

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

Patrick Y. Wu<sup>1</sup>, Jonathan Nagler<sup>1,2</sup>, Joshua A. Tucker<sup>1,2</sup>, and Solomon Messing<sup>1</sup>

<sup>1</sup>Center for Social Media and Politics, New York University

<sup>2</sup>Department of Politics, New York University

Version: October 30, 2023

[Click here for the latest version](#)

## Abstract

Existing text scaling methods require a large corpus, struggle with short texts, or require labeled data. We develop a text scaling method leveraging generative large language models (LLMs), setting texts against the backdrop of information from the near-totality of the web. Our approach, concept-guided chain-of-thought (CGCoT), utilizes expert-crafted prompts to generate a concept-specific breakdown for each text, akin to content analysis guidance. We then pairwise compare breakdowns using an LLM, and aggregate answers into a measure using a probability model. CGCoT transforms pairwise text comparisons from a reasoning problem to a pattern recognition task. We apply this approach to scale affective speech on Twitter, achieving stronger correlations with human judgments than methods like Wordfish. Besides a small pilot dataset to develop CGCoT prompts, our measures require no labeled data and produce predictions competitive with RoBERTa-Large fine-tuned on thousands of human-labeled tweets. We showcase the potential of integrating expertise with LLMs.

# Introduction

Text scaling methods are quantitative techniques used to analyze the latent positions of texts, such as tweets, documents, or speeches, along one or more dimensions. Popular text scaling methods, such as Wordfish, are used extensively in the social sciences (see, e.g., 1–5). However, these approaches require a large text corpus, struggle with short texts, and do not have a mechanism to precisely target the abstract concept of interest. Other approaches, such as fine-tuning pre-trained language models such as RoBERTa (6), require 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 scaling method 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, 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. 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 method 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.

Using the LLM to pairwise compare concept-specific breakdowns of the texts along a targeted concept, we then scale the LLM’s answers using the Bradley-Terry model (7). We

call the resulting scores for the texts CGCoT pairwise scores. Because we craft the CGCoT prompts and the basis of the pairwise comparisons, we can precisely target the abstract concept of interest. We rely on pairwise comparisons rather than attempting to use the LLM to adjudicate the presence or absence of concepts like aversion for a number of reasons. First, pairwise comparisons enable us to establish a scale where differences are meaningful; ordinal rankings indicate order without quantifying the gaps between ranked items. Second, at least for human tasks, pairwise comparisons are easier to complete and allow for more subtle distinctions, improving reliability over labeling tasks over single items (8).

These 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 (9–13). Wei *et al.* (10) 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.* (12) 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.” Wu *et al.* (13), for example, find that LLMs perform poorly on “counterfactual” tasks, which are variants of reasoning tasks that an LLM performs well on.

We build on these ideas 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 target 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., 14). 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 mak-

ing pairwise comparisons. Effectively, CGCoT shifts the pairwise comparisons of text from a reasoning problem to a pattern recognition problem. For example, if we are scaling 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/target of the text, and identifies whether aversion is expressed towards that target.

We apply the proposed approach to measuring affective polarization expressed in tweets (see, e.g., 15, 16). Affective polarization is the tendency for partisans to dislike or distrust members of the opposing party (17, 18). Affective polarization is typically studied using surveys but, except for Chen *et al.* (19), has not been extensively studied among political non-elites on social media. Chen *et al.* (19) label tweets for aversion against Republicans and aversion against Democrats in tweets, but these are binary labels. Expressions of aversion against a specific party are an inevitability on social media, but strong expressions of aversion on social media threaten productive discourse online, incentivize antidemocratic rhetoric, and, when analyzed in the aggregate, can be a sign of the declining health of democracy (20).

To validate the pairwise comparison approach with concept-specific tweet breakdowns, we calculate two affective polarization scales using a random sample of political tweets from Chen *et al.* (19): an aversion against Republicans scale and an aversion against Democrats scale. We develop a series of questions used as CGCoT prompts with the LLM—here, we use ChatGPT-3.5—that identify and describe aversion towards specific parties in a given tweet. Specifically, we use three prompts: the first prompt asks ChatGPT to summarize the tweet; the second prompt asks ChatGPT to identify the primary party that is the focus or target of the tweet; the third prompt asks ChatGPT to identify whether aversion is expressed towards the targeted party. The tweet’s concept-specific breakdown is the original tweet and all the LLM’s answers to these questions about the tweet. We then prompt ChatGPT-3.5 to select the concept-specific breakdown that exhibits greater aversion against a specific party, and scale its resulting answers using the Bradley-Terry model.

