

Large Language Models Can Be Used to Estimate the Latent Positions of Politicians

Patrick Y. Wu¹, Jonathan Nagler^{1,2}, Joshua A. Tucker^{1,2}, and Solomon Messing¹

¹Center for Social Media and Politics, New York University

²Department of Politics, New York University

Version: October 10, 2023

[Click here for the latest version](#)

Abstract

Existing approaches to estimating politicians' latent positions along specific dimensions often fail when relevant data is limited. We leverage the embedded knowledge in generative large language models (LLMs) to address this challenge and measure lawmakers' positions along specific political or policy dimensions. We prompt an instruction/dialogue-tuned LLM to pairwise compare lawmakers and then scale the resulting graph using the Bradley-Terry model. We estimate novel measures of U.S. senators' positions on liberal-conservative ideology, gun control, and abortion. Our liberal-conservative scale, used to validate LLM-driven scaling, strongly correlates with existing measures and offsets interpretive gaps, suggesting LLMs synthesize relevant data from internet and digitized media rather than memorizing existing measures. Our gun control and abortion measures—the first of their kind—differ from the liberal-conservative scale in face-valid ways and predict interest group ratings and legislator votes better than ideology alone. Our findings suggest LLMs hold promise for solving complex social science measurement problems.

Introduction

This paper outlines a novel approach to addressing challenges in measuring the latent positions of lawmakers using generative large language models (LLMs). Measuring latent positions along specific political or policy domains reduces the dimensionality of lawmakers’ complex actions and stances to a low-dimensional scale. When combined with other data, these measures allow us to assess core democratic functions: how well lawmakers represent their constituents [see, e.g., 1–5], whether enacted policies have broad support or are driven by one part of the ideological spectrum [see, e.g., 6–10], and how position-taking occurs outside of roll call voting [see, e.g., 11–13].

While there is broad agreement that lawmakers have positions in the space of ideology and other issue-specific dimensions, we cannot directly observe these positions—they exist in latent space and must be estimated. Existing approaches to estimating liberal-conservative ideology are based on either the behaviors or perceptions of lawmakers. Behavior-based estimates commonly use roll call votes to measure revealed preferences constrained by the legislative agenda [14–16]. Other measures of behavior, such as news media sharing behavior [17], are also measures of liberal-conservative ideology based on revealed preferences. Liberal-conservative measures based on campaign contributions [18] assume ideological homophily in campaign giving and are based on perceptions of the contributors. Each of these measures captures a different facet of liberal-conservative ideology in a different context—interpretive gaps can occur either from modeling assumptions or a lack of relevant data.

These approaches reduce the dimensionality of a complex political space to a single left-right dimension but do not tell us about lawmakers’ positions on specific issues. Specific positions on issues like gun control and abortion are difficult to measure using existing scaling approaches due to the absence of relevant data. For example, roll call votes cannot be used to measure lawmakers’ stances on gun control because most sessions of Congress have no such votes on this issue.

Generative LLMs are trained on massive corpora of internet and digitized media text,

embedding much information about politics, position-taking, and widely-held perceptions as reported by journalists and other content publishers. We propose leveraging this embedded information by prompting an LLM fine-tuned to follow instructions and engage in dialogue using reinforcement learning human feedback [19] to compare politicians on a relevant dimension. Specifically, we use ChatGPT-3.5 to pairwise compare the senators of the 116th U.S. Congress along three dimensions: liberal-conservative ideology, support of gun control support, and support of abortion rights support. We then use the Bradley-Terry Model [20] to estimate a unidimensional scale measuring latent political positions of interest, which we call **L**anguage **M**odel **P**airwise comparison (LaMP) scores.

These pairwise comparisons are made in a zero-shot learning setting because we do not provide ChatGPT with *any additional information* besides the senator’s name, party affiliation, and state represented, and we *do not* include any examples of pairwise comparisons. For the liberal-conservative ideology scale, we prompt ChatGPT to pick the senator who is more conservative (or liberal) for a given pair of senators. For the gun control scale, we prompt ChatGPT to pick the senator more likely to support gun control. For the abortion rights scale, we prompt ChatGPT to pick the senator more likely to be pro-choice (or pro-life). Figure 1 shows an overview of the proposed pairwise comparison approach.

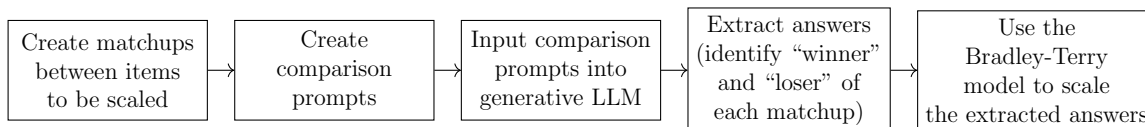


Figure 1: An overview of the proposed pairwise comparison approach with instruction/dialogue-tuned generative LLMs.

The liberal-conservative ideology scale has been extensively studied in the U.S. national legislature, providing a widely accepted and well-validated set of measures by which we can validate LLM-driven scaling and better understand its strengths. Our Ideology LaMP scores highly correlate with the first dimension of DW-NOMINATE, the most popular liberal-conservative scaling of senators [14–16]. However, Ideology LaMP scores also depart from DW-NOMINATE scores in important ways. For example, our approach places senators who

vote *against* their party for ideologically extreme reasons on the ends, while DW-NOMINATE places them more toward the center. We also find that Ideology LaMP scores predict human evaluations of the ideologies of senators better than other measures [21], including DW-NOMINATE.

The gun control scale has not been estimated in the political science literature because of a lack of data on the behaviors and perceptions of the senators concerning gun control. We find that our Gun Control LaMP scale places senators in sensible positions. It does not simply imitate Ideology LaMP scores. For example, Bernie Sanders, the most liberal senator on the Ideology LaMP scale, is placed in the middle of the Democratic senators on the Gun Control LaMP scale. Validating this measure, we find that Gun Control LaMP scores predict the grades given to each senator by the National Rifle Association (NRA) and individual votes on the 2022 Bipartisan Safer Communities Act, the latter of which is out-of-sample—the vote took place after ChatGPT-3.5’s training data ends.

Similarly, abortion rights scales have not been estimated in the political science literature for reasons related to absent or sparse data. Our Abortion Rights LaMP scores place senators in sensible positions, such as placing self-described pro-choice Republicans Susan Collins and Lisa Murkowski among the Democratic senators. The Abortion Rights LaMP scores also predict the grades given to each senator by NARAL Pro-Choice America better than DW-NOMINATE. These issue-specific scales offer advantages over using interest group ratings: they eliminate the issue of “lumpiness,” such as all Democratic senators receiving ‘Fs’ from the NRA, and provide a single measure of all senators, avoiding the need to combine interest group ratings that are based on different criteria [22].

Partial correlations analysis suggests that LaMP scores are a blend of textually available information about individual lawmakers. In other words, LaMP scores reflect both behaviors, such as votes, floor speeches, and commentary about other lawmakers, as well as perceptions of these lawmakers, such as news stories, blog posts, and editorials. Because of the black box nature of LLMs, we do not know how the LLM weighs the underlying textual

information given a prompt. At the same time, the scope of the LLM’s embedded knowledge base underpinning the measures is its strength, making up for interpretive gaps in existing measures of ideology and allowing us to measure politicians’ stances along specific issues.

We cannot simply prompt an LLM to return a list of lawmakers ranked along one of these dimensions for several reasons. First, ChatGPT, in particular, often does not consistently return a complete list of senators ranked from most liberal to most conservative: it will sometimes only return partial lists, generate inconsistent lists across repeated promptings, or refuse to create such lists. Pairwise comparisons also enable us to establish a scale where the differences are meaningful, whereas ordinal rankings indicate order without quantifying the gaps between ranks. For this reason, researchers have used pairwise comparisons extensively in social science scaling applications. They are also easier to complete. For example, Loewen, Rubenson, and Spirling [23] conduct a survey experiment with pairwise comparisons and the Bradley-Terry model to determine the most persuasive arguments. Carlson and Montgomery [24]’s `SentimentIt` R package conducts pairwise comparisons to label political texts. Hopkins and Noel [21], a closely related work to this study, use pairwise comparisons among political activists to scale senators of the 114th Congress and the 117th Congress along the liberal-conservative continuum.

In summary, we find that when prompted with pairwise comparisons, the LLM does not hallucinate; it does not simply parrot pre-existing measures of the ideologies of senators; and, most importantly, we can prompt it to evaluate pairwise comparisons that analysts can use to construct novel scales that would be impossible to do with existing scaling methods. Our validation and analysis of Ideology LaMP scores suggest that we can use a class of instruction/dialogue-tuned LLMs to create a scale that correlates with an amalgamation of information about policy positions, voting behaviors, campaign-giving patterns, and public perceptions about politicians. The validity of Gun Control LaMP scores and Abortion Rights LaMP scores suggest that complex issue-specific scales can be estimated using pairwise comparison scaling with generative LLMs. Overall, our proposed pairwise scaling approach can

give us a more comprehensive understanding of legislative behaviors and policy preferences.

Results

We use ChatGPT to make pairwise comparisons about liberal-conservative ideology, gun control, and abortion, and then scale its answers using the Bradley-Terry model, producing LaMP scores. We use the liberal-conservative ideology scale to better understand the strengths of LaMP scores. The gun control scale and abortion rights scale demonstrate how we can use the approach to estimate novel issue-specific scales. We call the pairwise comparisons “matchups.” There are 5,151 total matchups across all senators in the 116th Congress for each scale.

Liberal-Conservative Ideology Scaling

For liberal-conservative ideology scaling, the “winner” of each matchup was the senator ChatGPT answered as being more conservative in its response. “Winners” are assigned this way to intuitively place more conservative senators on the right side of the scale and more liberal senators on the left side of the scale. We call the resulting scores “Ideology LaMP scores.” We highlight interesting features of Ideology LaMP scores and analyze their relationship with pre-existing measures of these senators’ ideologies within and across parties. The Methods section details the prompts used for these pairwise comparisons.

Ideology LaMP scores highly correlate across repeated iterations

We ran the entire set of matchups across all senators three times. We look at the correlation of the Ideology LaMP scores generated by each set of complete matchups. Among the three iterations, the lowest correlation between any two iterations’ LaMP scores is 0.997. Breaking it down by party, the lowest correlation between any two iterations’ LaMP scores is 0.982 for Democratic senators and 0.972 for Republican senators. ChatGPT’s responses to pairwise

comparisons are highly consistent in repeat interactions. Given the high correlations, we use Ideology LaMP scores estimated using all matchups across all iterations for the rest of the analysis.

We also show, in Figure 9 in Supplementary Information, a plot of the Ideology LaMP scores of the senators with 95% confidence intervals derived from quasi-standard errors [25]. To better understand uncertainty specifically in the LLM’s responses to pairwise comparisons, we derive an approach, detailed in the Supplementary Information under the section “Quantifying Uncertainty in the LLM’s Pairwise Comparisons,” for estimating the probability of an LLM choosing a specific senator in a given pairwise comparisons and an entropy-based metric quantifying the uncertainty of this generated choice. The results of the entropy-based metric, illustrated in Figure 10, confirm the correlational study above: as the distance between the estimated ideology of two senators increases, the entropy of the LLM’s responses in pairwise comparisons quickly decreases.

Ideology LaMP scores highly correlate with DW-NOMINATE

DW-NOMINATE (Dynamic, Weighted NOMINAL Three-Step Estimation) is a multidimensional scaling approach that uses roll call voting patterns to estimate the ideological positions of legislators [6, 14–16]. It is the most widely used measure of legislator ideology [26]. The first dimension of DW-NOMINATE is typically interpreted as the liberal-conservative continuum in United States politics [6]. The overall correlation between Ideology LaMP scores and the first dimension of DW-NOMINATE is 0.967. They correlate at 0.838 among Democratic senators and correlate at 0.649 among Republican senators. Figure 2 compares the first dimension of DW-NOMINATE against Ideology LaMP scores.

