# The Risks of Using Large Language Models for Text Annotation in Social Science Research

Yongjun Zhang* and Hao Lin

Department of Sociology and Institute for Advanced Computational Science, Stony Brook University, New York, United States

* Both authors contributed to this paper equally. Direct correspondence to Yongjun Zhang; email: yongjun.zhang@stonybrook.edu

## Abstract

Generative artificial intelligence (AI) or large language models (LLMs) have revolutionized computational social science, particularly in automated textual analysis. In this paper, we conduct a systematic evaluation of the promises and risks of using LLMs for diverse coding tasks in social movement studies. We propose a framework for social scientists to adopt LLMs to text annotation, either as the primary coding decision-maker or as a coding assistant. Additionally, we discuss the associated epistemic risks related to validity, reliability, replicability, and transparency. We conclude by offering several practical guidelines for using LLMs in coding tasks.

## Introduction

Coding complex concepts from text corpora is one of the primary tasks for both quantitative and qualitative social scientists. Although human coding remains the gold standard, computer-aided coding has been proven to be capable of efficiently conducting sophisticated text annotation (Nelson, Burk, Knudsen, & McCall, 2021). For instance, in social movement studies, scholars have devoted considerable resources to building protest event databases by leveraging supervised machine learning techniques to distill protest information from large-scale news articles and social media posts (H. Zhang & Pan, 2019; Hanna, 2017). Recently, many scholars are applauding the potential that the swift advancement in generative artificial intelligence (AI)–particularly large language models (LLMs)–can transform how scholars conduct social research by automating the research pipeline including summarizing literature review, generating and testing hypotheses, and conducting data collection and analysis (Ziems et al., 2024; Messeri & Crockett, 2024; Manning, Zhu, & Horton, 2024; Bail, 2024). Yet, little research has been done to systematically examine its validity, reliability, transparency, and replicability in terms of coding complex concepts for social science research. Generative AI tools may carry great epistemic risks when social scientists treat them as knowledge-production partners (Messeri & Crockett, 2024). To fill this gap, this study asks

the crucial question of whether or under what circumstances we, as social scientists, can trust the results from generative AI models. More specifically, what are the risks of using LLMs for coding in social science research and how we can better mitigate and communicate the risks associated with the implementation of LLMs? By doing so, we explore and propose a schema for incorporating LLMs into text annotation in social science research in a reliable and valid manner.

Computer scientists have demonstrated the potential of generative AI to achieve high accuracy by benchmarking various large language models across a range of tasks, including multiple-choice questions (MMLU), reasoning (HellaSwag), coding (HumanEval benchmark), and mathematics (MATH benchmark). For instance, the benchmark dataset used for measuring multitask accuracy spans 57 tasks across STEM, humanities, social sciences, and more (Hendrycks et al., 2020). These evaluations offer the broader community a snapshot of how these LLMs perform in terms of knowledge acquisition and reasoning capabilities. For example, the average accuracy rates across benchmark datasets including MMLU, HellaSwag, HumanEval, BBHard, GSM-8K, and MATH are 79.45% for GPT-4, 80.08% for Gemini 1.5 Pro, and 84.83% for Claude 3 Opus [1]. In scientific contexts, generative AI has been lauded for its contributions to enhancing scientific discovery and objectivity, overcoming limitations related to scientists' time, attention, cognitive capacities, subjectivity, and bias (Messeri & Crockett, 2024; Ziems et al., 2024; Manning et al., 2024). However, generative AI also poses challenges to scientific understanding due to potential epistemic risks such as data leakage, hallucinations, strong alignment with social norms, and issues with replicability, transparency, and interpretability.

In this paper, we utilize the field of social movement studies as an illustrative case to outline several risks of using generative AI, specifically OpenAI's GPT-4, for coding tasks, and we propose solutions to address some of the associated epistemic risks. We examine the typical tasks conducted by social movement scholars in creating protest databases from newspaper articles, such as determining the relevance of articles to protests and extracting crucial information including participant numbers, objectives, tactics, and involved social movement organizations. In the social movement literature, a frequently utilized benchmark dataset for analyzing protests in the United States is the Dynamics of Collective Action (DoCA) project (Earl, Martin, McCarthy, & Soule, 2004; Hanna, 2017). This project annotated news articles from the New York Times covering the period 1960-1995 to construct a detailed event database containing information at both the article and event levels, such as the reporting year, article title, size of protests, claims, social movement organizations, locations, timing, and more [2]. As the DoCA dataset does not include the original NYT articles, we align the annotated NYT corpus (1987-2007) from the Linguistic Data Consortium (LDC) with the DoCA dataset and employ GPT-4 for coding various complex concepts. We conduct a series of experiments to evaluate the validity, reliability, transparency, and replicability of the outputs generated by GPT-4.

We compare the outputs of GPT-4 with the results of human coding as documented by the DoCA project. To summarize, the performance of LLMs decreased as the complexity of

[1] For more updated model comparisons, visit: `https://www.vellum.ai/llm-leaderboard`

[2] Further details can be found here: `https://web.stanford.edu/group/collectiveaction/`

coding tasks increases. GPT-4 performed pretty well in binary classification such as police presence, but relatively poor in mutilabel and multiclass classification. We also show that the reliability of the LLM ouputs varies across task complexity, with a relatively high level of reliability for simple task instead of complex ones. More importantly, we show that social scientists can use LLMs to output reasoning steps to increase the transparency in coding decision making processes.

We also stress that scholars need to be cautious when using LLMs as the primary coder. To achieve a high level of validity and reliability, scholars need to test the accuracy and stability of their prompts and find the optimal prompting strategy using a small development dataset and report the accuracy metrics by randomly selecting a small evaluation dataset with human evaluation after scaling the prompt to the whole dataset. Meanwhile, LLMs can serve as the secondary coding assistant by producing step-by-step reasoning results and potential classification for human coding.

There are several caveats when interpreting results. First, we only experimented zero-shot text classification with different prompting strategies. Scholars can further test additional strategies such as few-shot prompting or even fine-tuning their own LLMs. Second, with the rapid advancement of generative AI techniques, some of these limitations or risks might be mitigated in the near future. Third, we focus on the specific domain of social movement studies, and the performance of LLMs might vary across different domains. Finally, we only compared GPT-4 with LlaMa3-8B model but the latest release of Llama3.1-405B model has shown greater performance across different benchmarks.