We compare CGCoT pairwise scores with three alternative unsupervised text scaling approaches. The first two use Wordfish, a popular unsupervised Poisson text scaling model (2). We use Wordfish with the tweets and the tweet’s concept-specific breakdowns. The third text scaling approach is the resulting scale from using ChatGPT-3.5 to pairwise compare the raw tweets. We show that pairwise comparing the concept-specific tweet breakdowns generates scales that are more strongly associated with human judgments than all three alternative text scaling approaches. The results indicate that using both CGCoT prompts and pairwise comparisons are important for deriving high-quality scales.

We also show that our continuous scale, which only requires a small set of pilot data to develop the CGCoT prompts, is competitive with state-of-the-art supervised approaches. Binarizing the scales—treating all observations 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 against Democrats) or exceed (for aversion against Republicans) the performance metrics of a RoBERTa-Large (6) model fine-tuned on 3,000 human labeled tweets. Overall, the success of CGCoT in measuring affective polarization suggests that pairing substantive knowledge with LLMs can be immensely useful for solving social science text-as-data measurement problems.

## Results

**Data** We use ChatGPT-3.5 to pairwise compare political tweets from Chen *et al.* (19). Chen *et al.* (19) fine-tune a RoBERTa model that classifies whether tweets contain aversion against Republicans and/or aversion against Democrats. They used 3,000 tweets for fine-tuning, 500 tweets for validation, and 500 tweets as the test set. These tweets were selected using a set of political keywords and were sampled from all users on Twitter. Each tweet was labeled by 3 coders from SurgeAI. We create scales of each of these two concepts using the CGCoT method outlined below. We scale the test set tweets to make our results comparable

with theirs.

**Concept-Guided Chain-of-Thought** To create aversion against Republican-specific breakdowns, we created the following series of questions using definitions and concepts from the literature on affective polarization (e.g., 17–20):

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 against 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 the word “Democrats” in the fourth question. Taking advantage of the conversational aspect of generative LLMs, we prompt each question sequentially. The concept-specific breakdown is the concatenation of the original tweet and the answers to all four questions. In short, the process of creating these prompts is analogous to creating a codebook for human coders and qualitative content analysis (see, e.g., 14, 21). The Methods section contains greater details about the process we used to create these CGCoT prompts.

We apply this set of prompts to each tweet to obtain a concept-specific breakdown for each tweet. We then use ChatGPT-3.5 to pairwise compare these concept-specific breakdowns. The Methods section works through an example of one such pairwise comparison, including the prompts used for the pairwise comparison of concept-specific breakdowns.

## Face Validity of CGCoT Pairwise Scores

**The Affective Polarization CGCoT pairwise scores show face validity when examining the rank ordering of tweets**

We showcase three tweet examples from each of the affective polarization scales. The first has the highest, the second is around the mean, and the third shows the lowest CGCoT pairwise scores, demonstrating the scales' face validity. We begin with the aversion against 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 appears intuitively consistent with our concept of interest. There is a gradation of the intensity of aversion expressed in tweets from the highest to lowest CGCoT pairwise scores. The middling tweet expresses aversion against Republicans but is not nearly as intense as the tweets with the highest CGCoT pairwise scores.

Looking at the same ordering for the aversion against Democrats CGCoT pairwise scores, we get the following tweets.

- 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, again, a gradation 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.

