost of data annotation is often high, and the demographics of annotators can influence the objectivity of the annotations (Díaz et al. 2022). For specific annotation tasks, such as toxicity detection, annotators are exposed to harmful and offensive content. This exposure limits the pool of available annotators and restricts the volume of content they can reasonably review (Li et al. 2024).

The rise of large language models (LLMs) such as ChatGPT and Google PaLM is revolutionizing the field of computational social science (CSS), demonstrating strong performance on a wide range of tasks (Lan et al. 2024; Kumar, AbuHashem, and Durumeric 2024; Li et al. 2024; Breum et al. 2024; Hou, Leach, and Huang 2024; Ziems et al. 2024). A key advantage of LLMs widely discussed in the literature is their cost-effectiveness. Consequently, research has focused on LLM-generated annotations for fundamental tasks like sentiment analysis (Wang et al. 2023; Qin et al. 2023) to economically generate annotations at a large scale (Wang et al. 2021; Gilardi, Alizadeh, and Kubli 2023). Studies have also focused on tasks like toxicity detection (Kumar, AbuHashem, and Durumeric 2024; Li et al. 2024; Huang, Kwak, and An 2023) as LLMs offer a societal benefit by shielding human annotators, particularly those from vulnerable groups, from harmful content. Another major strength of LLMs lies in their reasoning and explainability capabilities (Zhang, Ding, and Jing 2022; Huang, Kwak, and An 2023; Liu et al. 2023a; Lan et al. 2024). For instance, Zhang, Ding, and Jing (2022) argue that stance detection can particularly benefit from LLM-generated explanations. Finally, some studies have also focused on using LLMs to generate annotations for tasks like news frame (Gilardi, Alizadeh, and Kubli 2023) that otherwise require expert annotations.

Most efforts in benchmarking LLMs' performance on various tasks have relied on a single prompt design. While some researchers (Wang et al. 2023; Huang, Kwak, and An 2023) experimented with multiple prompts, they often reported results based on a limited sample due to the high computational and financial costs of testing entire datasets on multiple prompts. In this paper, we address that gap by evaluating LLMs' performance on four highly impactful tasks (toxicity, sentiment, rumor stance, and news frames) where each task is annotated 48 different times, using 16 prompts and 4 LLMs. In the following sections, we provide more details on our selected tasks and how we designed our prompts.

Prompt Design

Prompts are a set of instructions designed to engage and guide the behavior of LLMs (White et al. 2023; Giray 2023). Typically, a prompt consists of four elements (DAIR 2023): (1) Instruction – a specific task for the model to perform, (2) Context – additional information, such as concept definitions, to help generate better responses, (3) Input data – the question or data for the model to respond to or annotate, and (4) Output indicator – the desired type or format of the response. When designed properly, prompts can vastly expand the range of tasks that LLMs can handle without requiring new training data or modifications to the underlying models (Zhang et al. 2021).

Researchers have explored a variety of prompting techniques to interact with LLMs, such as zero-shot (Xian, Schiele, and Akata 2017), few-shot (Brown et al. 2020), and chain-of-thought (CoT) (Wei et al. 2022). Amongst these, zero-shot prompting is most widely used as users can provide input instructions without needing additional labeled examples or training data (Wei et al. 2021). CoT prompting has recently gained attention due to its ability to elicit complex and multi-step reasoning by providing instructions in a step-by-step manner (Wei et al. 2022). Researchers have recently introduced more advanced prompt engineering techniques, such as tree-of-thought prompting (Yao et al. 2024), which organizes reasoning processes hierarchically, and graph prompting (Liu et al. 2023b), which leverages

graph structures to guide the model's understanding. These prompt designs have demonstrated their value across diverse domains, including medicine (Wang et al. 2024), education (Lee et al. 2024), and mathematical reasoning (Ranaldi and Freitas 2024).

Specific to CSS tasks, however, most prior work has used single-step prompts (Qin et al. 2023; Ziems et al. 2024; Gilardi, Alizadeh, and Kubli 2023; Wang et al. 2023; Zhang, Ding, and Jing 2022). For example, a recent study has assessed the zero-shot performance of 13 LLMs on 25 representative English CSS benchmarks, showing that LLMs can achieve fair levels of performance as compared to human annotators across various CSS tasks (Ziems et al. 2024). In addition, more complex frameworks like CoT are costly and challenging to scale for large-scale datasets (>1000 data points). Recent evidence also suggests that CoT does not improve performance on language reasoning tasks (Qin et al. 2023; Wang et al. 2023) to the same extent as for arithmetic tasks. Therefore, in line with prior work, our study also uses single-step prompts which are the most scalable and cost-effective for annotating large datasets.

Numerous guides have been released on formulating input prompts (DAIR 2023; Giray 2023; Akin 2023; Bach et al. 2022; White et al. 2024). Most guides suggest making the prompts as "descriptive and detailed" (DAIR 2023) as possible. Other guidelines include prompting with clear definitions to reduce the gap between humans and LLMs (Giray 2023; Akin 2023). While generally useful, the guides provide little empirical evidence to back their claims. For example, descriptive prompts can make annotations more expensive as LLM costs depend on the number of input tokens. However, it is unclear if re-writing prompts for concision still maintains accuracy.

Some researchers have also introduced prompt designs practically relevant to their task. For example, Li et al. (2024) introduced prompting for numerical scores on toxicity detection to select different thresholds for filtering toxic content. Zhang, Ding, and Jing (2022) introduced prompting for explanations on stance detection and underlined LLMs' potential to generate human-like annotations. Separate from the practical relevance of various prompt designs, there is little systematic evidence of the extent of LLMs' compliance with these input prompt instructions. It is important that LLMs conform to the input instructions as non-compliance wastes both time and money. Qin et al. (2023) report some examples where ChatGPT does not comply with the input prompt – despite explicit instructions to generate "positive" or "negative" sentiment labels only, ChatGPT generates "neutral" or "mixed" as the label. In the absence of any systemic evidence, however, it remains unclear whether certain prompt designs are more or less likely to generate compliant outputs.

In our study, we address this gap by separately measuring LLMs' *compliance* and *accuracy* in response to different prompt designs. We systematically vary our prompts along four highly relevant and practically beneficial dimensions (definition inclusion, output type, explanation, and prompt length) to separately measure the impact of each prompt dimension on the LLM's compliance and accuracy for various tasks. We explain each of our prompt design and its relevance in more detail below.

Experiment Details

In this section, we first explain the different tasks we experiment with and then describe the prompt designs and LLMs used for the experiment.

Datasets and Tasks

We conducted our experiments with 4 diverse tasks – toxicity, sentiment, rumor stance, and news frame detection. Each task is broadly relevant and has repeatedly featured in prior research (Li et al. 2024; Gilardi, Alizadeh, and Kubli 2023; Lan et al. 2024; Zhang, Ding, and Jing 2022; Wang et al. 2023; Huang, Kwak, and An 2023) on LLM-generated annotations indicating community interest in using LLMs for these tasks. Below, we describe each task and its relevance in more detail.

Toxicity: Manually annotating toxic content is expensive and causes psychological distress by exposing humans to harmful content. Using LLMs to generate toxicity annotations can provide both social and economic benefits. Prior works have already underlined the potential of using ChatGPT for annotating toxic content (Li et al. 2024) as well as providing meaningful explanations (Huang, Kwak, and An 2023). For our experiment, we used the HOT dataset (Wu et al. 2023) (i.e., the same dataset used in (Li et al. 2024)), consisting of 3480 social media comments annotated as toxic or not (i.e., **2 labels**) by crowdworkers.