# The Adoption of LLMs for Text Annotation

Transformer-based large language models such as BERT, RoBERTa, and Longformer with transfer learning have been the state-of-the-art techniques in natural language processing tasks for social scientists especially those with relatively limited training datasets (Vaswani et al., 2017; Devlin, Chang, Lee, & Toutanova, 2018; Liu & Salganik, 2019; Beltagy, Peters, & Cohan, 2020; Do, Ollion, & Shen, 2022; Wankmüller, 2022). Since social scientists often lack critical resources to label large-scale training examples to build a model from scratch, transfer learning allows scholars to transfer knowledge learned in the pretraining stage from other data sources to the learning process on the task in the target domain. To put it simply, social scientists can borrow transform-based models pretrained on large-scale textual or multimodal data and fine-tune them for downstream tasks with a limited number of training examples. The common practice is to (1) choose an appropriate model like BERT pretrained on a massive dataset of Wikipedia and Google Books Corpus and (2) to fine-tune the classification layer with human annotated training examples for domain specific tasks (e.g., binary or multiclass classification). For instance, scholars have used pre-trained transformer-based models to study a variety of topics such as populism (Bonikowski, Luo, & Stuhler, 2022), Sinophobia (Y. Zhang, Lin, Wang, & Fan, 2023), and propaganda (Lu & Pan, 2022). However, fine-tuning LLMs for domain-specific tasks requires certain knowledge of deep learning and programming which impedes its wide adoption in social sciences.

The recent advancement in generative AI has transformed the field of computational social

science, as scholars have begun to use LLMs to implement zero-shot or few-shots text classification without any prerequisite programming skills (Brown et al., 2020; Chae & Davidson, 2023; Ziems et al., 2024; Do et al., 2022). Generative AI refers to algorithms capable of producing realistic text, images, audios, videos, and other human-like outputs in response to prompts (Bail, 2024). Large language models, such as OpenAI's ChatGPT and Meta's LLaMa series, are pretrained on extensive textual datasets. These models learn the probability distribution of words within the training data, enabling them to conduct out-of-sample prediction and generate text based on an input sequence of tokens or the prompt.

Scholars are still debating whether social scientists should fine-tune transformer-based large models for information retrieval with transfer learning or simply use generative AI tools such as ChatGPT for zero-shot or few-shot classification. Although limited research indicates that generative AI tools do not surpass the performance of top fine-tuned models in text classification, LLMs such as ChatGPT offer a unique advantage in zero-shot annotation, frequently generating explanations of higher quality than those provided by crowd workers (Chae and Davidson 2023; Ziems et al. 2023). A recent study also shows that ChatGPT even outperforms crowd workers for annotation tasks based on four samples of tweets and news articles, including 25 percentage points higher in zero-shot accuracy and thirty times cheaper than MTurk (Gilardi, Alizadeh, & Kubli, 2023). Notably, LLMs are also widely used by crowd workers on MTurk to automate text annotation tasks, with a prevalence of approximately 30% (Veselovsky, Ribeiro, & West, 2023).

# The Framework of Text Annotation Workflow with Generative AI

Given the increasing use of LLMs for coding tasks, it is critical for social scientists to evaluate the workflows and standards governing these processes. We propose that LLMs can serve as two different roles in text annotation: as the primary coder with humans serving as evaluators for large-scale data, or as secondary assistants with humans as the final decision-makers for small/medium-scale data. Figure 1 proposes the workflow of using LLMs for coding, focusing on these two major roles.

For social scientists, the role of LLMs in text annotation — whether as primary coders or secondary assistants — depends on several factors, including the scale of the data (e.g., large versus small), task complexity (e.g., binary versus multiclass classification), and error tolerance (e.g., required accuracy rate). If researchers are dealing with a large-scale dataset with low task complexity and relatively high error tolerance, LLMs can serve as the primary coder while humans take on the evaluation role. To enhance the quality of coding quality, first we need to develop a coherent and clear codebook where all the concepts to be measured are well-defined. Second, researchers need to build a development dataset by purposefully selecting $100 \sim 200$ cases to iteratively test various prompting strategies for identifying the optimal prompt. Third, with the optimal prompt we scale it up to the entire dataset to obtain the coding output. After that, researchers need to further construct a separate evaluation dataset by randomly selecting an additional $100 \sim 200$ cases to report performance

metrics. Thus, the whole process takes both prompt engineering and data preparation into consideration. Note that selecting the optimal prompt is crucial to achieving the desired output in this process, and testing different prompts relies on the high-quality human-labeled development dataset.

The development set and the evaluation set are an archetype that we adapted from machine learning. The development set, also known as the validation set, is used to fine-tune the model's hyper-parameters in supervised machine learning and to select the optimal model. Here it is used to test and determine the optimal prompt. The evaluation set, sometimes referred to as the testing set, is used to evaluate the performance of the model. Here it is also used to assess the performance of the optimal prompt. Both datasets are separately sampled from the larger dataset, ideally ensuring they are distinct from each other because we cannot guarantee whether LLMs will incorporate the development set into their training process when used as input. This way, we can potentially avoid any inflated performance. The size of these datasets is flexible and depends on time and budget constraints, although we recommend a minimum of $100 \sim 200$ cases for each. Both sets require human labeling and serve as the ground truth.

However, if the annotation task is sophisticated, requiring strong reasoning capacity, the data scale is manageable within budget limits, and error tolerance is low, or the analytic approach is inductive, LLMs can serve as coding assistants. In this role, they help with text summarization and provide an initial round of reasoning for annotation. Similar to computer-assisted coding, LLM-powered coding allows scholars to efficiently digest key information from textual or multimodal data (e.g., images, audio) and make the final coding decisions.

# Prompt Engineering for Text Annotation Tasks

Given that the output of LLMs is highly dependent on the prompts developed by scholars for specific tasks, we further discuss common prompting strategies proposed by computer scientists and engineers. A prompt is a sequence of text input that directs LLMs to perform certain tasks, allowing scholars to bypass the fine-tuning of LLMs for downstream tasks. Prompt engineering is the process by which scholars refine these prompts to guide generative AI towards producing desirable outputs. Various strategies for prompt engineering have been developed, such as in-context learning, role-play prompting, chain-of-thought prompting, and tree-of-thought prompting, to optimize the effectiveness of LLMs.

