A More Coherent Measure of Party Cohesion Using Topic Models and Contextual Embeddings

Gechun Lin* Benjamin S. Noble†

Abstract

Measuring party cohesion is central to normative and positive political theories. While the limitations of roll-call–based measures are well known, we highlight concerns with increasingly popular text-based measures. Text offers the advantage of being multi-dimensional, yet scholars often measure a single dimension (e.g., ideology, text reuse) and treat it as a proxy for cohesion generally. We show how this assumption can fail and propose a novel approach for measuring rhetorical cohesion across multiple dimensions at scale by combining topic models with contextual embeddings. We validate our measure and apply it to House floor speeches (1995–2020) and e-newsletters (2010–2020). Our results indicate that intra-party rhetorical cohesion has not increased over this period and is not strongly correlated with roll-call cohesion. Moreover, venue matters: newsletters exhibit weaker cohesion than floor speeches, with surprising inter-party differences. Our measurement strategy has natural applications to related literatures on policy diffusion, administrative rulemaking, and media slant.

Word Count: 9,988

*Assistant Professor, Department of Political Science, Texas A&M University. lingechun.github.io. lingechun@tamu.edu.

†Assistant Professor, Department of Political Science, University of California, San Diego. benjaminnoble.org. b2noble@ucsd.edu.

We are especially grateful to Justin Gill and Zhengyu Xiao for research assistance. We thank Pamela Ban, Jaclyn Kaslovsky, Pamela McCann, Jacob Montgomery, Mollie Ritchie, Margaret E. Roberts, Amna Salam and participants at the USC PIPE and UCSD American Politics Workshops and Midwest Political Science Association Conference for helpful conversations and feedback on this project.

How cohesive are the modern congressional parties? The answer to this question is essential for positive and normative theories in political science. From a positivist perspective, valid measures of party cohesion can help us determine the causes (when used as an independent variable) and consequences (when used as an independent variable) of party unity—both historically and in our current polarized environment. Normatively, cohesion is core component of responsible parties and democratic accountability (APSA 1950; Schattschneider 1942). When parties advance coherent and differentiated policy agendas, voters can make informed electoral choices and sanction those that fail to uphold their promises.

To understand party unity, scholars and journalists typically turn to the roll-call record. There, we see that party cohesion has reached an historic level. Party unity scores have never been higher (Reynolds and Maehr 2024). The first-dimension NOMINATE distance between party medians is currently at its maximum while intra-party standard deviations have declined. And while these measures are the industry standard, others argue that they can give a misleading impression of party unity. For example, Lee (2018, 1464–1465) argues "roll-call voting data exaggerate party unity, especially in the contemporary era" where many divisive issues are kept off the floor and partisan messaging bills have proliferated. Further, the increasing use of omnibus vehicles and unorthodox procedures leave rank-and-file lawmakers with little input or discretion when casting votes (Curry and Lee 2020; Sinclair 2017), making roll-calls a poor indicator of cohesion. In an effort to better assess party cohesion, scholars have turned to text-based measures, which increase dimensionality and decrease constraint of the choice set (Proksch and Slapin 2012; Quinn et al. 2010). This shift has led to an explosion of text-based studies measuring concepts like ideology (e.g., Gaynor et al. 2025; Rheault and Cochrane 2020), partisanship (e.g., Gentzkow, Shapiro and Taddy 2019; Kaslovsky and Kistner 2025; Lauderdale and Herzog 2016), position-taking (e.g., Ban and Kaslovsky 2024; Grimmer 2013), emotionality (Gennaro and Ash 2022; Osnabrügge, Hobolt and Rodon 2021), and nostalgia (Müller

and Proksch 2023), among many others. What this discussion makes clear is that while text promises more dimensionality than a matrix of yea and nay votes, it also requires that researchers *select* a dimension of text on which to focus. However, the choice of dimension may have important implications for the conclusions we draw about party unity.

To the extent that scholars are testing hypotheses about a specific dimensions of rhetorical behavior in isolation (as in many of the above-cited papers), selection of the appropriate dimension naturally follows from the research question. However, a problem arises when scholars use a single rhetorical dimension—such as presidential references (Groeling 2010), topic selection (Hughes and Koger 2022), or text reuse (Gaynor 2025)—to make *general* statements about party cohesion. These statements are valid only to the extent that the chosen dimension reliably correlates with all other unmeasured dimensions—an assumption that is difficult to test and unlikely to hold. For example, suppose two Democrats give floor speeches about healthcare and use the phrase "healthcare is a human right." They may appear rhetorically cohesive or "on message," but it may also be the case that these same two lawmakers disagree about the underlying legislative proposal—with one supporting it (e.g., "healthcare is a human right, and I will vote for the bill") and the other saying it is not liberal enough (e.g., "healthcare is a human right, and this bill falls short of helping those most in need"). One could incorporate positioning into the measure, but then a new dimension, say, the framing, might become relevant. A general measure of rhetorical cohesion ostensibly requires an infinite set of dimensions.