Ideology LaMP scores do not simply parrot DW-NOMINATE

Figure 2 also illustrates interesting patterns in the Ideology LaMP scores. Notably, our method estimates Joe Manchin to be more conservative than Susan Collins. Such a placement

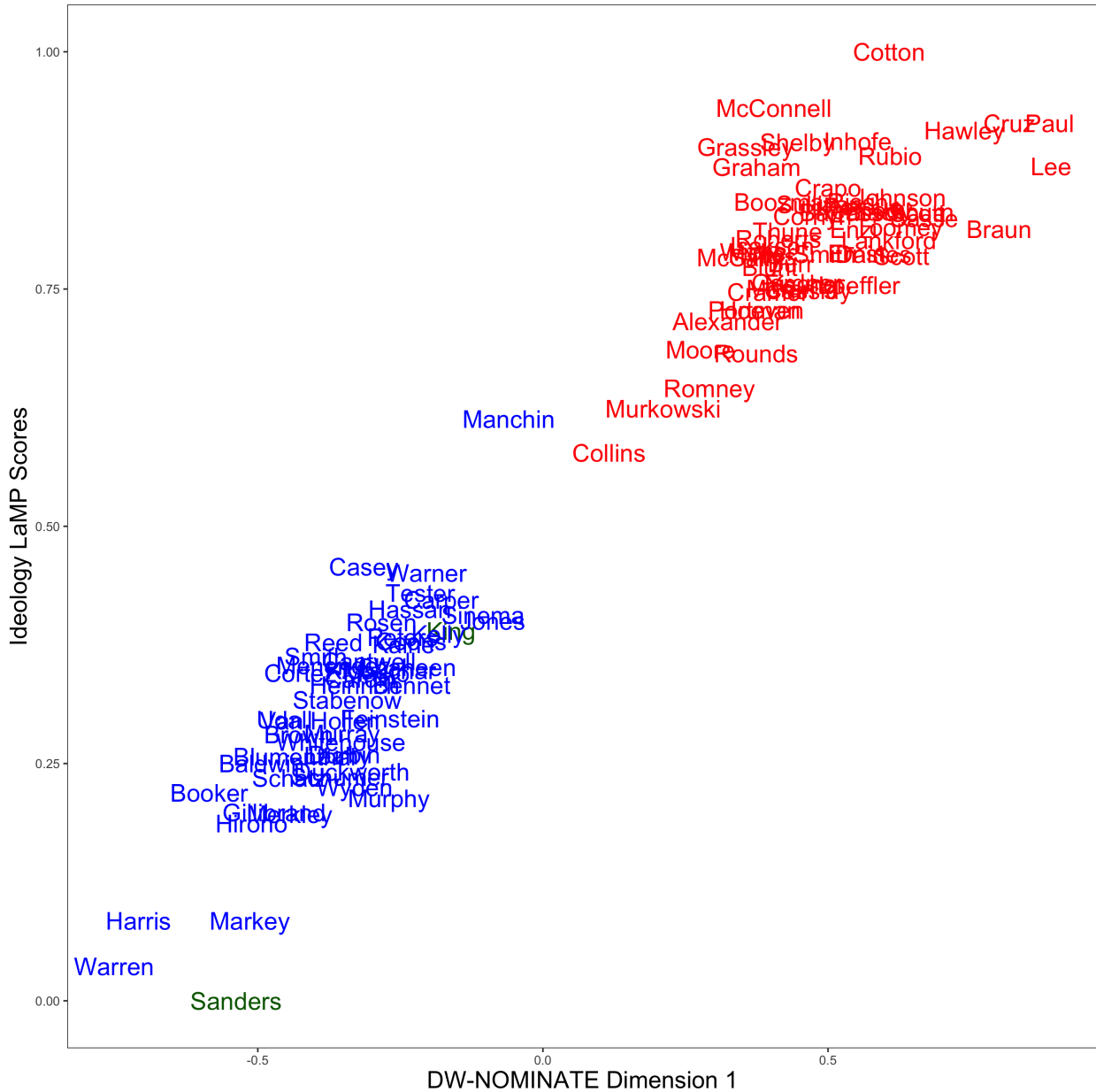


Figure 2: First Dimension of DW-NOMINATE vs. Ideology LaMP scores. Democratic senators are in blue, Republican senators are in red, and Independent senators are in green.

intuitively makes sense: for example, on the issue of abortion, Manchin is pro-life, while Collins is pro-choice. In contrast, there is no overlap between senators of opposing parties in the first dimension of DW-NOMINATE.

Looking at the extremes also indicates that ChatGPT is not simply recalling DW-NOMINATE scores for senators when responding to pairwise assessments. DW-NOMINATE

places Elizabeth Warren and Kamala Harris as the most liberal senators. Ideology LaMP scores, on the other hand, place Bernie Sanders and Elizabeth Warren as the most liberal senators. This comports with surveyed political activists, who also named Sanders and Warren the most liberal senators [21]. Sanders’ placement towards the center of the first dimension of DW-NOMINATE is the result of Sanders occasionally voting against the Democratic party [27]. LaMP scores, conversely, likely pick up signals from not only roll call votes but also sources such as, among other things, mainstream news articles that discuss Sanders’ left-leaning positions on various issues.

Among Republicans, Ideology LaMP scores rank Tom Cotton and Mitch McConnell as the most conservative senators; DW-NOMINATE ranks Mike Lee and Rand Paul as the most conservative. Surveyed political activists ranked Cruz and Cotton as the most conservative senator [21]¹; Cruz was the third most conservative senator according to Ideology LaMP scores.

Comparing the ordinal rankings of DW-NOMINATE and Ideology LaMP scores, senators differed, on average, by 8.31 positions. Some of the largest differences in ordinal rankings were Chuck Grassley (DW-NOMINATE: 48th most conservative; LaMP scores: 8th most conservative), Mitch McConnell (DW-NOMINATE: 38th most conservative; LaMP scores: 2nd most conservative), and Lindsey Graham (DW-NOMINATE: 45th most conservative; LaMP scores: 11th most conservative). Among the 10 senators with the largest differences in ordinal rankings, 9 are Republicans. These differences seem to be shaped by their public stances with respect to Donald Trump, his policies, and his nominees, such as Chuck Grassley’s vocal support for Brett Kavanaugh’s nomination to the Supreme Court of the United States. Again, DW-NOMINATE would not capture these public stances.

¹Sessions was ranked the second most conservative senator among political activists, but he was not in the 116th Senate.

LaMP scores highly correlate with alternative measures of ideology

Next, we compare Ideology LaMP scores to two alternative measures of ideology from the political science literature that are not based on roll call votes but on the perceptions of the senators: one scales senators based on political activists’ knowledge, and the other on patterns in campaign donations. Specifically, we look at the two following measures:

1. Perceived ideology scores [21]: Perceived ideology scores are estimated using political activists’ answers to pairwise comparisons of senators. Perceived ideology scores can be considered a separate but related measure of ideology: they capture the perceived ideological positions of politicians, which shape how people vote and interact with politicians. Perceived ideologies can differ from how politicians view themselves ideologically and what their revealed preferences are. Hopkins and Noel [21] included these pairwise comparisons of senators of the 117th Congress in a YouGov survey in April 2021. The authors had 1,110 activists answer these pairwise comparisons; they then scaled the activists’ answers using the Bradley-Terry model. The 11 senators who retired or did not secure a new term at the end of the 116th Congress were not included in their survey. These perceived ideology scores offer a way to compare a scale estimated using ChatGPT’s pairwise comparisons with a scale estimated using human-labeled pairwise comparisons.
2. Campaign Finance Scores [18]: Campaign Finance scores (CFscores) are a measure of the ideologies of politicians, donors, and interest groups. CFscores are estimated using a network that links all individual contributors to all political candidates who received donations. It assumes that individuals choose to give to candidates close to them in a latent ideological space and scales all actors in that space; in other words, it measures ideology based on the donors’ perceptions of lawmakers. We used each senator’s latest CFscore; the Database on Ideology, Money in Politics, and Elections have estimated CFscores up to the 2018 election cycle [28]. We look at the recipient CFscore, which

is the estimated ideology of the senator based on donations received. Tammy Baldwin, Mark Kelly, and Kelly Loeffler are missing recipient CFscores.

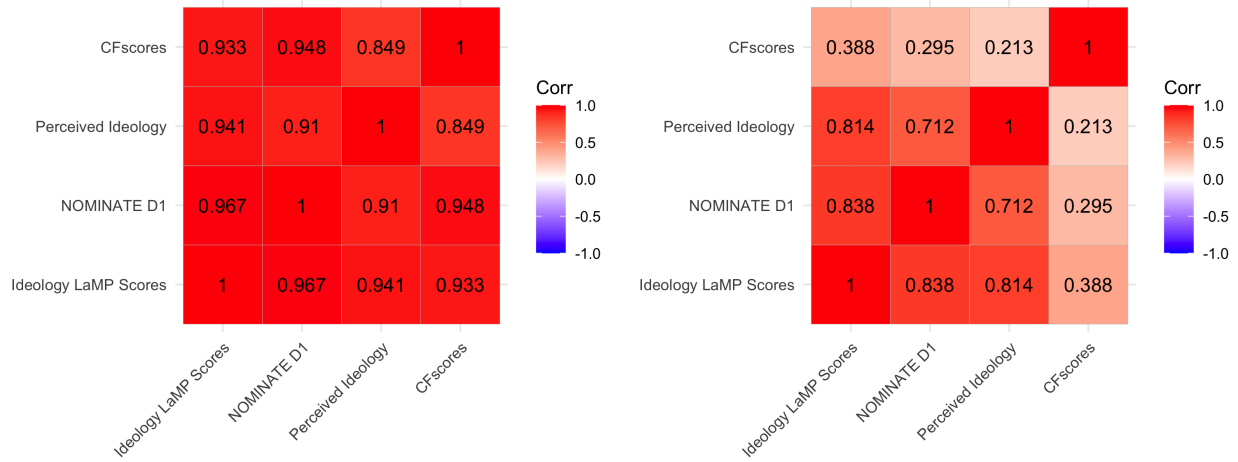
Figure 3 shows bivariate analyses across Ideology LaMP scores, DW-NOMINATE Dimension 1, perceived ideology scores, and CFscores. Across all senators, Ideology LaMP scores highly correlate with perceived ideology (0.941) and CFscores (0.933).

LaMP scores reflect both behaviors and perceptions

Table 1 shows the partial correlations between Ideology LaMP scores and the three measures of ideology as described above. DW-NOMINATE is a measure of ideology based on the behaviors of senators, while perceived ideology and CFscores are measures of ideology based on perceptions of the senators. For each cell, the partial correlation between the Ideology LaMP scores and the measure in the column title is calculated controlling for the other two measures of ideology. The partial correlations suggest that no single measure of ideology fully explains Ideology LaMP scores. Instead, the results indicate that Ideology LaMP scores reflect a measure of ideology based on both behaviors and perceptions of the senators. This interpretation holds when we look at the partial correlations across all senators, Democratic senators, and Republican senators, except for the partial correlation between Ideology LaMP scores and CFscores among Republican senators. To be clear, it does not suggest that these measures collectively capture everything related to behaviors and perceptions about the senators. However, these three measures of ideology capture a wide range of revealed preferences and perceptions of the senators.

