Concept-Guided Chain-of-Thought Prompting for Pairwise Comparison Scoring of Texts with Large Language Models

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker & Solomon Messing
Center for Social Media and Politics, New York University
{pyw230,jonathan.nagler,joshua.tucker,solomon.messing}@nyu.edu

Abstract

Existing text scoring methods require a large corpus, struggle with short texts, or require hand-labeled data. We develop a text scoring framework that leverages generative large language models (LLMs) to (1) set texts against the backdrop of information from the near-totality of the web and digitized media, and (2) effectively transform pairwise text comparisons from a reasoning problem to a pattern recognition task. Our approach, concept-guided chain-of-thought (CGCoT), utilizes a chain of researcher-designed prompts with an LLM to generate a concept-specific breakdown for each text, akin to guidance provided to human coders. We then pairwise compare breakdowns using an LLM and aggregate answers into a score using a probability model. We apply this approach to better understand speech reflecting aversion to specific political parties on Twitter, a topic that has commanded increasing interest because of its potential contributions to democratic backsliding. We achieve stronger correlations with human judgments than widely used unsupervised text scoring methods like Wordfish. In a supervised setting, besides a small pilot dataset to develop CGCoT prompts, our measures require no additional hand-labeled data and produce predictions on par with RoBERTa-Large fine-tuned on thousands of hand-labeled tweets. This project showcases the potential of combining human expertise and LLMs for scoring tasks.

1 Introduction

Text scoring methods are used to analyze the latent positions of texts, such as tweets, documents, or speeches, along one or more dimensions. Popular text scoring methods, such as Wordfish, are used extensively in the social and political sciences (see, e.g., Laver et al., 2003; Slapin & Proksch, 2008; Wu et al., 2019; Benoit et al., 2019; Rheault & Cochrane, 2020; Bailey, 2023). However, these approaches require a large text corpus, struggle with short texts, and do not have a mechanism to precisely target the latent concept of interest. Other approaches, such as fine-tuning pre-trained language models such as RoBERTa (Liu et al., 2019), require non-trivial quantities of hand-labeled data. These methods often fail with texts such as social media posts: they are short, it is often not clear how to pick documents for identification, and the usages of words rapidly shift over time.

In this article, we present a novel text scoring framework that leverages the embedded information and pattern recognition capabilities of generative large language models (LLMs). LLMs can set the texts against the backdrop of information accumulated from nearly the totality of the web and digitized media, giving greater context to short texts such as social media posts. The core idea is to use an LLM to conduct pairwise comparisons between two texts. In other words, we prompt an LLM to pick the text that reflects a greater quantity of some latent or abstract concept of interest, such as which text contains greater aversion to a particular political party. But instead of directly pairwise comparing texts, we compare concept-specific breakdowns of the texts. These concept-specific breakdowns are generated using an approach we call concept-guided chain-of-thought (CGCoT) prompting. CGCoT prompting is a framework that uses a series of researcher-crafted questions that examine the constituent parts of the concept of interest in a given text. The text and the LLM's answers

to the CGCoT prompts for that text form the text's concept-specific breakdown. These researcher-crafted prompts, akin to a codebook used to guide content analysis, are the same across all texts, making the concept-specific breakdowns directly comparable in pairwise comparisons.

We use the LLM to pairwise compare concept-specific breakdowns of the texts along a targeted concept. Following this, we score the LLM's answers using the Bradley-Terry model, a probabilistic model that predicts the outcome of pairwise comparisons based on the latent abilities of the items being compared (Bradley & Terry, 1952). We call the resulting "ability" scores of the texts CGCoT pairwise scores. Because we craft the CGCoT prompts and the basis of the pairwise comparisons, we can precisely target the latent concept of interest. We rely on pairwise comparisons rather than attempting to use the LLM to directly adjudicate the level, intensity, or scalar value of concepts like aversion for a number of reasons. First, pairwise comparisons enable us to establish a measure where differences are meaningful; ordinal rankings indicate order without quantifying the gaps between ranked items. Second, scalar values directly generated from the LLM lack transparency in their generation and calibration. In contrast, the Bradley-Terry model derives estimated scalar values from pairwise comparisons, providing a more interpretable and reliable basis for the score. Third, pairwise comparisons are easier to complete and allow for more subtle distinctions, improving reliability over labeling tasks over single items (Carlson & Montgomery, 2017).

We apply the proposed approach to better understand affective polarization on Twitter (see, e.g., Moran Yarchi & Kligler-Vilenchik, 2021; Nordbrandt, 2021). Affective polarization is the tendency for partisans to dislike or distrust members of the opposing party (Iyengar et al., 2012; Druckman et al., 2021). Affective polarization is typically studied using surveys but, except for Chen et al. (2022), has not been extensively studied in the context of political non-elites on social media. Chen et al. (2022) label tweets using a dichotomous classification for aversion to Republicans and then again for aversion to Democrats in tweets. Expressions of aversion to a specific party are an inevitability on social media, but strong expressions of aversion threaten productive discourse online, incentivize antidemocratic rhetoric, and, when analyzed in the aggregate, can be a sign of the declining health of democracy (Finkel et al., 2020).

We calculate two aversion scores using a random sample of political tweets from Chen et al. (2022): an aversion to Republicans score and an aversion to Democrats score. We develop a series of questions used as CGCoT prompts with the LLM—here, we use GPT-3.5—that identify and describe aversion to specific parties in a given tweet. Specifically, we use three prompts: the first prompt uses GPT to summarize the tweet; the second prompt uses GPT to identify the primary party that is the focus or target of the tweet; the third prompt uses GPT to identify whether aversion is expressed towards the targeted party. We apply these prompts to each tweet in the corpus: the tweet's concept-specific breakdown is the original tweet and all the LLM's responses to the CGCoT prompts about the tweet. Then, using a sample of pairwise comparisons between the concept-specific breakdowns, we prompt GPT-3.5 to select the breakdown that exhibits greater aversion to a specific party. Lastly, using the outcomes of these pairwise comparisons, we estimate an aversion score for each text using the Bradley-Terry model.