Sentiment analysis: Sentiment analysis is one of the most popular CSS tasks for understanding opinions and emotions expressed in large bodies of text. Unlike domain-specific models, LLMs can serve as universal sentiment analyzers across a wide range of domains (Wang et al. 2023). For our experiment, we used the SST5 dataset (Socher et al. 2013) for fine-grained sentiment analysis where each sentiment is assigned to one of the **5 labels**: very negative, somewhat negative, neutral, somewhat positive, or positive. We used the test set consisting of 2210 text movie reviews annotated by crowdworkers.

Rumor stance detection: Rumor stance detection is an important subtask for identifying online misinformation. Unlike identifying misinformation, rumor stance detection does not require LLMs to be up-to-date on the latest events. Rumor stance detection is also linguistically more complex than stance detection toward a fixed target, on which LLMs have achieved close to state-of-the-art performance (Zhang, Ding, and Jing 2022). For our experiment, we used the RumorEval dataset (Gorrell et al. 2019) containing 1675 tweet pairs where the relationship between tweets is annotated as support, query, comment, or deny **(4 labels)** by crowdworkers.

News frame identification: Automatic news frame identification can serve a wide range of applications, from uncovering media bias (Morstatter et al. 2018) to performing automated news curation (Atreja et al. 2023). Using LLMs for news frame identification can save both time and resources as manually annotating news frames often relies on

Prompt design	Δ(Num of words and fixed cost)
Adding definitions	+91.97%
Asking for explanation	+10.31%
Asking for numerical scores	+22.37%
Concise version	-53.97%

Table 1: Changes in prompt length and fixed annotation cost due to different prompt designs

Dataset (#labels)	#instances	#prompts	#LLMs	#annotations
Toxicity (2)	3,480	16	4	222,720
Sentiment (5)	2,210	8[1]	4	70,720
Rumour Stance (4)	1,675	16	4	107,200
News frame (9)	1,301	16	4	83,264
Total data				**483,904**

Table 2: Summary of annotations generated during the experiment

expert annotations or lengthy codebooks. We conducted our experiments on the most frequently used GVFC dataset (Liu et al. 2019) consisting of 1301 news headlines where expert scholars labeled the framing of the news article into one of the **9 frame classes**, such as gun rights, public opinion, etc.

In addition to selecting highly relevant tasks, we picked tasks that varied in terms of their linguistic complexity and number of output classes. We varied linguistic complexity as LLMs have been shown to match SOTA performance on simpler tasks (like binary sentiment annotations) but have struggled with more complex tasks. So, LLM compliance and accuracy may differ by linguistic complexity. We also varied the number of output classes as we suspected LLM compliance might suffer as prompts get more complex with more output classes. We selected tasks ranging from 2 (toxicity) to 9 (news frame) output classes. The class distribution for each dataset is also provided in Appendix Table 7.

Prompt Design

We designed 16 prompts for each task in our experiment. The prompts were varied along four dimensions – definition, output type, explanation, and prompt length. We grounded the prompt variations in prior work on prompting LLMs for our selected tasks. Yet, each prompt design is broadly relevant and has a practical significance for all CSS tasks, as we explain below.

Definition (yes or no): prompting with or without output class definitions. Providing class definitions can standardize interpretations for tasks where the output is subjective (e.g., toxicity) or can be defined in multiple ways (e.g., news frames). In our prompts, we used the same definitions provided to human raters when the datasets were first annotated in prior work. We excluded this variation for sentiment analysis as prior work did not include sentiment definitions. This variation was introduced by Gilardi, Alizadeh, and Kubli (2023) for stance and news frames, and Li et al. (2024) for toxicity.

Output type (label or score): prompting for a final output label or numerical (probabilistic) scores for individual labels. Prompting for numerical scores is crucial for setting decision thresholds and controlling the Precision/Recall of LLM-generated outputs. This design was introduced by Li et al. (2024) for toxicity.

Explanation (yes or no): prompting the model to provide an explanation in its output or not. Explanations can add useful context to the LLM's performance and errors but they can also introduce challenges in automated parsing of the output. This design was introduced by Huang, Kwak, and An (2023) for toxicity and Zhang, Ding, and Jing (2022) and Lan et al. (2024) for stance detection.

Prompt length (standard or concise): prompting with the standard prompt or its concise version. Standard prompts were descriptive and detailed (following prompting guides (DAIR 2023; Akin 2023)) to achieve the best performance. Concise prompts were paraphrased versions ($\sim$ 53% less words) of the standard prompt generated using GPT-3. Concise prompts are highly relevant for all tasks to reduce the fixed cost per annotation as LLM API costs are dependent on the number of input tokens. We manually verified every concise prompt to ensure that it contained all the information from the standard prompt.

We followed a complete factorial design between different dimensions (2*2*2*2) to isolate the effect of each prompt design and avoid compounded effects due to multiple variations. This resulted in a total of 16 prompts per task (8 in case of sentiment). Each prompt variation is of a different length and impacts the fixed cost per annotation. Table 1 shows the change in the number of words due to each prompt variation, which is indicative of the change in annotation costs. The complete list of prompts used in the experiment is provided in Appendix Table 10.

As explained under Section Related Work, each prompt design is a single-step prompt given their cost-effectiveness and scalability in annotating large datasets.

Models

We used four instruct-tuned LLMs in our experiment – **GPT-4o (gpt-4o-2024-08-06), GPT-3.5 (GPT3.5-turbo-0613)**[2], **PaLM2 (chat-bison-001)**, and **Falcon7b-instruct**, which represent different architectures, sizes, and costs.

- **GPT-4o** is OpenAI's most advanced and full-sized model.
- **GPT3.5-turbo** is OpenAI's older model. It is significantly less inexpensive than GPT-4o but still effective at performing most NLP tasks (Gilardi, Alizadeh, and Kubli 2023).
- **PaLM2** is a family of generative models launched by Google and shown to outperform human raters on many tasks (Suzgun et al. 2023; Sarkar, Feng, and Karmaker Santu 2023).

[1]We did not have a definition-based prompt variation for the sentiment task as prior work did not provide any definitions.

[2]the latest model available at the time of the experiment

- We picked **Falcon7b** as our third model to find out how smaller open-source LLMs compare against larger models. Falcon7b is part of the Falcon series of open source models[3] and has 7b parameters. Compared to its larger siblings, Falcon7b can be set up without a GPU. At the start of this study (June 2023), the Falcon series was ranked highest on Hugging Face's open source LLM leaderboard[4].

We set the temperature parameter to zero for all experiments to generate consistent outputs. All models were accessed via API endpoints. OpenAI and PaLM2 provide their own APIs while Falcon7b is available via Azure Cloud.

The original study only considered LLMs released as of June 2023 (at the start of the study). Given our emphasis on large-scale annotations for CSS tasks, we considered LLMs that are inexpensive and widely accessible to low-code users. We later added GPT-4o to validate our results on larger advanced models

Annotations Dataset

Table 2 shows the complete statistics of our collected annotations. We followed a complete factorial design between different prompt designs (2*2*2*2), LLMs (4), and datasets (4), resulting in a **total of 483,904 annotations**. The complete factorial design allowed us to isolate the effect of individual prompt design on each model and task combination, identify inconsistent effects due to particular tasks or models, and avoid any compounded effects due to multiple variations.

Evaluation

Parsing LLM output: We used simple string matching to extract potential labels from the LLM's raw output by matching against the list of labels provided in the input. In cases where the model was prompted to explain its output, any text after the keyword "explanation" was excluded from string matching. In cases where string matching returned multiple potential labels, we picked the first label in the order of appearance. If the LLM was prompted to provide numerical scores, we also extracted any floats between $[0, 1]$ from the output. The floats were matched to their corresponding labels based on the order in which they appeared.

Measuring compliance: When prompting for an output label, the LLM's output was considered compliant if a unique label matching the input labels was extracted from the output. For instance, a model compliant with the toxicity task returned "yes" or "no" labels. When prompting for numerical scores, the output was considered compliant if at least one valid label was extracted from the output and the sum of scores assigned to the extracted labels belonged to $[0.99, 1.01]$. We report our results by computing percentage compliance on the complete dataset.