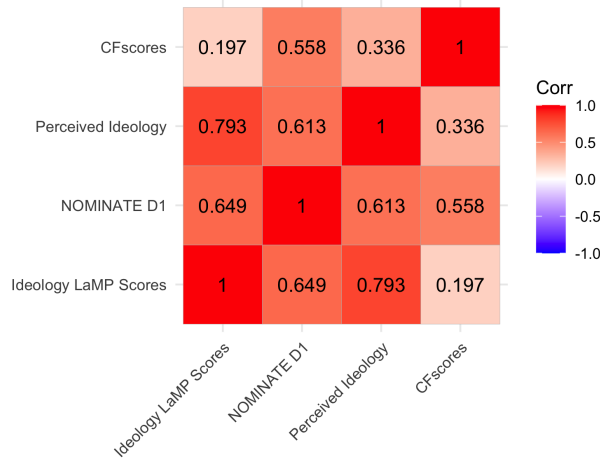
LaMP scores better predict human evaluations of ideology

Figure 3 indicates that the correlation between Ideology LaMP scores and perceived ideology scores is higher than the correlations between perceived ideology and the other measures of ideology when looking at all senators and senators by party. To more formally analyze the predictive power of Ideology LaMP scores on human evaluations of senators' ideologies,



(a) All Senators

(b) Democratic Senators



(c) Republican Senators

Figure 3: Correlation matrices of LaMP scores, the first dimension of DW-NOMINATE, Hopkins and Noel [21]’s perceived ideology scores, and Bonica [18]’s CFscores.

we compare the predictive power of Ideology LaMP scores and the first dimension of DW-NOMINATE, the second-highest correlating measure with perceived ideology scores. We use multivariate analyses to calculate how much the proportion of variance explained (R^2) in perceived ideology scores falls when we compare the full model, regressing perceived ideology

	DW-NOMINATE	Perceived Ideology	CFscores
Ideology LaMP Scores, All Senators	0.441***	0.621***	0.300**
Ideology LaMP Scores, Dem. Senators	0.578***	0.592***	0.334*
Ideology LaMP Scores, Rep. Senators	0.467**	0.678***	-0.360*

Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 1: Partial correlations between Ideology LaMP scores and the first dimension of DW-NOMINATE, perceived ideology, and CFscores. Each cell shows the partial correlations between Ideology LaMP scores and the measure in the column title, controlling for the other two measures of ideology. P-values are calculated using the t-statistic described in Kim [29].

scores on both the first dimension of DW-NOMINATE and Ideology LaMP scores, with reduced models, which only use Ideology LaMP scores or DW-NOMINATE as the predictor. Figure 4 shows how the proportion of variance explained in perceived ideology changes as we move from the full to reduced models.

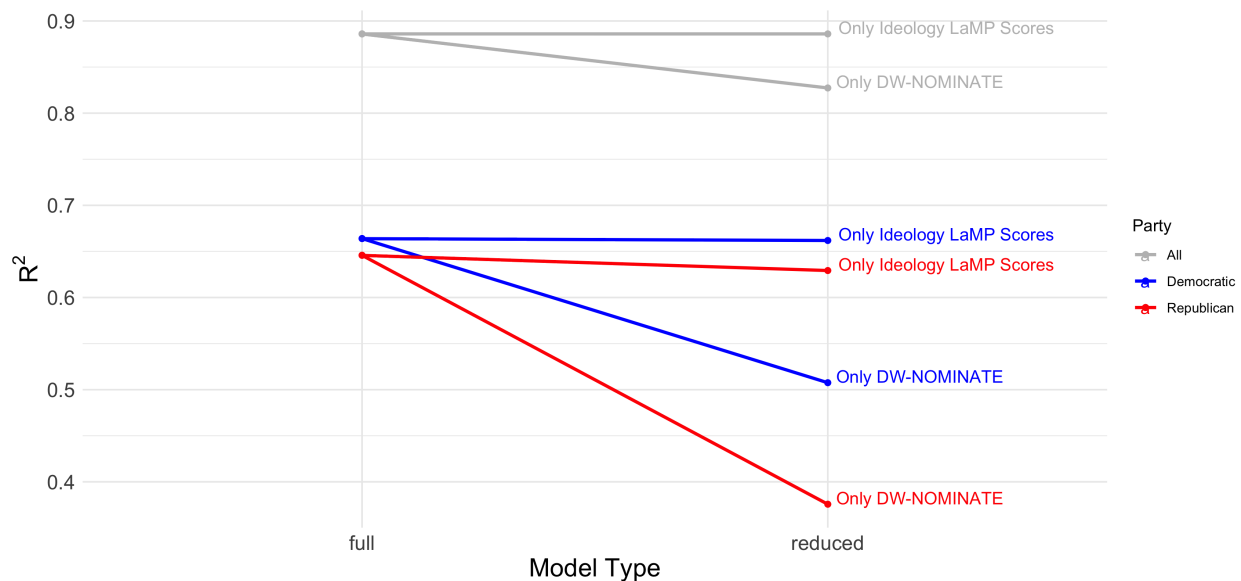


Figure 4: Comparing the proportion of variance, R^2 , between the full and reduced models. The full model regresses perceived ideology scores on both the first dimension of DW-NOMINATE and Ideology LaMP scores. The reduced models only use DW-NOMINATE (denoted as “Only DW-NOMINATE”) or Ideology LaMP scores (denoted as “Only Ideology LaMP scores”) as predictors. The proportion of variance explained, R^2 , is always lower when Ideology LaMP scores are removed as a predictor than when DW-NOMINATE is removed.

The proportion of variance explained in perceived ideology always falls more when we drop Ideology LaMP scores as a predictor. We statistically confirm these results using partial F-tests. When looking across all senators, only Democratic senators, and only Republican

senators, the partial F-test p-values for the full model and the reduced model with only DW-NOMINATE as a predictor is significant ($p < .0001$ for all F-test p-values), while the p-value for the full model and the reduced model with only Ideology LaMP scores as a predictor is not significant ($p > .16$ for all F-test p-values). It suggests that ChatGPT uses synthesized information that is highly correlated with how activists perceive these senators' ideologies. It also further indicates that Ideology LaMP scores go beyond measuring position-taking.

Gun Control Support Scaling

To scale senators along a gun control support continuum, we prompt the LLM to judge which senator is more likely to support gun control. We call the resulting scores “Gun Control LaMP scores.” We highlight features of Gun Control LaMP scores and analyze their relationship with interest group ratings and voting behaviors.

Gun Control LaMP scores highly correlate across repeated iterations

We ran the entire set of matchups across all senators three times. The lowest correlation among any two iterations is 0.993. Looking at these correlations by party, the lowest correlation between any two iterations' Gun Control LaMP scores is 0.970 for Democratic senators and 0.952 for Republican senators. Again, ChatGPT's responses to pairwise comparisons about gun control support are consistent despite this being a specific area of public policy.

To more closely analyze the uncertainty of the LLM's responses to pairwise comparisons of senators along the issue of gun control, we again use the entropy-based metric derived in the Supplementary Information. The results are in Figure 11 in the Supplementary Information. Again, the entropy-based metric confirms the correlational study: as the distance between LaMP scores increases, entropy in the LLM's pairwise comparison responses quickly decreases.

Gun Control LaMP scores differ from Ideology LaMP scores

The Gun Control LaMP scores of all senators of the 116th Congress are illustrated in Figure 5. This issue-specific scale differs from Ideology LaMP scores. We note that we expect negative correlations because of party alignment on the issue of gun control.

Overall, they correlate -0.943 , but this correlation is largely driven by party alignment on the issue of gun control. Within each party, the correlations are not as strong: among Republican senators, the correlation is -0.684 , and among Democratic senators (including the two independent senators), the correlation is -0.570 . We also note that there is no overlap between the senators of the two parties, unlike the Ideology LaMP scores.

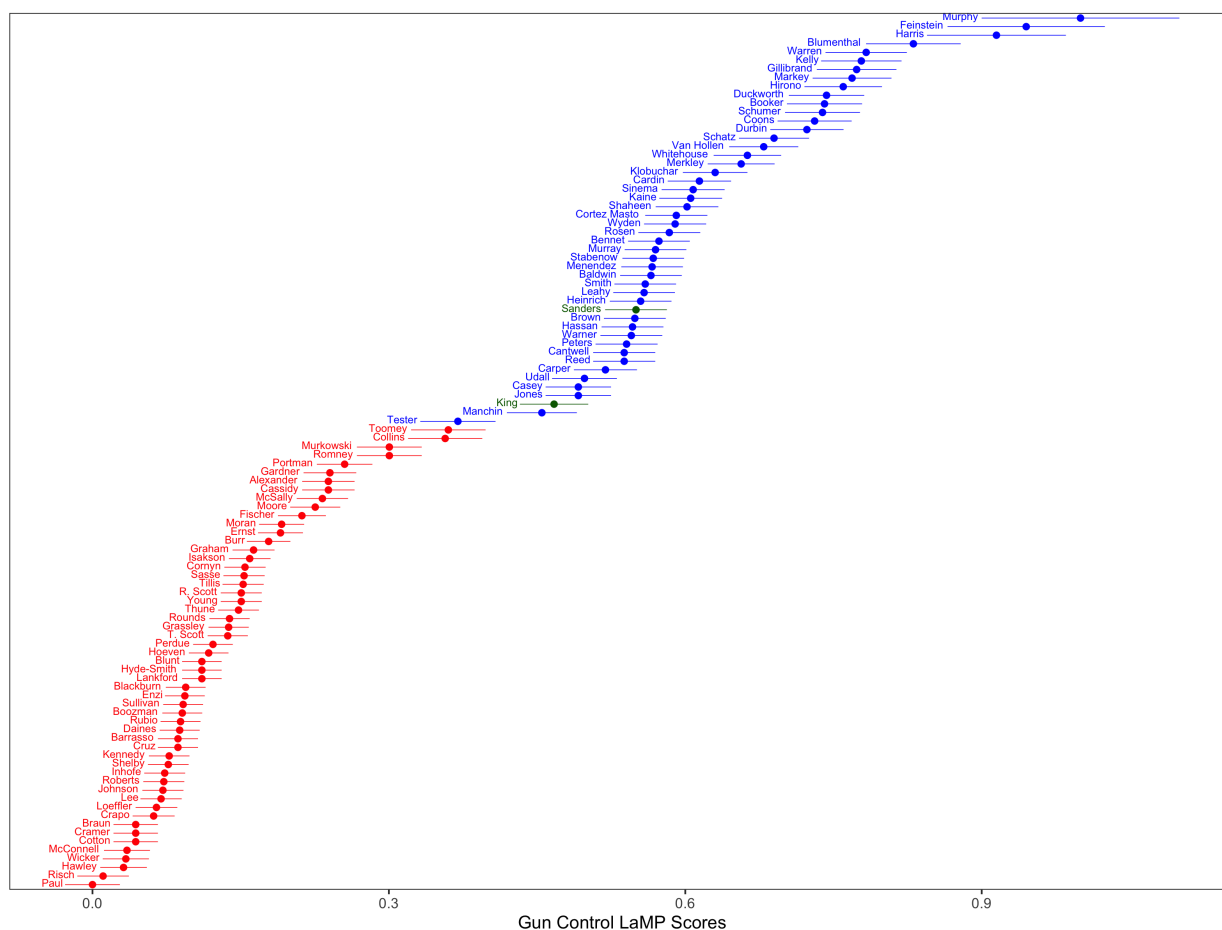


Figure 5: Gun Control LaMP scores across all senators with 95% confidence intervals derived from quasi-standard errors [25]. Democratic senators are in blue, Republican senators are in red, and Independent senators are in green.

Comparing Gun Control LaMP scores with Ideology LaMP scores, we find face validity with the Gun Control LaMP scores. For example, Ideology LaMP scores place Mark Kelly as a centrist Democratic senator, but Gun Control LaMP scores place him as one of the strongest gun control supporters. This aligns with his outspoken advocacy for gun control following the attempted assassination of his wife, former U.S. representative Gabby Giffords. On the other hand, Bernie Sanders, the most liberal Democratic senator based on Ideology LaMP scores, is placed in the middle among the Democratic senators on this issue-specific scale. Bernie Sanders often treads carefully on the issue of gun control, reflecting his support of the hunting traditions of his home state of Vermont. Pat Toomey, placed in the middle of the Republican Party on the overall liberal-conservative ideology scale, is placed as the Republican most supportive of gun control. Toomey, breaking from his party, has supported background checks and state red flag laws.