To validate the CGCoT pairwise scores, we compare CGCoT pairwise scores with three alternative unsupervised text scoring approaches. We show that pairwise comparing the concept-specific tweet breakdowns yields scores that are more strongly associated with human judgments than the three alternative text scoring approaches. The results indicate that using both CGCoT prompts and pairwise comparisons is important for deriving high-quality scores. We also show that our continuous score, which only requires a small set of pilot hand-labeled tweets to develop the CGCoT prompts, is competitive with state-of-the-art supervised approaches. Binarizing the scores—classifying all tweets with a CGCoT pairwise score above a cutoff as exhibiting aversion and all observations with a score below a cutoff as not exhibiting aversion—gives us a set of predictions that nearly match (for aversion to Democrats) or exceed (for aversion to Republicans) the performance of a RoBERTa-Large (Liu et al., 2019) model fine-tuned on 3,000 hand-labeled tweets. Overall,

the success of CGCoT in measuring aversion suggests that pairing substantive knowledge with LLMs can be immensely useful for solving social science text measurement problems.

2 Related Work

Our approach is situated in a rapidly growing literature on using generative LLMs for social science applications (see, e.g., Törnberg, 2023; Rathje et al., 2023; Argyle et al., 2023; Bisbee et al., 2023; Wu et al., 2023a). The works that study text, such as analyzing text along psychological constructs (Rathje et al., 2023), analyze the text as given. Our proposed text scoring approach breaks the text down into the concept of interest's constituent parts using prompts developed by substantive knowledge about the targeted latent concept; it involves substantive expert knowledge to a much greater extent than the other research studies that analyze text using LLMs.

This framework also speaks to a large body of text scoring methods. These text scoring methods roughly fall into unsupervised methods and supervised methods. Unsupervised methods such as Wordfish (Slapin & Proksch, 2008) and word embedding methods (Kozlowski et al., 2019; Wu et al., 2019; An et al., 2018; Kwak et al., 2021) typically require post hoc dimensional interpretation or selection of keywords to represent underlying concepts of interest and a large corpus. Supervised methods, such as WordScores (Laver et al., 2003; Lowe, 2008) and approaches that use pairwise comparisons of texts (Loewen et al., 2012; Simpson et al., 2019), rely on hand-labeled texts or manual pairwise comparisons of texts and typically focus on measuring one targeted concept within the corpus. Our approach minimizes the need for hand-labeling or pairwise comparisons of texts, can measure multiple targeted concepts within the same corpora, does not require a large corpus, relies on a transformers-based language model rather than bag-of-words, and leverages the researchers' substantive knowledge to precisely target the latent concept of interest instead of relying on post hoc dimensional interpretation.

Many recent works have also shown that generative LLMs can outperform crowd workers for text-annotation tasks (see, e.g., Gilardi et al., 2023; Törnberg, 2023). These papers usually focus on discrete classification tasks. The highly structured nature of these tasks, with clear gold standard comparisons, plays to the strengths of LLMs. However, it is less clear if these advantages hold with continuous, open-ended, and contentious concepts such as aversion to opposing parties. Some works have also directly queried scalar values from the LLM (see, e.g., O'Hagan & Schein, 2023). However, LLMs are not inherently designed to produce consistent and calibrated numbers. The generated scalar values may change with different prompts, training updates, and so on. Pairwise comparisons, on the other hand, are a task that is more aligned with the LLM's training: determining which of two items or texts have a greater quality is similar to other natural language tasks such as entailment, and is more compatible with the LLM's strengths in natural language tasks.

In our proposed framework, pairwise comparisons are made over *concept-specific breakdowns* of the texts rather than the texts themselves. It is well-documented that generative LLMs often make mistakes in problems that require intermediate reasoning steps (Liu et al., 2022; Wei et al., 2023; Kojima et al., 2023; Zhou et al., 2023; Wu et al., 2023b). Wei et al. (2023) propose "chain-of-thought," which prompts the LLM to explicitly generate its intermediate reasoning steps, leading to improved responses on problem-solving tasks such as arithmetic or question-answering. In related work, Zhou et al. (2023) use the LLM to automatically break a problem down into subproblems using few-shot examples, an approach they call least-to-most prompting; the LLM then solves each subproblem to solve a larger, harder problem. But despite these innovations, it is still unclear whether generative LLMs are able to "reason." For instance, Wu et al. (2023b) find that LLMs perform poorly on "counterfactual" tasks, which are variants of reasoning tasks that an LLM performs well on.

3 The Text Pairwise Comparison Framework using CGCoT

We build on previous chain-of-thought work by proposing CGCoT prompting. Rather than using the LLM to generate its own intermediate reasoning steps or its own breakdown of

the problem into subproblems, we leverage the researcher's substantive knowledge about the targeted latent concept to craft a sequence of questions that identify and describe the targeted concept and its constituent parts in the text. This approach is analogous to using a codebook for qualitative content analysis (see, e.g., Fonteyn et al., 2008). In other words, we use the LLM's pattern recognition capabilities in conjunction with researcher-guided prompts to generate the intermediate reasoning steps that we would ideally like the LLM to reason through for a targeted concept when making pairwise comparisons. CGCoT effectively shifts the pairwise comparisons of text from a reasoning problem to a pattern recognition task. For example, if we are scoring the level of aversion expressed towards a target in a text, we can use a series of prompts that summarizes the text, identifies the primary focus or target of the text, and identifies whether aversion is expressed towards that target.

To be more precise, generating the concept-specific breakdowns follows these steps:

1. Let t_i be a text, for some $i \in \{1, ..., n\}$, and let $(x_1, ..., x_m)$ be a set of m researcher-crafted concept-guided prompts to extract specific information from t_i
2. Sample a token sequence using an LLM with parameters θ, $s_{1,i} \sim p_\theta\left(s|x_1, t_i\right)$
3. Then, sample a token sequence $s_{2,i} \sim p_\theta\left(s|x_2, t_i, s_{1,i}, x_1\right)$
4. Repeat this iterative sampling approach with all prompts, such that the last token sequence sampled is $s_{m,i} \sim p_\theta\left(s|x_m, t_i, s_{m-1,i}, x_{m-1}, s_{m-2,i}, x_{m-2}, ..., s_{1,i}, x_1\right)$
5. Concatenate all sampled token sequences $s_{j,i}$ for $j \in \{1, ..., m\}$ and text t_i to form the concept-specific breakdown
6. For concept-guided prompt development, compare the concept-specific breakdown with a set of hand-labeled text data to assess if concepts and entities are correctly identified; if not, refine prompts $(x_1, ..., x_m)$

After generating the concept-specific breakdowns for each text in the corpus, we pairwise compare the breakdowns instead of the texts themselves. The pairwise comparison prompt is determined based on the application. We then use the outcomes of the pairwise comparisons with the Bradley-Terry model to estimate a scalar score for each text. The Appendix provides a technical overview of the Bradley-Terry model.

