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

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

Abstract

We propose a novel framework that can be used both to address challenges in measuring the latent positions of politicians and to assess and benchmark the capabilities of generative large language models (LLMs) along specific political dimensions. We prompt a generative 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 rights. 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 the internet and digitized media rather than memorizing existing measures. Our gun control and abortion rights 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. We compare results across multiple LLMs and demonstrate that the pairwise comparison method can also be used to evaluate the capabilities of LLMs.

1 Introduction

This paper outlines a novel framework to address challenges in measuring the latent positions of lawmakers along specified dimensions using generative large language models (LLMs). The framework can also be used to compare and benchmark the capabilities of generative LLMs along these dimensions.

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 can be used to potentially assess core democratic functions, such as how well lawmakers represent their constituents (see, e.g., Ansolabehere et al., 2001; Gerber and Lewis, 2004; Bartels, 2009; Thomsen, 2017; Caughey and Warshaw, 2018) and how position-taking occurs outside of roll call voting (see, e.g., Highton and Rocca, 2005; Boudreau et al., 2019; Russell, 2021).

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. Behavior-based estimates commonly use roll call votes to measure revealed preferences constrained by the legislative agenda (Poole and Rosenthal, 1985; Poole, 2005; Carroll et al., 2009). Other behaviors, such as news media sharing (Eady et al., 2020), can also be used to estimate ideology. Liberal-conservative measures based on campaign contributions (Bonica, 2014) assume ideological homophily in donations 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 reveal lawmakers' positions on specific issues. Positions on issues like gun control and abortion are difficult to measure using existing scaling approaches due to an absence of relevant data. For example, roll call votes cannot be used to measure stances on gun control because most sessions of Congress lack votes on this issue.

Generative LLMs are trained on massive corpora of internet and digitized media text, embedding 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 a generative LLM to compare politicians on a relevant dimension. Specifically, we use GPT-3.5 (Brown et al., 2020), Llama 2 13B, and Llama 2 7B (Touvron et al., 2023) to pairwise compare the

senators of the 116th U.S. Congress along three dimensions: liberal-conservative ideology, support of gun control, and support of abortion rights. We then use the Bradley-Terry model (Bradley and Terry, 1952) to estimate a unidimensional scale measuring latent political positions, which we call **La**nguage **M**odel **P**airwise comparison (LaMP) scores.

We also demonstrate how the framework can be used to compare, evaluate, and better understand the capabilities of LLMs in handling complex and often contentious concepts such as ideology.

Liberal-conservative ideology 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. It also provides us with a set of measures to benchmark and compare the capabilities of LLMs against. The gun control and abortion rights scales derived using the same approach have not been estimated in the literature because of a lack of data on the senators' behaviors and perceptions on these issue areas.

In summary, we find that LLMs, when prompted with pairwise comparisons, can be used to estimate novel scales that cannot be estimated using existing measurement methods. We also observe differences in LaMP scores across different LLMs. For example, we find that LaMP scores estimated using our largest model, GPT-3.5[1], align more closely with human perceptions of ideology. Our pairwise scaling framework provides a more comprehensive understanding of legislative behaviors and policy preferences and offers a new approach to evaluating and benchmarking the capabilities of LLMs.

2 Related Work

Our method is situated in a rapidly growing literature on using generative LLMs for social science applications. These works have studied how generative LLMs can be used for labeling purposes (Törnberg, 2023; Gilardi et al., 2023), analyzing text along psychological constructs (Rathje et al., 2023), reducing the divisiveness of online conversations (Argyle et al., 2023), and generating artificially politically extreme responses (Bisbee et al., 2023). Most of these works focus on generating answers about one item at a time and studying how the LLM's answers differ across items. On the other hand, our method examines how the LLM compares pairs of items and what novel continuous measures can be derived using the LLM's answers.

Pairwise comparisons have been extensively used in the social sciences, especially when measuring complex concepts such as congressional grandstanding (see, e.g., Loewen et al., 2012; Carlson and Montgomery, 2017; Benoit et al., 2019; Park, 2021). Pairwise comparisons have also been used to improve annotations of data (see, e.g., Bruyne et al., 2021; Narimanzadeh et al., 2023). In these papers, pairwise comparisons are made using human annotators.

Recent works have also explored prompting LLMs with pairwise comparisons. Qin et al. (2024) propose using LLMs to rank documents through an approach they call pairwise ranking prompting. Similarly, works such as Gao et al. (2023), Li et al. (2023), Chen et al. (2023), and Liusie et al. (2024), use pairwise comparisons with LLMs for natural language generation and summarization assessment. Unlike our framework, these approaches focus on ranking texts rather than estimating latent continuous scores. They also evaluate the quality of generated texts rather than how the LLM handles complex and contentious topics.

Our approach to scaling also speaks to a vast body of work on ideological scaling and ideal point estimation (see, e.g., Poole and Rosenthal, 1985, 1997; Heckman and Snyder, 1997; Martin and Quinn, 2002; Clinton et al., 2004; Slapin and Proksch, 2008; Carroll et al., 2009; Shor and McCarty, 2011; Lowe and Benoit, 2013; Bonica, 2014; Barberá, 2015; Temporão et al., 2018; Wu et al., 2019; Rheault and Cochrane, 2020; Eady et al., 2020; Hopkins and Noel, 2022; Duck-Mayr and Montgomery, 2023). 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 estimate scales of these latent dimensions of politics and policy.

O'Hagan and Schein (2024) have concurrent work using LLMs to estimate senators' ideologies. Instead of pairwise comparisons, they use individual prompts to score each senator's ideology. This

[1]Although the exact number of parameters is unknown, GPT-3.5 is based on a 175B parameter model.

approach requires contending with issues related to scaling, anchoring, and ordering of the prompts. The next section details the advantages of pairwise comparison prompts over individual prompts.

3 The Pairwise Comparison Framework

We propose a framework for placing lawmakers along specific latent dimensions. The framework is conceptually simple and flexible for a wide range of applications. We use prompts with a generative LLM to make pairwise comparisons of all pairs of lawmakers along a specific issue (e.g., who is more supportive of gun control). The outcomes of these pairwise comparisons are then scaled using the Bradley-Terry model (Bradley and Terry, 1952).

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." Defining $\alpha_i \equiv \exp(\lambda_i)$, the log-odds of i beating j is $\log\left[\frac{\Pr(i \text{ beats } j)}{\Pr(j \text{ beats } i)}\right] = \lambda_i - \lambda_j$. Intuitively, the larger the value of λ_i relative to λ_j, the more likely it is that player i will beat player j. We translate this into a contest between two lawmakers regarding who is more conservative, more likely to support gun control, etc.

More specifically, our framework is as follows.

1. Let $x_{i,j}$ be the pairwise comparison prompt between lawmakers i and j along a specific dimension (ideology, gun control, etc.)
2. Generate a token sequence $s_{i,j}$ using an LLM with parameters θ by $s_{i,j} \sim p_\theta\left(s|x_{i,j}\right)$
3. Use a deterministic mapping function $g\left(s_{i,j}\right)$ to extract from $s_{i,j}$ whether lawmaker i or j is, for example, more conservative
4. After calculating $g\left(s\right)$ for all pairs of lawmakers, scale the responses using the Bradley-Terry model

We use a deterministic mapping function to extract the LLM's answer in each response. We find that extracting the answer, rather than limiting the LLM to returning only the lawmaker's name with no other text, reduces issues such as ordering effects (i.e., which senator is named first). Previous works have also found that prompting LLMs to be concise can lower accuracy (Deng et al., 2024). In our applications, we use GPT-3.5 as the deterministic mapping function to extract the name of the lawmaker from each $s_{i,j}$. The Appendix details the prompts used. The Bradley-Terry model estimates are also rescaled to the unit interval, removing their dependence on a reference category.

Pairwise comparisons offer many advantages. First, we can adjust the prompt's wording to control what concept we are measuring. Second, they simplify the process by presenting only one task per prompt. Prompting LLMs to rank multiple items often yields incomplete rankings. Third, pairwise comparisons offer better interpretability. Another approach would be to directly prompt the LLM to return a number on a scale (see, e.g., O'Hagan and Schein, 2024). However, there are anchoring issues with the scores generated. Although we can keep all individual scoring prompts in the same "conversation" (context) to obtain relative scores, there are ordering biases with the prompts using this approach. Lastly, pairwise comparison outcomes can be compared across different LLMs.

3.1 Uncertainty

We quantify both the uncertainty of the Bradley-Terry model estimates and the pairwise comparisons themselves. For inference of Bradley-Terry model estimates, we use quasi-standard errors. Quasi-standard errors are reference-free and are comparable across all items (Firth and De Menezes, 2004). The bars shown in the graphs throughout the paper are 95% confidence intervals based on these quasi-standard errors.