Gun Control LaMP scores better predict NRA grades

We compare the predictive power of Gun Control LaMP scores and the first dimension of DW-NOMINATE on NRA grades. The NRA assigns grades each election cycle, although not all candidates receive one. We collected the latest available NRA grade for each senator from Vote Smart. We again use multivariate analyses to calculate how much the proportion of variance explained (R^2) in NRA grades falls when we compare the full model, regressing the NRA grade on both the first dimension of DW-NOMINATE and Gun Control LaMP scores, with reduced models, which only use DW-NOMINATE or the Gun Control LaMP scores as the predictor. Figure 6 shows how R^2 falls as we move from full to reduced models across all senators, across the Democratic senators, and across the Republican senators.

The proportion of variance explained in NRA grades always falls more when we drop Gun Control LaMP scores as a predictor than when dropping DW-NOMINATE as a predictor. Looking at the Republican senators in particular, the R^2 is near 0 when only using DW-NOMINATE to predict the NRA grades of Republicans. However, the R^2 is 0.214 when only

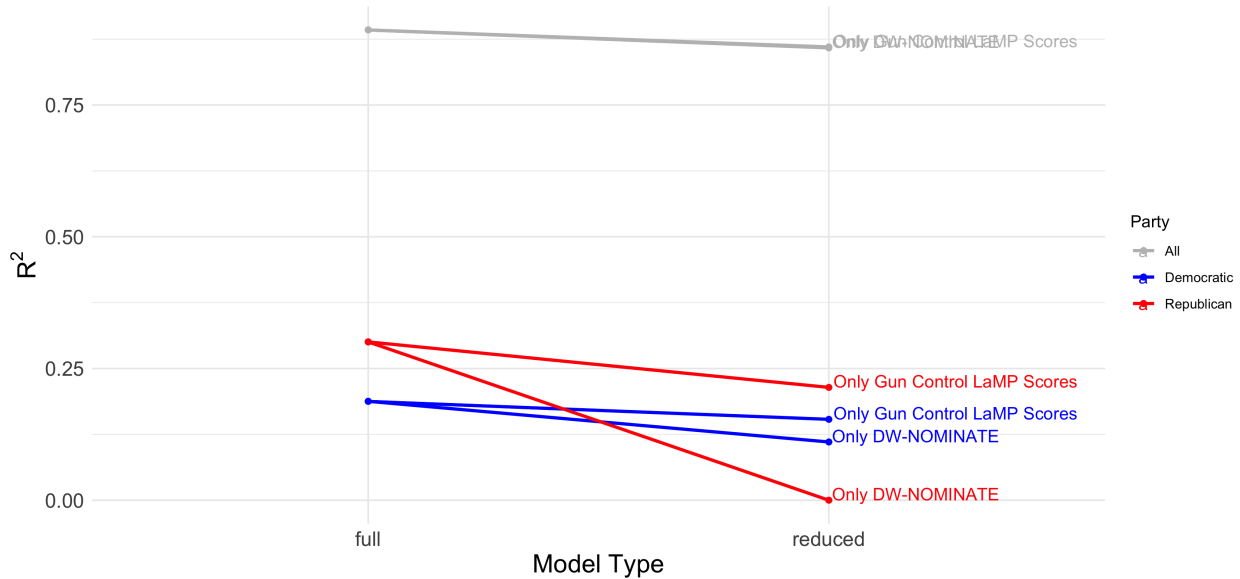


Figure 6: Comparing the proportion of variance, R^2 , between the full and reduced models. The full model regresses the NRA grades on both the first dimension of DW-NOMINATE and the Gun Control LaMP scores. The reduced models only use DW-NOMINATE or Gun Control LaMP scores as predictors. The R^2 is always lower when Gun Control LaMP scores are removed as a predictor than when DW-NOMINATE is removed.

using the Gun Control LaMP scores, indicating that the Gun Control LaMP scores measure a different latent construct than DW-NOMINATE. The same pattern holds for Democratic senators, although the difference is smaller. Because the NRA treats Democratic senators differently from Republican senators and gives nearly all of them scores close to 0—Tom Udall had the highest grade from the NRA among Democratic senators, at 21—the R^2 among Democratic senators is less interpretable because there is so little variance to be explained. Among all senators, the difference in R^2 is negligible, largely because positions on gun control are still aligned with the two parties.

Gun Control LaMP scores predict Republican votes on the 2022 Bipartisan Safer Communities Act

ChatGPT’s training data could explain the predictive power of Gun Control LaMP scores on NRA grades. To evaluate the scale’s external validity, we predict Republican votes on the 2022 Bipartisan Safer Communities Act, which could not have been used in ChatGPT’s

training data because the model was only trained on data through September 2021. 15 Republican senators voted alongside all Democratic senators to pass the bill; thus, we exclude the votes of Democratic senators. We used a logistic regression predicting Republican votes on the bill using Gun Control LaMP scores and the first dimension of DW-NOMINATE; results are in Table 2. Gun Control LaMP scores are a statistically significant predictor of Republican votes on the 2022 Bipartisan Safer Communities Act, even when controlling for the first dimension of DW-NOMINATE.

	Voted Yea
Gun Control LaMP scores	21.484** (8.258)
DW-NOMINATE	-9.863 (5.447)
Constant	0.882 (2.745)
Observations	45
Log Likelihood	-15.202
<i>Note:</i>	*p<0.05; **p<0.01; ***p<0.001

Table 2: We used logistic regression to predict Republican votes on the 2022 Bipartisan Safer Communities Act using the Gun Control LaMP scores and the first dimension of DW-NOMINATE as predictor variables. Gun Control LaMP scores are statistically significantly predictive of Republican votes on this bill. Democratic votes were excluded because all Democratic senators voted in favor of the bill.

Abortion Rights Support Scaling

To scale senators along an abortion rights support continuum, the “winner” of each matchup was the senator the LLM answered as being more pro-choice (or more pro-life). We call these “Abortion Rights LaMP scores.”

Abortion Rights LaMP scores highly correlate across repeated iterations

Similar to the other applications, we find that the lowest correlation among any two iterations is 0.996. Looking at these correlations by party, the lowest correlation between any two iterations' Abortion Rights LaMP scores is 0.968 for Democratic senators and 0.952 for Republican senators. It again demonstrates ChatGPT's consistency in its responses to pairwise comparisons despite being asked about a specific policy area.

To more closely analyze the uncertainty of the LLM's responses to pairwise comparisons of senators along this specific issue, we again use the entropy-based metric derived in the Supplementary Information. The results in Figure 12 in the Supplementary Information show that as the distance between LaMP scores increases, the entropy in the LLM's pairwise comparison responses decreases.

Abortion Rights LaMP scores have high face-validity

The Abortion Rights LaMP scores of all senators of the 116th Congress are illustrated in Figure 7. There is face validity with these scores. For example, it correctly separates the moderately pro-choice Republicans, Lisa Murkowski and Susan Collins. They are the only Republicans who describe themselves as pro-choice, although they often vote to confirm pro-life nominees. It also correctly separates Bob Casey and Joe Manchin, who self-describe themselves as pro-life and are endorsed by the Democrats for Life of America, a PAC that seeks to elect anti-abortion Democratic candidates.

Abortion Rights LaMP scores better predict NARAL grades

We compare the predictive power of Abortion Rights LaMP scores and the first dimension of DW-NOMINATE on NARAL Pro-Choice America grades. NARAL assigns these grades using a set of votes on motions, bills, and confirmations that are related to abortion rights in some way. We used the NARAL grades from the end of 2020. We again use multivariate analyses to calculate how much the proportion of variance explained in NARAL grades falls

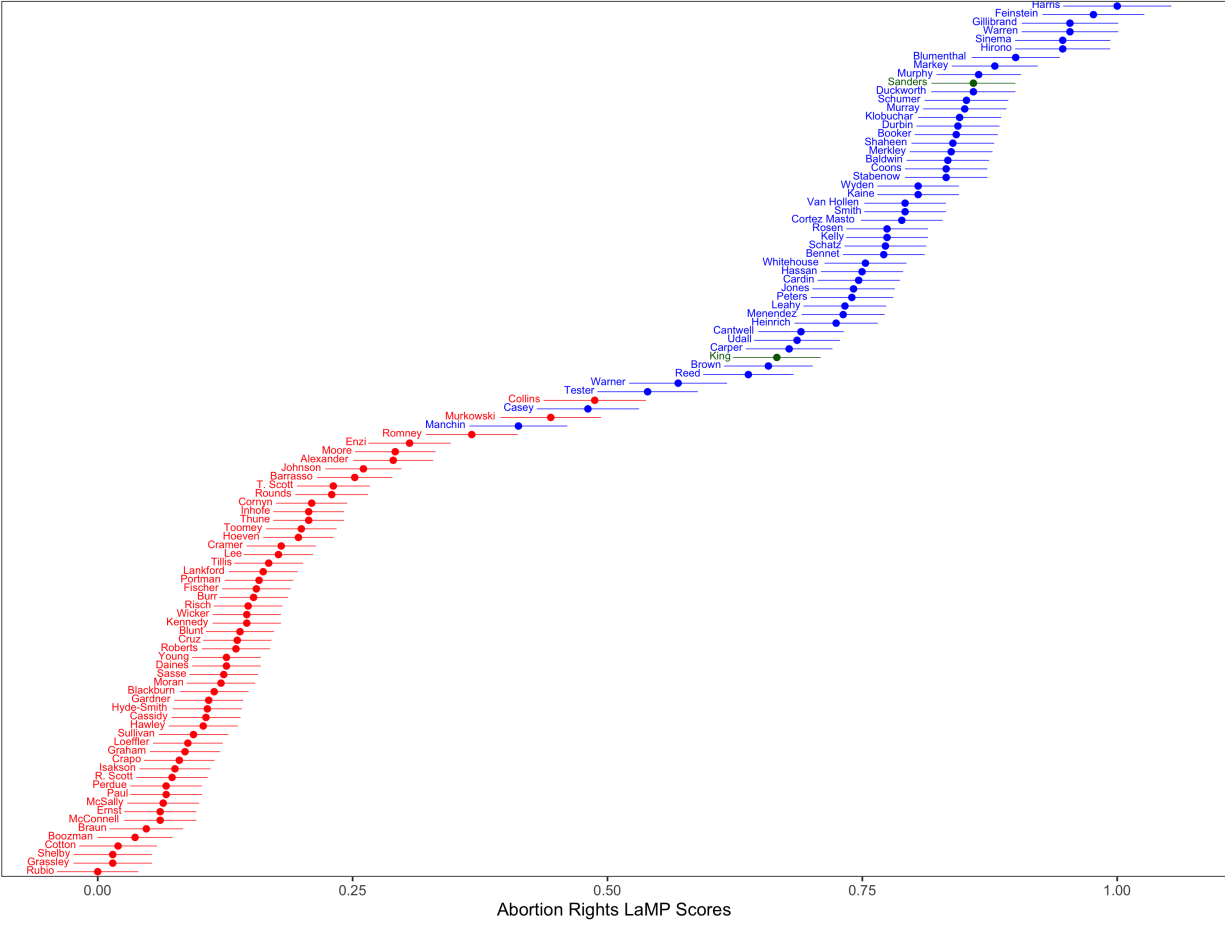


Figure 7: Abortion Rights LaMP scores across all senators with 95% confidence intervals derived from quasi-standard errors [25]. Democratic senators are in blue, Republican senators are in red, and Independent senators are in green.

when we compare the full model, regressing the NARAL grade on both the first dimension of DW-NOMINATE and Abortion Rights LaMP scores, with reduced models, which only use DW-NOMINATE or the Abortion Rights LaMP scores as the predictor. Figure 8 shows how the R^2 drops as we move from full to reduced models across all senators, across the Democratic senators, and across the Republican senators.