## **Validating CGCoT Pairwise Scores**

### **Comparing CGCoT with human-coded labels**

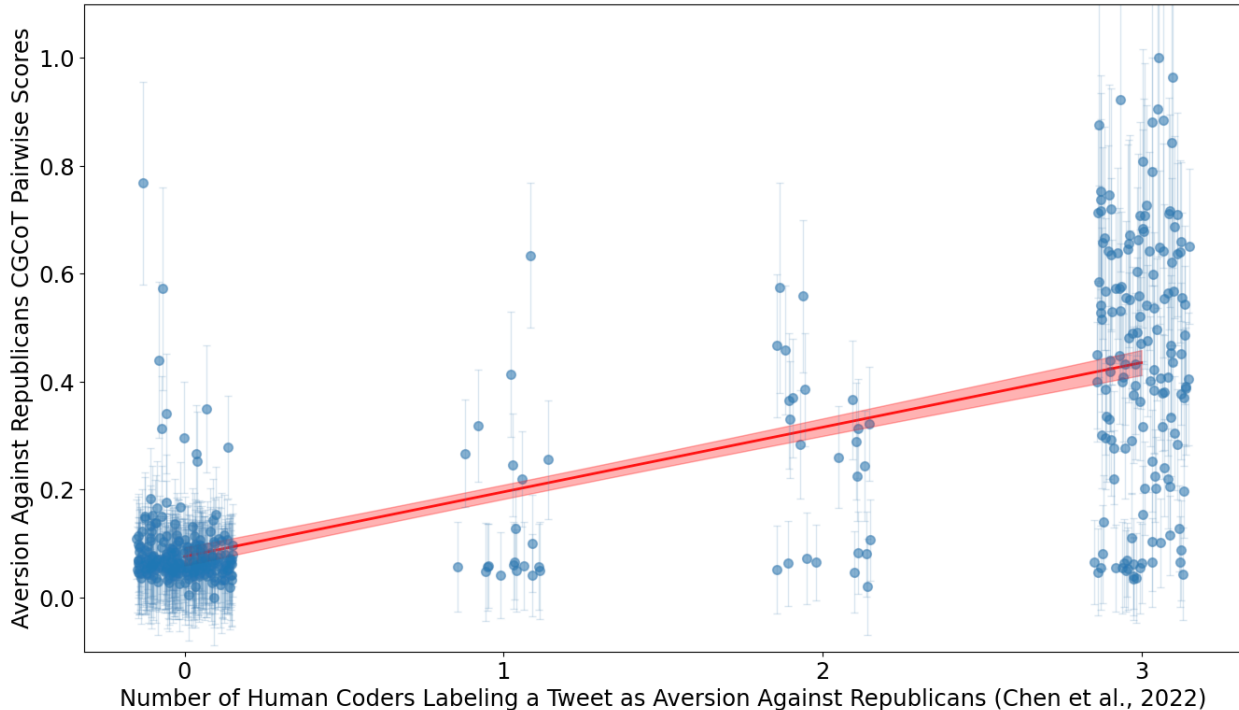
We first compare the aversion against Republicans CGCoT pairwise scores against the number of human coders that labeled a tweet as containing aversion against Republicans (Figure 1). The less ambiguity in text expressing aversion, the more agreement we expect among human coders. We note that the wider distribution of CGCoT pairwise scores for tweets with two or three coders labeling a tweet as containing aversion against Republicans is a function of the pairwise comparisons: tweets that do not contain any aversion against Republicans tend to tie with each other in matchups, resulting in scores “clumping.” In the Supplementary Materials, we examine the tweets labeled by three coders as containing aversion against Republicans with the lowest aversion against Republicans CGCoT pairwise scores.

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

### **CGCoT pairwise scores outperform other unsupervised text scaling methods**

We compare Spearman’s correlations between human labels and an array of text scaling methods, including CGCoT pairwise scores. We do this for both aversion against Democrats and Republicans for the following methods:



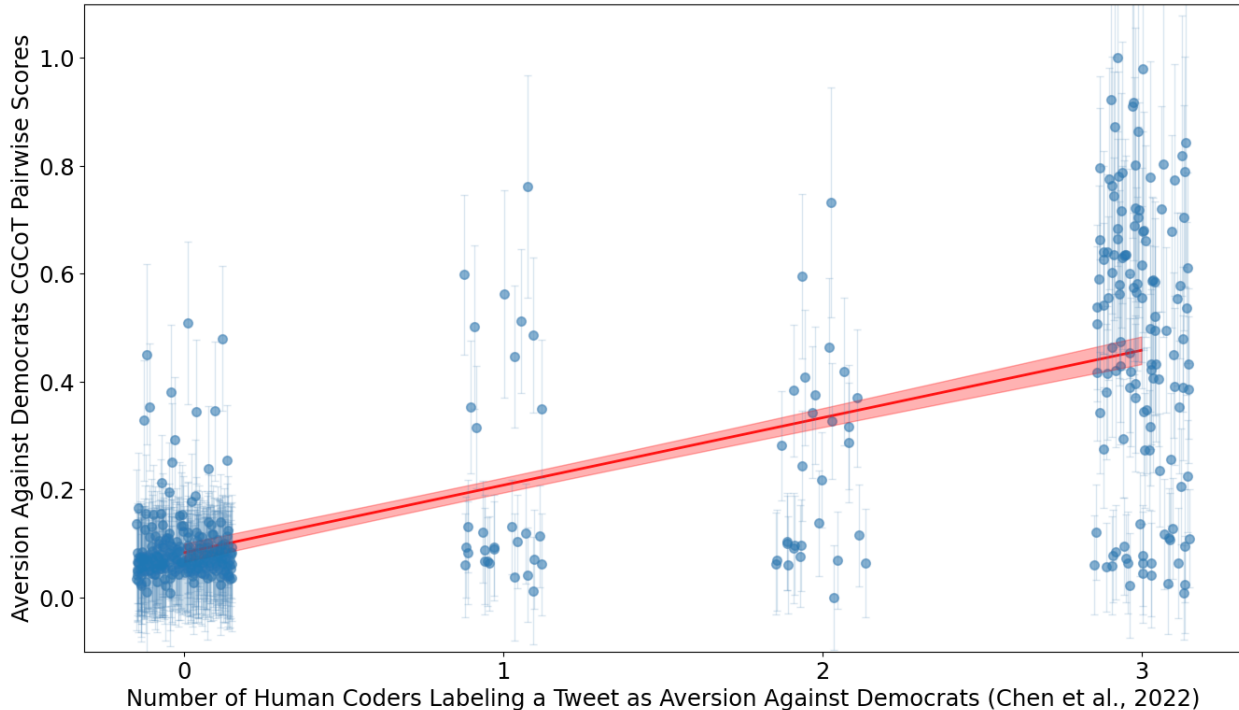


**Fig. 1. Aversion against Republicans CGCoT pairwise scores are strongly associated with the number of coders applying an aversion against Republicans label** (data from Chen *et al.* (19)). CGCoT pairwise score estimates are shown with 95% confidence intervals derived from quasi-standard errors (22). A linear regression line is drawn through the points with a 95% confidence band.

1. Using Wordfish with the tweets only
2. Using WordFish with the tweets’ concept-specific breakdowns
3. Using the pairwise comparison approach with ChatGPT-3.5 with the tweets only (“non-CGCoT tweets-only pairwise scores”)
4. CGCoT pairwise scores

Details about how we fit the Wordfish model are in the Methods section. Table 1 shows these correlations for both aversion against Republicans and Democrats.

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



**Fig. 2. Aversion against Democrats CGCoT pairwise scores are strongly associated with the number of coders applying an aversion against Democrats label** (data from Chen *et al.* (19)). CGCoT pairwise score estimates are shown with 95% confidence intervals derived from quasi-standard errors (22). A linear regression line is drawn through the points with a 95% confidence band.

	Aversion Against Republicans, $\rho$	Aversion Against Democrats, $\rho$
Wordfish Using Tweets Only	0.031	0.035
Wordfish Using Concept-Specific Breakdowns	0.545	0.223
Non-CGCoT Tweets-Only Pairwise Scores	0.554	0.560
CGCoT Pairwise Scores	0.638	0.611

**Table 1. CGCoT pairwise scores outperform other scaling methods**, based on Spearman’s rank correlation coefficient with the number of human coders labeling a tweet as containing aversion against Republicans and Democrats.