In-context learning is a prompting strategy that provides LLMs with a few training examples to learn from analogy. Scholars create prompts by incorporating several examples as the demonstration context, enabling LLMs to discern hidden patterns in these examples to correctly perform downstream tasks. Notably, in-context learning does not update the LLMs' parameters, which significantly reduces computational costs and can be easily applied to social science problems (Dong et al., 2023). For instance, to classify whether a news article is related to a protest, movement scholars can include several protest-related articles as training examples when formulating the prompt.

Role-play prompting is a novel strategy that assigns a specific role to the LLM, enabling it to

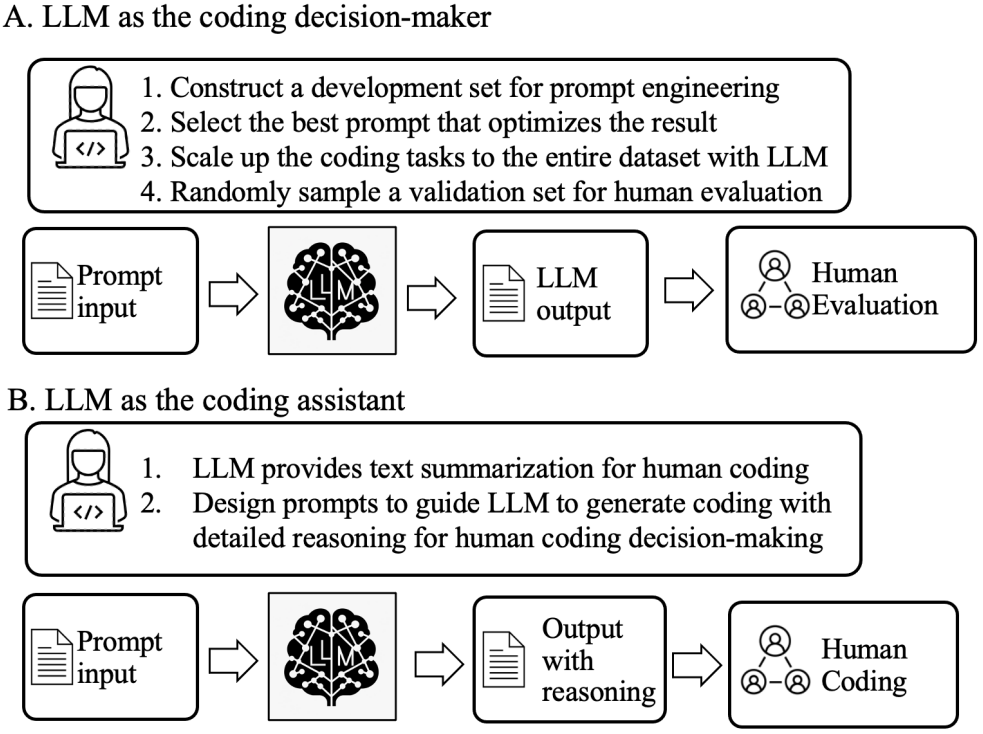


Figure 1: The Workflow of Using LLM for Text Annotation. Note: the LLM icon was created using GPT-4o.

produce outputs tailored to the LLM's identity and persona given by the researcher (Kong et al., 2024). For instance, researchers can assign the LLM the role of an expert movement scholar role who excels at extracting protest information from news articles.

The chain-of-thought (CoT) prompting strategy directs LLMs to generate the intermediate reasoning steps, improving their ability to perform complex reasoning (Wei et al., 2022). CoT allows LLMs to decompose multi-step problems into several intermediate steps, offering an interpretable window into the models' decision-making processes. CoT can be combined with in-context learning by providing several examples (few-shot CoT) and can also be triggered by simply adding *let us think step by step* at the end of the prompt (zero-shot CoT) (Wei et al., 2022; X. Wang et al., 2023; B. Wang et al., 2023).

Instead of modeling the human thought process as a linear chain, the Tree-of-Thoughts (ToT) approach decomposes multi-step problems into a tree structure, where each thought represents an intermediate step in problem-solving (Yao et al., 2023; Long, 2023). This method allows language models to self-evaluate multiple intermediate thoughts through a deliberate reasoning process (Yao et al., 2023). Hulbert (2023) applies the ToT framework to guide LLMs in evaluating intermediate thoughts within a single prompt. For instance, the sample prompt used by Hulbert (2023) to trigger ToT is as follows: *"Imagine three different experts are answering this question. All experts will write down one step of their thinking, then share it with the group. Then all experts will proceed to the next step, and so on. If any expert realizes they're wrong at any point, they will leave. The question is..."*

Next we show several examples how we can adopt some of these common prompting strategies for coding tasks in social science research, particularly social movement studies.

Standard Zero-Shot Prompting with Role Play

**System:** *You are an excellent AI language model trained to extract protest information from news articles.*
**User:** *You need to determine whether it discusses peace movement in the following article: <fulltext >. Please structure your response as follows: <numeric answer (Yes=1; No=0) >*
**Assistant:** *1-9012028*

Zero-Shot CoT Prompting with Role Play

**System:** *You are an excellent AI language model trained to extract protest information from news articles.*
**User:** *You need to determine whether it discusses peace movement in the following article: <fulltext >. Let us think step by step. Please structure your response as follows: <numeric answer (Yes=1; No=0) <Step-by-step reasoning >>*
**Assistant:** *1;*
*The article mentions the arrest of eleven people, including antiwar activist Philip Berrigan, who were part of a demonstration against war. They scaled a fence at the White House and threw red ink into a fountain as a form of protest.*
*The demonstrators are described as being affiliated with an antiwar group, specifically named the D.C. Coalition Against the U.S. in the Middle East.*
*This indicates that the protest was related to a peace movement, as it involved a group actively opposing war in the Middle East.*