We find that the proportion of variance explained in NARAL grades always falls more when we drop Abortion Rights LaMP scores as a predictor than when dropping DW-NOMINATE as a predictor. This holds for both Democratic and Republican senators. The partial F-test p-values for the full model and the reduced model with only DW-NOMINATE

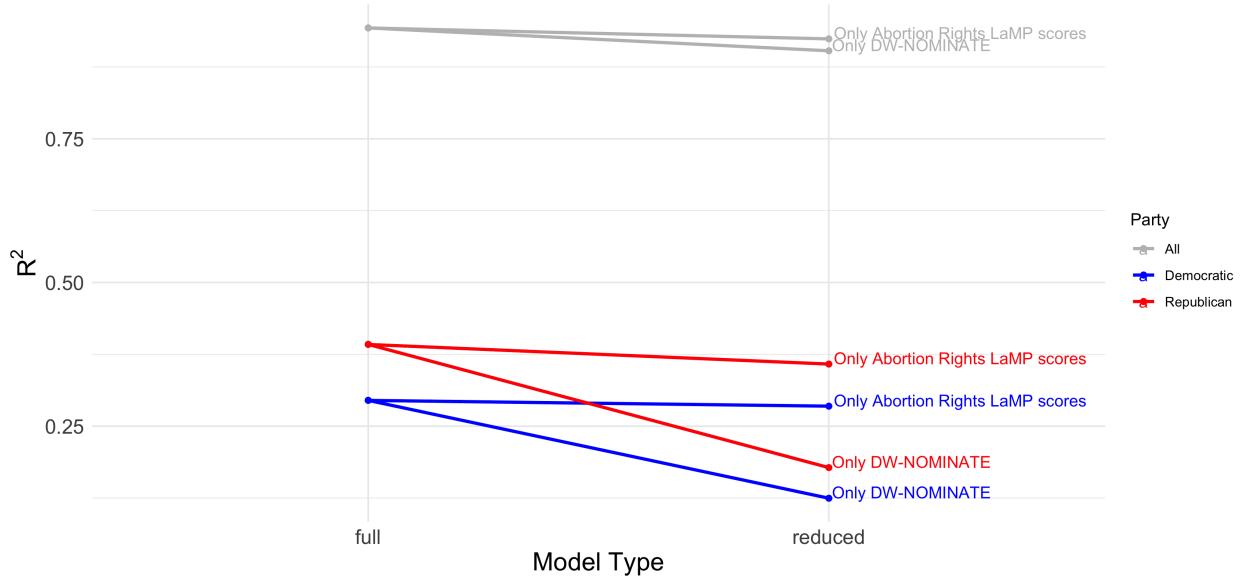


Figure 8: Comparing the proportion of variance explained, R^2 , between the full and reduced models. The full model regresses the NARAL grades on both the first dimension of DW-NOMINATE and the Abortion Rights LaMP scores. The reduced models only use DW-NOMINATE or the Abortion Rights LaMP scores as predictors. The R^2 is always lower when Abortion Rights LaMP scores are removed as a predictor compared to when DW-NOMINATE is removed.

as a predictor are significant ($p < .001$) for both parties, while the p-value for the full model and the reduced model with only Abortion Rights LaMP scores are not ($p > .099$) for both parties. The difference in R^2 for models across all senators is negligible because stances on abortion still broadly fall along party lines.

Discussion

Our findings suggest that generative large language models can be useful for measuring the latent positions of lawmakers, especially on specific issues such as gun control and abortion. We find that the LLM is not hallucinating in these pairwise comparisons, LaMP scores are stable between repeated sets of matchups, and the LLM is not simply parroting existing scales such as DW-NOMINATE or interest group ratings. In other words, pairwise comparisons with an LLM yield sensible scales of lawmakers. Our evidence is consistent with the idea that the LLM synthesizes a great deal of information about lawmakers to evaluate latent

constructs in predictable and sensible ways, agreeing with existing scales and predicting lawmaker behaviors like votes.

Our method is situated in a rapidly growing literature on using generative LLMs for social science applications. For example, these works have studied how generative LLMs can be used for labeling purposes [30], analyzing text along psychological constructs [31], reducing the divisiveness of online conversations [32], and generating artificially politically extreme responses [33]. Most of these works focus on generating answers about one item at a time and studying how the LLM answers change across different items. On the other hand, our method examines how the LLM compares pairs of items and what kind of novel continuous measures can be derived using the LLM’s answers.

Our approach to scaling also speaks to a vast body of work on ideological scaling and ideal point estimation [see, e.g., 6, 9, 14, 16–18, 21, 27, 34–42]. Estimation of ideology and stance has usually focused on behavior, such as how lawmakers vote in roll call votes or what specific words Twitter users use in tweets; alternative measures have focused on the perceptions of lawmakers, such as campaign donations, following-follower behavior on Twitter, and political activists’ opinions. Our approach uses the embedded knowledge and analytical capabilities of LLMs, elicited through pairwise comparisons, about lawmakers’ ideologies and their stances on public policy issues such as gun control and abortion to create continuous, unidimensional scales of these latent dimensions of politics and policy.

Our approach has limitations due to the black box nature of deep neural networks: we do not know how ChatGPT is formulating its response to pairwise comparison prompts. We do know what texts were used to train ChatGPT-3.5 [43]—most notably Wikipedia and the Common Crawl, a corpus created from crawling the entire web—and we know that the existing measures discussed in this paper are present in the training data, as evidenced by its ability to define such measures correctly when prompted. One could attempt to “understand” why ChatGPT chooses a particular senator in each pairwise comparison using its outputted explanation. However, it may also use the answer as its context to generate

the explanation. In other words, we still would not know what factors the LLM uses, why the LLM places greater weight on certain factors, nor would we know to what degree or what direction its explanation is related to its answer. The LLM cannot “reason” or “deliberate” about ideology or politics; it is, instead, a pattern recognition model.

However, our analysis shows that the liberal-conservative ideology scale produced using an LLM’s response to pairwise comparison prompts correlates well with multiple existing measures and reflects factors beyond position-taking. In other words, the findings above, such as the partial correlations and the Ideology LaMP scores’ strong predictive power of activists’ perceptions, suggest that LaMP scores reflect a blend of textual information: the scales are influenced by any number of factors that exist in its training corpora such as existing measures of ideology, floor speeches about certain public policy areas, news stories about senators, and people’s expressed perceptions of these senators. LaMP scores can offset interpretive gaps in existing measures of ideology, such as dealing with lawmakers who vote against their own party for ideological reasons. In short, the LLM generates reasonable answers based on its training data.

What this means is that we can use LLMs with pairwise comparisons to estimate novel measures along specific political or policy dimensions, such as support for gun control or abortion rights, that could not be estimated using conventional scaling methods because of an absence of data related to behaviors or perceptions of the lawmakers. Estimating support for gun control among the same senators, we find that the scale not only differs in meaningful and intuitive ways from liberal-conservative scales but also that the scale *predicts* votes on a gun control bill that is not in ChatGPT’s training data. We find a similar pattern with the support of abortion rights.

Despite some of its limitations, our proposed approach creates scales with both face validity and external validity. Our applications have been limited to American politics applications at the federal level, but one could extend the method to estimating the ideology of legislators in parliamentary systems where members of parliament vote strictly along

party lines. A limitation of these comparisons, however, is that the LLM may not have enough information about every member of parliament. In such cases, a solution could be to pairwise compare text, such as campaign materials or tweets. Our approach can also be used to automatically evaluate whether the LLM’s scaling of concepts of interest aligns with human judgments; this has potential applications for alignment research in natural language processing. In short, there are significant prospective future research contributions using our pairwise comparison approach for both the social sciences and natural language processing.

Methods

A Brief Overview of ChatGPT

ChatGPT stands for Chat **G**enerative **P**retrained **T**ransformer. Given an initial text input called the prompt, ChatGPT generates a response. ChatGPT is built on GPT-3 [43], a complex neural network (specifically, a decoder-only transformer) that predicts what token appears next given the set of existing tokens. It is trained on a massive corpus of text, which includes a filtered version of Common Crawl, WebText2, Books1, Books2, and Wikipedia. Alammar [44] explains and illustrates how decoder-only transformer large language models are trained in much greater detail. In short, its ability to produce coherent textual responses comes from the massive training corpora, the sheer size of the neural network (175 billion parameters), and the self-attention mechanism. The self-attention mechanism, in effect, allows the model to dynamically upweight and downweight certain parts of the input sequence [45]. It achieves this by assigning a weight to each element of the input sequence that reflects the importance of that element relative to other elements.

ChatGPT is specifically trained to be a chatbot using a reinforcement learning technique called **reinforcement learning from human feedback**, or RLHF [see, e.g., 46]. RLHF works by first obtaining the generated text from the model. Human annotators then rank the output from most preferred to least preferred. A reinforcement learning algorithm—specifically,

proximal policy optimization [47]—is then used to update the large language model. For a more detailed discussion about RLHF, see Lambert et al. [48]. The addition of RLHF training enables ChatGPT to generate human-like responses. It is important to note that ChatGPT can still produce incorrect responses (and sound extremely confident in doing so!), a phenomenon referred to as “hallucination.”

It is also well-known that GPT-3 can generate text with biases, negative stereotypes, and unfair associations [see, e.g., 49]. These biases can be potentially leveraged for social science purposes. For example, Argyle et al. [50] find that biases in GPT-3 are fine-grained and demographically correlated and can be used to emulate partisan responses from a wide variety of human subgroups.

Using ChatGPT to Measure the Ideologies of Politicians

We analyze the senators of the 116th Congress, which convened on January 3, 2019 and ended on January 3, 2021. We use this particular Congress because ChatGPT is trained on information up to 2021. We do not look at previous Congresses in order to prevent ChatGPT’s newest information from leaking into assessments of the ideologies of members of previous Congresses. We obtain this list of senators from Voteview [51]. We keep Martha McSally (R-AZ) and Kelly Loeffler (R-GA) on the list of senators. Martha McSally was appointed to the Senate following interim Senator Jon Kyl’s resignation. She then ran in Arizona’s special election to finish the remainder of the Senate term but lost to Mark Kelly. Similarly, Kelly Loeffler was appointed to the Senate following Johnny Isakson’s resignation for health reasons at the end of 2019.

We pairwise compare all senators of the 116th Congress. We call these pairwise comparisons “matchups.” We input the following prompt into ChatGPT for matchups between Democratic senators and matchups between a Democratic senator and a Republican senator:

Based on past voting records and statements, which senator is more liberal: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator

2 party abbrev]-[senator 2 state abbrev]))?

For matchups between Republican senators, we use a similar prompt:

Based on past voting records and statements, which senator is more conservative: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])?

We change the wording for matchups between Republican senators strictly because of a quirk of ChatGPT (and illustrative of its inability to “reason” about politics): when asked which senator in each pair is more liberal when comparing two conservative Republicans, it will often reply that neither senator is “more liberal” because they are both conservative Republicans. We use the default temperature parameter of 1 for ChatGPT-3.5.

We use the “more liberal” prompt when comparing a Democratic senator and a Republican senator. Running pairwise comparisons using the “more conservative” prompt instead when comparing a Democratic senator and a Republican senator, the two Ideology LaMP scores correlate at 0.997. Using “more liberal” or “more conservative” for those particular matchups does not affect the LaMP scores.

We record the name of the senator that ChatGPT considers to be more conservative (liberal) in each matchup. More specifically, we take ChatGPT’s answer to the comparison prompt and use it in another prompt to extract the senator’s name from the answer. See the Supplementary Information for more information about the prompt we used to extract the name of the more conservative senator from ChatGPT’s response. Ties are allowed; this is when the LLM cannot assess who is more liberal (or more conservative) between the two senators.