### CGCoT pairwise scores are competitive with supervised learning approaches

To further analyze the validity of the affective polarization CGCoT pairwise scores, we create binary labels using cutoffs in the two aversion scales. For each scale, 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 scale 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 scale generated using ChatGPT pairwise comparisons of the tweets only (i.e., non-CGCoT tweets-only comparisons).

We compare CGCoT pairwise scores with a RoBERTa-Large model fine-tuned using Chen *et al.* (19)’s 3,000 labeled political tweets. Table 2 contains the performance metrics of the two cutoff classifiers and the RoBERTa-Large model.

<i>Classifier</i>	<i>Aversion against</i>	F1 Score	Precision	Recall
Non-CGCoT Tweets-Only Pairwise Scores Cutoff Classifier	Republicans	0.696	0.643	0.760
	Democrats	0.668	0.579	0.791
CGCoT Pairwise Scores Cutoff Classifier	Republicans	0.836	0.890	0.789
	Democrats	0.789	0.837	0.747
Fine-Tuned RoBERTa-Large Model	Republicans	0.808	0.818	0.797
	Democrats	0.814	0.816	0.811

**Table 2. The F1 score, precision, and recall of the non-CGCoT tweets-only pairwise scores cutoff classifier, the CGCoT pairwise scores cutoff classifier, and RoBERTa-Large classifier fine-tuned on Chen *et al.* (19)’s aversion labels.**

The performance metrics show that CGCoT pairwise scores outperform non-CGCoT tweets-only pairwise scores on all metrics, except for recall for aversion against Democrats. The metrics show that the aversion against Republicans CGCoT pairwise scores outperform the fine-tuned RoBERTa-large model on F1 score and precision and are nearly equivalent on recall. Similarly, the aversion against Democrats CGCoT pairwise scores are comparable with the fine-tuned RoBERTa-Large on F1 score. 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 human-labeled tweets, except a small set of pilot tweets that were used to develop the CGCoT prompts for the concept-specific breakdowns.

## Discussion

We develop a novel text scaling method that leverages pairwise comparisons and an expert-crafted prompting procedure called concept-guided chain-of-thought (CGCoT), which creates concept-specific breakdowns of the texts. We prompt ChatGPT to make pairwise decisions between the concept-specific breakdowns along an abstract concept of interest. The resulting scales are called CGCoT pairwise scores. We apply the method to derive affective polarization scales: we use ChatGPT to pick the concept-specific breakdowns of tweets that exhibit greater aversion towards a specific political party in pairwise comparisons.

We find that the scales largely correlate with how humans interpret aversion against Republicans and aversion against Democrats on Twitter. We also find that using CGCoT and pairwise comparisons with LLMs are crucial, as scales that do not use one or both of these methods are demonstrably worse. We also show that using a cutoff with the scale can yield 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 scales but can also be useful for generating discrete classifications with high accuracy and performance without the use of any labeled data. In other words, our approach can generate a scale that is competitive with both unsupervised and supervised approaches.

Our method is situated in a rapidly growing literature on using generative LLMs for social science applications (see, e.g., 23–27). The works that have analyzed text, such as analyzing text along psychological constructs (24), directly analyze the text. Our proposed text scaling method breaks the text down into the concept of interest’s constituent parts using prompts developed by substantive knowledge about the targeted concept; it involves substantive expert knowledge to a much greater extent than the other research studies that analyze text using LLMs.

This method also speaks to a large body of text scaling methods, such as WordScores (1, 28), Wordfish (2), and word embedding methods (3, 29). Our method minimizes the need

for hand-labeling of text, 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 abstract concept of interest instead of relying on post hoc dimensional interpretation.

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 conflict within the presented texts. Again, this capability stems from its training on various instances of political conflict on 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 ChatGPT.

The need for CGCoT prompting when making out-of-sample text pairwise comparisons also indicates the limitations of current generative LLM models. As shown in previous works, even the best generative LLM models make errors when solving “out-of-sample” reasoning tasks. These errors are reduced when the intermediate reasoning steps are made explicit. But rather than allowing the LLM to generate its own reasoning steps, we accomplish this by engineering a specific set of questions that examine the concept of interest in the text. It shows that the LLM cannot effectively “reason” about these concepts. Using substantive knowledge to generate the concept-specific breakdowns changes the pairwise comparisons of text from a reasoning problem to a pattern recognition problem, greatly increasing the effectiveness of pairwise comparisons for scaling purposes.