4 Application: Analyzing Affective Polarization on Social Media

We apply the proposed framework to better analyze affective polarization on social media. Affective polarization is defined as the tendency for partisans to dislike or distrust members of the opposing party (Iyengar et al., 2012; Druckman et al., 2021). The topic has commanded increasing interest because of its potential contributions to democratic backsliding and political violence (Finkel et al., 2020). In this paper, we specifically focus on aversion expressed towards specific parties.

We use GPT-3.5, with its default temperature and nucleus sampling hyperparameter values, to pairwise compare political tweets from Chen et al. (2022). Chen et al. (2022) use 3,000 hand-labeled tweets to fine-tune a RoBERTa model that classifies tweets as containing aversion to Republicans and/or aversion to Democrats in a multilabel setting. 500 tweets are used for validation, and 500 tweets are set aside as a test set. These tweets were selected using a set of political keywords from the Twitter Decahose. Each tweet was labeled by 3 coders from Surge AI. A brief overview of their hand-labeling approach is provided in the Appendix. We estimate measures using the CGCoT approach outlined in the previous section. We score the test set tweets to make our results comparable with the results from Chen et al. (2022).

To generate aversion to Republican-specific breakdowns, we created the following concept-guided prompts using definitions and concepts from the literature on affective polarization (e.g., Iyengar et al., 2012; Finkel et al., 2020; Druckman et al., 2021; Chen et al., 2022):

```
1. Summarize the Tweet.
```

2. We broadly define Republicans to include any member of the Republican Party/GOP, the Republican Party/GOP generally, conservatives, right-wingers, anyone that supports MAGA, or the alt-right. We broadly define Democrats to include any member of the Democratic Party, the Democratic Party generally, liberals, leftists, or progressives. Using these definitions, does the Tweet primarily focus on Republicans (or a Republican) or Democrats (or a Democrat)? The focus can be on a specific member of a party.
3. If the Tweet primarily focuses on Republicans based on your above answer, does the Tweet express aversion, dislike, distrust, blame, criticism, or negative sentiments of Republicans (or a Republican)? If the Tweet primarily focuses on Democrats based on your above answer, does the Tweet express aversion, dislike, distrust, blame, criticism, or negative sentiments of Democrats (or a Democrat)? If the Tweet focuses on neither party, answer ‘‘N/A.’’
4. Using only your answer immediately above, does the Tweet express aversion, dislike, distrust, blame, criticism, or negative sentiments of Republicans (or a Republican)?

We used a similar set of prompts to create aversion to Democrats-specific breakdowns: the first three questions are the same, except we flipped the order of the definitions, and we replaced the word "Republicans" with "Democrats" in the fourth question. Flipping the order of the definitions controls for potential order effects in the prompts used to generate the breakdown for each aversion to a specific party. The last prompt provides information on which party is the subject of any aversion. Taking advantage of the conversational aspect of generative LLMs and as detailed in the previous section, we prompt each question sequentially. The concept-specific breakdown is the concatenation of the original tweet and the LLM's responses to all four prompts.

The process of developing these prompts is analogous to creating a codebook for human coders and qualitative content analysis (see, e.g., Elo & Kyngäs, 2008; Fonteyn et al., 2008). To "make sense of the data and whole," we labeled 50 tweets as containing aversion to Republicans, 50 tweets as containing no aversion to Republicans, 50 tweets as containing aversion to Democrats, and 50 tweets as containing no aversion to Democrats from Chen et al. (2022)'s training set; we did not use their labels (Elo & Kyngäs, 2008). We then iterated on an initial set of CGCoT prompts and examined outputs from GPT-3.5 until the summaries and party identifications aligned with expectations across labeled tweets. In short, we combined content analysis techniques with prompt engineering methods such as changing specific words, repeatedly providing definitions, and splitting up complex questions.

We apply this set of prompts to each tweet to obtain a concept-specific breakdown. To pairwise compare these concept-specific breakdowns, we prompt GPT-3.5 to pick the breakdown that expresses a greater aversion to a specific party. The Appendix details the exact pairwise comparison prompt used for each of the two types of aversion measured and provides an example of one such pairwise comparison.

There are a total of 124,750 potential matchups. To reduce the total number of matchups, we sample 20 matchups per tweet ID for a total of 10,000 matchups. The Bradley-Terry model does not require complete matchups to estimate scores for each tweet.[1] The scores are then rescaled to a 0-1 range, making each score independent of any reference tweet. We also estimate 95% confidence intervals based on quasi-standard errors, which are also independent of any reference tweet (Firth & De Menezes, 2004).

4.1 Face Validity of the Aversion CGCoT Pairwise Scores

The aversion CGCoT pairwise scores show face validity when examining the rank ordering of tweets. We showcase three tweet examples with different aversion scores. The first has the highest score, the second has a score closest to the mean of the measure, and the third

[1] In the Appendix, we show that CGCoT pairwise scores estimated using 5, 10, and 15 matchups per tweet ID highly correlate for both aversion scores.

has the lowest score. We first examine the tweets associated with these specific aversion to Republicans CGCoT pairwise scores.

- Highest: "#ArrestTrump and all associates immediately. Try them for treason. PUBLIC CAPITAL PUNISHMENT. They're all traitors & murderers. NEVER FORGET all the blood they have on their hands #Trump #Republicans #1776RestorationMovement #January6thCommittee #Jan6Hearings"
- Middling: "@newsmax America doesn't have to worry about socialism. The threat to American democracy is not socialism, it is the radicalized and unhinged Republicans."
- Lowest: "House Republican leader Kevin McCarthy calls the situation at the southern border a humanitarian and national security crisis."

The ordering of these tweets are intuitively consistent with our concept of interest. The intensity of aversion expressed in tweets decreases from the highest to the lowest CGCoT pairwise scores. The middling tweet expresses aversion to Republicans but is not nearly as intense as the tweet with the highest CGCoT pairwise score.

We then examine the tweets associated with the same ordering of the aversion to Democrats CGCoT pairwise scores.

- Highest: "Very true. I keep wondering the same. The Demonrats have sunk deeper into the swamp! They've been exposed, but that's only let them do their dirty deeds defiantly out in the open. Why has no one been stopped or paid for their crimes?"
- Middling: "I kind of like blocking the liberals and only seeing common sense comments."
- Lowest: "Get over it, scoop! It's obvious he's waiting to put it all together with the DNC this week. Then I bet he'll be holding pressers with his policies and all. But, this incessant reporter whining is for the birds."