Between 102 senators, there are a total of 5,151 unique matchups. Each senator is compared to all other senators three times to study the consistency of ChatGPT’s answers. The final comparison table contains each matchup with the number of wins (times a senator was deemed more conservative in that specific matchup) and losses (times a senator was deemed more liberal in that specific matchup) for that particular matchup. We consider ties

0.5 wins for both senators in the matchup; the section on the Bradley-Terry Model describes why we chose this approach for ties.

Using ChatGPT to Measure Gun Control Support

To analyze the senators of the 116th Congress specifically for gun control support, we use the following prompt for all possible combinations of senators (5,151 comparisons).

Based on past voting records and statements, who is more likely to support gun control: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])?

We then extract the names using a separate prompt; the Supplementary Information contains more information about the prompt we use to extract the name of the senator who is more likely to support gun control. Each senator is compared to all other senators three times to study the consistency of ChatGPT's answers; again, there are a total of 15,453 pairwise comparisons. We handle the wins, losses, and ties in the same way as the liberal-conservative ideology scale.

Using ChatGPT to Measure Abortion Rights Support

To analyze the senators of the 116th Congress specifically for support of abortion rights, we use the following prompt for matchups between Democratic senators and matchups between a Democratic senator and a Republican senator.

Based on past voting records and statements, which senator is more pro-choice: [senator 1] ([senator 1 party-abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])?

For matchups between Republican senators, we used a similar prompt:

Based on past voting records and statements, which senator is more pro-life: [senator 1] ([senator 1 party-abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])?

Again, we change the wording for matchups between Republican senators because ChatGPT will often not compare two Republican senators on a pro-choice comparisons basis. It will often only make a pairwise comparison when asked which Republican senator is more pro-life.

Using the Bradley-Terry Model to Estimate 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 α_i/α_j , where α_i and α_j are positive-valued parameters that indicate latent “ability” [20]. 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 senators regarding who is more conservative, more likely to support gun control, or more pro-choice. We use the liberal-conservative ideology scale as the running example in this section. The λ parameters are measures of the senators’ latent liberal-conservative ideology; the estimated parameters $\hat{\lambda}$ are the Ideology LaMP scores we describe in the Results section. We denote the more conservative senator in each matchup as the “winner” so that conservative senators have higher scores, intuitively matching the liberal-conservative political spectrum. For ties, we consider these 0.5 wins for both senators in the matchup. Turner and Firth [52] 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 responses to pairwise comparisons to estimate our scales of interest [52]. The estimated parameters $\hat{\lambda}$ are relative to a reference senator (for all scales, we use Lisa Murkowski). However, this choice is unimportant because we rescale the estimated parameters to the unit interval to

get the LaMP scores.

We can also calculate standard errors for the parameter estimates. The standard errors are derived from the inverse of the observed information matrix. However, the interpretation of these standard errors is relative to the “reference” senator. Firth and De Menezes [25] derive quasi-variances, which can be interpreted as if they refer to reference-free estimates of the variance of the underlying “ability” of interest of senators. This means that any pair of senators can be directly compared. These quasi-variances can then be used to derive standard errors and 95% confidence intervals. We use the `qvcalc` package to calculate quasi-standard errors [53]. The 95% confidence intervals in the results of this paper are derived from these quasi-standard errors.

Acknowledgements and Funding Sources

We gratefully acknowledge that the Center for Social Media and Politics at New York University is supported by funding from the John S. and James L. Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, the William and Flora Hewlett Foundation, the Siegel Family Endowment, and the Bill and Melinda Gates Foundation. We thank the members of the Center for Social Media and Politics for their helpful comments when workshopping this paper. We would also like to thank Maggie Macdonald and Megan Brown for their helpful comments throughout the paper-writing process.

References

- [1] Stephen Ansolabehere, James M. Snyder, and Charles Stewart. “Candidate Positioning in U.S. House Elections”. In: *American Journal of Political Science* 45.1 (2001), pp. 136–159.
- [2] Elisabeth R. Gerber and Jeffrey B. Lewis. “Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation”. In: *Journal of Political Economy* 112.6 (2004), pp. 1364–1383.
- [3] Larry M. Bartels. “Economic Inequality and Political Representation”. In: *The Unsustainable American State*. Oxford University Press, Oct. 2009.
- [4] Danielle M. Thomsen. *Opting Out of Congress: Partisan Polarization and the Decline of Moderate Candidates*. Cambridge University Press, 2017.
- [5] Devin Caughey and Christopher Warshaw. “Policy Preferences and Policy Change: Dynamic Responsiveness in the American States, 1936–2014”. In: *American Political Science Review* 112.2 (2018), pp. 249–266.
- [6] Keith T. Poole and Howard Rosenthal. *Ideology and Congress: A Political Economic History of Roll Call Voting*. New Haven, CT: Yale University Press, 1997.
- [7] Michele L. Swers. “Are Women More Likely to Vote for Women’s Issue Bills than Their Male Colleagues?” In: *Legislative Studies Quarterly* 23.3 (1998), pp. 435–448.
- [8] Keith Krehbiel. *Pivotal Politics: A Theory of U.S. Lawmaking*. Chicago: University of Chicago Press, 1998.
- [9] Joshua Clinton, Simon Jackman, and Douglas Rivers. “The Statistical Analysis of Roll Call Data”. In: *American Political Science Review* 98.2 (2004), pp. 355–370.
- [10] Gary W. Cox and Mathew D. McCubbins. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge University Press, 2005.
- [11] Benjamin Highton and Michael S. Rocca. “Beyond the Roll-Call Arena: The Determinants of Position Taking in Congress”. In: *Political Research Quarterly* 58.2 (2005), pp. 303–316.
- [12] Cheryl Boudreau, Christopher S. Elmendorf, and Scott A. MacKenzie. “Racial or Spatial Voting? The Effects of Candidate Ethnicity and Ethnic Group Endorsements in Local Elections”. In: *American Journal of Political Science* 63.1 (2019), pp. 5–20.
- [13] Annelise Russell. “Minority Opposition and Asymmetric Parties? Senators’ Partisan Rhetoric on Twitter”. In: *Political Research Quarterly* 74.3 (2021), pp. 615–627.
- [14] Keith T. Poole and Howard Rosenthal. “A Spatial Model for Legislative Roll Call Analysis”. In: *American Journal of Political Science* 29.2 (1985), pp. 357–384.
- [15] Keith T. Poole. *Spatial Models of Parliamentary Voting*. Analytical Methods for Social Research. Cambridge University Press, 2005.
- [16] Royce Carroll, Jeffrey B. Lewis, James Lo, Keith T. Poole, and Howard Rosenthal. “Measuring Bias and Uncertainty in DW-NOMINATE Ideal Point Estimates via the Parametric Bootstrap”. In: *Political Analysis* 17.3 (2009), pp. 261–275.

- [17] Gregory Eady, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler. “News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians”. Nov. 2020.
- [18] Adam Bonica. “Mapping the Ideological Marketplace”. In: *American Journal of Political Science* 58.2 (2014), pp. 367–386.
- [19] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.
- [20] Ralph Allan Bradley and Milton E. Terry. “Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons”. In: *Biometrika* 39.3-4 (Dec. 1952), pp. 324–345.
- [21] Daniel J. Hopkins and Hans Noel. “Trump and the Shifting Meaning of “Conservative”: Using Activists’ Pairwise Comparisons to Measure Politicians’ Perceived Ideologies”. In: *American Political Science Review* 116.3 (2022), pp. 1133–1140.
- [22] Nolan McCarty. “Measuring Legislative Preferences”. English (US). In: *The Oxford Handbook of the American Congress*. Publisher Copyright: © The several contributors 2011. All rights reserved. United Kingdom: Oxford University Press, May 2011.
- [23] Peter John Loewen, Daniel Rubenson, and Arthur Spirling. “Testing the power of arguments in referendums: A Bradley–Terry approach”. In: *Electoral Studies* 31.1 (2012). Special Symposium: Germany’s Federal Election September 2009, pp. 212–221.
- [24] David Carlson and Jacob M. Montgomery. “A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts”. In: *American Political Science Review* 111.4 (2017), pp. 835–843.
- [25] David Firth and Renée X. De Menezes. “Quasi-variances”. In: *Biometrika* 91.1 (Mar. 2004), pp. 65–80.
- [26] Devin Caughey and Eric Schickler. “Substance and Change in Congressional Ideology: NOMINATE and Its Alternatives”. In: *Studies in American Political Development* 30.2 (2016), pp. 128–146.
- [27] JBrandon Duck-Mayr and Jacob Montgomery. “Ends Against the Middle: Measuring Latent Traits when Opposites Respond the Same Way for Antithetical Reasons”. In: *Political Analysis* (2023), pp. 1–20.
- [28] Adam Bonica. *Database on Ideology, Money in Politics, and Elections*. Available at <https://data.stanford.edu/dime>. 2016.
- [29] Seongho Kim. “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients”. eng. In: *Communications for Statistical Applications and Methods* 22.6 (2015), pp. 665–674.
- [30] Petter Törnberg. *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. 2023.
- [31] Steve Rathje, Dan-Mircea Mirea, Iliia Sucholutsky, Raja Marjeh, Claire Robertson, and Jay J Van Bavel. *GPT is an effective tool for multilingual psychological text analysis*. May 2023.

- [32] 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.
- [33] James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. *Synthetic Replacements for Human Survey Data? The Perils of Large Language Models*. May 2023.
- [34] James J. Heckman and James M. Snyder. “Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators”. In: *The RAND Journal of Economics* 28 (1997), S142–S189.
- [35] Andrew D. Martin and Kevin M. Quinn. “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999”. In: *Political Analysis* 10.2 (2002), pp. 134–153.
- [36] Jonathan B. Slapin and Sven-Oliver Proksch. “A Scaling Model for Estimating Time-Series Party Positions from Texts”. In: *American Journal of Political Science* 52.3 (2008), pp. 705–722.
- [37] Boris Shor and Nolan McCarty. “The Ideological Mapping of American Legislatures”. In: *The American Political Science Review* 105.3 (2011), pp. 530–551.
- [38] Will Lowe and Kenneth Benoit. “Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark”. In: *Political Analysis* 21.3 (2013), pp. 298–313.
- [39] Pablo Barberá. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data”. In: *Political Analysis* 23.1 (2015), pp. 76–91.
- [40] Mickael Temporão, Corentin Vande Kerckhove, Clifton van der Linden, Yannick Dufresne, and Julien M. Hendrickx. “Ideological Scaling of Social Media Users: A Dynamic Lexicon Approach”. In: *Political Analysis* 26.4 (2018), pp. 457–473.
- [41] 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.
- [42] Ludovic Rheault and Christopher Cochrane. “Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora”. In: *Political Analysis* 28.1 (2020), pp. 112–133.
- [43] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language Models are Few-Shot Learners”. 2020.
- [44] Jay Alammar. *The Illustrated GPT-2 (Visualizing Transformer Language Models)*. Retrieved from <https://jalamar.github.io/illustrated-gpt2/>. 2019.

- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems* 30 (2017). Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett.
- [46] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. “Fine-Tuning Language Models from Human Preferences”. 2019.
- [47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. *Proximal Policy Optimization Algorithms*. 2017.
- [48] Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. “Illustrating Reinforcement Learning from Human Feedback (RLHF)”. In: *Hugging Face Blog* (2022). <https://huggingface.co/blog/rlhf>.
- [49] Li Lucy and David Bamman. “Gender and Representation Bias in GPT-3 Generated Stories”. In: *Proceedings of the Third Workshop on Narrative Understanding* (June 2021), pp. 48–55.
- [50] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. “Out of One, Many: Using Language Models to Simulate Human Samples”. In: *Political Analysis* (2023), pp. 1–15.
- [51] Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. *Voteview: Congressional Roll-Call Votes Database*. <https://voteview.com/>. 2021.
- [52] Heather Turner and David Firth. “Bradley-Terry Models in R: The BradleyTerry2 Package”. In: *Journal of Statistical Software* 48.9 (2012), pp. 1–21.
- [53] David Firth. *qvalc: Quasi Variances for Factor Effects in Statistical Models*. R package version 1.0.3. 2023.
- [54] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. “Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation”. In: *The Eleventh International Conference on Learning Representations*. 2023.
- [55] Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. *Evaluating the Moral Beliefs Encoded in LLMs*. 2023.
- [56] Lauren R. Johnson, Deon McCray, and Jordan M. Ragusa. “#NeverTrump: Why Republican members of Congress refused to support their party’s nominee in the 2016 presidential election”. In: *Research & Politics* 5.1 (2018).