Measuring accuracy: To calculate accuracy, we compared the LLM-generated labels with the human annotations provided for each dataset. We reported both F1 score (macro) and percentage accuracy in our results. We included only compliant outputs when measuring accuracy.

[3]https://falconllm.tii.ae/falcon.html

[4]https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

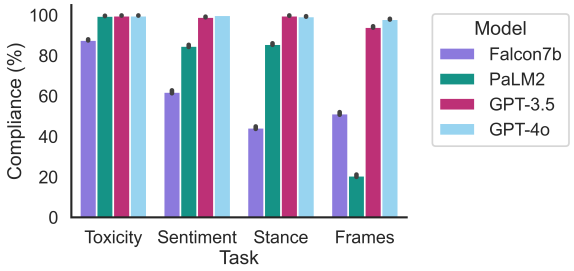

Figure 2: Percentage compliance for different tasks and LLMs.

	Toxicity		Sentiment		Stance		Frames	
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Falcon7b	0.28	0.28	0.29	0.29	0.07	0.07	0.38	0.23
PaLM2	**0.82**	**0.73**	0.53	0.48	0.61	**0.44**	0.62	0.54
GPT-3.5	0.71	0.65	0.41	0.40	0.62	0.38	0.61	0.51
GPT-4o	0.77	0.71	**0.55**	**0.54**	**0.63**	**0.44**	**0.75**	**0.71**

Table 3: Percentage accuracy and F1 (macro) score for different tasks and LLMs

Results

First, we compare the overall accuracy and compliance for GPT-4o, GPT-3.5, Falcon7b, and PaLM2 on each task, and then present a breakdown of their compliance and accuracy for different prompt designs.

Comparing LLMs

Compliance We find the two GPT models, GPT-3.5 and GPT-4o, are most compliant for all tasks (see Figure 2). The smallest model, Falcon7b, is the least compliant on all tasks except news frame identification, for which PaLM2 is the least compliant due to frequent refusals. Figure 3 shows PaLM2's refusal to perform news frame identification and other examples of noncompliance. Falcon7b often ignored instructions and generated arbitrary outputs. In other cases, LLMs generated incorrect class labels or failed to comply with numerical rules.

Accuracy Table 3 shows the overall accuracy and F1 score (macro) for all LLMs and tasks. Comparing both accuracy and F1 score, GPT-4o is the best-performing model on all tasks but toxicity, for which PaLM2 is the best performing model. Falcon7b is significantly less accurate than the two GPT models and PaLM2.

Comparing Prompts

Compliance **Overview:** Table 4 presents a breakdown of the compliance of LLMs for different prompt designs. For toxicity annotations, a task with only two class labels, compliance is high for all prompt and model combinations. However, for other tasks, compliance varies by prompt and

LLM	Task	LLM output	Cause of noncompliance
PaLM2	News frames	I'm not able to help with that, as I'm only a language model. If you believe this is an error, please send us your feedback.	Refusal to generate output
Falcon7b	Rumor stance	Tweets 1 and 2 have opposing stances. The sentiment of Tweet 2 is more likely to be categorized as supporting. The sum of the probability scores for these two tweets is 0.5.	Arbitrary output
Falcon7b	Rumor stance	Mike Pence is a hypocrite.	Arbitrary output
Falcon7b	Toxicity	trump2024, trump2020, trump2022, trump2021, trump2023, trump2025, trump2026....	Arbitrary output
Falcon7b	Sentiment	class1 0.8 class2 0.6 class3 0.4 class4 0.2 class5 0.8	Incorrect class names
PaLM2	Sentiment	negative	Incorrect class names
PaLM2	Sentiment	Positive. The text is positive because it is praising the movie...	Incorrect class names
Falcon7b	News frames	Politics: 0.8, Public opinion: 0.8, Society and culture: 0.8, Gun rights: 0.8, Gun control: 0.8, Mental health: 0.8...	Failed numerical parsing
PaLM2	Rumor stance	comment 0.0 0.0 0.0 1.0	Failed numerical parsing
ChatGPT	Rumor stance	Tweet 2's stance towards Tweet 1: - Support: 0% - Deny: 0% - Question: 90% - Comment: 10% Explanation: <>	Failed numerical parsing

Figure 3: Examples demonstrating LLM noncompliance

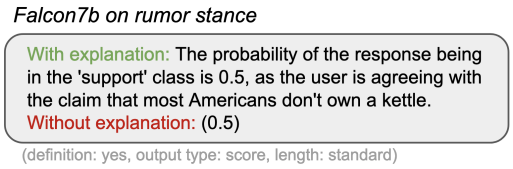

Figure 4: Falcon7b's response on the same data when prompted with/without explanation

model. We also find that models exhibit differing levels of compliance for different prompts, with GPT-3.5 and GPT-4o being the most consistent across prompts. Below, we highlight a few key observations for individual prompt designs.

Definition (yes/no): Prompting with class definitions reduces compliance for both Falcon7b and PaLM2 on rumor stance (4 classes) and news frame annotations (9 classes). Multiple class definitions make input instructions more complex and difficult to follow. We do not observe this relationship between definitions and compliance for GPT-3.5 and GPT-4o which have demonstrated superior reasoning and dialogue capability compared to other LLMs (Qin et al. 2023).

Explanation (yes/no): Prompting for an explanation in the output increases compliance for some task and model combinations: PaLM2 (sentiment), ChatGPT (news frames), and Falcon7b (sentiment, rumor stance, and news frames). Specific examples (see Figure 4) suggest that LLMs are less likely to respond with nonexistent or missing class labels when prompted to explain their output.

Prompt length (standard/concise): Prompting GPT-3.5 and GPT-4o with concise prompts has little impact on their compliance. This offers a significant cost advantage as their costs depend on the number of input tokens, and the concise version of a prompt on average contains 40% fewer tokens. We do not observe this relationship between prompt length and compliance for other LLMs.

Output type (label/score): Prompting for numerical scores instead of label class decreases compliance for Falcon7b (sentiment and rumor stance), PaLM2 (rumor stance), and the GPT models (news frame). Noncompliance is often due to LLMs assigning the same score to each label, or providing scores with sum greater than 1 (Figure 3). This is expected given LLMs' limitations in understanding numerical rules (Zhao et al. 2023).

While prompting PaLM2 for numerical scores increases compliance for sentiment annotations, examples show that PaLM2 sometimes responds with coarse sentiment labels (instead of fine-grained) leading to noncompliance. This, however, is less likely to happen when PaLM2 is prompted for numerical scores (detailed example provided in Appendix Table 8).

Accuracy Table 5 presents a breakdown of LLMs' accuracy (F1 scores provided in the Appendix Table 9) for different prompt designs. We find that the impact of prompt design on accuracy is highly model and task dependent. Below, we highlight a few key observations for individual prompt designs.

Definition (yes/no): Prompting with class definitions increases the accuracy of news frames annotations for GPT-3.5, GPT-4o, and PaLM2[5]. News frames can be defined in multiple ways (Nicholls and Culpepper 2021). Providing their definitions can standardize the interpretations, leading to more accurate outputs from LLMs. We do not observe this trend for the smaller model, Falcon7b.

Prompt length (standard/concise): Prompting with concise prompts results in sentiment annotations with the same or higher accuracy for all LLMs. This is advantageous as concise prompts can reduce the costs of annotation. However, for toxicity annotations, concise prompts lead to lower accuracy for all LLMs, highlighting a tradeoff between cost and quality. While concise prompts can be efficient and cost-saving for some tasks, more detailed prompts may be necessary for achieving higher accuracy on complex tasks, such as toxicity.