There is a similar pattern in the intensity of the aversion expressed in the tweets over the CGCoT pairwise scores. The middling tweet is a jab at liberals, but it is not the same type of criticism leveled at Democrats as the tweet with the highest CGCoT pairwise score. We also note that "Demonrats" is correctly identified as an intense insult to Democrats, an example of the potential contextualizing capabilities of generative LLMs.

4.2 Comparing CGCoT Pairwise Scores with Human Coders

We compare both scores with the number of human coders that labeled a tweet as containing aversion to Republicans or aversion to Democrats. Despite the differences in granularity between the discrete labels assigned by human coders and the continuous measures estimated using our proposed approach, we can gain insights into the overall prevalence and intensity of aversion expressed in the tweets by comparing the number of coders labeling each tweet as containing aversion to the distribution of tweets along the estimated scores.

We first compare the aversion to Republicans CGCoT pairwise scores against the number of human coders that labeled a tweet as containing aversion to Republicans (Figure 1). We find a positive correlation between these two measures. We note that the wider distribution of CGCoT pairwise scores for tweets with two or three coders labeling a tweet as containing aversion to Republicans is a function of the pairwise comparisons: tweets that do not contain any aversion to Republicans tend to tie with each other in matchups, resulting in scores "clumping." In the Appendix, we examine the tweets with the lowest aversion to Republicans CGCoT pairwise scores among tweets labeled by three coders as containing aversion to Republicans.

We repeat this exercise for aversion to Democrats. We again find a positive correlation when comparing aversion to Democrats CGCoT pairwise scores against the number of human coders that labeled a tweet as containing aversion to Democrats. Figure 2 illustrates this comparison. Again, in the Appendix, we examine the tweets with the lowest aversion to Democrats CGCoT pairwise scores among tweets labeled by three coders as containing aversion to Democrats.

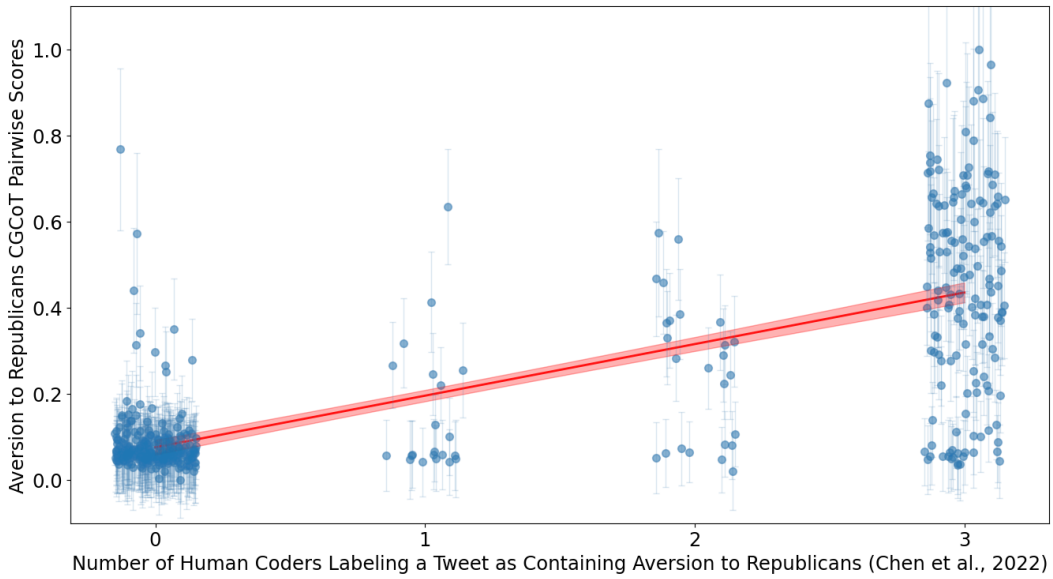

Figure 1: Aversion to Republicans CGCoT pairwise scores are strongly associated with the number of coders applying an aversion to Republicans label. CGCoT pairwise score estimates are shown with 95% confidence intervals derived from quasi-standard errors. A linear regression line is drawn through the points with a 95% confidence band.

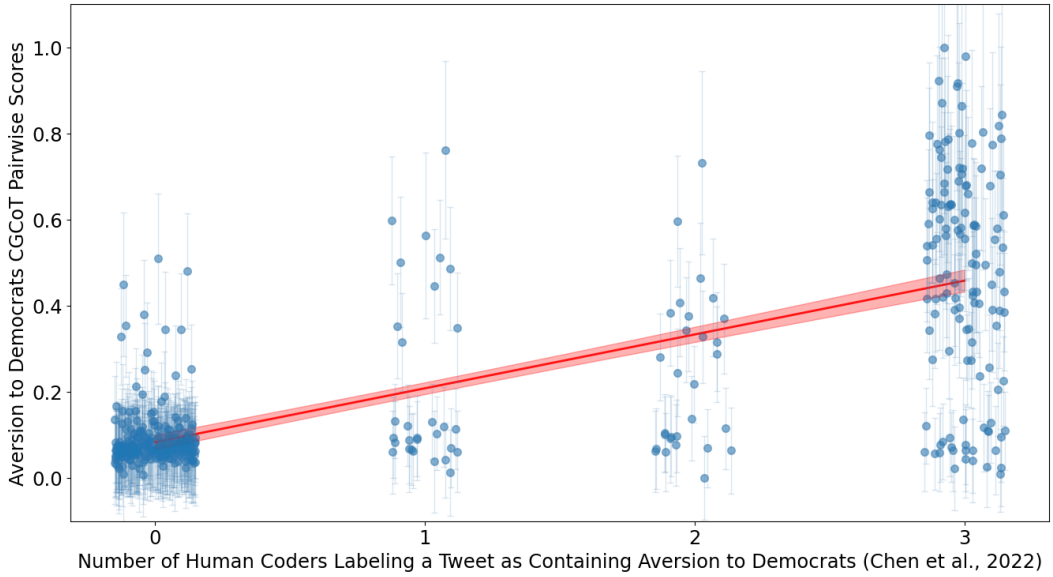

Figure 2: Aversion to Democrats CGCoT pairwise scores are strongly associated with the number of coders applying an aversion to Democrats label. CGCoT pairwise score estimates are shown with 95% confidence intervals derived from quasi-standard errors. A linear regression line is drawn through the points with a 95% confidence band.