Building on previous work from Kuhn et al. (2023) and Scherrer et al. (2023), we develop an entropy-based metric to measure the uncertainty in the LLM's responses for each pairwise comparison. This is used to analyze if answers are more consistent as the pairwise comparisons are more "obvious" (e.g., comparing a very conservative Republican senator with a very liberal Democratic senator). Let $r_{i,j,i}$ be the event where lawmaker i is chosen to be more conservative than lawmaker j in a pairwise comparison, and let $x_{i,j}$ be the matchup prompt between lawmakers i and j. Then, the likelihood that the LLM p_θ picks lawmaker i over j is $p_\theta(r_{i,j,i}|x_{i,j}) = \sum_{s \in c(r_{i,j,i},x_{i,j})} p_\theta\left(s|x_{i,j}\right)$, where $c\left(r_{i,j,i}, x_{i,j}\right)$ is the set of all token sequences that are semantically equivalent in encoding i as more conservative than j. The entropy function is then defined as $H_\theta\left[R_{i,j}|x_{i,j}\right] = \sum_{k \in \{i,j\}} -p_\theta\left(r_{i,j,k}|x_{i,j}\right) \log\left(p_\theta\left(r_{i,j,k}|x_{i,j}\right)\right)$.

However, this entropy function is intractable, as we cannot calculate the set of all token sequences that are semantically equivalent in encoding, for example, i as more conservative than j. Instead,

we sample M token sequences $\{s_{i,j,1}, ..., s_{i,j,M}\}$ from the LLM by $s_{i,j,k} \sim p_\theta(s|x_{i,j})$. We then use the deterministic mapping function $g(s_{i,j})$ to determine if lawmaker i or j is more conservative. The likelihood is estimated using $\hat{p}_\theta(r_{i,j,i}|x_{i,j}) = \frac{1}{M}\sum_{k=1}^{M} \mathbb{1}[g(s_{i,j,k}) = r_{i,j,i}]$. We can then plug in this estimated likelihood to calculate the entropy.

4 Estimating the Latent Positions of Senators of the 116th U.S. Congress

We use GPT-3.5 and Llama 2 to make pairwise comparisons about liberal-conservative ideology, gun control, and abortion for the senators of the 116th U.S. Congress, which was the Congress in session from 2019 to 2021. We use the liberal-conservative ideology scale to better understand the strengths of LaMP scores and compare the capabilities of the LLMs. The gun control scale and abortion rights scale demonstrate how we can use the approach to estimate novel issue-specific scales.

There are 5,151 pairwise comparisons across all senators in the 116th Congress for each scale. To estimate the entropy in pairwise comparisons, we repeated the set of pairwise comparisons three times for each measure. Separately, to analyze the consistency in scores, we estimated the LaMP scores for each set of comparisons and analyzed their correlations. The correlations are reported in the Appendix. The LaMP scores in the following results use the outcomes of pairwise comparisons from all three iterations. We used default hyperparameters (e.g., temperature) for all LLMs.

4.1 Ideology LaMP Scores

The LLM selected the more conservative senator in each pairwise comparison. This intuitively places more conservative senators on the right side of the scale and more liberal senators on the left side. The scores are called Ideology LaMP scores. We used the following 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])?` The Appendix contains further details about the prompts used.

4.1.1 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 (Poole and Rosenthal, 1985, 1997; Poole, 2005; Carroll et al., 2009). It is the most widely used measure of legislator ideology (Caughey and Schickler, 2016). The first dimension of DW-NOMINATE is interpreted as the liberal-conservative continuum in United States politics (Poole and Rosenthal, 1997). Figure 1 compares the first dimension of DW-NOMINATE against Ideology LaMP scores estimated using GPT-3.5.

Table 1 reports the correlations between Ideology LaMP scores estimated using different LLMs and the first dimension of DW-NOMINATE. While correlations are high across the LLMs, the correlations, especially when looking at the parties individually, fall as the model size decreases.

Model	All	Dems	GOP
GPT-3.5	0.97	0.84	0.65
Llama 2 13B	0.96	0.77	0.64
Llama 2 7B	0.91	0.60	0.25

Table 1: Correlations between Ideology LaMP scores estimated using different LLMs and the first dimension of DW-NOMINATE.

4.1.2 Ideology LaMP scores do not simply parrot DW-NOMINATE

Figure 1 illustrates interesting patterns in GPT-3.5's Ideology LaMP scores. Notably, our method estimates Joe Manchin to be more conservative than Susan Collins. This placement intuitively makes sense: for example, 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 GPT-3.5 is not merely recalling DW-NOMINATE. DW-NOMINATE ranks Elizabeth Warren and Kamala Harris as the most liberal senators, whereas Ideology LaMP scores identify Bernie Sanders and Warren as the most liberal. This aligns with political activists' views, who also named Sanders and Warren as the most liberal senators (Hopkins and Noel, 2022). Sanders' placement towards the center in DW-NOMINATE stems from his occasional votes against the Democratic Party (Duck-Mayr and Montgomery, 2023), while LaMP scores likely reflect his left-leaning positions as highlighted in roll call votes, news articles, and statements.

When comparing the ordinal rankings of DW-

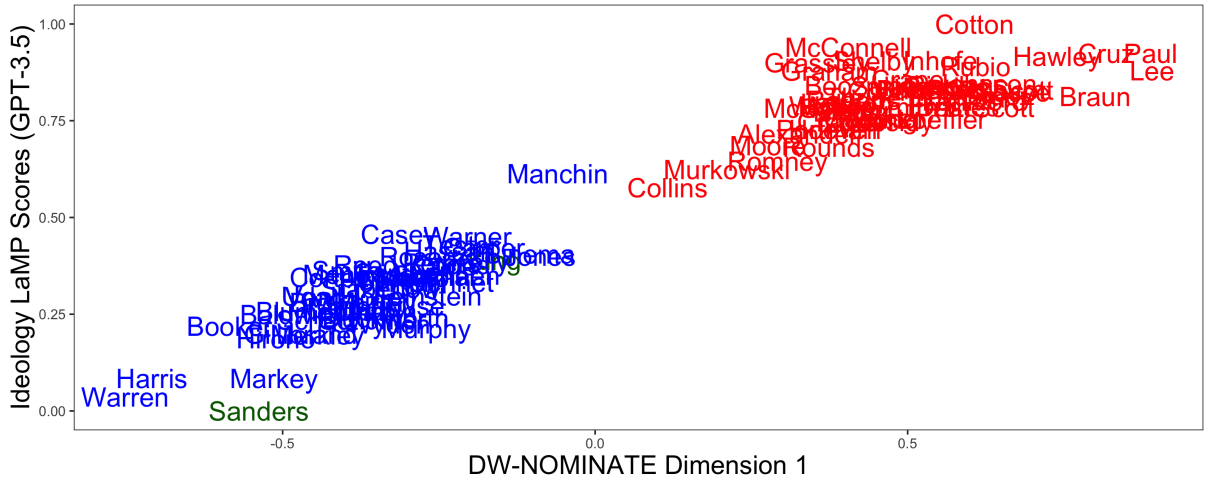

Figure 1: First Dimension of DW-NOMINATE vs. Ideology LaMP scores estimated using GPT-3.5. Democratic senators are in blue, Republican senators are in red, and Independent senators are in green.

NOMINATE and Ideology LaMP scores, senators differed, on average, by 8.31 positions. Nine of the 10 senators with the largest differences in ordinal rankings are Republicans. These differences appear to be shaped by their public stances regarding Donald Trump. For example, Lindsey Graham, who strongly supports Trump, is ranked as the 45th most conservative Republican by DW-NOMINATE but 11th by LaMP scores. Again, DW-NOMINATE would not capture these public stances.

The Ideology LaMP scores estimated with Llama models lack some of the face validity seen with GPT-3.5. For instance, Llama 2 7B places Sanders towards the center, and Llama 2 13B ranks centrist Republican Lisa Murkowski in the middle of Republicans. In other words, scores estimated using smaller models miss nuances captured by our largest model. The Appendix contains plots of the Llama-based Ideology LaMP scores.

4.1.3 LaMP scores also highly correlate with other measures of ideology

We compare Ideology LaMP scores with two alternative measures of ideology from the political science literature that are based on the perceptions of the senators. The first is perceived ideology scores (Hopkins and Noel, 2022), which are estimated using the Bradley-Terry model with political activists' answers to pairwise comparisons of senators. These scores reflect how these activists perceive politicians, which can differ from how politicians view themselves ideologically.

The second is Campaign Finance Scores (Bonica, 2014), or CFscores, which 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. Put another way, it measures ideology based on the donors' perceptions of lawmakers.