Even if we could solve the dimensionality issue, the problem of measuring cohesion is exacerbated when we consider that rhetorical behavior varies across venues like floor speeches, press releases, and social media (Gaynor et al. 2025; Green et al. 2024)—and constituent impressions are formed from some mixture of venue-specific rhetoric. Thus, we cannot truly understand party cohesion in a rhetorical context without a measure that (i) solves for multi-dimensionality and (ii) can easily be applied and standardized across venues. Developing such a measure is important given the positive and normative

implications of party cohesion not just in the U.S. Congress, but across the U.S. political system and representative democracies broadly.

We address both of these critiques with a novel, scalable, multi-dimensional measure of pairwise text similarity, which we call *rhetorical cohesion*. Our approach takes advantage of topic modeling and contextual embeddings to assess paired text similarity across an arbitrary set of dimensions. We describe our measurement workflow and apply the methodology to House floor speeches in the Congressional Record delivered between 1995–2020 (Gaynor et al. 2025; Gentzkow, Shapiro and Taddy 2019). Working with an RA, we validate a ground-truth sample of speech-pairs, which confirms that our measure accurately captures a general, multi-dimensional understanding of "similarity." Then, we demonstrate that our measure outperforms other conventional approaches for measuring rhetorical cohesion in the literature, including text reuse, text-based ideal point scaling (TBIP), and static GloVe embeddings. Then, we show that our measure of rhetorical cohesion exhibits modest positive correlation with these existing measures, as expected, but that these measures do not correlate well with one another. Taken together, these results indicate that our measure accounts for a broader understanding of similarity while illustrating the risks of using unidimensional measures to proxy for a general measure of cohesion.

After validating our measure, we turn to two applications. First, we examine the relationship between rhetorical and roll-call cohesion. In the aggregate, we show roll-call cohesion has been increasing since the 1990s, yet rhetorical cohesion is no higher today than it was three decades ago. At the pair-level, these two measures are positively correlated, but the relationship is weak, suggesting rhetorical and roll-call cohesion serve different purposes. This result is consistent with the idea that explanations are constituency-tailored to post-hoc justify partisan roll-call voting behavior (cf. Grose, Malhotra and Parks Van Houweling 2015) or shape constituent impressions (Grimmer, Westwood and Messing 2014; Hassell, Heseltine and Reuning N.d.). Second, we apply

our methodology to e-newsletters (2011-2020) and descriptively compare the results to cohesion on the floor. Consistent with the procedural constraints of, and leader influence on, the floor, we show that cohesion is consistently higher in the chamber than it is online. However, we also observe unexpected differences between parties. Democrats' floor cohesion is higher than Republicans', but the relationship is the opposite for e-newsletters. These results suggest that Republicans are more naturally cohesive (cf. Grossman and Hopkins 2016), but Democratic leaders use floor procedure more effectively. This latter result contrasts with other studies using unidimensional measures, which find Republicans to be the more disciplined party (Gaynor 2025; Russell 2018). We argue that existing methods have captured a form of *partisan* messaging, at which Republicans excel. By contrast, our measure is more general and speaks to the underlying level of policy disagreement within the Republican party (Lee 2018). Ultimately, our research provides a new workflow for measuring text similarity, raises new questions about congressional party cohesion, and has natural applications in research areas like legislative hitchhiking (Casas, Denny and Wilkerson 2020), policy diffusion (Desmarais, Harden and Boehmke 2015), and media slant (Gentzkow and Shapiro 2010).

Why Party Cohesion?

Like many of his contemporaries, John Adams hated political parties. As the U.S. fought the British for independence, he penned a letter to a friend, arguing that "a division of the republic into two great parties...is to be dreaded as the great political evil" (Adams 1780). Although a popular opinion in the 1780s, this dismal view of parties is not shared by modern political theorists. The theory of Responsible Party Government asserts the opposite: that parties are essential for democratic governance (APSA 1950; Schattschneider 1942). Policy is made collectively, and thus, parties are needed to collectively hold policymakers to account (Aldrich 1995). Yet, the existence of parties is

necesary but not sufficient. As Fiorina (1980, 26) writes, "responsibility requires cohesive parties." Parties can only be held accountable by voters when they advance clear, unified, and differentiated platforms. To evaluate whether parties are fulfilling these normative ideals, theorists need methods to measure party cohesion.

Beyond its normative importance, party cohesion is thought to benefit legislators. Lawmakers' electoral and policy goals are linked to their party's electoral success. And electoral success depends, to an extent, on parties' collective reputations (Aldrich 1995; Cox and McCubbins 2005; Lee 2009). When party members advance the same arguments and vote the same way on roll-calls, they maximize the likelihood that the media will echo, and that voters will receive and understand, what parties stand for and what they would do if given institutional power (Sellers 2009; Lee 2016). To empirically test these theories, scholars require quantitative measures of party cohesion over time.