Zero-Shot ToT Prompting with Role Play

**System:** *You are an excellent AI language model trained to extract protest information from news articles.*
**User:** *Imagine three different social movement experts are answering this question. All experts will write down one step of their thinking, then share it with the group. Then all experts will proceed to the next step, and so on. If any expert realizes they are wrong at any point, they will leave. The question is that You need to determine whether the following article discusses peace movement: <fulltext >. Please structure your response as follows: <Step-by-step reasoning ><numeric answer (Yes=1; No=0) >*
**Assistant:**
*Step 1: Identify the main activity described in the article.*
*- Expert 1: The article describes a group of people, including a known antiwar activist, who were arrested for scaling a fence at the White House and throwing red ink into a fountain.*
*- Expert 2: The demonstrators are affiliated with an antiwar group, indicating their primary motive is related to opposing war.*
*- Expert 3: The act of throwing red ink to simulate blood suggests a symbolic protest against violence or war.*
*Step 2: Determine the nature of the group involved in the activity.*
*- Expert 1: The group is identified as the D.C. Coalition Against the U.S. in the Middle East, which clearly positions itself against U.S. policies or actions in the Middle East.*
*- Expert 2: The involvement of Philip Berrigan, a notable antiwar activist, supports the interpretation that this is an antiwar protest.*
*- Expert 3: The name of the group and the nature of the protest both align with typical peace movement activities, which often involve symbolic actions to draw attention to the costs of war.*
*Step 3: Assess if the described actions align with those typical of a peace movement.*
*- Expert 1: Peace movements often use dramatic, non-violent actions to make a political point, which is consistent with the described protest.*
*- Expert 2: The use of red ink as a symbol for blood is a strong visual metaphor against war violence, aligning with peace movement strategies.*
*- Expert 3: The target of the protest, the White House, is a strategic location for making a bold statement against national policies on war.*
*Conclusion:*
*- Expert 1: The article describes a protest by an antiwar group using symbolic non-violent action, typical of peace movements.*
*- Expert 2: The involvement of known peace activists and the nature of the demonstration clearly point to a peace movement.*
*- Expert 3: The overall context and details provided in the article align with the characteristics of a peace movement protest.*
*Numeric Answer:1*

# Limitations of Using Generative AI for Text Annotation

Next, we discuss the limitations of using generative AI for text annotation. It is important to note that we are not AI-Luddites; as the technology advances, we anticipate improvements that will mitigate these limitations and challenges, enabling more effective and ethical use of LLMs in social science research. We will use social movement studies as an illustrative example to demonstrate the potential epistemic risks in using generative AI for coding. We want to recognize that these epistemic risks are particularly critical when scholars use generative AI as the primary coder instead of secondary coding assistant. Drawing on studies examining the risks and limitations of LLMs, we discuss how to incorporate those considerations into the prompt development process to eliminate such risks.

## Are Generative AI coding results valid?

Integrating generative AI into social science research, particularly for constructing protest event datasets in social movement studies, presents unique challenges. Movement scholars demand high accuracy due to the complexity and nuance of the protests involved. The primary concern among social scientists when adopting generative AI is whether it can match human-expert level precision in coding complex concepts, since these LLMs may suffer from hallucination and biases (Srinivasan & Chander, 2021; Bail, 2024). Recent studies indicate that LLMs like ChatGPT and other similar technologies can outperform crowd workers in tasks such as text summarization (Gilardi et al., 2023). However, a notable gap remains in achieving expert-level accuracy for domain-specific tasks. We benchmark generative AI-produced results with the DoCA-LDC dataset to systematically assess whether LLMs can achieve expert-level accuracy. We consider different levels of task complexity for classification tasks, including binary (police presence), multi-class (protest size), and multilabel classifications (protest activities). For demonstration, we only use several common strategies including standard zero-short with role-play, zero-short CoT, and zero-shot ToT (see Appendix A for example prompts) by querying OpenAI-GPT4-turbo model using the matched DoCA-LDC dataset.

To demonstrate the potential variations in the relationship between prompting strategies and task complexity, we conducted experiments using the same prompt question, incorporating different triggers to initiate various strategies. In real settings, researchers should have a development set to test the performance of strategies and choose the optimal one. Table 1 presents the accuracy and F1 scores across a variety of text classification tasks. *Police presence* is a binary variable, indicating whether police involvement was mentioned at the event. *Protest size* is a categorical variable scaled from 1 to 6, where a score of 1 denotes a small group of 1-9 individuals, and a score of 6 corresponds to gatherings of 10,000 or more participants. The *protest activity* involves a multi-label classification problem where LLMs are tasked with identifying up to four activities from a predefined list, which includes bannering, candle-lighting, petitioning, holding signs, camping, sit-ins, physical attacks, looting, and more. For assessing the classification performance on protest activities, we adopt the *subset accuracy*. This metric is defined as the proportion of instances where all relevant protest codes listed in DoCA are completely encompassed within the predictions made by

GPT-4. Consequently, a prediction is deemed accurate only if every protest activity code generated by GPT-4 for an article is inclusive of all corresponding codes in DoCA. We show that the coding performance of GPT-4 varies across different levels of task complexity and prompting strategies. Overall, GPT-4 performs pretty well in less complex tasks but not in those multiclass and multilabel classification tasks. Note that we only ran experiments using zero-shot prompts, and results might be different for other prompting strategies such as few-shot prompts.

| Task Complexity | Zero-shot | Zero-shot CoT | Zero-Shot ToT |
|---|---|---|---|
| Binary (Police) | 0.84,0.63 | 0.85, 0.64 | 0.82,0.61 |
| Multiclass (Size) | 0.499, 0.464 | 0.484, 0.449 | 0.461, 0.426 |
| Multilabel (Activity) | 0.548, - | 0.409, - | 0.564, - |

Table 1: GPT-4 Performance across Different Prompting Strategies and Tasks. All methods are with role play prompting.

## Are Generative AI coding results reliable?

Another widely shared concern among scholars is the reliability of using LLMs in social science research. A critical question is whether Generative AI produces consistent results when using the same prompting strategy and parameter settings. Our proposed approach involves running the same prompt multiple times (at least 25) on a development dataset and calculating the intraclass correlation (ICC) to measure inter-rater consistency. Each run is treated as a separate rater. According to Koo and Li's guideline (2016), ICC values below 0.5 indicate poor reliability, values between 0.5 and 0.75 suggest moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values above 0.90 indicate excellent reliability.