We also calculate Spearman's rank correlations between human labels and an array of text scoring methods, including CGCoT pairwise scores. We do this for both types of aversion using (1) Wordfish with the tweets only; (2) WordFish with the tweets' concept-specific breakdowns; (3) the pairwise comparison approach with GPT-3.5 using the tweets only ("non-CGCoT tweets-only pairwise scores"); and (4) CGCoT pairwise scores. Wordfish treats word j's count in document i as a Poisson-distributed variable. Parameter λ_{ij} depends on

the document's position on a unidimensional latent dimension and the word's importance in discriminating between positions (Slapin & Proksch, 2008). Wordfish is one of the most popular unsupervised text scoring methods in the social and political sciences and has been used in many recent works (see, e.g., Bailey, 2023). Details about how we fit the Wordfish models are in the Appendix. Table 1 shows these correlations for both aversion to Republicans and Democrats.

	Aversion to...	
	Republicans	**Democrats**
Wordfish Using Tweets Only	0.03	0.04
Wordfish Using Concept-Specific Breakdowns	0.55	0.22
Non-CGCoT Tweets-Only Pairwise Scores	0.55	0.56
CGCoT Pairwise Scores	0.64	0.61

Table 1: CGCoT pairwise scores outperform other unsupervised text scoring methods based on Spearman's rank correlation coefficient with the number of human coders labeling a tweet as containing aversion to Republicans and Democrats.

The results demonstrate the utility of both CGCoT and pairwise comparisons, with notable gains in correlation when moving from the use of tweets to concept-specific breakdowns of the tweets, and from Wordfish to pairwise comparisons. Our proposed procedure of using CGCoT with LLM pairwise comparisons yields a measure that most closely aligns with human judgments compared to other text scoring approaches.

4.3 CGCoT Pairwise Scores are Competitive with Supervised Learning Approaches

To further analyze the validity of the aversion CGCoT pairwise scores, we create binary labels using cutoffs in the two aversion measures. For each measure, we label all tweets with CGCoT pairwise scores above the mean of the CGCoT pairwise scores as 1, and all tweets with CGCoT pairwise scores below the mean as 0. While binarizing the CGCoT pairwise scores by labeling observations above the mean as 1 and those below the mean as 0 is slightly arbitrary, this approach is also guided by a principled decision to use the central tendency of the measure as the threshold. Future work will use a training set to choose a more accurate cutoff, which would almost certainly improve results. We repeat this process for the measure generated using GPT-3.5 pairwise comparisons of the tweets only (i.e., non-CGCoT tweets-only comparisons). We also compare CGCoT pairwise scores with a RoBERTa-Large model; the model was fine-tuned using Chen et al. (2022)'s training set of 3,000 hand-labeled political tweets with hyperparameters chosen using the validation set. Table 2 contains the performance metrics of the two cutoff classifiers and the RoBERTa-Large model.

Classifier	**Aversion to**	**F1**	**Precision**	**Recall**
Tweets-Only Pairwise Scores Cutoff Clf.	Republicans	0.70	0.64	0.76
	Democrats	0.67	0.58	0.79
Fine-Tuned RoBERTa-Large Model	Republicans	0.81	0.82	0.80
	Democrats	0.81	0.82	0.81
CGCoT Pairwise Scores Cutoff Clf.	Republicans	0.84	0.89	0.79
	Democrats	0.79	0.84	0.75

Table 2: The F1 score, precision, and recall of the non-CGCoT tweets-only pairwise scores cutoff classifier, the RoBERTa-Large classifier fine-tuned on Chen et al. (2022)'s training set, and the CGCoT pairwise scores cutoff classifier.

The performance metrics show that CGCoT pairwise scores outperform non-CGCoT tweets-only pairwise scores on all metrics, except for recall for aversion to Democrats. The metrics also show that the aversion to Republicans CGCoT pairwise scores outperform the fine-tuned RoBERTa-Large model on F1 and precision and are nearly equivalent on recall.

Similarly, the aversion to Democrats CGCoT pairwise scores are comparable with the fine-tuned RoBERTa-Large model on F1. The former performs better on precision and the latter performs better on recall. Again, the CGCoT pairwise comparison cutoff classifier's predictions were calculated using no hand-labeled tweets, except for a small set of 200 hand-labeled pilot tweets used to develop the CGCoT prompts for the concept-specific breakdowns. In other words, our approach drastically reduces the need for training coders and hand-labeling data while still retaining the expertise needed to analyze this nuanced and complex concept expressed in social media posts.

5 Conclusion

We develop a novel text scoring framework that leverages pairwise comparisons and a prompting procedure called concept-guided chain-of-thought (CGCoT), which creates concept-specific breakdowns of the texts. We then prompt GPT-3.5 to make pairwise decisions between the concept-specific breakdowns along a latent concept. We call the resulting measures CGCoT pairwise scores. We apply the approach to better understand affective polarization on social media and derive two novel latent measures of aversion to specific political parties in tweets.

We find that the measures largely correlate with how humans interpret aversion to Republicans and aversion to Democrats on Twitter. We also find that using *both* CGCoT and pairwise comparisons with LLMs is crucial, as scores that do not use one or both of these techniques are demonstrably worse. Moreover, our approach can estimate scores that are competitive with or outperform both unsupervised and supervised approaches. We show that using a cutoff with the score yields binary classifications that are highly competitive with a RoBERTa-Large model fine-tuned on *thousands* of human-labeled tweets. Our findings suggest that using substantive knowledge with generative LLMs can not only be useful for calculating high-quality continuous measures but can also be useful for generating discrete classifications with high performance with the use of very little labeled data.

Our findings align with the notion that the LLM synthesizes information about complex concepts such as affective polarization, allowing it to reliably and coherently evaluate latent constructs, abstract concepts, stances, and sentiments within texts using its pattern recognition capabilities. While we provide information about what constitutes a "Republican" and "Democrat" being targeted in our CGCoT prompts, we still assume that the LLM is able to identify Republican or Democratic figures and organizations, such as Donald Trump, Joe Biden, and the DCCC. Additionally, we assume that the LLM possesses the capability to recognize aversion within the presented texts. Again, this capability stems from the presence of many instances of political contention in social media and other forms of content. However, the precise impact of this training on both CGCoT and the pairwise comparisons remains obscured due to the black box nature of GPT and requires further investigation.

It is also well-known that there is a significant mental toll on people identifying attacks against individuals/groups and harmful content for data labeling and content moderation purposes. Our approach, which rivals the binary prediction performance of language models fine-tuned on thousands of hand-labeled social media posts, can help avoid having human coders label thousands of posts containing potentially harmful or sensitive content.