Our proposed approach, in spite of its limitations, creates scales that largely agree with human judgments of the texts along different dimensions of interest. Lastly, it is well-

known that there is a significant mental toll on people identifying harmful materials 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 material.

## Materials and Methods

### Using CGCoT Pairwise Scores to Measure Affective Polarization

There are a total of 4,000 tweets from Chen *et al.* (19). The training set has 3,000 tweets, the validation set has 500 tweets, and the test set has 500 tweets. We scale the 500 tweets of the test set to have comparable performance metrics when binarizing the CGCoT pairwise scores.

We estimate two versions of the pairwise comparison scores for each scale. One version compares the tweets directly, which we call non-CGCoT tweets-only comparisons. The other version compares the concept-specific breakdowns generated using concept-guided chain-of-thought, which we call aversion CGCoT pairwise scores. After obtaining the concept-specific breakdowns, we compare pairs of the concept-specific breakdowns using the following prompt for aversion against Republicans:

Tweet Description 1: [concept-specific breakdown for Tweet 1]

Tweet Description 2: [concept-specific breakdown for Tweet 2]

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” in the prompt. 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.” Again, we use another prompt to extract the answers from these pairwise comparisons. The Supplementary Materials contains the prompts we use to extract these answers.

Because we are comparing 500 tweets in a pairwise fashion, there are a total of 124,750 possible 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, detailed later in this Methods section, does not require complete matchups to obtain scores for each tweet. In the Supplementary Materials, we compare the CGCoT pairwise scores when we sample 5, 10, and 15 matchups per tweet ID for both affective polarization scales.

## **Developing the Concept-Guided Chain-of-Thought Prompts**

As the Results section indicated, using concept-specific breakdowns in pairwise comparisons is essential to creating a scale that more closely matches human judgments. As previous works have shown, ChatGPT and other generative LLMs often make errors in intermediate reasoning steps; these errors are reduced when the prompt asks the LLM to explicitly provide such reasoning steps (10). We build on this idea by using a set of concept-specific questions that first force the LLM to generate “intermediate reasoning steps” in the form of concept-specific breakdowns; these concept-specific breakdowns are then used in pairwise comparisons with the LLM.

Creating the set of questions used for concept-guided chain-of-thought is not trivial. There is no defined algorithm for creating the set of questions or the number of questions, as the series of questions is dependent on the concept of interest. Instead, the prompts are created in a similar fashion to developing a codebook for human coders or for qualitative content analysis. We use approximately 50 tweets labeled as aversion against Republicans, 50 tweets labeled as not aversion against Republicans, 50 tweets labeled as aversion against Democrats, and 50 tweets labeled as not aversion against Democrats. These tweets came

from Chen *et al.* (19)’s training set. Rather than relying on the labels from Chen *et al.* (19), we labeled these tweets ourselves in order to familiarize ourselves with the text that we were working with. We iterate on an initial set of CGCoT prompts and examine outputs from ChatGPT-3.5 until the summaries and party identifications align with our expectations across the labeled tweets. This is akin to the steps described in Elo and Kyngäs (21): we first “make sense of the data and whole” by labeling the tweets ourselves and then apply deductive content analysis steps to summarize and identify key aspects of the text. We combine these steps with prompt engineering tactics such as changing specific words, repeatedly providing definitions, and splitting up complex questions can all cause the LLM to provide better responses about the concept of interest (see, e.g., 30). It is an example of how having a substantive understanding of the concept of interest is still important, even with the recent development of generative LLMs that provide human-like responses. However, the inspiration we took from qualitative content analysis can be applied to the design of prompts for other concepts of interest.

We also found that asking the LLM to first summarize the tweet always resulted in better responses for subsequent questions. The conversational nature of the recent generative LLMs comes from using the current prompt *and* the previous prompts and LLM responses in order to generate a response to the current prompt. Thus, asking for the LLM to first summarize a tweet forces the LLM to “explain the tweet in its own words,” which yielded improved answers to subsequent questions.