To demonstrate, we used the OpenAI GPT-4 model to classify a subset of the DoCA-LDC dataset, focusing on two specific tasks: police presence and protest size. To mimic a development set, we randomly selected 200 articles and queried the OpenAI API 25 times using the same prompt and parameter settings. Figure  illustrates the trend of average accuracy across these 25 classifications. Overall, the accuracy scores fluctuate for both measures when repeated 25 times. We also use intraclass correlation to evaluate the inter-rater consistency (if we treat each run as a rater). The ICC score for police presence is 0.934 (95% CI : [0.921,0.946]), indicating excellent reliability. While the ICC score for protest size is 0.518 (95% CI :[0.463,0.577]), a moderate reliability which means the results are not very reliable. It is quite evident that results are more reliable or consistent when performing simple tasks than complex ones (i.e., binary versus categorical variable).

## Are Generative AI coding results replicable?

The potential lack of replicability is another hurdle that lowers social scientists' enthusiasm to adopt LLMs in their research (Barrie, Palaiologou, & Törnberg, 2024). A recent work conducted by Barrie et al. (2024) shows that the prompt stability of outputs generally

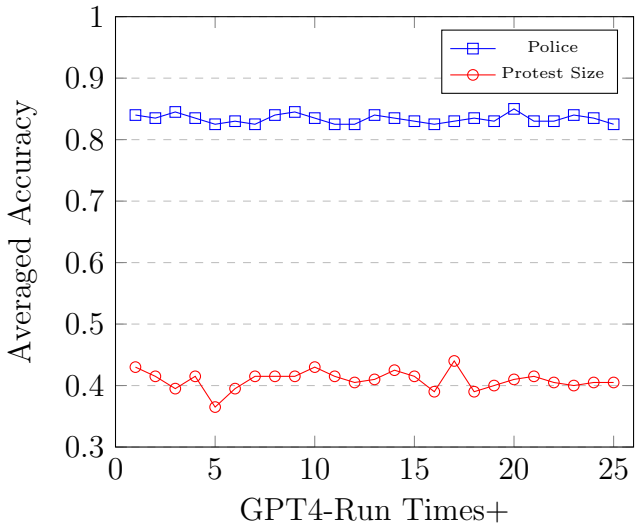


Figure 2: The Consistency of 25 Repeated Outcomes.

decreases with increasing temperature due to the degradation in the quality of prompts. Datasets exhibiting the least prompt stability typically contain data that are either ill-suited or have imprecisely defined outcome constructs. Thus, can other researchers replicate generative AI results using the same method and data? The answer is very complicated. The replicability depends on a variety of factors, including the specific prompt, parameter setting, specific type of model, etc.

*The alchemy of prompt engineering.* The performance of coding tasks is highly sensitive to prompt engineering. In this paper, we tested several common prompting strategies, including zero-shot with role play, CoT, and ToT. Our experiments show that LLM performance varies across different types of prompting strategies. We show that the simple zero-shot with role play performs pretty well across different tasks. But for binary classification, we show that the zero-shot CoT outperforms others, while for multiclass classification, zero-shot ToT works better. This is not to suggest that scholars need to use CoT or ToT in their research but to show that different prompting strategies may lead to different results.

*Model drifting.* The rapid development in generative AI brings the unprecedented challenges to social scientists in tandem with opportunities. New advanced models offer scholars more power to handle multimodal data with better performance. However, this also brings uncertainty and replicability issues. For instance, OpenAI offers a series of different models, such as GPT-4, GPT-4o, and GPT-3.5, for scholars to handle different tasks with different pricing and speed. When some of those legacy models are not accessible to scholars, others cannot replicate their results due to model drifting.

*Personalization and memory.* The alignment of LLMs with human values and norms via reinforcement learning from human feedback was used to maintain the safety of LLMs, but the alignment norms were often dictated by fewer developers or researchers, leading to the exclusion or under-representation issues. To better reflect individual preferences, these proprietary models such as OpenAI GPT-4 models can be tailored to individuals via

the memory or personalization feature. Personalized LLMs can reduce cultural hegemony and improve usefulness, but also bring risks pertaining to privacy infringement and bias reinforcement (Kirk, Vidgen, Röttger, & Hale, 2024). Although personalization is the new frontier in LLM advancement, it poses particular challenges for social scientists regarding replicability. This is particular the case when researchers use these proprietary models to produce coding with memory enabled.

## Are Generative AI text annotation results transparent?

LLMs are like black-box, especially the proprietary models. Scholars have warned the danger of using proprietary LLMs in social science research in part due to the transparency issues (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Ollion, Shen, Macanovic, & Chatelain, 2024). One strategy for social scientists to improve the coding transparency is to ask LLMs to output the reasoning steps which might help us understand the potential logic. In so doing, human evaluators can take advantage of the outputs for cross-validation. The below is the example using zero-shot CoT to ask GPT-4 to determine whether the article discusses peace movement. GPT-4 reasons output the following steps:

1. *The article mentions the arrest of eleven people, including antiwar activist Philip Berrigan, who were part of a demonstration against war. They scaled a fence at the White House and threw red ink into a fountain as a form of protest.*
2. *The demonstrators are described as being affiliated with an antiwar group, specifically named the D.C. Coalition Against the U.S. in the Middle East.*
3. *This indicates that the protest was related to a peace movement, as it involved a group actively opposing war in the Middle East.*

These steps detail how GPT-4 evaluates whether this article discusses a peace movement based on several key information. However, scholars should be cautious about the step-by-step reasoning results because of model hallucination. LLMs may just simply produce results conforming to the prompt instead of learning what LLMs have seen (B. Wang et al., 2023). Thus, we selected 100 news articles to ask GPT4 and LLaMa3 to use zero-shot CoT to output the step by step reasoning. We then manually examine whether these outputs are factually correct. Based on our human evaluation, GPT-4 correctly reasoned these 100 cases as we manually assessed the materials extracted from articles. However, we did find some errors for LlaMa outputs. Below are some examples:

Examples of Llama' Wrong Reasoning - The Inference Goes too Far

**Example 1**

**Output:** *1-Yes*

*Here is my step-by-step reasoning:*

** The article describes an event where David Duke takes the oath of office as a member of the Louisiana Legislature.*

** The event took place in the House chamber, which suggests that it was held at the Louisiana State Capitol building or possibly another government building.*

** Later in the article, it mentions "lawmakers" and "Representative C.D. Jones", who were present at the event and reacted to Mr. Duke's swearing-in.*

** From this context, I infer that police officers would likely be present at such an event, especially considering the controversy surrounding David Duke's appointment.*

** Therefore, I conclude that police were indeed present at the event mentioned in the news article, which is why my answer is 1-Yes.*

**Where the mistake lies:** *The news article does not mention the presence of police, but Llama infer from the context that usually police officers would likely be present at such an event.*

**Example 2**

**Output:** *1-Yes;*

*Step-by-step reasoning: The article mentions police only once, in the following sentence: "Those departments that fail to do so will be required to show why they failed and to undergo an outside review of their recruiting procedures." This suggests that there might have been a mention of police presence at the event, but it does not explicitly state the presence of police.*

**Where the mistake lies:** *The news article does not mention the presence of police, but Llama infer from an irrelevant statement.*

**Example 3**

**Output:** *1-Yes;*

*Step by step reasoning:*

*The article mentions that the student editors have protested the rule by handing out their publications across campus, which could prompt the university to disband their groups. This implies that there may be some interaction or presence of police at this event.*

**Where the mistake lies:** *The news article does not mention the presence of police, but Llama infer from the context that police officers would likely be involved in such situation.*

Another wide concern among social scientists when using LLMs for coding is on the confidence of model output. How confident is the Generative AI classification? When proprietary LLMs do not output the distribution of token probability, then it poses greater epistemic risks as scholars are not informed about the confidence of the results. OpenAI GPT-4 provides a feature that allows scholars to access the *logprob* parameter and calculate the probability. Note that log probability can provide researchers with important insight into the confidence of the model regarding each token it predicts. Here we use police presence as an example to show how we can obtain the token probability based on *logprobs* parameter. We use the

below prompt with setting *logprobs* as True: *You need to determine if it mentions police at the event in the following article: <fulltext >. Please structure your response (1-Yes; 0-No) as follows: <numeric answer >.*

Figure 3 shows the extracted predicted token probability (i.e., our classification result). The majority of these cases reported a value of over 95%. For the ease of presentation, we only report the 137 cases with a value less than 95%. These cases show a relatively low level of probability or confidence in terms of generating the predicted token. These cases oftentimes warrant further investigation.

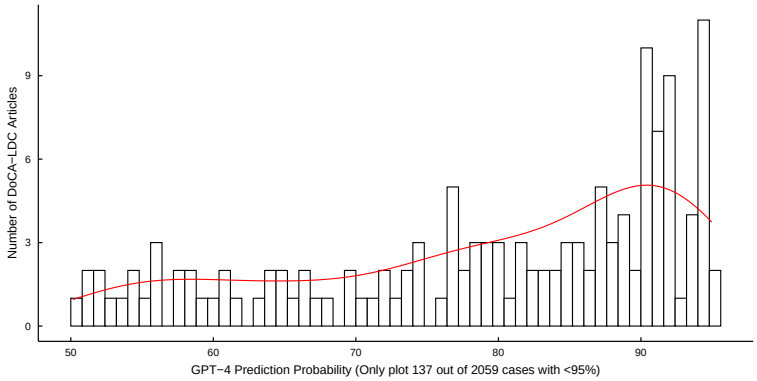


Figure 3: Histogram of Reported Probability for Police Presence

# The Future of Text Annotation with Generative AI

### Ascertaining the role of LLMs for text annotation in social science research

LLMs can serve as invaluable tools in social science research, functioning either as primary coders or as secondary coding assistants. These models are particularly beneficial for handling extensive texts that require significant effort for human coders to process. Scholars can use LLMs to perform intermediate steps such as generating text summarizations, which human coders can then assess for further analysis in downstream tasks. This application proves especially useful when the volume of text is substantial. To obtain detailed and reasoned coding results, researchers need to craft specific prompts, such as the Chain of Thought (CoT), to guide the LLMs effectively. In scenarios where high accuracy is paramount, the dataset is relatively small, and the budget allows for hiring additional research assistants, LLMs typically assume a secondary assistant role to support human coders in their decision-making processes. In scenarios where the error tolerance is relatively high, the dataset is large, and the coding task is not complicated, LLMs can function as primary coders. In other words, scholars fully depend on these models to complete coding tasks such as classification and regression problems. Of course, scholars can use LLMs as primary coders to generate labeled training datasets for fine-tuning their own large language models for large-scale datasets. Given the divergence in the workflow, scholars need to ascertain the role of LLMs in their specific coding tasks.

### Enhancing the quality of AI results with prompt engineering

Our experiments indicate that the quality of generative AI coding outputs is highly sensitive to the specific prompts used. Consequently, scholars need to pay particular attention to prompt engineering—finding the optimal prompt for their specific coding tasks. For LLMs serving as secondary coding assistants, we recommend that researchers experiment with the chain-of-thought prompting strategy. This approach guides LLMs to produce step-by-step reasoning, providing scholars with detailed information for human evaluation and decision-making in coding tasks. For LLMs that serve as primary coders, it is crucial for scholars to create an development dataset to systematically test the performance of different prompting strategies. This testing phase ensures that the chosen prompt candidate is effective before its application to the entire dataset. We also propose that scholars can report on the log probability of predicted tokens and select the cases with low confidence to diagnose the potential issues leading to low accuracy. In this paper, we only showed several common prompt strategies including zero-shot, role play, chain of thought, and tree of thought. Other prompting strategies such as in-context learning with few examples and graph of thoughts. Notably, with sufficient budget, scholars can also fine-tune LLMs for specific downstream tasks with own training datasets.

### Decomposing complex tasks into simple ones such as binary classification

We tested several common coding tasks in social science research, particularly in social movement studies. Our experiments show that GPT-4 performs better in simple binary classification task than multilable and multiclass classification tasks. As the task complexity increases, LLMs' performance worsens. We recommend that scholars should decompose complex coding tasks into simple ones. For instance, scholars can modify multilable or multiclass tasks into a series of binary classification.