Output type (label/score): Prompting for numerical

[5]Although PaLM2's accuracy is measured on a very small subset of the complete data (<1%) on which the model is compliant

	Toxicity				Sentiment				Rumor stance				News Frames			
	Falcon7b	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**
Definition (yes)	87.77	99.68	99.71	99.65	—	—	—	—	42.73	80.63	99.68	99.15	37.74	0.73	**98.53**	96.79
Definition (no)	87.68	99.58	99.91	99.95	62.05	84.78	99.06	100.00	**45.93**	**90.77**	99.90	99.61	**64.98**	**40.43**	89.78	**99.23**
Explanation (yes)	87.07	99.55	99.71	99.60	**64.80**	**91.65**	98.17	100.00	**54.90**	86.29	99.62	99.32	**63.90**	21.14	**95.80**	97.67
Explanation (no)	88.38	99.72	99.91	100.00	59.31	77.91	99.95	100.00	33.76	85.10	99.96	99.44	38.83	20.02	92.51	98.35
Output Type (label)	87.09	99.57	99.89	100.00	**83.22**	77.01	100.00	100.00	**71.85**	**97.37**	99.98	100.00	**83.29**	19.05	**99.93**	**100.00**
Output Type (score)	88.35	99.70	99.73	99.60	40.88	**92.55**	98.12	100.00	16.81	74.02	99.60	98.76	19.43	22.11	88.37	96.02
Length (standard)	88.11	99.58	99.99	100.00	**63.52**	**95.97**	99.85	100.00	42.20	74.43	99.94	99.57	50.73	**26.30**	93.51	98.38
Length (concise)	87.34	99.69	99.63	99.60	60.59	73.59	98.27	100.00	**46.46**	**96.96**	99.63	99.19	52.00	14.86	94.79	97.65

Table 4: LLM percentage compliance for different prompt designs. The more compliant variant of a prompt feature is highlighted in bold ($\Delta > 2\%$)

	Toxicity				Sentiment				Rumor stance				News Frames			
	Falcon7b	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**	**Falcon7b**	**PaLM2**	**GPT 3.5**	**GPT-4o**
Definition (yes)	24.20	81.13	**73.11**	**74.52**	—	—	—	—	7.27	60.87	61.51	62.77	38.06	**76.32**	**67.77**	**80.21**
Definition (no)	**31.36**	82.70	69.45	79.56	28.78	53.23	41.20	54.99	7.35	61.65	61.83	63.87	37.71	62.05	53.18	70.14
Explanation (yes)	23.23	81.13	71.22	77.46	**34.11**	**55.70**	36.21	**56.37**	7.28	60.04	**63.72**	64.29	**41.86**	**68.00**	54.88	75.17
Explanation (no)	**32.25**	82.70	71.34	76.62	22.96	50.33	**46.11**	53.62	7.37	**62.55**	59.62	62.36	31.21	56.29	**66.96**	75.05
Output Type (label)	17.68	**85.09**	**73.99**	**78.24**	**31.00**	**57.52**	**46.39**	**56.90**	7.25	**67.44**	**69.87**	**65.37**	37.70	**63.64**	**67.49**	**78.88**
Output Type (score)	**37.73**	78.75	68.56	75.84	24.27	49.66	35.91	53.09	7.55	53.18	53.44	61.25	38.43	61.15	53.26	71.18
Length (standard)	**31.85**	**85.35**	**73.38**	**79.31**	27.12	51.45	39.46	55.67	7.57	61.62	**63.38**	**65.16**	33.38	61.64	**63.57**	75.39
Length (concise)	23.67	78.48	69.17	74.77	**30.53**	**55.56**	**42.97**	54.32	7.07	61.03	59.95	61.48	**42.18**	63.48	58.10	74.83

Table 5: LLM percentage accuracy for different prompt designs. The more accurate variant of a prompt feature is highlighted in bold ($\Delta > 2\%$)

	Label	**Explanation (yes)**	**Explanation (no)**	Δ
GPT-3.5 on Sentiment	very positive	1.43	5.05	-3.62
	somewhat positive	16.74	33.25	-16.51
	neutral	54.37	19.68	34.69
	somewhat negative	19.57	26.12	-6.55
	very negative	7.89	15.90	-8.01
Falcon7b on Toxicity	True	91.59	78.42	13.17
	False	8.41	21.58	-13.17

Table 6: Percentage distribution of generated labels when prompting LLMs with or without explanations

scores instead of label class decreases the accuracy for all LLMs on all tasks (except for Falcon7b on toxicity).

Explanation (yes/no): Prompting for an explanation in the output has a mixed impact on the accuracy of annotations depending on tasks and LLMs. In particular, prompting GPT-3.5 for explanation reduces the accuracy of sentiment and news frame annotations. Prompting Falcon7b for explanation also reduces accuracy of toxicity annotations. This undesirable impact undermines the potential of LLMs at generating human-like explanations (Huang, Kwak, and An 2023).

Further investigation shows that the impact of prompting with explanations on accuracy can be attributed to a major change in the distribution of LLM-generated labels. Table 6 shows two examples. For sentiment labels generated by GPT-3.5 and toxicity labels generated by Falcon7b, the class distributions differed significantly depending on whether the model was prompted to provide an explanation or not.

Discussion

We empirically assess the impact of prompt design on the quality of LLM-generated annotations using multiple LLMs and a diverse set of tasks. Our analysis reveals evidence-driven best practices for designing effective prompts. Our findings also uncover inconsistencies in the impact of prompt design, highlighting the need for researchers to carefully consider their prompt choices. In the sections that follow, we unpack some of the implications of our work and highlight potential directions for future research.

Prompt Design Implications

Prompting with definitions For certain CSS tasks, output labels can be subjective (e.g., toxicity (Wu et al. 2023)) or defined in multiple ways (e.g., news frames (Nicholls and Culpepper 2021)). In these cases, practitioners may prefer to prompt LLMs using their own definitions to produce more accurate annotations. However, having multiple class definitions can sometimes reduce compliance, as more complex instructions become harder to follow. Our findings indicate this is not an issue for GPT-3.5 and GPT-4o. Prompting GPT models with definitions improved their accuracy in identifying news frames without reducing compliance.

Prompting for numerical scores Prompting LLMs to generate numerical scores rather than categorical outputs offers practitioners greater control over LLM-generated annotations. For example, content moderators could apply different thresholds to toxicity scores, automatically removing highly toxic content while flagging moderately toxic con-

tent for manual review (Kumar, AbuHashem, and Durumeric 2024; Li et al. 2024). However, our results show that asking LLMs to provide numerical scores reduces compliance across all models and tasks. This is not surprising given LLMs' limited numerical reasoning capabilities (Zhao et al. 2023). Nonetheless, it is crucial to acknowledge this trade-off as addressing non-compliance will require additional resources.

Cost of prompting The cost of using LLMs is directly tied to the length of input prompts. In many cases, longer prompts are required due to the context or complexity of the task. For example, prompts that include definitions are nearly twice as long as those without (see Table 1). To reduce costs for advanced models like GPT-4o, prompts can be rewritten to be more concise while still conveying the same information.

Prompting for explanations Prompting LLMs for explanations enhances transparency and builds trust in the model's output. Explanations also help clarify the context of LLM errors (Zhang, Ding, and Jing 2022). For example, when evaluating rumor stance annotations, understanding the LLM's reasoning can lead to follow-up responses with more information to refine the LLM's output (Lan et al. 2024). Interestingly, prompting for explanations can lead to more compliant outputs since LLMs are more likely to mention the correct class labels. However, prompting for explanation can also cause significant shifts in the distribution of LLM-generated labels, which we discuss further in Section .