Supplementary Information for “Large Language Models Can Be Used to Estimate the Latent Positions of Politicians”

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

Ideology LaMP Scores with 95% Confidence Intervals

Figure 9 shows the Ideology LaMP scores across all senators with 95% confidence intervals that are derived from quasi-standard errors. We use quasi-standard errors because the standard errors derived directly from the Bradley-Terry model are calculated relative to a “reference” senator. Quasi-standard errors can be interpreted as if they refer to independent estimates of ideology for each senator [25, 52]. This allows for comparisons between individual senators.

Quantifying Uncertainty in the LLM’s Pairwise Comparisons

We obtain quasi-standard errors for each senator’s estimated ideology or stance on a specific issue. These quasi-standard errors are, in turn, derived from standard errors calculated using the inverse of the observed information matrix of the Bradley-Terry maximum likelihood estimates. These quasi-standard errors are derived from all three iterations—recall that each pairwise comparison is made 3 times to the LLM. However, we do not know what proportion of these quasi-standard errors are from uncertainty in the LLM’s responses to pairwise comparisons. To better understand uncertainty in the LLM’s responses to pairwise comparisons, we derive an approach based on related work in Kuhn, Gal, and Farquhar [54] and Scherrer et al. [55].

Without loss of generality, assume we are interested in pairwise comparing senators on the basis of liberal-conservative ideology. Let $r_{i,j,i}$ be the event where representative i is chosen to be more conservative than representative j in a pairwise comparison between representatives i and j ; define $r_{i,j,j}$ similarly except representative j is chosen to be more conservative than representative i . We first define the likelihood of the LLM picking representative i over

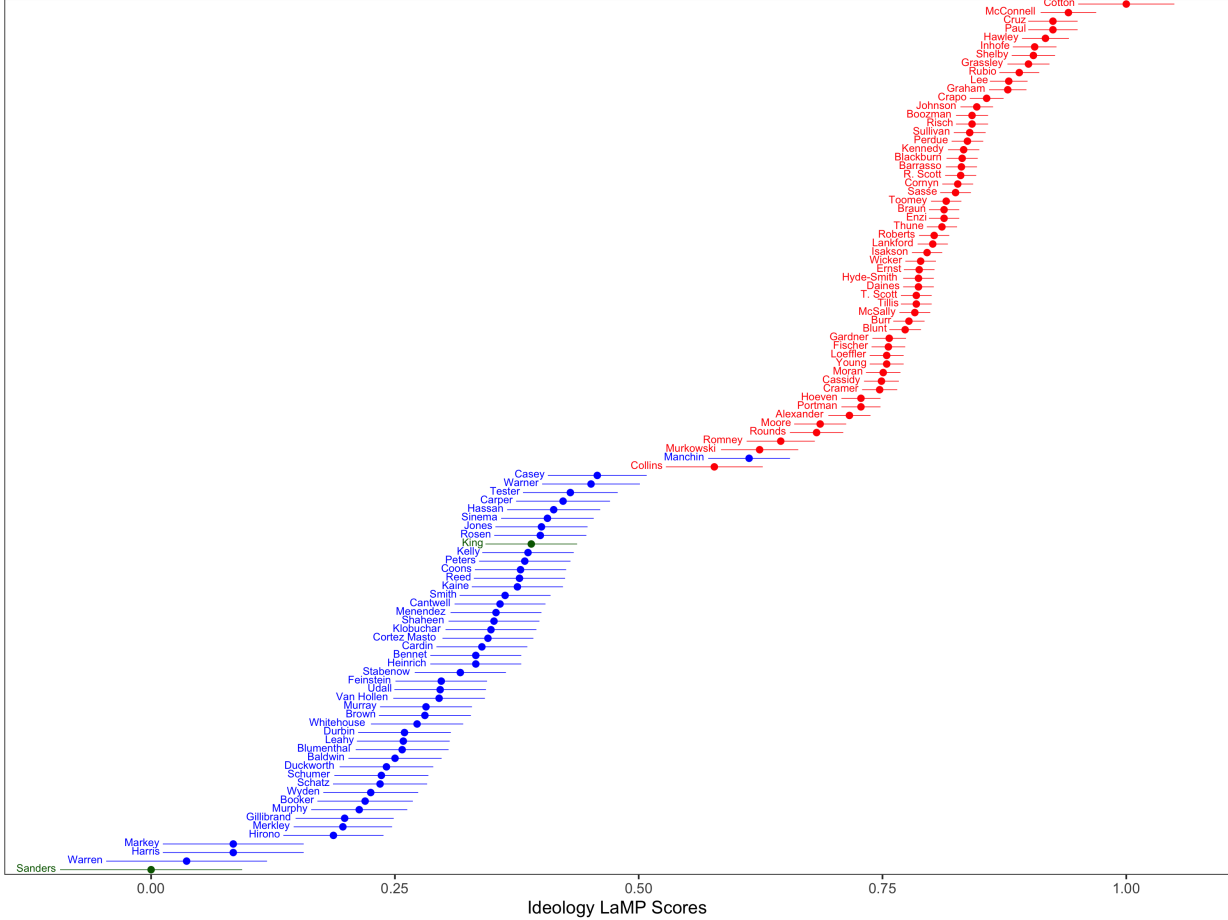


Figure 9: Ideology LaMP scores across all senators with 95% confidence intervals derived from quasi-standard errors [25]. Democratic senators are in blue, Republican senators are in red, and Independent senators are in green.

representative j .

Definition 1 *The likelihood of representative i being picked as being more conservative than representative j using a large language model p_θ in a comparison between representative i and j using prompt $x_{i,j}$ is defined as*

$$p_\theta(r_{i,j,i}|x_{i,j}) = \sum_{s \in c(r_{i,j,i}, x_{i,j})} p_\theta(s|x_{i,j}) \quad (1)$$

where $c(r_{i,j,i}, x_{i,j})$ is the set of all token sequences s that are semantically equivalent in encoding a preference for representative i in a pairwise comparison between representative i

and j . This is the probability that an LLM “chooses” a representative i given the matchup prompt $x_{i,j}$.

We now define a semantic entropy measure, introduced in Kuhn, Gal, and Farquhar [54], that quantifies the uncertainty of the LLM.

Definition 2 *The entropy of a large language model p_θ on a pairwise comparison $x_{i,j}$ is defined as*

$$H_\theta [R_{i,j}|x_{i,j}] = -p_\theta(r_{i,j,i}|x_{i,j}) \log(p_\theta(r_{i,j,i}|x_{i,j})) - p_\theta(r_{i,j,j}|x_{i,j}) \log(p_\theta(r_{i,j,j}|x_{i,j})) \quad (2)$$

Calculating Equation 1 is intractable because of the large number of possible token sequences s that are semantically equivalent. Instead, we propose a sampling estimator based on a similar estimator from Scherrer et al. [55]. We first sample M token sequences $\{s_1, \dots, s_M\}$ from the LLM by $s_k \sim p_\theta(s|x_{i,j})$. We then map each of these token sequences s to either representative i or representative j being chosen as the more conservative representative using a deterministic mapping function $g : (s, x_{i,j}) \rightarrow R_{i,j}$. Then, the likelihood in Equation 1 is

$$\hat{p}_\theta(r_{i,j,i}|x_{i,j}) = \frac{1}{M} \sum_{k=1}^M \mathbb{1}[g(s_k) = r_{i,j,i}] \quad (3)$$

Here, our mapping function g uses the LLM with another prompt to extract the answer. In future work, we will study the uncertainty of this particular mapping function. We can plug Equation 3 into Equation 2 to estimate the entropy for each matchup.

To analyze uncertainty in the model’s responses, we look at the estimated entropy, using the estimated version of Equation 2, for each matchup as a function of the difference in the LaMP scores for each matchup. Figure 10 shows the relationship between entropy and the difference in Ideology LaMP scores for each pairwise comparison, Figure 11 shows the relationship between entropy and the difference in Gun Control LaMP scores for each pairwise comparison, and Figure 12 shows the relationship between entropy and the difference in Abortion Rights LaMP scores for each pairwise comparison.

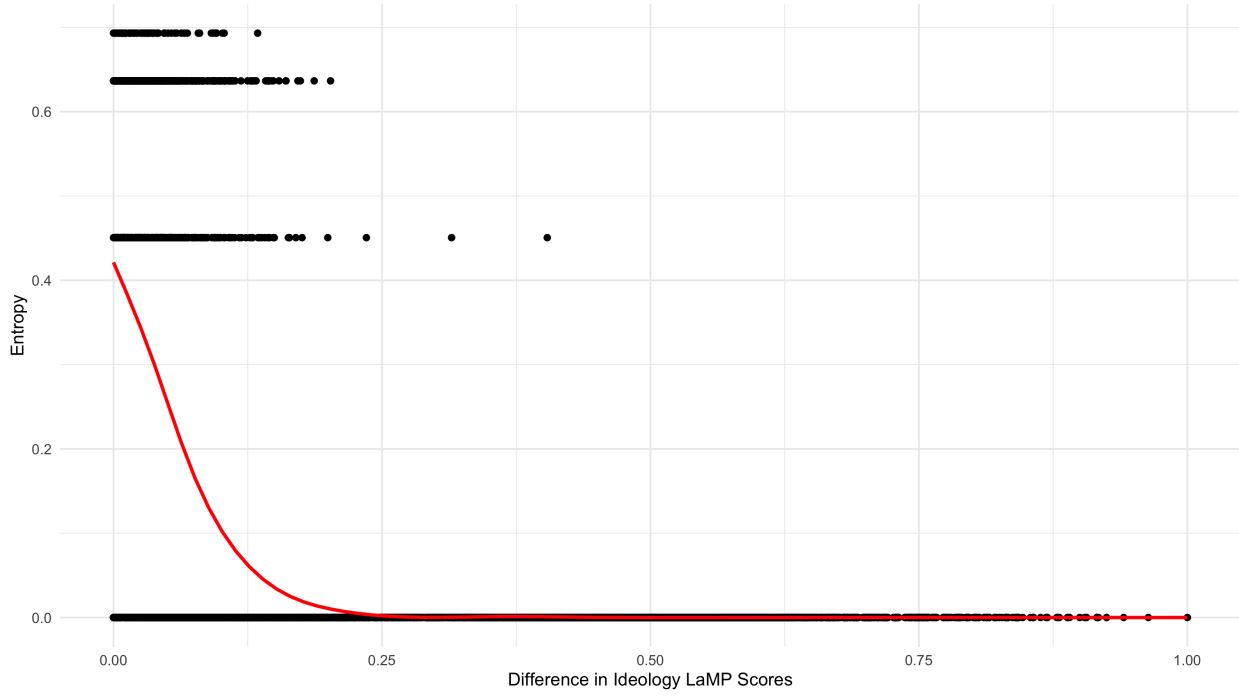


Figure 10: Entropy with respect to the difference in Ideology LaMP scores for each pairwise comparison between senators. The red line is a generalized additive model line of best fit.