In future work, we aim to more systematically design the CGCoT prompts. There are currently no frameworks for sampling semantically equivalent text and prompts. Such a framework could allow us to find a prompt that is semantically equivalent to our current prompt while yielding concept-guided breakdowns that more closely align with the intention of our analysis. We have ongoing work that aims to develop an MCMC-like algorithm for sampling semantically equivalent text.



## 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 against 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 against 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 ChatGPT-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 ChatGPT-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 scale. The answers elicited from ChatGPT-3.5 are then scaled using the Bradley-Terry model to estimate the aversion against Republicans and aversion against Democrats scales.

## Using the Bradley-Terry Model to Estimate the Affective Polarization Scales

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  $\alpha_i/\alpha_j$ , where  $\alpha_i$  and  $\alpha_j$  are positive-valued parameters that indicate latent “ability” ( $\gamma$ ). 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  $\lambda_i$  compared to  $\lambda_j$ , 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 affective polarization scales as our example, the estimated  $\lambda$  parameters are the

measures of the level of aversion towards a specific party. We denote the concept-specific breakdown exhibiting greater aversion towards a specific party as the “winner.” For ties, we considered these 0.5 wins for both tweets in the matchup. Turner and Firth (31) 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 ChatGPT’s answers to pairwise comparisons to estimate the level of aversion towards a specific party. These scores are the affective polarization CGCoT pairwise scores described in the previous sections. The estimated  $\lambda$  parameters are relative to a reference tweet, but this choice is unimportant because we scale the parameters to the unit interval.

We can also estimate standard errors for the estimated  $\lambda$  parameters. These standard errors are interpreted relative to a reference tweet. We calculate quasi-variances, which can be interpreted as if they refer to reference-free estimates of the variance of the underlying “ability” of each tweet. This means that any pairs of tweets can be directly compared. We use the `qvcalc` package to calculate quasi-standard errors (32). The 95% confidence intervals in the Results section are derived from these quasi-standard errors.

## Using Wordfish to Estimate the Affective Polarization Scales

The primary goal of Wordfish (2) 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

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

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

where  $y_{ij}$  is the count of word  $j$  in tweet  $i$ ,  $\alpha$  is the set of tweet fixed effects,  $\psi$  is the set of word fixed effects,  $\beta$  is an estimate of a word-specific weight that reflects the importance of

word  $j$  in discriminating between positions, and  $\omega_i$  is tweet  $i$ 's position. We fit this model using `quanteda` R package (33).

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 against Republicans, a word had to be used at least 4 times across at least 4 tweets. For aversion against Democrats, a word had to be used at least 5 times across at least 5 tweets. We applied this to both the tweets and the concept-specific breakdowns.

## References

1. M. Laver, K. Benoit, J. Garry, Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review* **97**, 311–331 (2003).
2. J. B. Slapin, S.-O. Proksch, A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* **52**, 705–722 (2008).
3. P. Y. Wu, W. R. Mebane Jr., L. Woods, J. Klaver, P. Due, “Partisan Associations of Twitter Users Based on Their Self-Descriptions and Word Embeddings”, Presented at APSA 2019, 2019.
4. L. Rheault, C. Cochrane, Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis* **28**, 112–133 (2020).
5. M. Bailey, *Measuring Candidate Ideology from Congressional Tweets and Websites*, 2023.
6. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019.
7. R. A. Bradley, M. E. Terry, Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika* **39**, 324–345 (Dec. 1952).
8. D. Carlson, J. M. Montgomery, A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts. *American Political Science Review* **111**, 835–843 (2017).
9. J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, H. Hajishirzi, presented at the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3154–3169.
10. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 2023.
11. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, *Large Language Models are Zero-Shot Reasoners*, 2023.
12. D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, E. Chi, *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*, 2023.
13. Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, Y. Kim, *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks*, 2023.
14. M. E. Fonteyn, M. Vettese, D. R. Lancaster, S. Bauer-Wu, Developing a codebook to guide content analysis of expressive writing transcripts. *Applied Nursing Research* **21**, 165–168 (2008).
15. C. B. Moran Yarchi, N. 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**, 98–139 (2021).