Table 2 shows correlations across the three LLMs and across all senators, only Democrats, and only Republicans. Ideology LaMP scores estimated using GPT-3.5 generally have higher correlations with these two measures of ideology compared to the Llama-based scores. Ideology LaMP scores estimated using Llama 2 13B have higher correlations with CFscores compared with the other two LLMs. It suggests that Ideology LaMP scores estimated using different LLMs may highlight different aspects of ideology.

Party	Model	Perc. Ideo.	CFscores
All	GPT-3.5	0.94	0.93
All	Llama 2 13B	0.90	0.94
All	Llama 2 7B	0.87	0.91
Dems	GPT-3.5	0.81	0.39
Dems	Llama 2 13B	0.68	0.51
Dems	Llama 2 7B	0.61	0.43
GOP	GPT-3.5	0.79	0.20
GOP	Llama 2 13B	0.50	0.25
GOP	Llama 2 7B	0.35	0.09

Table 2: Correlations between Ideology LaMP scores estimated using different LLMs and two perceptions-based measures of ideology: perceived ideology scores (perc. ideo.) and CFscores.

We also calculated partial correlations between Ideology LaMP scores and DW-NOMINATE, perceived ideology scores, and CFscores. We calculated each partial correlation controlling for the other two measures of ideology. These results are in the Appendix.

Across the three LLMs, correlations and partial correlations suggest that no single existing 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.

4.1.4 Ideology LaMP scores better predict human evaluations of ideology

We use linear regression analyses to compare the effectiveness of Ideology LaMP scores and DW-NOMINATE in capturing human perceptions of political ideology. We calculate the proportion of variance explained (R^2) in perceived ideology scores using a full model with both DW-NOMINATE and Ideology LaMP scores as predictors and two reduced models, each with only one of the predictors. Comparing the R^2 values of the full and reduced models reveals the explanatory power lost when omitting either predictor.

When we exclude Ideology LaMP scores estimated using GPT-3.5 from the full model, R^2 drops 42% when the data is limited to Republican senators, 23% when limited to Democratic senators, and 7% when including all senators. In contrast, dropping DW-NOMINATE only reduces R^2 by 3% for Republican senators, 0% for Democratic senators, and 0% for all senators. Partial F-tests support these findings: across all senators, Democratic senators, and Republican senators, partial F-tests indicate significant differences ($p < .0001$ for all p-values) when comparing the full model to the reduced model with only DW-NOMINATE, but no significant differences ($p > .16$ for all p-values) when comparing the full model to the reduced model with only Ideology LaMP scores. These results indicate that Ideology LaMP scores are a stronger predictor of human evaluations of ideology than DW-NOMINATE.

However, DW-NOMINATE better predicts human evaluations of ideology compared to the Llama-based Ideology LaMP scores. This indicates that the larger LLM captures nuances of ideology that align more closely with human evaluations.

4.1.5 Smaller LLMs have higher average entropy

Figure 2 plots the entropy of each pairwise comparison using the three LLMs against the difference in Ideology LaMP scores within each pairwise comparison. We find that entropy decreases for all LLMs as the difference in Ideology LaMP scores increases within each pairwise comparison. This indicates greater consistency in responses as the senators are farther apart on the ideology scale. The smoothed line is a generalized additive model with a cubic regression spline. We find that as the size of the language model decreases, the average entropy increases and the rate at which entropy decreases becomes slower.

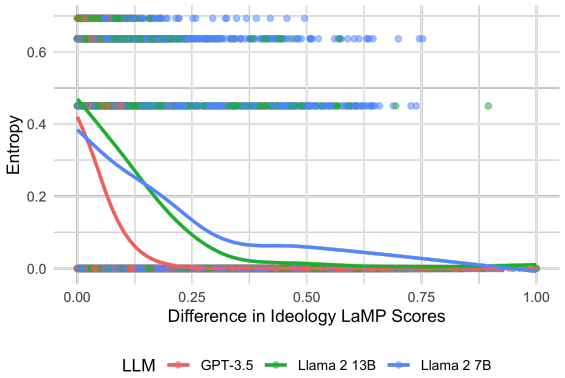

Figure 2: Entropy of ideology pairwise comparisons across the three LLMs.

4.2 Gun Control LaMP Scores

In each pairwise comparison, the LLM selected the senator more supportive of gun control. The prompt used was the following: `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])?`.

4.2.1 Gun Control LaMP scores differ from Ideology LaMP scores

The Gun Control LaMP scores of all senators estimated using GPT-3.5 are illustrated in Figure 3. All versions of Gun Control LaMP scores, whether estimated using GPT-3.5, Llama 2 7B, or Llama 2 13B, capture the partisan divide on this issue: all Republicans are placed on the left side of the scale and all Democrats are placed on the right side. Llama-based scores are illustrated in the Appendix.

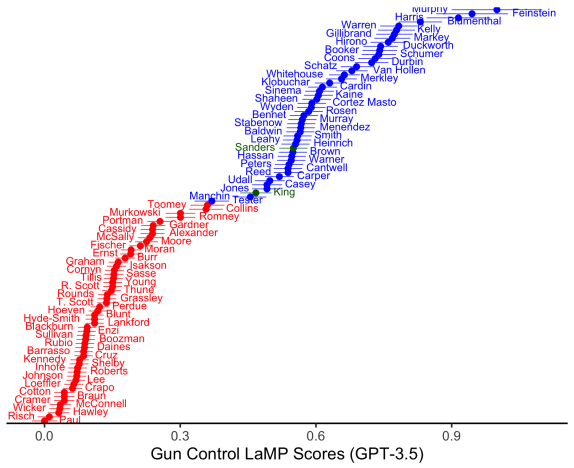

Figure 3: Gun Control LaMP scores of senators estimated using GPT-3.5 with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

Comparing Gun Control LaMP scores with Ideology LaMP scores further suggests the face validity of the former. For example, Ideology LaMP scores place Mark Kelly as a centrist Democrat, 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 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. Sanders often treads carefully on the issue of gun control, reflecting his support of Vermont hunting traditions. Pat Toomey, placed in the middle of the Republicans based on Ideology LaMP scores, is placed as the Republican most supportive of gun control. Toomey has supported background checks and state red flag laws. These patterns held true across all three LLMs used to estimate the Gun Control LaMP scores.

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

To evaluate the scale's external validity, we predict Republican votes on the 2022 Bipartisan Safer Communities Act, which was passed in June 2022. This falls outside GPT-3.5's pretraining cutoff of September 2021; however, it is within Llama 2's pretraining, which has a cutoff of September 2022.

15 Republican senators voted alongside all Democratic senators; we excluded Democratic senators' votes. Table 3 shows the coefficients of a logistic regression predicting Republican votes on the bill using Gun Control LaMP scores estimated using GPT-3.5, DW-NOMINATE, and each senator's last National Rifle Association (NRA) grade up until 2020. The NRA assigns grades to each senator based on their votes on gun control bills and their expressed gun control opinions.[2] The Gun Control LaMP scores are the only statistically significant predictor of Republican votes on the bill. However, Llama-based Gun Control LaMP scores are not statistically significant predictors using the same logistic regression specification.

	Voted Yea
Gun Control LaMP scores	19.6^{**}
DW-NOMINATE	-12.0
NRA grades	-0.03
Constant	5.0

Table 3: Logistic regression predicting Republican votes ($n = 45$) on the 2022 Bipartisan Safer Communities Act using the Gun Control LaMP scores estimated using GPT-3.5, the first dimension of DW-NOMINATE, and NRA grades as predictor variables. ** indicates $p < .01$; all other p-values $> .05$.

4.2.3 Entropy of Gun Control LaMP Scores

Figure 4 plots the entropy of each pairwise comparison using the three LLMs against the difference in Gun Control LaMP scores within each pairwise comparison. We find that as the difference in Gun Control LaMP scores increases within each pairwise comparison, entropy decreases for all LLMs, indicating greater consistency in responses. Again, the smoothed line is a GAM with a cubic regression spline. We again find that as the size of the LLM decreases, the average entropy increases and the rate at which entropy decreases becomes slower.

4.3 Abortion Rights LaMP Scores

To estimate the Abortion Rights LaMP scores, we identified the more pro-choice (or pro-life) senator in each matchup using an LLM. The validation outcomes largely follow the same patterns of validation for the Gun Control LaMP scores. Complete details of the Abortion Rights LaMP scores validation can be found in the Appendix.

The Abortion Rights LaMP scores reflect the partisan divide on the issue. Scores estimated us-

[2] Additional analyses on how well Gun Control LaMP scores predict NRA grades are in the Appendix.