There are still many open questions around this framework. We apply the approach using only one generative LLM to one substantive problem. We have also not yet analyzed the consistency of pairwise comparisons over repeated promptings, and we have not yet considered how the timing of social media posts comports with the LLM's training data. Ongoing work aims to answer many of these open questions, including expanding the framework to use recently developed techniques such as retrieval-augmented generation (see, e.g., Lewis et al., 2020). In spite of the approach's limitations, it estimates scores that agree with human judgments of the texts along different dimensions of interest. It lends a better understanding of how human-machine collaboration can be used to improve the quantification of complex latent concepts.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2450–2461, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1228. URL https://aclanthology.org/P18-1228.

Lisa P. Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. Ai chat assistants can improve conversations about divisive topics, 2023.

Michael Bailey. Measuring candidate ideology from congressional tweets and websites, 2023. URL https://ssrn.com/abstract=4350550.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774, 2018. doi: 10.21105/joss.00774. URL https://quanteda.io.

Kenneth Benoit, Kevin Munger, and Arthur Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2): 491–508, 2019. doi: https://doi.org/10.1111/ajps.12423. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12423.

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. Synthetic replacements for human survey data? the perils of large language models, 5 2023. URL osf.io/preprints/socarxiv/5ecfa.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345, 12 1952. ISSN 0006-3444. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

David Carlson and Jacob M. Montgomery. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review*, 111 (4):835–843, 2017. doi: 10.1017/S0003055417000302.

Haohan Chen, Zhanna Terechshenko, Patrick Y. Wu, Richard Bonneau, and Joshua A. Tucker. Detecting political sectarianism on social media: A deep learning classifier with application to 2020-2022 tweets. 2022.

James N. Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. Affective polarization, local contexts and public opinion in america. *Nature Human Behaviour*, 5(1):28–38, 2021. ISSN 2397-3374. doi: 10.1038/s41562-020-01012-5. URL https://doi.org/10.1038/s41562-020-01012-5.

Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008. doi: https://doi.org/10.1111/j.1365-2648.2007.04569.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2007.04569.x.

Eli J. Finkel, Christopher A. Bail, Mina Cikara, Peter H. Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C. McGrath, Brendan Nyhan, David G. Rand, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James N. Druckman. Political sectarianism in america. *Science*, 370(6516):533–536, 2020. doi: 10.1126/science.abe1715. URL https://www.science.org/doi/abs/10.1126/science.abe1715.

David Firth. *qvcalc: Quasi Variances for Factor Effects in Statistical Models*, 2023. URL https://CRAN.R-project.org/package=qvcalc. R package version 1.0.3.

David Firth and Renée X. De Menezes. Quasi-variances. *Biometrika*, 91(1):65–80, 03 2004. ISSN 0006-3444. doi: 10.1093/biomet/91.1.65. URL https://doi.org/10.1093/biomet/91.1.65.

Marsha E. Fonteyn, Margaret Vettese, Diane R. Lancaster, and Susan Bauer-Wu. Developing a codebook to guide content analysis of expressive writing transcripts. *Applied Nursing Research*, 21(3):165–168, 2008. ISSN 0897-1897. doi: https://doi.org/10.1016/j.apnr.2006.08.005. URL https://www.sciencedirect.com/science/article/pii/S0897189706000991.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2305016120.

Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly*, 76(3):405–431, 09 2012. ISSN 0033-362X. doi: 10.1093/poq/nfs038. URL https://doi.org/10.1093/poq/nfs038.

Gary King, Patrick Lam, and Margaret E. Roberts. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4): 971–988, 2017. doi: https://doi.org/10.1111/ajps.12291. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12291.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019. doi: 10.1177/0003122419877135. URL https://doi.org/10.1177/0003122419877135.

Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. Frameaxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644, 2021. doi: 10.7717/peerj-cs.644.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003. doi: 10.1017/S0003055403000698.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3154–3169, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.225. URL https://aclanthology.org/2022.acl-long.225.

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

Peter John Loewen, Daniel Rubenson, and Arthur Spirling. Testing the power of arguments in referendums: A Bradley–Terry approach. *Electoral Studies*, 31(1):212–221, 2012. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2011.07.003. URL https://www.sciencedirect.com/science/article/pii/S0261379411000953. Special Symposium: Germany's Federal Election September 2009.

Will Lowe. Understanding wordscores. *Political Analysis*, 16(4):356–371, 2008. doi: 10.1093/pan/mpn004.

Christian Baden Moran Yarchi and Neta Kligler-Vilenchik. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1-2):98–139, 2021. doi: 10.1080/10584609.2020.1785067. URL https://doi.org/10.1080/10584609.2020.1785067.

Maria Nordbrandt. Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media & Society*, 0(0), 2021. doi: 10.1177/14614448211044393. URL https://doi.org/10.1177/14614448211044393.

Sean O'Hagan and Aaron Schein. Measurement in the age of llms: An application to ideological scaling, 2023.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J Van Bavel. Gpt is an effective tool for multilingual psychological text analysis, 5 2023. URL psyarxiv.com/sekf5.

Ludovic Rheault and Christopher Cochrane. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133, 2020. doi: 10.1017/pan.2019.26.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5716–5728, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1572. URL https://aclanthology.org/P19-1572.

Jonathan B. Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008. doi: https://doi.org/10.1111/j.1540-5907.2008.00338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2008.00338.x.

Heather Turner and David Firth. Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 48(9):1–21, 2012. doi: 10.18637/jss.v048.i09. URL https://www.jstatsoft.org/index.php/jss/article/view/v048i09.

Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Patrick Y. Wu, Walter R. Mebane, Jr., Logan Woods, Joseph Klaver, and Preston Due. Partisan associations of twitter users based on their self-descriptions and word embeddings. Presented at APSA 2019, 2019.

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. Large language models can be used to estimate the latent positions of politicians, 2023a.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2023b.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.

A Appendix

A.1 Bradley-Terry Model

The Bradley-Terry model assumes that in a contest between two players i and j, the odds that i beats j in a matchup are α_i/α_j, where α_i and α_j are positive-valued parameters that indicate latent "ability" (Bradley & Terry, 1952). We can define $\alpha_i \equiv \exp(\lambda_i)$. Then, the log-odds of i beating j is

$$\log\left[\frac{\Pr(i \text{ beats } j)}{\Pr(j \text{ beats } i)}\right] = \lambda_i - \lambda_j$$

The intuition is that the larger the value of λ_i compared to λ_j, the more likely it is for player i to beat player j.