Broader Implications of Using LLMs for CSS Tasks

The success of LLM prompting has led to many applications in social science research (Ziems et al. 2024), such as monitoring public opinion (Zhang, Ding, and Jing 2022) and quantifying online toxic content (Li et al. 2024). Our results indicate that conclusions drawn from such research are highly dependent on the prompt design. For example, when annotating sentiment labels without prompting for an explanation, ChatGPT annotated $\sim$19% of the data as neutral. However, when prompted to explain its output, ChatGPT labeled over 54% of the data as neutral (see Table 6). Given the widespread applications of sentiment analysis for monitoring public opinion, such large systemic shifts in reported sentiment labels can lead to significant over- or underrepresentation of opinions. Furthermore, models trained using LLM-generated datasets will likely perpetuate these shifts in distribution. For instance, Falcon7b annotated 13% more content as toxic when prompted to explain its output. Training a content moderation model on this dataset, or using Falcon7b directly, could result in more content being filtered or removed, exacerbating concerns of overmoderation (Kumar, AbuHashem, and Durumeric 2024; Ferrara 2023).

The reasons behind these shifts are unclear and could be due to differences in model architecture or the nature of training, including reinforcement learning from human feedback (RLHF). Nevertheless, social scientists should be cautious about the downstream impact of prompt variations on understanding social phenomena. When using LLMs, rather than using a single prompt unquestioningly, researchers should carefully evaluate several prompting strategies within their domain of interest, and potentially use several prompts for robustness when making claims.

Implications for other CSS tasks Our results have implications for tasks not covered in our experiment as well. For example, we find that LLM compliance decreases as the number of output classes increases from two (toxicity) to nine (news frames). When using LLMs for annotating tasks with many output classes (e.g., persuation and counterspeech techniques), practitioners may want to model the task as a binary annotation task and use separate prompts for each output class to achieve high LLM compliance.

In addition to the number of output classes, the linguistic complexity of a task can also impact prompt design. For simple tasks, such as sentiment analysis, practitioners can rewrite prompts for concision to significantly lower prompting costs (up to 50%). However, for more complex tasks, such as implicit hate or misinformation detection, detailed and descriptive prompts may still be necessary to achieve high accuracy.

Directions for future work Future work should explore whether our findings hold for a wider range of tasks by varying the number of output classes and linguistic complexity. Our findings show that even larger and advanced models like GPT-4o are susceptible to prompt-induced variations although the effect size is smaller compared to GPT-3.5. Future work should investigate whether the findings hold for different model families, like Claude. Another direction of research should explore potential strategies to improve the robustness of LLM-generated annotations. One potential approach is to combine results across several prompts, similar to how crowdsourcing involves independent raters annotating the same content to improve the quality of annotations. Recent research (Echterhoff et al. 2024) shows that LLMs are capable of self-help debiasing to mitigate cognitive biases in input prompts. Therefore, exploring whether models can identify and self-correct large shifts in the distribution of generated labels induced by prompt variations could be a promising direction.

Limitations

Conducting rigorous evaluation of LLMs is challenging because we cannot determine whether these models have been exposed to our chosen datasets during their training phases, particularly popular datasets like SST-5. However, the fact that LLMs in our study fail to surpass existing baselines and show performance variability with different prompt designs suggests that any potential data leakage had limited impact on our findings.

Due to limited API availability and compute resources, we restricted our study to three LLMs. We only considered LLMs that were released as of June 2023. Later on, we added a larger and more advanced model, GPT-4o. Given our findings that the impact of prompt design depends on individual models, including their architecture and train-

ing data, we caution readers from generalizing our findings to other LLMs. Nevertheless, our inclusion of a smaller model, Falcon7b, reveals that its compliance decreases drastically as input prompts become more complex, such as when prompting for numerical scores. This finding underscores the need for future research to investigate alternative prompting techniques better suited for smaller, more affordable models.

Our study, with 16 prompts, is an extensive comparison of LLMs. However, the space of prompting is vast, and many variations for each prompt aspect are possible. Our results analyzed one prompt variation at a time, leaving open the possibility that different prompt variations may interact and produce different impacts.

Despite these limitations, our research provides a foundation for future studies to explore additional prompt designs and investigate the interactions between different design choices.

Conclusion

The success of LLM prompting has inspired many uses of LLMs to generate text annotations. Our study is the first to conduct a multi-prompt, multi-LLM experiment, providing empirical evidence on the best prompting strategies for CSS tasks. Using a multifactorial design, we reveal inconsistencies in LLM-generated annotations caused by prompt design and emphasize the importance of carefully choosing prompts. We hope our work will serve as a foundation for further research into developing more effective and nuanced prompts for using LLMs across different fields.

Acknowledgements

Special thanks to members of the Hemphill Research Group for their generous feedback that shaped this work. This project was made possible in part by the Institute of Museum and Library Services LG-256652-OLS-24. This material is based upon work supported by the National Science Foundation under Award No. 1928434.

References

Akin, F. K. 2023. The Art of ChatGPT Prompting: A Guide to Crafting Clear and Effective Prompts. https://app.gumroad.com/d/a1d2e54db0ad8bb888072d8ce2a3dceb. Accessed: 2024-5-11.

Atreja, S.; Srinath, S.; Jain, M.; and Pal, J. 2023. Understanding Journalists' Workflows in News Curation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13.

Bach, S.; Sanh, V.; Yong, Z. X.; Webson, A.; Raffel, C.; Nayak, N. V.; Sharma, A.; Kim, T.; Bari, M. S.; Févry, T.; et al. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 93–104.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.

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

DAIR. 2023. Elements of a Prompt. https://www.promptingguide.ai/introduction/elements. Accessed: 2023-12-11.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Echterhoff, J.; Liu, Y.; Alessa, A.; McAuley, J.; and He, Z. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Ferrara, E. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-17.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120.

Giray, L. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, 1–5.

Gorrell, G.; Kochkina, E.; Liakata, M.; Aker, A.; Zubiaga, A.; Bontcheva, K.; and Derczynski, L. 2019. SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, 845–854. Association for Computational Linguistics.

Hou, H.; Leach, K.; and Huang, Y. 2024. ChatGPT Giving Relationship Advice–How Reliable Is It? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 610–623.

Huang, F.; Kwak, H.; and An, J. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, 294–297.

Huynh, J.; Bigham, J.; and Eskenazi, M. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.

Kumar, D.; AbuHashem, Y. A.; and Durumeric, Z. 2024. Watch Your Language: Investigating Content Moderation with Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 865–878.

Lan, X.; Gao, C.; Jin, D.; and Li, Y. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 891–903.

Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; and Zhai, X. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6: 100213.

Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2): 1–36.

Liu, S.; Guo, L.; Mays, K.; Betke, M.; and Wijaya, D. T. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 504–514.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Cheng, R. G. H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023a. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374*.

Liu, Z.; Yu, X.; Fang, Y.; and Zhang, X. 2023b. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, 417–428.

Morstatter, F.; Wu, L.; Yavanoglu, U.; Corman, S. R.; and Liu, H. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2): 1–18.

Nicholls, T.; and Culpepper, P. D. 2021. Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 38(1-2): 159–181.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Ranaldi, L.; and Freitas, A. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1812–1827.

Sarkar, S.; Feng, D.; and Karmaker Santu, S. K. 2023. Zero-Shot Multi-Label Topic Inference with Sentence Encoders and LLMs. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16218–16233. Singapore: Association for Computational Linguistics.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q.; Chi, E.; Zhou, D.; and Wei, J. 2023. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13003–13051. Toronto, Canada: Association for Computational Linguistics.

Wang, L.; Chen, X.; Deng, X.; Wen, H.; You, M.; Liu, W.; Li, Q.; and Li, J. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1): 41.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4195–4205.

Wang, Z.; Xie, Q.; Feng, Y.; Ding, Z.; Yang, Z.; and Xia, R. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; and Schmidt, D. C. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

White, J.; Hays, S.; Fu, Q.; Spencer-Smith, J.; and Schmidt, D. C. 2024. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In *Generative AI for Effective Software Development*, 71–108. Springer.