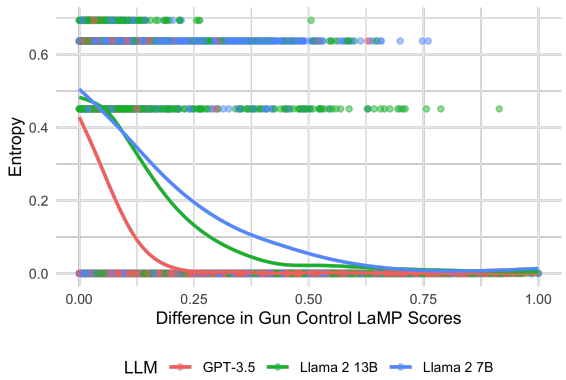

Figure 4: Entropy of gun control pairwise comparisons across the three LLMs.

ing GPT-3.5 identify some overlap between the parties across pro-choice Republicans and pro-life Democrats. However, the smaller Llama models lack the nuance of the larger GPT-3.5 model in estimating latent positions, either misidentifying senators' stances or perfectly splitting the parties.

No standalone legislation on abortion was passed in the time period after the pretraining data for our LLM models of interest. We do find that the GPT-3.5-based Abortion Rights LaMP scores are a better predictor of NARAL (a pro-choice interest group) grades than DW-NOMINATE, but this does not hold true for Llama-based scores.

5 Discussion and Conclusion

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, across three LLMs of different sizes, the LLMs are not hallucinating in pairwise comparisons and the LLMs are not simply parroting existing scales such as DW-NOMINATE or interest group ratings. In other words, pairwise comparisons with LLMs, even of different sizes, yield sensible scales of lawmakers along ideology and specific issues. Our evidence is consistent with the idea that LLMs synthesize a great deal of information about lawmakers to evaluate latent constructs in predictable and sensible ways, agreeing with existing scales and predicting lawmaker behaviors such as votes.

We find that LaMP scores can provide a clearer understanding of lawmakers' ideologies, especially in cases where traditional measures struggle with behaviors that violate the assumptions of the underlying measurement model, such as when lawmakers vote against their own party for ideological reasons. It also means 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.

At the same time, our approach can also be used to compare and benchmark LLMs. Pairwise comparisons force the LLM to evaluate specific attributes or preferences between two options rather than recalling a memorized label from its training data. This approach also enables more fine-grained distinctions, helping us better understand the subtleties in how the model evaluates items, especially on complex or contentious issues. Across ideology, gun control, and abortion rights, we find that the largest LLM we evaluate, GPT-3.5, yields estimates that better correlate with existing measures of ideology, better predict human evaluations of ideology, better predict out-of-sample votes on legislation, have better face validity, and are more consistent in repeated iterations compared with Llama 2 7B and Llama 2 13B.

Our current application is to American politics at the federal level; future work will extend the method to settings outside of the U.S., such as estimating the ideology of legislators in parliamentary systems where members of parliament vote strictly along party lines. However, a challenge arises when we shift away from the very populous and media-rich United States, which is also the home country of the LLMs analyzed in this paper, and also potentially away from the English language: the LLMs may not have enough information about every member of parliament. In such cases, a solution could be to pairwise compare text produced by individual legislators, such as campaign materials or tweets. Future work will also look at how our approach can be used to automatically evaluate whether the LLM's scaling of concepts of interest aligns with human judgments both within and outside the domain of politics, which has potential extensions to alignment research. In short, our pairwise comparison framework holds significant prospective contributions for both social sciences and natural language processing.

Limitations

As discussed in Section 5, our application concerns United States senators. Senators are widely covered and discussed in the media and the political science literature. Consequently, there is extensive information about them and their political positions in the LLMs' training corpora. We have not tried this approach with state or local politicians nor have we tried this approach with politicians outside the United States. In both cases, there may be significantly less information about these politicians in the training data for LLMs. This may result in greater hallucinations in the pairwise comparisons. We have ongoing work that examines how the entropy metric used in this paper can detect higher levels of model uncertainty in pairwise comparisons when there is less information available about the politicians or items of interest.

Relatedly, we make pairwise comparisons in English, the language LLMs are strongest in (Zhang et al., 2023). It is unknown how well pairwise comparisons, both within and outside of the domain of ideology and latent position estimation, would work in settings outside of English.

Another limitation of the paper is that we rely on black box LLMs to assess pairwise comparisons. Consequently, we do not know what sources the LLMs are "relying" on to generate their answers to pairwise comparisons. The responses to the pairwise comparisons always come with an explanation; however, the LLM may use the answer as context to generate the explanation. Recent work by Anthropic on extracting interpretable features from LLMs is likely a fruitful path to better understanding how LLMs respond to pairwise comparisons (Templeton et al., 2024).

We also use the LLMs out-of-the-box: they were not fine-tuned on any additional data. Thus, the responses depend on the type of data the model was pre-trained on and the cutoff date of the data. Possibilities for future work can involve fine-tuning the model or using recent techniques such as retrieval-augmented generation to expand the knowledge base of the LLM (Lewis et al., 2020).

We have also not used Llama 2 70B, one of the most advanced open models available, because of resource constraints. Quantizing the weights of Llama 2 70B yielded poor performance in pairwise comparisons. In future work, we hope to acquire more computational resources that will allow us to use Llama 2 70B.

Ideology and issue positions are not static measures. They can evolve over time. In this paper, we do not specify a time period to make pairwise comparisons. Ongoing work looks at how limiting pairwise comparisons to a specific timeframe can capture the evolution of ideological and issue positions over time.

Lastly, we have not experimented with how the outcomes of pairwise comparisons would change if using different but semantically equivalent prompts. This adds another dimension of uncertainty to the outcomes of pairwise comparisons. In this paper, we developed prompts that would generate meaningful responses from the LLMs. For example, without including the "Based on past voting records and statements" part of the prompt, GPT-3.5 would often not generate responses to pairwise comparisons. We also included state and party information for each senator in the prompts to avoid potential ambiguities with other federal, state, or local politicians with the same names. We will study how closely LaMP scores correlate using different but semantically equivalent pairwise comparison prompts in future work.

References

Stephen Ansolabehere, James M. Snyder, and Charles Stewart. 2001. Candidate positioning in U.S. house elections. *American Journal of Political Science*, 45(1):136–159.

Lisa P. Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023. Ai chat assistants can improve conversations about divisive topics. *Preprint*, arXiv:2302.07268.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Larry M. Bartels. 2009. Economic Inequality and Political Representation. In *The Unsustainable American State*. Oxford University Press.

Kenneth Benoit, Kevin Munger, and Arthur Spirling. 2019. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508.

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. Synthetic replacements for human survey data? the perils of large language models.

Adam Bonica. 2014. Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386.

Adam Bonica. 2016. Database on ideology, money in politics, and elections. Available at https://data.stanford.edu/dime.

Cheryl Boudreau, Christopher S. Elmendorf, and Scott A. MacKenzie. 2019. Racial or spatial voting? the effects of candidate ethnicity and ethnic group endorsements in local elections. *American Journal of Political Science*, 63(1):5–20.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345.

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. 2020. Language models are few-shot learners.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Annotating affective dimensions in user-generated content. *Language Resources and Evaluation*, 55(4):1017–1045.

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

Royce Carroll, Jeffrey B. Lewis, James Lo, Keith T. Poole, and Howard Rosenthal. 2009. Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Political Analysis*, 17(3):261–275.

Devin Caughey and Eric Schickler. 2016. Substance and change in congressional ideology: Nominate and its alternatives. *Studies in American Political Development*, 30(2):128–146.

Devin Caughey and Christopher Warshaw. 2018. Policy preferences and policy change: Dynamic responsiveness in the american states, 1936–2014. *American Political Science Review*, 112(2):249–266.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2024. Rephrase and respond: Let large language models ask better questions for themselves. *Preprint*, arXiv:2311.04205.

JBrandon Duck-Mayr and Jacob Montgomery. 2023. Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons. *Political Analysis*, page 1–20.

Gregory Eady, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler. 2020. News sharing on social media: Mapping the ideology of news media content, citizens, and politicians.

David Firth and Renée X. De Menezes. 2004. Quasi-variances. *Biometrika*, 91(1):65–80.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *Preprint*, arXiv:2304.02554.

Elisabeth R. Gerber and Jeffrey B. Lewis. 2004. Beyond the median: Voter preferences, district heterogeneity, and political representation. *Journal of Political Economy*, 112(6):1364–1383.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

James J. Heckman and James M. Snyder. 1997. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics*, 28:S142–S189.

Benjamin Highton and Michael S. Rocca. 2005. Beyond the roll-call arena: The determinants of position taking in congress. *Political Research Quarterly*, 58(2):303–316.

Daniel J. Hopkins and Hans Noel. 2022. Trump and the shifting meaning of "conservative": Using activists' pairwise comparisons to measure politicians' perceived ideologies. *American Political Science Review*, 116(3):1133–1140.