16. M. 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** (2021).
17. S. Iyengar, G. Sood, Y. Lelkes, Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly* **76**, 405–431 (Sept. 2012).
18. J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, J. B. Ryan, Affective polarization, local contexts and public opinion in America. *Nature Human Behaviour* **5**, 28–38 (2021).
19. H. Chen, Z. Terechshenko, P. Y. Wu, R. Bonneau, J. A. Tucker, “Detecting Political Sectarianism on Social Media: A Deep Learning Classifier with Application to 2020-2022 Tweets”, 2022.
20. E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. V. Bavel, C. S. Wang, J. N. Druckman, Political sectarianism in America. *Science* **370**, 533–536 (2020).
21. S. Elo, H. Kyngäs, The qualitative content analysis process. *Journal of Advanced Nursing* **62**, 107–115 (2008).
22. D. Firth, R. X. De Menezes, Quasi-variances. *Biometrika* **91**, 65–80 (Mar. 2004).
23. P. Törnberg, *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*, 2023.
24. S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjeh, C. Robertson, J. J. Van Bavel, *GPT is an effective tool for multilingual psychological text analysis*, May 2023.
25. L. P. Argyle, E. Busby, J. Gubler, C. Bail, T. Howe, C. Rytting, D. Wingate, *AI Chat Assistants can Improve Conversations about Divisive Topics*, 2023.
26. J. Bisbee, J. Clinton, C. Dorff, B. Kenkel, J. Larson, *Synthetic Replacements for Human Survey Data? The Perils of Large Language Models*, May 2023.
27. P. Y. Wu, J. Nagler, J. A. Tucker, S. Messing, *Large Language Models Can Be Used to Estimate the Latent Positions of Politicians*, 2023.
28. W. Lowe, Understanding Wordscores. *Political Analysis* **16**, 356–371 (2008).
29. A. C. Kozlowski, M. Taddy, J. A. Evans, The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review* **84**, 905–949 (2019).
30. E. Saravia, Prompt Engineering Guide. <https://github.com/dair-ai/Prompt-Engineering-Guide> (Dec. 2022).
31. H. Turner, D. Firth, Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software* **48**, 1–21 (2012).
32. D. Firth, *qvcalc: Quasi Variances for Factor Effects in Statistical Models*, R package version 1.0.3.
33. K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, A. Matsuo, quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* **3**, 774 (2018).

# Supplementary Materials for “Concept-Guided Chain-of-Thought Prompting for Pairwise Comparison Scaling of Texts with Large Language Models”

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing

## **Tweets with the Lowest CGCoT Pairwise Scores that were Labeled by Three Coders as Containing Aversion Against Republicans**

Looking at just the tweets labeled by three coders as containing aversion against Republicans, we examined the three tweets with the lowest aversion against Republicans CGCoT pairwise scores. The text of these tweets are 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>”

ChatGPT 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, ChatGPT interpreted a vague headline describing something backfiring against the Trump campaign as not expressing aversion against Republicans, an arguably correct interpretation.

## **Tweets with the Lowest CGCoT Pairwise Scores that were Labeled by Three Coders as Containing Aversion against Democrats**

Looking at just the tweets labeled by three coders as containing aversion against Democrats, we again examine the three tweets with the lowest aversion against Democrats CGCoT

pairwise scores. The text of these tweets are 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 doesn't 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 can't just get rehired at the next department over." Wtf does this mean? Proof your statements, Dems.

Here, ChatGPT-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 against 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 ChatGPT-3.5 did not interpret as an insult or criticism of Democrats.

### **Extracting the Tweet that Exhibits Greater Aversion Against Republicans and Democrats in Each Matchup**

For each pairwise comparison prompt, ChatGPT typically returns a small paragraph explaining its choice. We concatenate the model's answer 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."

We then prompt ChatGPT with the concatenated text. Ties and answers that deviated from either "Tweet Description 1," "Tweet Description 2," or "Tie" are manually fixed. We used the extract same prompt for extracting answers from non-CGCoT comparisons, except we replaced the "Tweet Description" with just "Tweet."



### Correlations Between Scores with Differing Number of Matchups

Results in the main text 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 scales. 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 against 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 against Democrats.