Wu, S.; Schöpke-Gonzalez, A.; Kumar, S.; Hemphill, L.; and Resnick, P. 2023. HOT Speech: Comments from Political News Posts and Videos that were Annotated for Hateful, Offensive, and Toxic Content.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4582–4591.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zhang, B.; Ding, D.; and Jing, L. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Zhang, N.; Li, L.; Chen, X.; Deng, S.; Bi, Z.; Tan, C.; Huang, F.; and Chen, H. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*.

Zhao, Y.; Long, Y.; Liu, H.; Nan, L.; Chen, L.; Kamoi, R.; Liu, Y.; Tang, X.; Zhang, R.; and Cohan, A. 2023. DocMath-Eval: Evaluating Numerical Reasoning Capabilities of LLMs in Understanding Long Documents with Tabular Data. *arXiv preprint arXiv:2311.09805*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. We hope that our work will serve as a foundation for further research into developing more effective and nuanced prompts for using LLMs in CSS
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see Section on Experiment Details.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? Yes, see Section on Limitations
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see Discussion subsection on "Broader Implications of prompting"
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we provide complete information about model versions used and values of experimental parameters for reproducibility (see Section on Experiment Details)
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have included all the model versions and parameters to reproduce the results. A URL to the code is masked and will be included in the camera-ready version
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the section on Models under Experiment Design
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see section on Evaluation under Experiment Design
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see section on "Broader Implications" under Discussion
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see the section on Dataset and Tasks under Experiment Details
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? None of the datasets contain personally identifiable information, while one of the datasets is explicitly labeled as "toxicity dataset"
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Note on news frame identification task

The input prompt for news frame identification instructed LLMs to provide up to two frame classes in their outputs. This prompt design is based on the instructions given to human annotators when the dataset was initially annotated in prior work. However, the original paper's authors noted that less than 10% of the data was assigned more than one class. Consequently, the computational model introduced in the paper, and subsequent models trained on this data, treated the task as a single-class label and measured performance using only the first label provided by the human annotators. For consistency, we followed the same approach in our analysis and considered news frame identification as a single-label task.

Table 7: Class distribution in original datasets used for our experiment

Dataset	#instances	#prompts	#LLMs	#annotations
Toxicity	**3480**	**16**	**3**	**167,040**
no	2885			
yes	595			
Sentiment	**2210**	**8**	**3**	**53,040**
very positive	399			
somewhat positive	510			
neutral	389			
somewhat negative	633			
very negative	279			
Rumour Stance	**1675**	**16**	**3**	**80,400**
comment	1405			
support	104			
deny	100			
query	66			
News frame	**1301**	**16**	**3**	**62,448**
Politics	380			
Public opinion	237			
Gun control	215			
School and public space	130			
Race and ethnicity	114			
Economic consequences	81			
Mental health	65			
Society and culture	41			
Gun rights	38			
Total				**362,928**

Table 8: Example output generated by PaLM2 on sentiment annotation. PaLM2 is more likely to generate (compliant) fine-grained sentiment labels when prompted for numerical scores.

data_id	output type	raw output	is_compliant
1820	label	negative	0
1820	label	negative	0
1820	label	negative, The text is negative because it is about a fish dying.	0
1820	label	Negative. The text is about a fish out of water, which is a metaphor for someone who is not in their element and is therefore likely to fail.	0
1820	score	very negative 0.99999999999999999 very negative 0.99999999999999999 very negative 0.99999999999999999 very negative 0.99999999999999999	1
1820	score	very negative 0.9999999999999999, somewhat negative 0.0, neutral 0.0, somewhat positive 0.0, very positive 0.0	1
1820	score	very negative 0.99; The text is very negative because it is about a fish that is out of water and will die.	1
1820	score	very negative 0.99 , The text is about a fish out of water. Fish out of water usually die. So the text is very negative.	1

Table 9: LLM F1 scores for different prompt designs

	Toxicity			Sentiment			Rumor stance			News frames		
	Falcon7b	**PaLM2**	**ChatGPT**	**Falcon7b**	**PaLM2**	**ChatGPT**	**Falcon7b**	**PaLM2**	**ChatGPT**	**Falcon7b**	**PaLM2**	**ChatGPT**
Definition (yes)	0.24	0.72	0.67	—	—	—	0.06	0.47	0.38	0.25	0.53	0.62
Definition (no)	0.31	0.74	0.64	0.29	0.48	0.40	0.07	0.42	0.37	0.21	0.54	0.47
Explanation (yes)	0.23	0.72	0.65	0.34	0.51	0.34	0.07	0.45	0.39	0.26	0.60	0.48
Explanation (no)	0.32	0.74	0.65	0.19	0.46	0.44	0.05	0.44	0.37	0.19	0.46	0.62
Output type (label)	0.16	0.76	0.67	0.31	0.53	0.42	0.06	0.50	0.42	0.22	0.56	0.58
output type (score)	0.37	0.70	0.63	0.23	0.44	0.36	0.08	0.38	0.33	0.23	0.53	0.44
Length (standard)	0.32	0.76	0.67	0.24	0.46	0.38	0.07	0.44	0.38	0.21	0.52	0.52
Length (concise)	0.23	0.70	0.64	0.30	0.51	0.41	0.07	0.45	0.37	0.24	0.60	0.51

Table 10: Set of prompts used in the experiment

	Prompt key	Prompt string
0	toxicity-no_definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. In your output, I want you to respond with a yes if the text is toxic, otherwise respond with a no. Only respond with a yes or a no. Do not provide any other outputs or any explanation for your output. Text: <>
1	toxicity-no_definition-label-no_explanation-concise	Perform data annotation by determining if the text is toxic or not. Respond with only a yes or no. No other outputs or explanations needed. Text: <>
2	toxicity-no_definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. In your output, I want you to respond with a yes if the text is toxic, otherwise respond with a no. Only respond with a yes or a no. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
3	toxicity-no_definition-label-explanation-concise	Perform a data annotation task by determining if the provided text is toxic. Respond with a yes or no and provide an explanation for your response in two lines. Text: <>
4	toxicity-no_definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. In your output, I want you to provide a number between 0 and 1 where the number represents the probability of the text being toxic. A higher score indicates a higher probability of the text being toxic. Only respond with a number between 0 and 1. Do not provide any other outputs or any explanation for your output. Text: <>
5	toxicity-no_definition-score-no_explanation-concise	Perform data annotation by determining if the text is toxic. Provide a number between 0 and 1 indicating the probability of toxicity, with a higher score indicating higher probability. Only provide the number, no other outputs or explanations. Text: <>
6	toxicity-no_definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. In your output, I want you to provide a number between 0 and 1 where the number represents the probability of the text being toxic. A higher score indicates a higher probability of the text being toxic. Only respond with a number between 0 and 1. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
7	toxicity-no_definition-score-explanation-concise	Perform data annotation by reading the text and determining its toxicity level with a number between 0 and 1. Higher scores indicate higher toxicity. Include an explanation for your response in the output. Text: <>
8	toxicity-definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. For this task, I also want you to consider the following definition of toxicity: A rude, disrespectful, or unreasonable comment that is likely to make individuals leave a discussion. Annotate the text as toxic only if it satisfies the definition provided above, otherwise do not annotate it as toxic. In your output, I want you to respond with a yes if the text is toxic, otherwise respond with a no. Only respond with a yes or a no. Do not provide any other outputs or any explanation for your output. Text: <>
9	toxicity-definition-label-no_explanation-concise	Perform data annotation task. Read text and determine toxicity based on definition: rude, disrespectful, or unreasonable comment likely to make individuals leave discussion. Respond with yes if toxic, otherwise no. No other outputs or explanations. Text: <>