Seongho Kim. 2015. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6):665–674.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2021. Voteview: Congressional roll-call votes database. https://voteview.com/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. *Preprint*, arXiv:2307.02762.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Peter John Loewen, Daniel Rubenson, and Arthur Spirling. 2012. Testing the power of arguments in referendums: A Bradley–Terry approach. *Electoral Studies*, 31(1):212–221. Special Symposium: Germany's Federal Election September 2009.

Will Lowe and Kenneth Benoit. 2013. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, 21(3):298–313.

Andrew D. Martin and Kevin M. Quinn. 2002. Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953–1999. *Political Analysis*, 10(2):134–153.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G. Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Sean O'Hagan and Aaron Schein. 2024. Measurement in the age of llms: An application to ideological scaling. *Preprint*, arXiv:2312.09203.

Ju Yeon Park. 2021. When do politicians grandstand? measuring message politics in committee hearings. *The Journal of Politics*, 83(1):214–228.

Keith T. Poole. 2005. *Spatial Models of Parliamentary Voting*. Analytical Methods for Social Research. Cambridge University Press.

Keith T. Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.

Keith T. Poole and Howard Rosenthal. 1997. *Ideology and Congress: A Political Economic History of Roll Call Voting*. Yale University Press, New Haven, CT.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. *Preprint*, arXiv:2306.17563.

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

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

Annelise Russell. 2021. Minority opposition and asymmetric parties? Senators' partisan rhetoric on twitter. *Political Research Quarterly*, 74(3):615–627.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms. *Preprint*, arXiv:2307.14324.

Boris Shor and Nolan McCarty. 2011. The ideological mapping of American legislatures. *The American Political Science Review*, 105(3):530–551.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Mickael Temporão, Corentin Vande Kerckhove, Clifton van der Linden, Yannick Dufresne, and Julien M. Hendrickx. 2018. Ideological scaling of social media users: A dynamic lexicon approach. *Political Analysis*, 26(4):457–473.

Danielle M. Thomsen. 2017. *Opting Out of Congress: Partisan Polarization and the Decline of Moderate Candidates*. Cambridge University Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *Preprint*, arXiv:2304.06588.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Additional Notes about the Senators of the 116th U.S. Congress

We obtained the list of senators from Voteview (Lewis et al., 2021). We kept 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.

A.2 Pairwise Comparison Prompts

A.2.1 Ideology LaMP Scores

We inputted the following prompt into the LLMs 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 used 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 changed the wording for matchups between Republican senators strictly because of a quirk of the LLMs (and illustrative of their inability to "reason" about politics): when asked which senator in each pair is more liberal when comparing two conservative Republicans, the LLM would often generate a response that states neither senator is "more liberal" because they are both conservative Republicans.

While we used the "more liberal" prompt when comparing a Democratic senator and a Republican senator, we also ran pairwise comparisons using GPT-3.5 with the "more conservative" prompt instead when comparing a Democratic senator and a Republican senator. We found that the two Ideology LaMP scores correlated at 0.997.

A.2.2 Gun Control LaMP Scores

We did not use different prompts depending on the pairwise comparison type; all pairwise comparisons on gun control used the same prompt as described in the main text.

A.2.3 Abortion Rights LaMP Scores

To analyze the senators of the 116th Congress specifically for support of abortion rights, we used

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 changed the wording for matchups between Republican senators because the LLMs would often not compare two Republican senators on a pro-choice comparison basis. It would usually only make a pairwise comparison when asked which Republican senator is more pro-life.

A.3 The Deterministic Mapping Function

For each pairwise comparison prompt, the LLM returns a paragraph explaining its choice rather than returning only the name of the senator. For example, in a comparison between Ed Markey and Cory Booker on ideology, GPT-3.5 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.

From the above, we need to extract the name of the more liberal (or conservative) senator. The function used to extract the name is what we call the deterministic mapping function, described in step 3 of the pairwise comparison framework detailed in Section 3.

The deterministic mapping function used was a prompt made to GPT-3.5 containing the text and a query on which senator from the given text was more liberal/conservative, more supportive of gun control, or more pro-choice/pro-life.

Punctuation and titles (such as "Senator" at the beginning of the name) were automatically removed using a Python function. Answers that deviate from names or "Tie" were 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 appear to be a pattern in the mistakes. We also found that these mistakes were not repeated in repeated iterations of matchups.

A.3.1 Ideology LaMP Scores

If the pairwise comparison was based on who was more liberal, we extracted the name of the more liberal senator using the following prompt concatenated with the text returned from the LLM:

> `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 the LLM 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."

A.3.2 Gun Control LaMP Scores

We concatenate the LLM'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.