We translate the above matchup into a contest between two concept-specific breakdowns. Using the aversion measures as our example, the estimated λ parameters are the measures of the level of aversion to a specific party. We denote the concept-specific breakdown exhibiting greater aversion to a specific party as the "winner." For ties, we considered these 0.5 wins for both tweets in the matchup. Turner & Firth (2012) find that this approach yields ability parameter estimates that highly correlate with more complex approaches that explicitly deal with ties. We use the bias-reduced maximum likelihood estimation approach implemented in the `BradleyTerry2` R package with GPT's answers to pairwise comparisons to estimate the level of aversion expressed towards a specific party in each tweet. These scores are the aversion CGCoT pairwise scores. The estimated λ parameters are relative to a reference tweet, but this choice is unimportant because we rescale the parameters to the unit interval.

We also estimate standard errors for the estimated λ parameters. These standard errors are interpreted relative to a reference tweet. We calculate quasi-variances from these standard errors, which can be interpreted as reference-free estimates of the variance of the score of each tweet. Confidence intervals derived from these quasi-variances can be directly compared. We use the `qvcalc` package to calculate quasi-standard errors (Firth, 2023). The 95% confidence intervals of the estimated scores are derived from these quasi-standard errors.

A.2 A Brief Overview of Hand-Labeling Tweets for Aversion to Republicans and Democrats Used in Chen et al. (2022)

Chen et al. (2022) used Surge AI to label 4,000 political tweets. 3,000 tweets formed the training set, 500 tweets formed the validation set, and 500 tweets formed the test set. These tweets were selected from the Twitter Decahose using an expanded keyword filtering process (King et al., 2017). Coders were given the following instructions and examples:

> Aversion or dislike/distrust toward members of political party (Democratic or Republican). Tweets that contain aversion tend to express strong negative sentiment towards the opposing party, its elites and/or supporters often expressed in hateful speech.
>
> Examples of aversion:
>
> @TheDTrain3 @nytimes Democrats are not the reason police are corrupt. Go into any red area in the country and your police is the "good ole boy system". The difference is the democrats want change and republicans could care less like the bootlickers they are...also republicans run the poorest areas!
>
> Yes, the Senate will probably let people starve. Just to own the libs & then try to blame it all on the Biden administration. The GOP is just disgusting. McConnell is the worst of the lot.

Each tweet was labeled by three coders. Labels were not mutually exclusive: tweets could express both aversion to Republicans and aversion to Democrats. 33.9% of the tweets were labeled as expressing aversion to Republicans, and 31.2% of the tweets were labeled as expressing aversion to Democrats. The average Cohen's κ was 0.795 for aversion to Republicans and 0.794 for aversion to Democrats.

A.3 Pairwise Comparison Prompt

We pairwise compare the concept-specific breakdowns for aversion to Republicans using the following prompt:

```
Tweet Description 1: [concept-specific breakdown for the
first tweet above]
Tweet Description 2: [concept-specific breakdown for the
second tweet above]
Based on these two Tweet Descriptions, which Tweet
Description expresses greater aversion, dislike, distrust,
blame, criticism, or negative sentiments of Republicans:
Tweet Description 1 or Tweet Description 2? If both
equally express or do not express aversion, distrust, blame,
criticism, or negative sentiments of Republicans, reply with
‘‘Neither’’ or ‘‘Tie.’’
```

We use this exact same prompt for aversion against Democrats, except we replace the word "Republicans" with "Democrats." We also use the same prompt to directly compare tweets, except we replace all instances of "Tweet Description" with "Tweet" and all instances of "Tweet Descriptions" with "Tweets." We use another prompt to extract the answers from these pairwise comparisons, detailed in the next section.

A.4 Extracting the Tweet that Exhibits Greater Aversion to Republicans and Democrats in Each Matchup

For each pairwise comparison prompt, GPT-3.5 typically returns a small paragraph explaining its choice. Instead of restricting GPT-3.5's answers to only "Tweet Description 1" or "Tweet Description 2," we use a separate prompt to extract the answers. We find that this two-step process improves GPT's responses in pairwise comparisons. To extract the answer from the response, we first concatenate the model's response to the pairwise comparison with the following text:

```
In the above Text, which Tweet Description is described to
be expressing greater aversion, dislike, distrust, blame,
criticism, or negative sentiments of Republicans: Tweet
Description 1 or Tweet Description 2? Return only ‘‘Tweet
Description 1’’ or ‘‘Tweet Description 2’’. If neither Tweet
Descriptions are described to be more likely to be expressing
greater aversion, dislike, distrust, blame, criticism, or
negative sentiments of Republicans, reply with ‘‘Tie.’’
```

This is then used as a prompt for GPT-3.5. Ties and answers that deviate from either "Tweet Description 1," "Tweet Description 2," or "Tie" are manually fixed. We use the same prompt for extracting answers from non-CGCoT tweets-only comparisons, except we replace the "Tweet Description" with just "Tweet."

A.5 Example of Applying CGCoT on Two Tweets and Pairwise Comparing the Concept-Specific Breakdowns

We walk through an example of one pairwise comparison using concept-specific breakdowns. The first tweet reads: "He's such a lowlife piece of shit. Isn't this slander? Can

he be sued? And the GOP are silent. What crap they are. They did this. They own this. #NeverForget #VoteBlueToSaveAmerica." The concept-specific breakdown for this tweet, using the CGCoT prompts for aversion to Republicans, is the following.

```
(0) Original Tweet: ‘‘He’s such a lowlife piece of shit.
Isn’t this slander? Can he be sued? And the GOP are
silent. What crap they are. They did this. They own this.
#NeverForget #VoteBlueToSaveAmerica’’
(1) The tweet contains a link and expresses anger and
frustration towards an individual, calling them a derogatory
term. The author questions whether the person’s actions
constitute slander and suggests the possibility of a lawsuit.
They also criticize the Republican Party for their silence,
blaming them for the situation and urging others to vote for
Democratic candidates to protect America. The hashtags used
are #NeverForget and #VoteBlueToSaveAmerica.
(2) The Tweet primarily focuses on criticizing Republicans
(or the Republican Party/GOP). It expresses dissatisfaction
with their silence and holds them responsible for a certain
issue. It also includes a hashtag (#VoteBlueToSaveAmerica)
encouraging support for the Democratic Party as an
alternative.
(3) The Tweet expresses aversion, dislike, distrust, blame,
criticism, and negative sentiments towards Republicans (or
the Republican Party/GOP). It criticizes their silence and
assigns responsibility to them, using negative language to
describe their actions.
(4) Yes, the Tweet expresses aversion, dislike, distrust,
blame, criticism, and negative sentiments towards Republicans
(or the Republican Party/GOP).
```