10	toxicity-definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. For this task, I also want you to consider the following definition of toxicity: A rude, disrespectful, or unreasonable comment that is likely to make individuals leave a discussion. Annotate the text as toxic only if it satisfies the definition provided above, otherwise do not annotate it as toxic. In your output, I want you to respond with a yes if the text is toxic, otherwise respond with a no. Only respond with a yes or a no. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
11	toxicity-definition-label-explanation-concise	Perform a data annotation task by determining if the text below is toxic. Use the definition of toxicity as a rude, disrespectful, or unreasonable comment likely to make individuals leave a discussion. Respond with a yes or no and provide an explanation for your response. Text: <>
12	toxicity-definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. For this task, I also want you to consider the following definition of toxicity: A rude, disrespectful, or unreasonable comment that is likely to make individuals leave a discussion. Annotate the text as toxic only if it satisfies the definition provided above, otherwise do not annotate it as toxic. In your output, I want you to provide a number between 0 and 1 where the number represents the probability of the text being toxic. A higher score indicates a higher probability of the text being toxic. Only respond with a number between 0 and 1. Do not provide any other outputs or any explanation for your output. Text: <>
13	toxicity-definition-score-no_explanation-concise	Perform data annotation by determining if the text is toxic. Use the definition of toxicity as a rude, disrespectful, or unreasonable comment likely to make individuals leave a discussion. Annotate as toxic only if it meets this definition. Provide a number between 0 and 1 indicating the probability of toxicity, with a higher score indicating a higher probability. Do not provide any other outputs or explanations. Text: <>
14	toxicity-definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read the text provided below and determine whether the text is toxic or not. For this task, I also want you to consider the following definition of toxicity: A rude, disrespectful, or unreasonable comment that is likely to make individuals leave a discussion. Annotate the text as toxic only if it satisfies the definition provided above, otherwise do not annotate it as toxic. In your output, I want you to provide a number between 0 and 1 where the number represents the probability of the text being toxic. A higher score indicates a higher probability of the text being toxic. Only respond with a number between 0 and 1. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
15	toxicity-definition-score-explanation-concise	Perform data annotation by determining if the text is toxic. Use the definition of toxicity as a rude, disrespectful, or unreasonable comment likely to make individuals leave a discussion. Annotate as toxic only if it meets this definition. Provide a probability score between 0 and 1, with a higher score indicating a higher probability of toxicity. Include an explanation for the score in the output. Text: <>
16	stance-no_definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. In your output, only respond with the name of the class: support, deny, question, or comment, depending on the relation you identify between tweet 1 and tweet 2. Do not respond with any other output. Do not provide any other outputs or any explanation for your output. Tweet 1: <>Tweet 2: <>

17	stance-no_definition-label-no_explanation-concise	Perform data annotation by reading two tweets and identifying the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Only provide the name of the class in your output without any additional explanation. Tweet 1: <>Tweet 2: <>
18	stance-no_definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. In your output, only respond with the name of the class: support, deny, question, or comment, depending on the relation you identify between tweet 1 and tweet 2. Do not respond with any other output. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Tweet 1: <>Tweet 2: <>
19	stance-no_definition-label-explanation-concise	Perform a data annotation task by reading two tweets and identifying the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Only provide the class name in your output and include an explanation in the second line. Tweet 1: <>Tweet 2: <>
20	stance-no_definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. In your output, I want you to provide a probability score for each of the 4 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 4 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. Do not provide any other outputs or any explanation for your output. Tweet 1: <>Tweet 2: <>
21	stance-no_definition-score-no_explanation-concise	Perform data annotation by reading 2 tweets and determining tweet 2's stance towards tweet 1. Choose from 4 classes: support, deny, question, or comment. Provide probability scores for each class, with higher numbers indicating higher probability. Total probability scores should equal 1. Output only class names and scores, no explanations. Tweet 1: <>Tweet 2: <>
22	stance-no_definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. In your output, I want you to provide a probability score for each of the 4 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 4 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Tweet 1: <>Tweet 2: <>
23	stance-no_definition-score-explanation-concise	Perform a data annotation task by analyzing two tweets and determining the stance of tweet 2 towards tweet 1. Categorize tweet 2 as supporting, denying, questioning, or commenting on tweet 1 and provide a probability score for each class. The sum of the probability scores should be 1. Output the name of the class and its probability score for each line. Additionally, provide an explanation for your response in the second line of the output. Tweet 1: <>Tweet 2: <>

24	stance-definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. For this task, respond with support if the reply supports the claim, respond with deny if the reply disagrees with the claim, respond with question if the reply is asking for additional evidence in relation to the claim, and respond with comment if the reply is making its own claim without a clear contribution to assessing the veracity of the claim. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, only respond with the name of the class: support, deny, question, or comment, depending on the relation you identify between tweet 1 and tweet 2. Do not respond with any other output. Do not provide any other outputs or any explanation for your output. Tweet 1: <>Tweet 2: <>
25	stance-definition-label-no_explanation-concise	Perform a data annotation task by reading two tweets and determining the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Only respond with the name of the class that aligns with the instructions. Do not provide any other output or explanation. Tweet 1: <>Tweet 2: <>
26	stance-definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. For this task, respond with support if the reply supports the claim, respond with deny if the reply disagrees with the claim, respond with question if the reply is asking for additional evidence in relation to the claim, and respond with comment if the reply is making its own claim without a clear contribution to assessing the veracity of the claim. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, only respond with the name of the class: support, deny, question, or comment, depending on the relation you identify between tweet 1 and tweet 2. Do not respond with any other output. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Tweet 1: <>Tweet 2: <>
27	stance-definition-label-explanation-concise	Perform data annotation by reading two tweets and identifying the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Respond with only the class name and provide an explanation for your choice in the second line. Follow the instructions and do not deviate from them. Tweet 1: <>Tweet 2: <>
28	stance-definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. For this task, respond with support if the reply supports the claim, respond with deny if the reply disagrees with the claim, respond with question if the reply is asking for additional evidence in relation to the claim, and respond with comment if the reply is making its own claim without a clear contribution to assessing the veracity of the claim. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, I want you to provide a probability score for each of the 4 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 4 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. Do not provide any other outputs or any explanation for your output. Tweet 1: <>Tweet 2: <>

29	stance-definition-score-no_explanation-concise	Perform a data annotation task by reading two tweets and determining the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Respond with the name of the class and a probability score between 0 and 1 for each class. The sum of the probability scores should be 1. Follow the instructions carefully and do not provide any additional outputs or explanations. Tweet 1: <>Tweet 2: <>
30	stance-definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read two tweets and determine the stance of tweet 2 in response to tweet 1. Your response must belong to one of the four classes, depending on whether tweet 2 supports, denies, questions, or comments on tweet 1. For this task, respond with support if the reply supports the claim, respond with deny if the reply disagrees with the claim, respond with question if the reply is asking for additional evidence in relation to the claim, and respond with comment if the reply is making its own claim without a clear contribution to assessing the veracity of the claim. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, I want you to provide a probability score for each of the 4 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 4 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Tweet 1: <>Tweet 2: <>
31	stance-definition-score-explanation-concise	Perform data annotation by analyzing two tweets and determining the stance of tweet 2 towards tweet 1. Choose from four classes: support, deny, question, or comment. Respond with a probability score for each class, where higher numbers indicate higher probability. The sum of all probability scores should be 1. Follow the instructions carefully and provide an explanation for your response. Tweet 1: <>Tweet 2: <>
32	sentiment-no_definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text and identify the polarity of the sentiment that is conveyed. Your response must belong to one of the five classes, depending on whether the text is very positive, somewhat positive, neutral, somewhat negative, or very negative. In your output, only respond with the name of the class: very positive, somewhat positive, neutral, somewhat negative, or very negative, depending on the sentiment that is conveyed in the text. Do not provide any other outputs or any explanation for your output. Text: <>
33	sentiment-no_definition-label-no_explanation-concise	Perform data annotation by identifying sentiment polarity as very positive, somewhat positive, neutral, somewhat negative, or very negative. Only provide the name of the class in your output, without any additional explanation. Text: <>
34	sentiment-no_definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read the text and identify the polarity of the sentiment that is conveyed. Your response must belong to one of the five classes, depending on whether the text is very positive, somewhat positive, neutral, somewhat negative, or very negative. In your output, only respond with the name of the class: very positive, somewhat positive, neutral, somewhat negative, or very negative, depending on the sentiment that is conveyed in the text. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
35	sentiment-no_definition-label-explanation-concise	Perform a data annotation task by identifying the sentiment polarity of the text as very positive, somewhat positive, neutral, somewhat negative, or very negative. Provide only the class name and an explanation for your response in the first and second lines of your output, respectively. Text: <>