A.3.3 Abortion Rights LaMP Scores

We concatenate the LLM's answer 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."
```

A.4 Correlations in LaMP Scores Across Iterations

We ran the entire set of pairwise comparisons across all senators three times. We looked at the correlations of the LaMP scores estimated using each individual iteration.

A.4.1 Ideology LaMP Scores

Table 4 reports the lowest correlation between any two Ideology LaMP scores estimated using each iteration. The results indicate that the estimated Ideology LaMP scores highly correlate between repeated iterations. Again, we find that the largest LLM has the most consistent scores both within and across parties.

Model	Party	Lowest ρ
GPT-3.5	All	0.997
GPT-3.5	Democratic	0.982
GPT-3.5	Republican	0.972
Llama 2 13B	All	0.988
Llama 2 13B	Democratic	0.922
Llama 2 13B	Republican	0.891
Llama 2 7B	All	0.988
Llama 2 7B	Democratic	0.938
Llama 2 7B	Republican	0.933

Table 4: Lowest Pearson correlation coefficient between any two iterations for Ideology LaMP scores, broken down by model and party.

A.4.2 Gun Control LaMP Scores

Table 5 reports the lowest correlation between any two Gun Control LaMP scores estimated using each iteration. Similarly, the results indicate that the estimated Gun Control LaMP scores highly correlate between repeated iterations. Moreover, we again find that the largest LLM has the most consistent scores both within and across parties.

Model	Party	Lowest ρ
GPT-3.5	All	0.993
GPT-3.5	Democratic	0.970
GPT-3.5	Republican	0.951
Llama 2 13B	All	0.985
Llama 2 13B	Democratic	0.922
Llama 2 13B	Republican	0.826
Llama 2 7B	All	0.979
Llama 2 7B	Democratic	0.917
Llama 2 7B	Republican	0.814

Table 5: Lowest Pearson correlation coefficient between any two iterations for Gun Control LaMP scores, broken down by model and party.

A.4.3 Abortion Rights LaMP Scores

Table 6 reports the lowest correlation between any two Abortion Rights LaMP scores estimated using each iteration. Similar to the previous two measures, the results indicate that the estimated Abortion Rights LaMP scores highly correlate between repeated iterations. Moreover, we again find that the largest LLM has the most consistent scores both within and across parties.

Model	Party	Lowest ρ
GPT-3.5	All	0.996
GPT-3.5	Democratic	0.968
GPT-3.5	Republican	0.952
Llama 2 13B	All	0.990
Llama 2 13B	Democratic	0.938
Llama 2 13B	Republican	0.931
Llama 2 7B	All	0.986
Llama 2 7B	Democratic	0.936
Llama 2 7B	Republican	0.936

Table 6: Lowest Pearson correlation coefficient between any two iterations for Abortion Rights LaMP scores, broken down by model and party.

A.5 Llama-Based Ideology LaMP Scores

Figure 5 shows the Ideology LaMP scores of senators estimated using Llama 2 13B, and Figure 6

shows the Ideology LaMP scores of senators estimated using Llama 2 7B.

A.6 Ideology LaMP Scores: Partial Correlations

Table 7 shows the partial correlations between Ideology LaMP scores and DW-NOMINATE, perceived ideology scores, and CFscores. 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. P-values are calculated using the t-statistic described in Kim (2015).

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. We find that these patterns hold across the three different LLMs, although many of the partial correlations are not significant when using Llama 2 7B.

A.7 Llama-Based Gun Control LaMP Scores

Figure 7 shows the Gun Control LaMP scores of senators estimated using Llama 2 13B, and Figure 8 shows the Gun Control LaMP scores of senators estimated using Llama 2 7B.

A.8 GPT-3.5-Based Gun Control LaMP Scores Better Predict NRA Grades

We compare the predictive power of Gun Control LaMP scores and DW-NOMINATE on NRA grades. The NRA assigns grades using a set of votes on motions, bills, and confirmations that are related to gun rights or gun control in some way. They also use public statements on the issue. We use each senator's most recent NRA grade up until 2020. We calculate the proportion of variance explained (R^2) in NRA grades using a full model with both predictors—Gun Control LaMP scores and DW-NOMINATE—and two reduced models, each with one predictor. Similar to the analysis with Ideology LaMP scores, comparing the R^2 values of the full and reduced models reveals the explanatory power lost when omitting either predictor.

We find that dropping the GPT-3.5-based Gun Control LaMP scores from the full model leads to R^2 dropping 99% when the data is limited to Republican senators, 41% when the data is limited to Democratic senators, and 4% when including all senators. In contrast, dropping DW-NOMINATE reduces R^2 by 29% for Republican senators, 18% for Democratic senators, and 4% for all senators. Partial F-tests do find significant differences when comparing the full model with the reduced models in all cases except when looking at just Democratic senators: the partial F-test p-value when comparing the full model to the reduced model with Gun Control LaMP scores is 0.18, while the p-value when comparing the full model to the reduced model with DW-NOMINATE is 0.04. Although all reduced models are significantly different except when looking at just Democratic senators, the reduction in R^2 is always greater when we drop Gun Control LaMP scores over DW-NOMINATE.