### Improving the interpretability by outputting the AI underlying reasoning and coding decision

One of the major concerns of using LLMs in social science coding is its black box nature that lacks interpretability. We propose that scholars can tailor prompts to guide LLMs to produce the underlying reasoning for human evaluation. Instead of outputting the final result, LLMs can create the step-by-step reasoning that explains their coding decision making processes. But scholars should be cautious when relying on generative AI as it may not reflect what it has learn from CoT and the underlying reasoning could be hallucination or simply conform to the guidance of the prompts (B. Wang et al., 2023). But our human evaluation of 100 cases show that the underlying reasoning is reliable in our case of coding police presence in protests. We did find that in some cases, LLMs are hallucinating as they might infer too far.

### Increasing the replicability by making codes and outputs transparent

To improve the replicability of using generative AI social science research, we encourage scholars to deposit and share codes in public repository such as GitHub and SocArxiv, documenting the specific prompt, the specific model, and the specific parameter setting. If

possible, we also call for scholars to share their outputs from LLMs. We acknowledge that scholars should anonymize/remove some of these input/output data that contain sensitive information (Li, Dohan, & Abramson, 2021).

**Prioritizing open-source models over proprietary LLMs for data security and responsible use**

Social scientists should call for responsible and ethical use of generative AI in their own research. One of the primary concerns is related to sharing proprietary and sensitive data. Social scientists often engage with proprietary and sensitive data, such as the New York Times news articles, voter records, interview transcripts, and survey results. When individual researchers use proprietary LLMs from leading companies such as OpenAI and Google, they have to upload their data, which may cause potential data leakage and privacy violation. One common task of social movement scholars is to extract key protest information from news articles. Is it okay to upload these news articles to proprietary LLMs for classification tasks? One common practice is to ascertain the proprietary LLMs do not use your uploaded data for training their own models or share them with third parties. Other qualitative scholars may interview protesters on a series of sensitive questions and use generative AI to transcribe and analyze data but is it okay to share these data with LLMs? One common research protocol is to get permission from participants to use these generative AI tools to analyze their data.

The alternative to use proprietary LLMs is to prioritize open sourced LLMs in social science research, such as LLaMA family models. You can run these open source models with langchain, ollama, or huggingface transformers on your own computers or remote servers in a secure environment. In this study, we evaluated the performance of the Meta Llama3 model, specifically we experimented with the 4-bit quantized version known as Meta-Llama-3-8B, across the three tasks. We applied the same zero-shot prompt technique used for the GPT-4 model at a temperature of 0.2. We deployed the model in our local computer and employed the ollama interface. The completion times for each task were as follows: 1.6 hours for the binary task assessing police presence, 2.2 hours for the multiclass task determining the number of participants, and 5.5 hours for the multilabel task identifying protest activities. The results of Llama3 is shown in Table 2.

| Task Complexity | Zero-shot CoT |
|---|---|
| Binary (Police) | 0.762,0.566 |
| Multiclass (Size) | 0.367,0.284 |
| Multilabel (Activity) | 0.099, - |

Table 2: LLaMa3 Performance based on CoT.

# Conclusion and Discussion

Generative AI has been applauded since it can improve social science research including automated text analysis (Ziems et al., 2024; Bail, 2024). But can we replace human coders

with LLMs in text analysis? Our findings indicate both promise and pitfalls. In this paper, we propose a framework to integrate LLMs in social science coding by assuming two roles: LLMs as secondary coding assistant and LLMs as primary coding decision-maker. When LLMs are the secondary coding assistants, human coders can use the information extracted by prompted LLMs to assist their decision making processes. The epistemic risk associated with this role is relatively low since LLMs perform pretty well in information retrieval and human coders make the final decision. However, when LLMs assume the primary coder role, scholars should be cautious of replacing human coders with LLMs as they tend to misportray marginalized groups as more like outgroup imitations instead of ingroup representations, flatten demographic groups with ignoring within-group heterogeneity, and essentialize identities with amplifying stereotypes (Dillion, Tandon, Gu, & Gray, 2023). To better communicate the epistemic risks associated with using LLMs in social science coding, we propose several practice guidelines, including the before-prompt engineering and after-evaluation pipeline, the explainable coding by outputting underlying reasoning, improving transparency by sharing codes with specific prompts, model parameters, and returned outcomes, prioritizing open sourced LLMs for data security, and obtain permission for using proprietary LLMs to analyze sensitive data.

# References


Bail, C. A. (2024, May). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. Retrieved 2024-05-11, from https://www.pnas.org/doi/10.1073/pnas.2314021121 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2314021121

Barrie, C., Palaiologou, E., & Törnberg, P. (2024). *Prompt stability scoring for text annotation with large language models.* Retrieved from https://arxiv.org/abs/2407.02039

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Virtual Event Canada: ACM. Retrieved 2023-10-27, from https://dl.acm.org/doi/10.1145/3442188.3445922 doi: 10.1145/3442188.3445922

Bonikowski, B., Luo, Y., & Stuhler, O. (2022, November). Politics as Usual? Measuring Populism, Nationalism, and Authoritarianism in U.S. Presidential Campaigns (1952–2020) with Neural Language Models. *Sociological Methods & Research*, *51*(4), 1721–1787. Retrieved 2023-10-27, from http://journals.sagepub.com/doi/10.1177/00491241221122317 doi: 10.1177/00491241221122317

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. Retrieved 2023-10-27, from https://proceedings.neurips.cc/paper_files/paper/2020/hash/

1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

Chae, Y., & Davidson, T. (2023). Large Language Models for Text Classification: From Zero-Short Learning to Fine-Tuning. Retrieved 2023-10-27, from https://files.osf.io/v1/resources/sthwk/providers/osfstorage/64e61d8f9dbc3f068f681622?format=pdf&action=download&direct&version=2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*. Retrieved 2023-10-27, from https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(23)00098-0 (Publisher: Elsevier)

Do, S., Ollion, , & Shen, R. (2022, December). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, 004912412211345. Retrieved 2023-10-27, from http://journals.sagepub.com/doi/10.1177/00491241221134526 doi: 10.1177/00491241221134526