The second tweet reads: "Conservative stupidity will do that." The concept-specific breakdown for this tweet, using the same CGCoT prompts for aversion to Republicans, is the following.

```
(0) Original Tweet: ‘Conservative stupidity will do that.’
(1) The tweet criticizes conservative individuals by
suggesting that their stupidity is responsible for certain
negative consequences.
(2) The tweet primarily focuses on conservatives, which
includes Republicans, right-wingers, and individuals who
support MAGA (Make America Great Again).
(3) The tweet expresses criticism and negative sentiments
towards conservatives, including Republicans and individuals
who align with right-wing ideologies.
(4) Yes, the tweet expresses criticism and negative sentiments
towards conservatives, including Republicans.
```

We then pairwise compare these two concept-specific breakdowns using GPT-3.5 using the following prompt:

```
Tweet Description 1: [concept-specific breakdown for the
first tweet above]

Tweet Description 2: [concept-specific breakdown for the
second tweet above]
```

```
Based on these two Tweet Descriptions, which Tweet
Description expresses greater aversion, dislike, distrust,
blame, criticism, or negative sentiments of Republicans:
Tweet Description 1 or Tweet Description 2?  If both
equally express or do not express aversion, distrust, blame,
criticism, or negative sentiments of Republicans, reply with
‘‘Neither’’ or ‘‘Tie.’’
```

Inputting this prompt into GPT-3.5, the model returns the following answer.

```
Tweet Description 1 expresses greater aversion, dislike,
distrust, blame, criticism, or negative sentiments of
Republicans compared to Tweet Description 2.
```

We repeat this exercise for each matchup between the tweets' concept-specific breakdowns. Because of the sheer number of pairwise comparisons, we randomly sample 20 matchups per tweet, giving us a total of 10,000 matchups for each aversion measure. The answers elicited from GPT-3.5 are used with the Bradley-Terry model to estimate the aversion to Republicans and aversion to Democrats measures.

A.6 Correlations Between Scores with Differing Number of Matchups

The reported results are from 20 matchups per tweet ID, for a total of 10,000 matchups. We analyze how the CGCoT pairwise scores calculated using 5, 10, 15, and 20 matchups per tweet ID correlate for both measures. Results are in Tables 3 and 4. Across all configurations, correlations are greater than 0.90.

	5	10	15	20
5	1.000	0.937	0.940	0.953
10	0.937	1.000	0.964	0.979
15	0.940	0.964	1.000	0.986
20	0.953	0.979	0.986	1.000

Table 3: Pearson correlation between CGCoT pairwise scores 5, 10, 15, and 20 matchups per tweet ID for aversion to Republicans.

	5	10	15	20
5	1.000	0.934	0.929	0.943
10	0.934	1.000	0.958	0.976
15	0.929	0.958	1.000	0.984
20	0.943	0.976	0.984	1.000

Table 4: Pearson correlation between CGCoT pairwise scores 5, 10, 15, and 20 matchups per tweet ID for aversion to Democrats.

A.7 Tweets with the Lowest CGCoT Pairwise Scores that were Labeled by Three Coders as Containing Aversion

A.7.1 Aversion to Republicans

Looking at just the tweets labeled by three coders as containing aversion to Republicans, we examined the three tweets with the lowest aversion to Republicans CGCoT pairwise scores. The text of these tweets is as follows.

1. Dear @POTUS: Are those forgotten men & women who you say never protest the same people showing up at statehouses armed w/military-style weaponry? Are those same forgotten ones the same who call themselves Boogaloo? Or are they those very fine Nazis you favor? All of the above?

2. Cadet bone-spur & tribe be innocent then they should welcome investigations clearing their good name besmirched by furtive conniving fake news liberals.
3. Trump campaign thought their 'huge news' on pre-existing conditions had Democrats cornered – but it backfired spectacularly https://t.co/Ue8ugrCzRF

GPT-3.5 makes mistakes in the interpretation of certain phrases in two of these tweets: it misinterprets a message addressed to the @POTUS account as directed towards President Biden, not President Trump, and it did not recognize "Cadet bone-spur" to be a derisive nickname for Trump. In the third tweet, GPT interpreted a vague headline describing something backfiring against the Trump campaign as not expressing aversion to Republicans, an arguably correct interpretation.

A.7.2 Aversion to Democrats

Looking at just the tweets labeled by three coders as containing aversion to Democrats, we again examine the three tweets with the lowest aversion to Democrats CGCoT pairwise scores. The text of these tweets is as follows.

1. Maybe if Yang and Tulsi were running the party, I might have a much better opinion of the Democratic Party. However as it stands, I cannot. I do respect Yang for wanting to help fighters get paid fairly. My brother boxed for 20 years and doesnt have anything to show for it.
2. This is what the MSM and libs won't ever tell you. And folks reply to this with all sorts of whataboutism, that it's perfectly fine these cops were injured by violent protesters. So that means all Jan 6 defendants should never have been arrested? AmIright?
3. Regular law enforcement like any other regulated profession so that bad cops cańt just get rehired at the next department over." Wtf does this mean? Proof your statements, Dems.

Here, GPT-3.5 interpreted the text differently from humans. For example, it did not interpret someone describing how they do not respect the Democratic Party as expressing aversion to Democrats; in another tweet, it did not interpret "libs" as an insult towards liberals. In the third tweet, the author asked Democrats to "proof your statements," which GPT-3.5 did not interpret as an insult or criticism of Democrats.

A.8 Using Wordfish to Estimate the Aversion Measures

The primary goal of Wordfish (Slapin & Proksch, 2008) is to estimate the position of a document along a single dimension. The assumption is that the rate that tweet i mentions word j is drawn from a Poisson distribution. The functional form of the model is

$$y_{ij} \sim Poisson(\lambda_{ij})$$

$$\lambda_{ij} = \exp\left(\alpha_i + \psi_j + \beta_j \omega_i\right)$$

where y_{ij} is the count of word j in tweet i, α is the set of tweet fixed effects, ψ is the set of word fixed effects, β is an estimate of a word-specific weight that reflects the importance of word j in discriminating between positions, and ω_i is tweet i's position. We fit this model using the `quanteda` R package (Benoit et al., 2018).

We used standard preprocessing steps: we removed symbols, numbers, and punctuation. We also stemmed the words. Lastly, we had to impose minimum word counts and word-document frequency counts to prevent non-convergence. For aversion to Republicans, a word had to be used at least 4 times across at least 4 tweets. For aversion to Democrats, a word had to be used at least 5 times across at least 5 tweets. We used these configurations when using Wordfish with both the tweets and the concept-specific breakdowns.