We find similar patterns with the Gun Control LaMP scores estimated using Llama 2 13B: dropping the Llama 2 13B-based Gun Control LaMP scores from the full models leads to R^2 dropping 99% when the data is limited to Republican senators, 38% when the data is limited to Democratic senators, and 5% when including all senators. In contrast, dropping DW-NOMINATE reduces R^2 by 3% for Republican senators, 26% for Democratic senators, and 4% for all senators. The results, however, do not hold if we use Llama 2 7B. For Democratic senators and across all senators, R^2 falls more if we drop DW-NOMINATE from the full model compared to dropping Gun Control LaMP scores, indicating that DW-NOMINATE is more predictive in this case. When looking at just Republican senators, R^2 still falls more if we drop Gun Control LaMP scores: R^2 falls by 99% when dropping Gun Control LaMP scores from the full model compared to R^2 falling by 9% when dropping DW-NOMINATE from the full model. Again, this indicates that smaller LLMs do not capture many of the nuances of lawmakers' issue positions compared to larger LLMs.

A.9 Abortion Rights LaMP Scores

We estimate the Abortion Rights LaMP scores using the prompts described in Section A.2.

A.9.1 Abortion Rights LaMP scores differ from Ideology LaMP scores

The Abortion Rights LaMP scores of all senators estimated using GPT-3.5, Llama 2 13B, and Llama 2 7B are illustrated in Figures 9, 10, and 11, re-

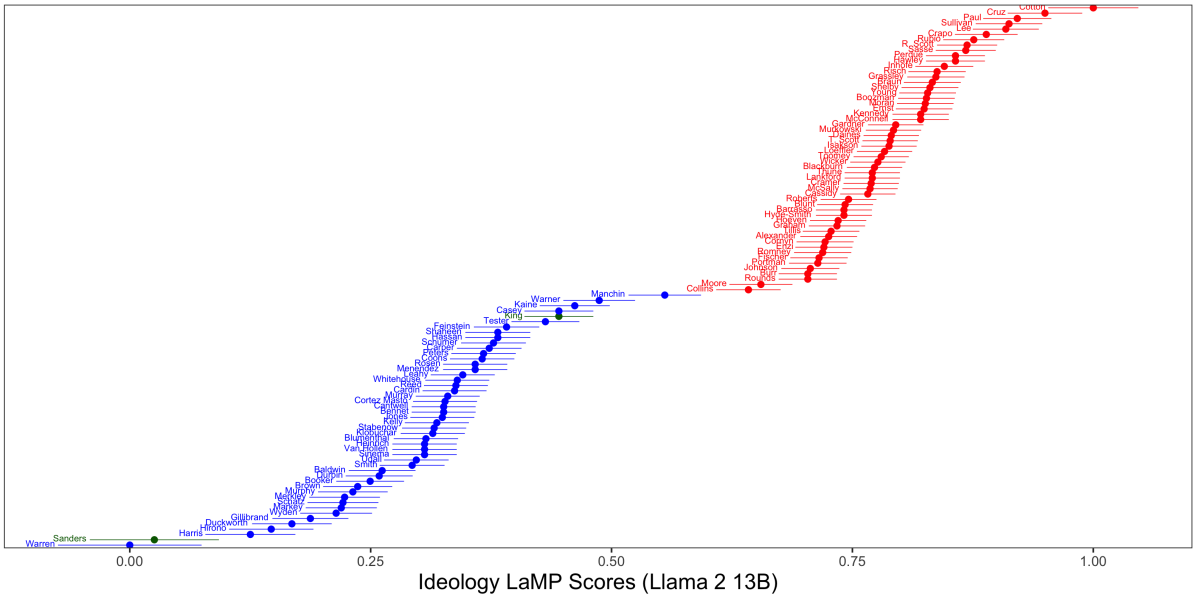

Figure 5: Ideology LaMP scores of senators estimated using Llama 2 13B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

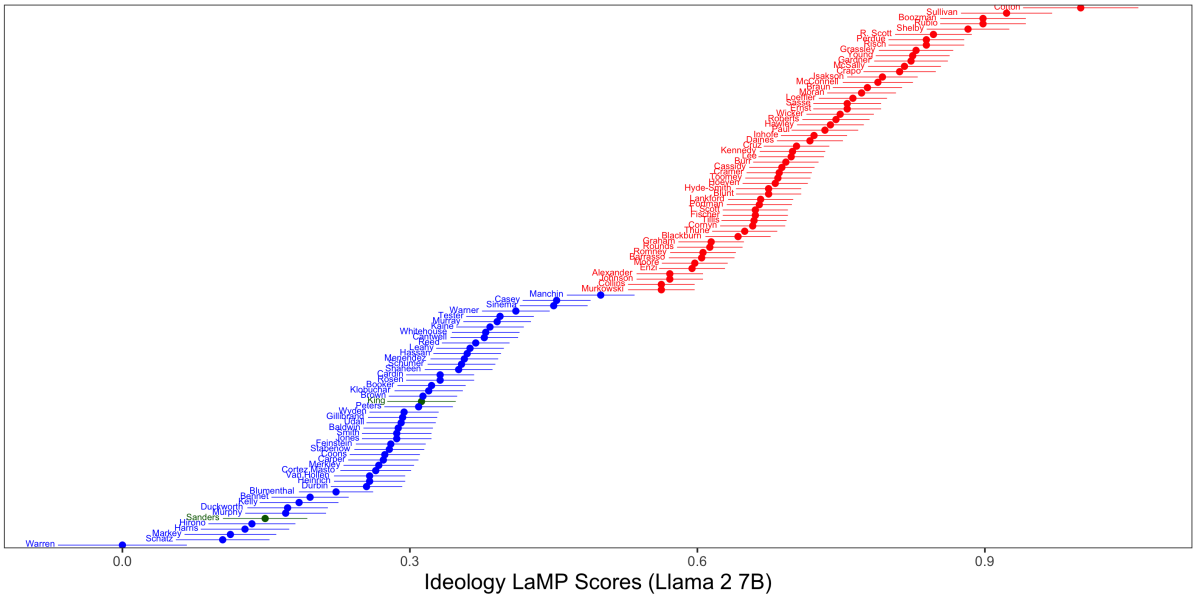

Figure 6: Ideology LaMP scores of senators estimated using Llama 2 7B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

spectively. On this scale, the likelihood of senators supporting abortion rights increases from left to right.

As noted in Section 4, 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. However, Abortion Rights LaMP scores estimated using the Llama 2 models do not capture these patterns. Scores estimated using Llama 2 7B incorrectly identify Lamar Alexander as the Republican senator with the most pro-choice stance—he was endorsed by the National Right to Life Committee—and scores estimated using Llama 2 13B find perfect separation between the two parties.

Party	Model	NOMINATE	Perceived Ideology	CFscores
All	GPT-3.5	0.44^{***}	0.62^{***}	0.30^{**}
All	Llama 2 13B	0.46^{***}	0.28^{**}	0.34^{**}
All	Llama 2 7B	0.19	0.29^{**}	0.30^{**}
Dems	GPT-3.5	0.58^{***}	0.59^{***}	0.33^{*}
Dems	Llama 2 13B	0.55^{***}	0.30^{*}	0.45^{**}
Dems	Llama 2 7B	0.31^{*}	0.32^{*}	0.31^{*}
GOP	GPT-3.5	0.47^{**}	0.68^{***}	-0.36^{*}
GOP	Llama 2 13B	0.53^{***}	0.17	-0.23
GOP	Llama 2 7B	0.16	0.23	-0.17

Note: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Partial correlations between Ideology LaMP scores estimated using different LLMs and DW-NOMINATE, perceived ideology scores, 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 (2015).

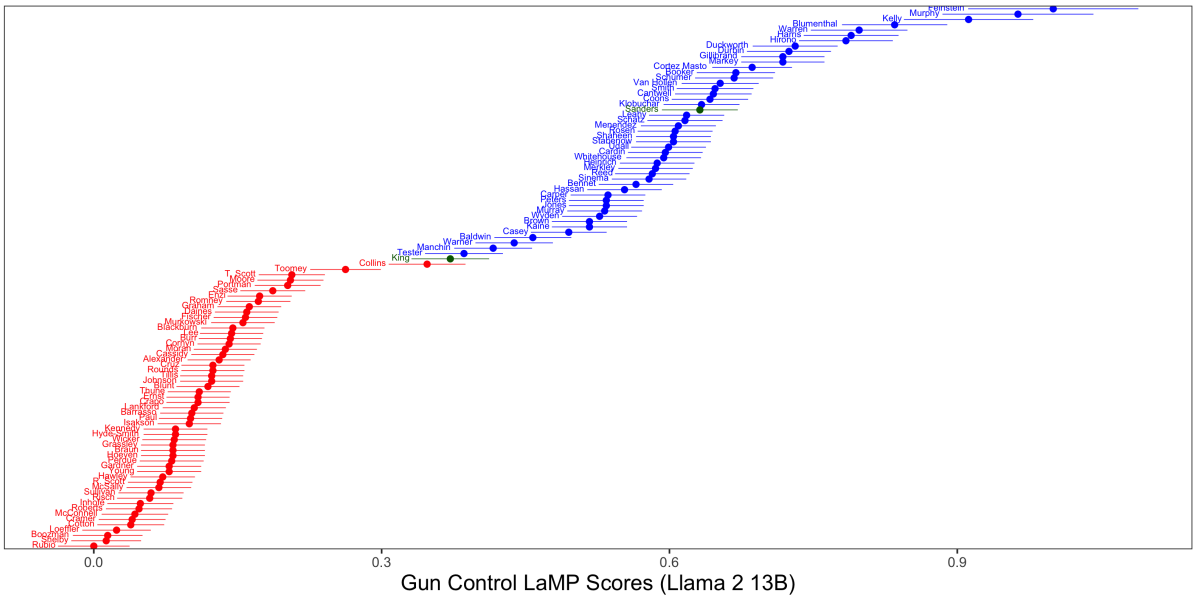

Figure 7: Gun Control LaMP scores of senators estimated using Llama 2 13B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

A.9.2 Abortion Rights LaMP scores estimated using GPT-3.5 better predict NARAL grades than DW-NOMINATE

We compare the predictive power of Abortion Rights LaMP scores and 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 2021. We calculate the proportion of variance explained (R^2) in NARAL grades using a full model with both predictors—Abortion Rights LaMP scores and DW-NOMINATE—and two reduced models, each with only one predictor. Again, comparing the R^2 values of the full and reduced models reveals the explanatory power lost when omitting either predictor.

We find that when we drop Abortion Rights LaMP scores estimated using GPT-3.5 from the full model, R^2 drops 55% when the data is limited to Republican senators, 58% when the data is limited to Democratic senators, and 4% when including all senators. In contrast, dropping DW-NOMINATE reduces R^2 by only 9% for Republican senators, 3% for Democratic senators, and 2% for all senators. Partial F-tests, again, confirm these results. For Republican senators, the partial F-test shows no significant difference when comparing the full model to the reduced model with Abortion Rights

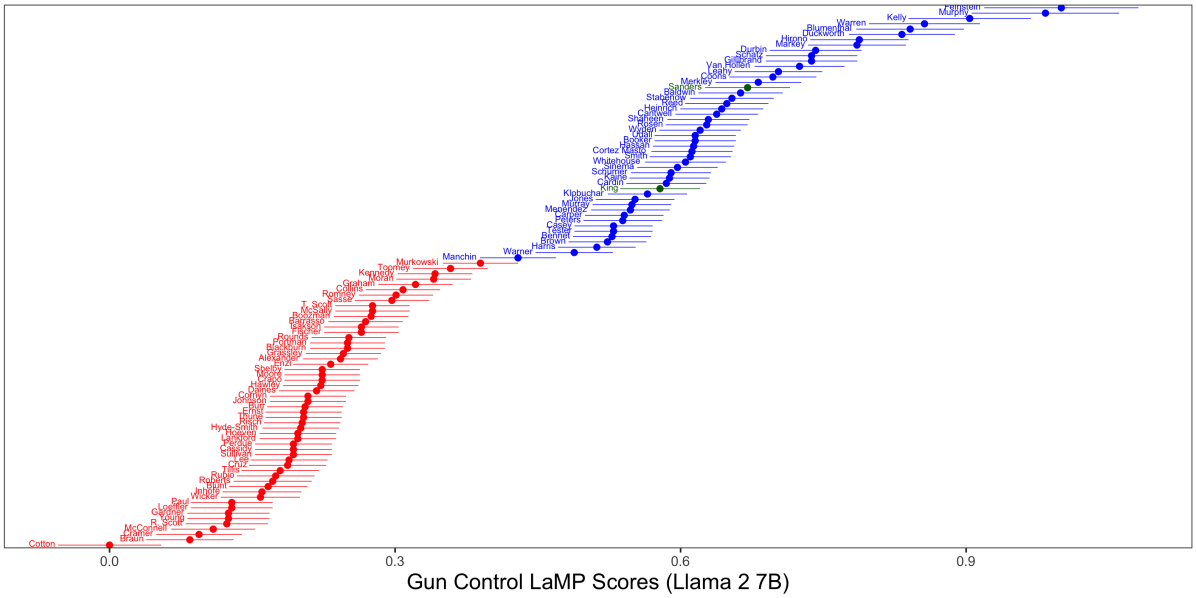

Figure 8: Gun Control LaMP scores of senators estimated using Llama 2 7B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