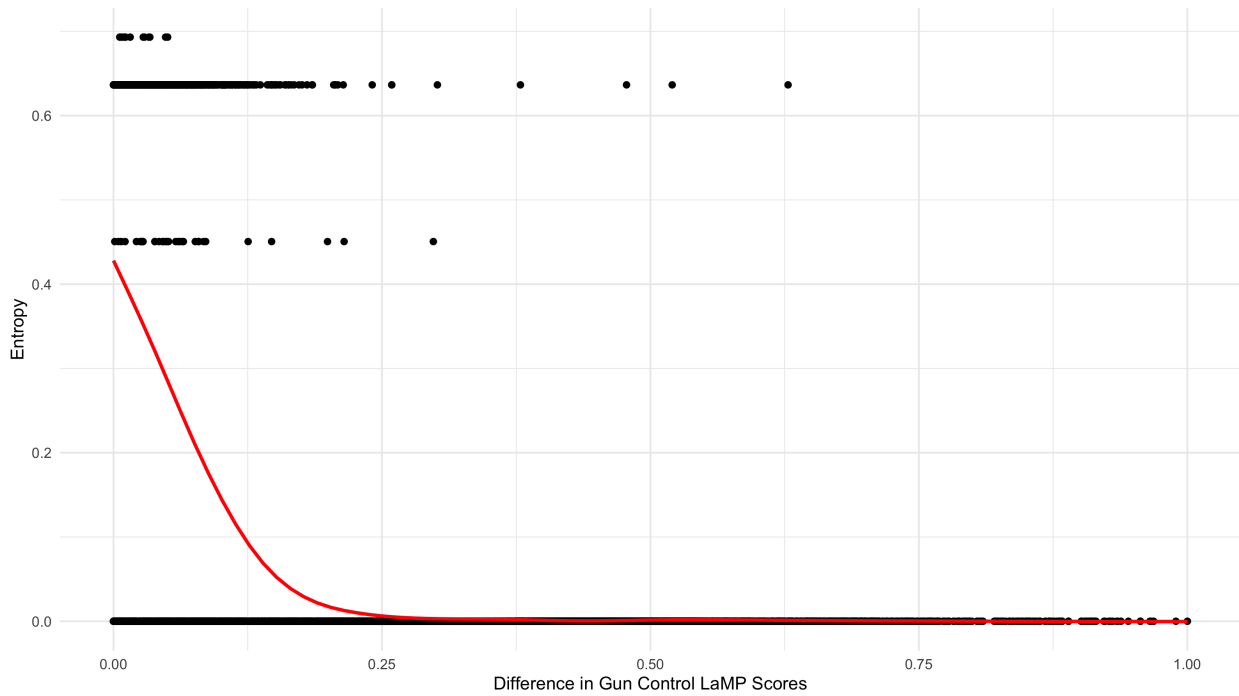


Figure 11: Entropy with respect to the difference in Gun Control LaMP scores for each pairwise comparison between senators. The red line is a generalized additive model line of best fit.

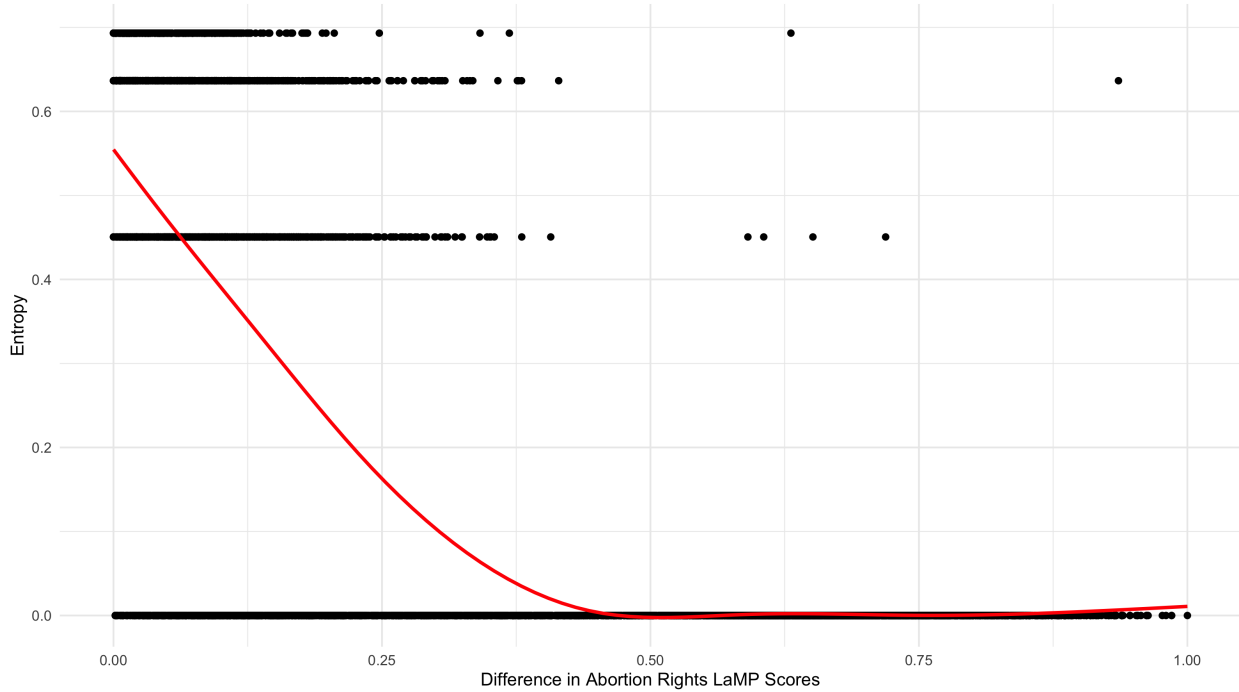


Figure 12: Entropy with respect to the difference in Abortion Rights LaMP scores for each pairwise comparison between senators. The red line is a generalized additive model line of best fit.

All three plots largely follow the same pattern: on average, the entropy falls as the difference in LaMP scores increases. The rate of decrease in entropy for Abortion Rights LaMP scores is lower than the rate of decrease in entropy for Ideology and Gun Control LaMP scores, indicating that ChatGPT has relative higher uncertainty, overall, about pairwise comparisons of senators along the issue of abortion rights.

Comparing Ideology LaMP Scores to the Second Dimension of DW-NOMINATE

We also compare Ideology LaMP scores with the second dimension of DW-NOMINATE, which is typically interpreted as politicians' stance over cross-cutting social issues and sectional division [6, 16]. In other words, the second dimension is interpreted as the stance of the major political parties over slavery, currency, nativism, civil rights, and lifestyle issues in various parts of American history. In recent Congresses, it has been interpreted as a measure capturing the anti-establishment and pro-establishment split [56].

The correlation among all senators between Ideology LaMP scores and the second di-

mension of DW-NOMINATE is 0.388. This is approximately the same as the correlation between perceived ideology scores and the second dimension of DW-NOMINATE (0.412). It is also higher than the correlation between CFScores (0.317) and the first dimension of DW-NOMINATE (0.317) with the second dimension of DW-NOMINATE.

The correlation between LaMP scores and the second dimension of DW-NOMINATE is greater among Democratic senators (0.583) compared to Republican senators (0.027). If we interpret the second dimension of DW-NOMINATE to currently measure antiestablishmentarianism [56], it suggests that ChatGPT’s responses correlate with the stances of Democratic senators towards the current party and government structures. On the other hand, ChatGPT’s responses do not seem to correlate with anti-establishment or pro-establishment views of Republican senators.

Extracting the Name of the More Conservative Senator in Each Matchup

For each pairwise comparison prompt, ChatGPT typically returns a small paragraph explaining its choice rather than returning only the name of the senator. For example, in a comparison between Ed Markey and Cory Booker, ChatGPT returned this answer:

Based on past voting records and statements, Ed Markey (D-MA) has generally been considered more liberal than Cory Booker (D-NJ).

Ed Markey is known for his progressive stances on various issues such as climate change, renewable energy, and healthcare. He co-authored the Green New Deal resolution, a comprehensive plan to address climate change and create jobs. Markey has also been a vocal advocate for net neutrality, consumer protection, and stricter gun control laws. His voting record reflects consistent support for liberal causes and aligns with the Democratic Party’s progressive wing.

Cory Booker, on the other hand, is often seen as a moderate within the Democratic Party. While he has supported liberal policies and causes, such as criminal justice reform and affordable housing, his stance has occasionally been more moderate than that of Ed Markey. Booker has emphasized issues related to economic inequality, racial justice, and expanding access to education. However, his positions on certain issues, such as healthcare, have been more centrist compared to some other progressive senators. It’s important to note that political stances can evolve over time, and individual senators may take different positions on

different issues. Therefore, it’s always a good idea to refer to the most recent information and statements from the senators themselves to get the most accurate understanding of their current positions.

We need to extract the names of the more liberal senators for Democratic senator matchups and Democratic-Republican senator matchups and the more conservative senators for Republican senator matchups. To do this, we ask ChatGPT to extract the name. Specifically, we concatenate the above answer with the following prompt:

In the above Text, who is described to be the more liberal, more progressive, or less conservative senator: [senator 1] or [senator 2]? Return only the full name without party affiliation or state information. If one senator is described as more conservative, return the other senator’s name. If one senator is described as more moderate, return the other senator’s name. If neither senators are described to be more liberal, more progressive, less conservative, more conservative, or more moderate, reply with “Tie.”

For matchups where we prompt ChatGPT to return the name of the more conservative senator, we concatenate that answer with the following text:

In the above Text, who is described to be the more conservative or less liberal senator: [senator 1] or [senator 2]? Return only the full name without party affiliation or state information. If one senator is described as more liberal, return the other senator’s name. If one senator is described as more moderate, return the other senator’s name. If neither senators are described to be more conservative, less liberal, more liberal, or more moderate, reply with “Tie.”

We then prompt ChatGPT with the concatenated text. ChatGPT usually returns the full name of the more liberal or more conservative senator. Punctuation and titles (such as appending “Senator” to the beginning of the name) were automatically removed using a Python function. Ties and answers that deviate from names are manually fixed. We also manually reviewed a sample of the answers that were given. There were occasional mistakes in the names extracted from the answers, but there did not seem to be a pattern in the mistakes. We also found that these mistakes were not repeated in repeated iterations of matchups.

Extracting the Name of the Senator More Likely to Support Gun Control in Each Matchup

Again, for each pairwise comparison prompt, ChatGPT typically returns a small paragraph explaining its choice rather than returning only the name of the senator. We concatenate the model’s answer with the following prompt:

In the above Text, which senator is described to be more likely to support gun control: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])? If one senator is described as being less likely to support gun control, return the name of the other senator. If one senator is described as more likely to support gun rights, return the name of the other senator. If neither senator is described to be more likely to support gun control, neither senator is described to be less likely to support gun rights, neither senator is less likely to support gun control, or neither senator is more likely to support gun rights, reply with “Tie.” Return only the full name without party affiliation or state information. Ignore any language about viewpoints changing.

We then prompt ChatGPT with the concatenated text. Again, ties and answers that deviated from the names in the pairwise comparisons are manually fixed.

Extracting the Name of the More Pro-Choice Senator in Each Matchup

We concatenate the model’s answers with the following prompt for comparisons between Democratic senators or comparisons between a Democratic senator and a Republican senator in order to obtain the name of the senator who is more pro-choice in each matchup:

In the above Text, which senator is described to be more pro-choice: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])? If one senator is described to be less pro-choice, return the name of the other senator. If one senator is described to be more pro-life, return the name of the other senator. Ignore any language about viewpoints changing. Return only the full name without party affiliation or state information. If both senators are described to be equally pro-choice, reply with “Tie.”

For matchups between Republican senators, the following prompt is used to extract the name of the senator who was more pro-life from the model’s output:

In the above Text, which senator is described to be more pro-life: [senator 1] ([senator 1 party abbrev]-[senator 1 state abbrev]) or [senator 2] ([senator 2 party abbrev]-[senator 2 state abbrev])? If one senator is described to be less pro-life, return the name of the other senator. If one senator is described to be more pro-choice, return the name of the other senator. Ignore any language about viewpoints changing. Return only the full name without party affiliation or state information. If both senators are described to be equally pro-life, reply with "Tie."