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., ... Sui, Z. (2023, June). *A Survey on In-context Learning.* arXiv. Retrieved 2024-05-18, from http://arxiv.org/abs/2301.00234 (arXiv:2301.00234 [cs]) doi: 10.48550/arXiv.2301.00234

Earl, J., Martin, A., McCarthy, J. D., & Soule, S. A. (2004). The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.*, *30*, 65–80.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023, July). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. Retrieved 2024-04-23, from https://www.pnas.org/doi/10.1073/pnas.2305016120 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2305016120

Hanna, A. (2017). Mpeds: Automating the generation of protest event data.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hulbert, D. (2023, May). *Using tree-of-thought prompting to boost chatgpt's reasoning.* https://github.com/dave1010/tree-of-thought-prompting. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.10323452 doi: 10.5281/ZENODO.10323452

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 1–10.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., ... Dong, X. (2024, March). *Better Zero-Shot Reasoning with Role-Play Prompting.* arXiv. Retrieved 2024-05-18, from http://arxiv.org/abs/2308.07702 (arXiv:2308.07702 [cs]) doi: 10.48550/arXiv.2308.07702

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163.

Li, Z., Dohan, D., & Abramson, C. M. (2021). Qualitative coding in the computational era:

A hybrid approach to improve reliability and reduce effort for coding ethnographic interviews. *Socius*, *7*, 23780231211062345.

Liu, D. M., & Salganik, M. J. (2019, January). Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World*, *5*, 237802311984980. Retrieved 2023-10-27, from http://journals.sagepub.com/doi/10.1177/2378023119849803 doi: 10.1177/2378023119849803

Long, J. (2023). *Large language model guided tree-of-thought.*

Lu, Y., & Pan, J. (2022). The pervasive presence of chinese government content on douyin trending videos. *Computational Communication Research*, *4*(1).

Manning, B. S., Zhu, K., & Horton, J. J. (2024, April). *Automated Social Science: Language Models as Scientist and Subjects.* arXiv. Retrieved 2024-04-22, from http://arxiv.org/abs/2404.11794 (arXiv:2404.11794 [econ, q-fin])

Messeri, L., & Crockett, M. J. (2024, March). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. Retrieved 2024-04-22, from https://www.nature.com/articles/s41586-024-07146-0 (Publisher: Nature Publishing Group) doi: 10.1038/s41586-024-07146-0

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

Ollion, , Shen, R., Macanovic, A., & Chatelain, A. (2024, January). The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, *6*(1), 4–5. Retrieved 2024-04-23, from https://www.nature.com/articles/s42256-023-00783-6 (Publisher: Nature Publishing Group) doi: 10.1038/s42256-023-00783-6

Srinivasan, R., & Chander, A. (2021, August). Biases in AI systems. *Communications of the ACM*, *64*(8), 44–49. Retrieved 2023-08-11, from https://dl.acm.org/doi/10.1145/3464903 doi: 10.1145/3464903

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2023-03-29, from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Veselovsky, V., Ribeiro, M. H., & West, R. (2023). *Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks.*

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2023, June). *Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters.* arXiv. Retrieved 2024-05-18, from http://arxiv.org/abs/2212.10001 (arXiv:2212.10001 [cs]) doi: 10.48550/arXiv.2212.10001

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... Zhou, D. (2023, March). *Self-Consistency Improves Chain of Thought Reasoning in Language Models.* arXiv. Retrieved 2024-05-18, from http://arxiv.org/abs/2203.11171 (arXiv:2203.11171 [cs]) doi: 10.48550/arXiv.2203.11171

Wankmüller, S. (2022, December). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological Methods & Research*, 00491241221134527. Retrieved 2024-01-06, from https://doi.org/10.1177/

00491241221134527 (Publisher: SAGE Publications Inc) doi: 10.1177/00491241221134527

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023, December). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models.* arXiv. Retrieved 2024-05-18, from http://arxiv.org/abs/2305.10601 (arXiv:2305.10601 [cs]) doi: 10.48550/arXiv.2305.10601

Zhang, H., & Pan, J. (2019). Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*(1), 1–57.

Zhang, Y., Lin, H., Wang, Y., & Fan, X. (2023, August). Sinophobia was popular in Chinese language communities on Twitter during the early COVID-19 pandemic. *Humanities and Social Sciences Communications*, *10*(1), 1–12. Retrieved 2023-12-06, from https://www.nature.com/articles/s41599-023-01959-6 (Number: 1 Publisher: Palgrave) doi: 10.1057/s41599-023-01959-6

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 1–55.

# Appendix

## Appendix A

Below are a list of example prompts we used for GPT-4 to get protest size.

**Zero-shot**: You need to extract the number of participants and tell me which category: 1 – Small, few handful (1-9 people); 2 – Group, committee (10-49 people); 3 – Large, gathering (50-99 people); 4 – Hundreds, mass, mob (100-999 people); 5 – Thousands (1000-9999 people); 6 – Tens of thousands (10000 or more people) in the following article: $< fulltext >$. Please structure your numeric response (1-6) as follows: <numeric answer >

**Zero-shot CoT:** You need to extract the number of participants and tell me which category: 1 – Small, few handful (1-9 people); 2 – Group, committee (10-49 people); 3 – Large, gathering (50-99 people); 4 – Hundreds, mass, mob (100-999 people); 5 – Thousands (1000-9999 people); 6 – Tens of thousands (10000 or more people) in the following article: $< fulltext >$. Please structure your numeric response (1-6) as follows: <numeric answer >. Let us think step by step.

**Zero-shot ToT:** Imagine three different social movement experts are answering this question. All experts will write down one step of their thinking, then share it with the group. Then all experts will proceed to the next step, and so on. If any expert realizes they're wrong at any point, they will leave. The question is that You need to extract the number of participants and tell me which category: 1 – Small, few handful (1-9 people); 2 – Group, committee (10-49 people); 3 – Large, gathering (50-99 people); 4 – Hundreds, mass, mob (100-999 people); 5 – Thousands (1000-9999 people); 6 – Tens of thousands (10000 or more

people) in the following article: $< fulltext >$. Please structure your numeric response (1-6) as follows: <numeric answer >