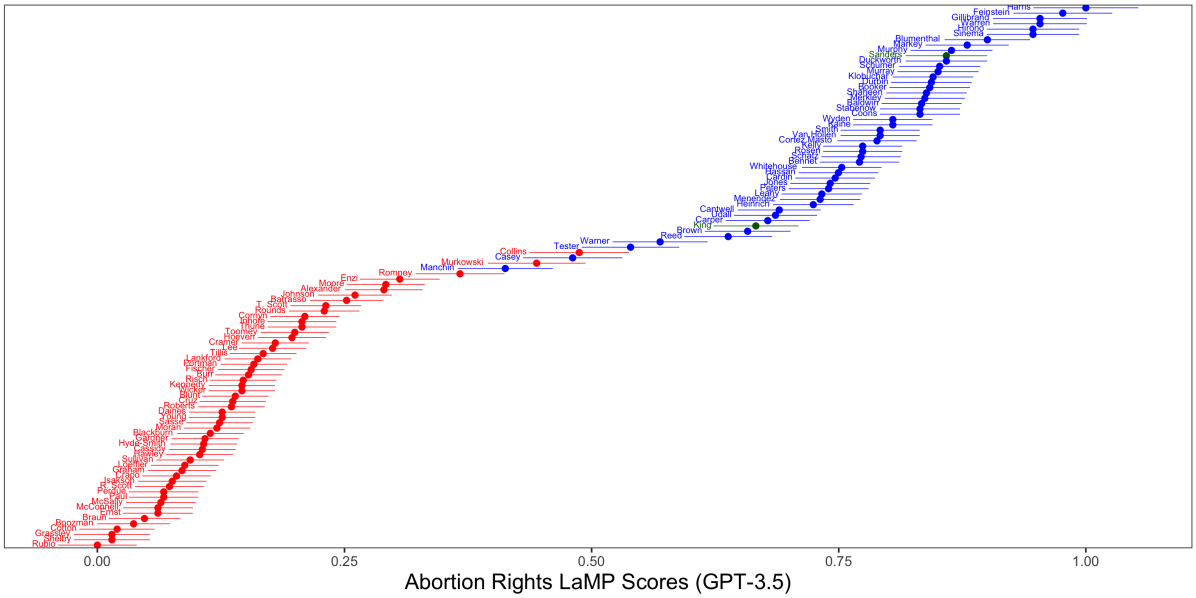

Figure 9: Abortion Rights LaMP scores of senators estimated using GPT-3.5 with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

LaMP scores ($p = 0.10$) and a significant difference when comparing the full model to the reduced model with DW-NOMINATE ($p = 0.0001$). Likewise, for Democratic senators, the partial F-test shows no significant difference when comparing the full model to the reduced model with Abortion Rights LaMP scores ($p = 0.43$) and a significant difference when comparing the full model to the reduced model with DW-NOMINATE ($p = 0.002$). Across all senators, the partial F-tests show significant differences when comparing the full model to the reduced models for both Abortion Rights LaMP scores and DW-NOMINATE ($p < .0001$).

These results, however, do not hold up when estimating the Abortion Rights LaMP scores using the Llama 2 models. With these scores, DW-NOMINATE is a stronger predictor of NARAL grades than the Llama-based Abortion Rights LaMP scores. This finding is consistent with the findings from the rest of the paper: smaller models yield scores that have less explanatory power and less nuance.

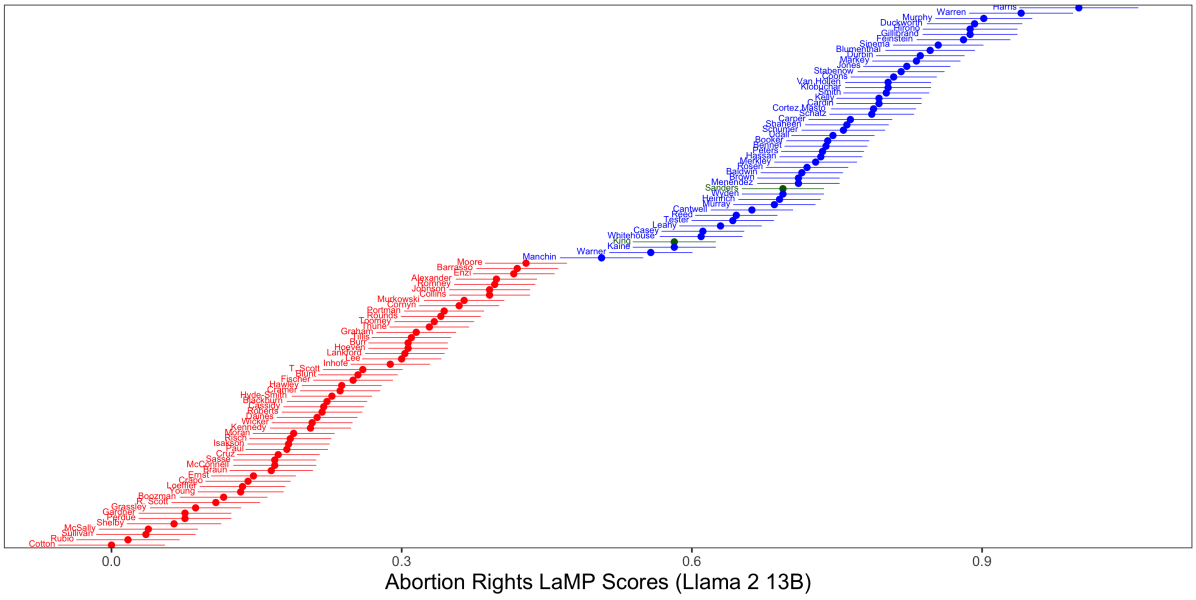

Figure 10: Abortion Rights LaMP scores of senators estimated using Llama 2 13B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

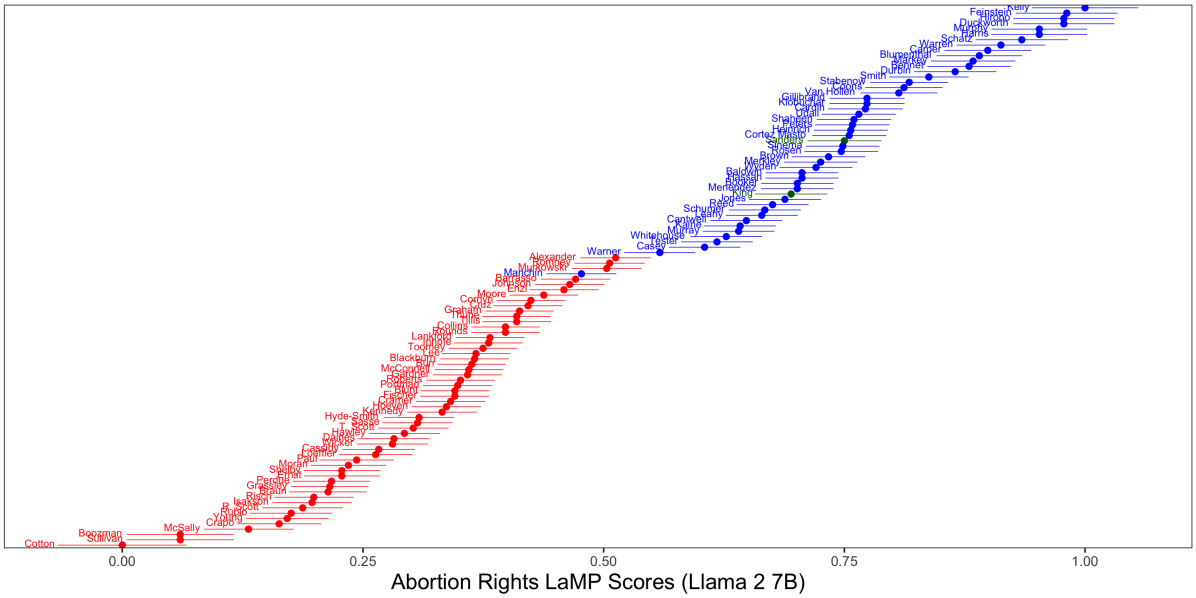

Figure 11: Abortion Rights LaMP scores of senators estimated using Llama 2 7B with 95% confidence intervals based on quasi-standard errors. Democrats are in blue, Republicans are in red, and Independents are in green.

A.9.3 Entropy of Abortion Rights LaMP Scores

Figure 12 plots the entropy of each pairwise comparison using the three LLMs against the difference in Abortion Rights LaMP scores within each pairwise comparison. Again, we find that as the difference in Abortion Rights LaMP scores increases within each pairwise comparison, entropy decreases for all LLMs. Comparing the entropy across LLMs, we find that the entropy's rate of decrease is slower as the model size decreases.

A.10 Datasets and License Information

We used three liberal-conservative ideology datasets to evaluate Ideology LaMP scores. DW-NOMINATE data can be downloaded from https://www.voteview.com/. The website itself is under an MIT License; it is unclear what is the dataset's license. The perceived ideology scores dataset can be downloaded from https://dataverse.harvard.edu/dataset.xhtml?

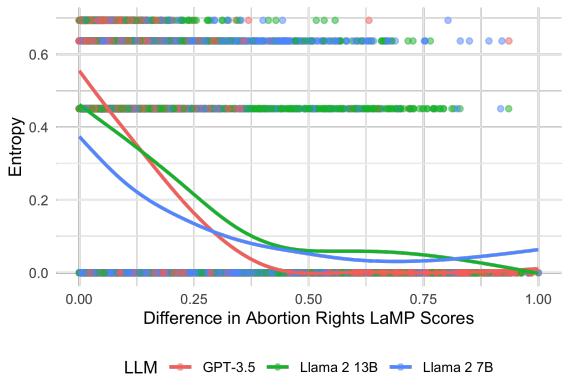

Figure 12: Entropy of abortion rights pairwise comparisons across the three LLMs.

persistentId=doi:10.7910/DVN/MGJBON and is under a CC0 1.0 license. CFscores can be downloaded from https://data.stanford.edu/dime and is under an ODC-BY 1.0 license. None of these datasets had specified intended uses, and none of this data contains identifying information except for public figures (in this case, American politicians at the federal level).

NRA grades were obtained from https://justfacts.votesmart.org/, which in turn, took its data from the NRA. NARAL grades were obtained from https://reproductivefreedomforall.org/resources/congressional-records/. Both sites have language allowing the use of their content for non-commercial purposes. Using this data also constitutes fair use. Votes on legislation are from the public record.

A.10.1 Perceived Ideology Scores

The perceived ideology scores dataset from Hopkins and Noel (2022) included pairwise comparisons of senators of the 117th Congress conducted in a YouGov survey in April 2021. Hopkins and Noel (2022) 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.

A.10.2 Campaign Finance Scores (CFscores)

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 (Bonica, 2016). We looked at the recipient CFScore for each senator, which is the estimated ideology of the senator based on donations received. Tammy Baldwin, Mark Kelly, and Kelly Loeffler are missing recipient CFscores.

A.11 Computational Resources Used

We defined "one set of pairwise comparisons" as the set of pairwise comparisons across all senators. In the paper, we had three sets of pairwise comparisons for all models for each scale (ideology, gun control, and abortion rights). For the Llama models, we used one A100 GPU. It took approximately 60 minutes to iterate through one set of pairwise comparisons with Llama 2 7B, and it took approximately 150 minutes to iterate through one set of pairwise comparisons with Llama 2 13B. It was approximately $5 (at most) to iterate through one set of pairwise comparisons using GPT-3.5-turbo with OpenAI's latest pricing. It was approximately $3 (again, at most) to extract the answers from the text using GPT-3.5-turbo.

For GPT-3.5, we used the OpenAI API. For the Llama models, we used the Hugging Face transformers package (Wolf et al., 2020).