36	sentiment-no_definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the text and identify the polarity of the sentiment that is conveyed. Your response must belong to one of the five classes, depending on whether the text is very positive, somewhat positive, neutral, somewhat negative, or very negative. In your output, I want you to provide a probability score for each of the 5 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 5 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. Do not provide any other outputs or any explanation for your output. Text: <>
37	sentiment-no_definition-score-no_explanation-concise	Perform data annotation by identifying sentiment polarity as very positive, somewhat positive, neutral, somewhat negative, or very negative. Provide probability scores for each class, with higher numbers indicating higher probability. Sum of probability scores for all classes should be 1. Respond with class name and probability score for each line, without any additional output or explanation. Text: <>
38	sentiment-no_definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read the text and identify the polarity of the sentiment that is conveyed. Your response must belong to one of the five classes, depending on whether the text is very positive, somewhat positive, neutral, somewhat negative, or very negative. In your output, I want you to provide a probability score for each of the 5 classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are only 5 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Text: <>
39	sentiment-no_definition-score-explanation-concise	Perform a data annotation task by identifying the sentiment polarity of a given text. Choose from five classes: very positive, somewhat positive, neutral, somewhat negative, or very negative. Provide a probability score for each class, ranging from 0 to 1, with the sum of all scores equaling 1. Output the name of each class followed by its probability score, along with an explanation for your response. Text: <>
40	frames-no_definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. In your output, respond with the frame class the headline belongs to. In your response, you may provide one additional class if you believe the headline belongs to multiple classes. Do not respond with more than 2 classes. Only respond with the name of the classes. Do not respond with any other output. Do not provide any other outputs or any explanation for your output. Headline: <>
41	frames-no_definition-label-no_explanation-concise	Perform data annotation by assigning news headlines to one or more of 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Provide only the name of the class(es) without any additional output or explanation. Do not assign more than 2 classes per headline. Headline: <>

42	frames-no_definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. In your output, respond with the frame class the headline belongs to. In your response, you may provide one additional class if you believe the headline belongs to multiple classes. Do not respond with more than 2 classes. Only respond with the name of the classes. Do not respond with any other output. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Headline: <>
43	frames-no_definition-label-explanation-concise	Perform a data annotation task by assigning news headlines to one or more of 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Provide only one or two class names in your output and an explanation for your choice. Headline: <>
44	frames-no_definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. In your output, I want you to provide a probabilistic response for each of the 9 frame classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are 9 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. Do not provide any other outputs or any explanation for your output. Headline: <>
45	frames-no_definition-score-no_explanation-concise	Perform data annotation by assigning news headlines to 1 or more of 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Provide a probabilistic response for each class, with a score between 0 and 1. The sum of scores for all classes should be 1. Output only the class name and its probability score for each line. No additional output or explanation needed. Headline: <>
46	frames-no_definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. In your output, I want you to provide a probabilistic response for each of the 9 frame classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are 9 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Headline: <>
47	frames-no_definition-score-explanation-concise	Perform data annotation by assigning news article frames. Read headlines and assign one or more of 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Output probabilistic response for each class, with a number between 0 and 1. Sum of probability scores should equal 1. Provide name of class and probability score for each line. Explain output in first and second line of response. Headline: <>

48	frames-definition-label-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Annotation guidelines: For this task, additional instructions for each of the frame class are provided below: 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, respond with the frame class the headline belongs to. In your response, you may provide one additional class if you believe the headline belongs to multiple classes. Do not respond with more than 2 classes. Only respond with the name of the classes. Do not respond with any other output. Do not provide any other outputs or any explanation for your output. Headline: <>
49	frames-definition-label-no_explanation-concise	Perform data annotation by assigning news headlines to one or more of 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Follow provided guidelines for each class. 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. Provide only the name of the class(es) and do not exceed 2 classes. Headline: <>

50	frames-definition-label-explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Annotation guidelines: For this task, additional instructions for each of the frame class are provided below: 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, respond with the frame class the headline belongs to. In your response, you may provide one additional class if you believe the headline belongs to multiple classes. Do not respond with more than 2 classes. Only respond with the name of the classes. Do not respond with any other output. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Headline: <>
51	frames-definition-label-explanation-concise	Perform data annotation by assigning news article frames. Assign one or more of the following 9 frame classes to each headline: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Follow the provided guidelines for each class. 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. Provide only one or two classes in your response and an explanation for your choice. Do not provide any other output. Headline: <>

52	frames-definition-score-no_explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Annotation guidelines: For this task, additional instructions for each of the frame class are provided below: 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, I want you to provide a probabilistic response for each of the 9 frame classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are 9 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. Do not provide any other outputs or any explanation for your output. Headline: <>
53	frames-definition-score-no_explanation-concise	Perform data annotation by assigning news headlines to one or more of the 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Follow the provided guidelines for each class. 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. Provide a probabilistic response for each class, with a number between 0 and 1 representing the probability of that class. The sum of all probabilities should be 1. Output the name of the class followed by its probability score for each line. No other output or explanation is required. Headline: <>

54	frames-definition-score-explanation	I want you to perform a data annotation task. Your task is to carefully read the headline of a news article and determine the frame(s) of the news article. Each news headline must be assigned one or more of the following 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Annotation guidelines: For this task, additional instructions for each of the frame class are provided below: 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. You must follow the instructions mentioned above when providing your response. Do not provide a response that does not align with the instructions. In your output, I want you to provide a probabilistic response for each of the 9 frame classes. The probability of each class should be a number between 0 and 1 where higher numbers represent a higher probability of that class. Since there are 9 possible classes, the sum of their probability scores should always be equal to 1. For each class, respond with the name of the class followed by its probability score in each line. In your output, I also want you to provide an explanation for the output. Provide your response in the first line and provide the explanation for your response in the second line. Headline: <>
55	frames-definition-score-explanation-concise	Perform data annotation by assigning news headlines to one or more of the 9 frame classes: Politics, Public opinion, Society and culture, Economic consequences, Gun rights, Gun control, Mental health, School and public space safety, Race and ethnicity. Follow the provided guidelines for each class. 1) Gun rights: The story is related to the Constitution, the second amendment, and protection of individual liberty and gun ownership as a right, 2) Gun control: The story is about issues related to regulating guns through legislation and other institutional measures. 3) Politics: The story is mainly about the political issues around guns and shootings. 4) Mental health: The story is about issues related to individuals' mental illnesses or emotional well-being, or the mental health system as a whole. 5) School and public space safety: Issues related to institutional and school safety 6) Race and ethnicity: The story is about gun issues related to certain ethnic group(s) 7) Public opinion: The story is about the public's, including a certain community's reactions to gun-related issues. 8) Society and culture: Societal-wide factors that are related to gun violence. 9) Economic consequences: The story is about financial losses or gains, or the costs involved in gun-related issues. Provide a probabilistic response for each class, with a number between 0 and 1 representing the probability of that class. The sum of all probabilities should be 1. Output the name of the class followed by its probability score in each line. Provide an explanation for the output. Headline: <>