Party unity is an essential concept in normative and positive political science. Yet, there is widespread debate about how party cohesion ought to be measured and what our existing measures can tell us about party cohesion.

Limitations of Roll-Call Based Measures of Party Cohesion

The roll-call record is the standard data source for understanding party cohesion in the popular press and the academic literature. Some measures use the roll-call record directly. For example, the Rice index accounts for the intra-party differences between the yeas and nays, averaged across a series of votes. Party unity scores measure the proportion of lawmakers who vote with their party on bills where a majority of one party opposes the majority of another. Cox and McCubbins (2005) focus on the "roll rate," the proportion of bills on which a majority of one party votes against a bill that ultimately passes. Beyond these direct summaries, NOMINATE has become a workhorse method to generate unidimensional legislator ideal points from roll-call voting patterns (Poole and Rosenthal 1997). Others (e.g., Aldritch, Berger and Rohde 2002; Rohde 1991) have

then taken these ideal points to compute summary statistics measuring polarization (the difference between party median ideal points) and cohesion (the standard deviation of a party's ideal points). In general, these roll-call based measures have shown increasing party polarization and cohesion since the mid-1990s, which have reached historic highs in the mid-2020s.

While these measures are direct, interpretable, and convenient, they raise three methodological concerns when it comes to understanding party cohesion. First, roll-call votes are not placed on the agenda at random or across all issues. Selection is a persistent problem given that many issues are kept off the agenda to ensure the majority party avoids an embarrassing roll (Cox and McCubbins 2005) or because a party may promote unifying messaging bills that will never become law (Lee 2016, 2018). In general, "Party members (and leaders) have an incentive to request [roll-call votes] when their party is going to have a noticeably higher level of unity than the opposition" (Ainsley et al. 2020, 704), biasing these measures. Second, a bill is a bundled package designed by a coalition leader to attract sufficient support (Arnold 1990). These packages present lawmakers with a binary choice between all items they contain and the status quo. Setting aside issues of selection, even a randomly chosen bundle of policies can only tell us so much about unity. Two members may vote in the same direction for very different reasons (Duck-Mayr and Montgomery 2023). Or, two lawmakers may vote for a package but vote against one another were the items unbunbled. Third, the roll-call record is a poor tool to understand *why* lawmakers are (or are not) unified. Although Rohde (1991) shows that parties have become more internally cohesive over time, these changes may be driven by either a change in preferences or an increase in leader-enforced discipline. Methods to infer party influence from the roll call record (Krehbiel 1993; McCarty, Poole and Rosenthal 2001; Sinclair 2002; Snyder and Groseclose 2000), face an endogeneity problem in that "these votes are partly a cause and partly a consequence of the very things the theories seek to explain" (Clinton 2007, 457). This problem is difficult to solve when measuring unity from

a single source of data given that there is no cross-sectional variation in leader influence.

Limitations of Text-Based Measures of Party Cohesion

Given these concerns, some scholars have turned to text-based measures of party cohesion which are generally "more varied and less amenable to disciplinary actions by party leaders than votes" (Schwarz, Traber and Benoit 2017, 394). Even so, they come with their own set of assumptions and limitations. For example, one popular approach is to look at *what* lawmakers talk about, given that "speech offers a direct measure of the issues that MCs prioritize" (Witko et al. 2021). While this methodology can help us understand when parties cohere around the same topics (Hughes and Koger 2022; Quinn et al. 2010), new challenges arise. First, topics are insufficiently granular to measure cohesion in the way political scientists often desire. Typically, researchers are not simply interested in whether lawmakers are talking about healthcare, but rather, the positions lawmakers are taking, the ideological valence of their stance, the sentiment or emotionality of their arguments, etc. While topic models can be used to detect framing, doing so often requires narrowing the lens to a small number of congresses or a specific macro-topic. Second, topics are often downstream of leader or majority-driven agenda setting, especially in parliamentary contexts like the House floor where speeches must be germane. That partisans discuss healthcare on a given day may simply follow from the House agenda and rules, rather than party cohesion.

Beyond the inductive search for topics, scholars have used deductive methods to measure rhetorical party cohesion. For example, Groeling (2010) uses party discussion of the president as a proxy for general party cohesion. In theory, a cohesive party publicly supports its own party's president and attacks the opposition party's president. While this is a form of party cohesion, it represents just one dimension of intra-party conflict. Using this measure as a proxy for generalized party cohesion requires that presidential discussion is representative of the myriad other ways parties might cohere or diverge. However,

existing literature suggests that assumption is unlikely to hold. The president is a unique cue (Hopkins 2018; Lee 2009; Nicholson 2012; Noble 2024), and like voting against one's party on a roll-call, vocally speaking out against one's president represents a particularly salient and binary choice. There is no reason to think that a party that is unified in supporting its president (or opposing the opposition president) is unified when it comes to its rhetorical stance across all policies. For example, the dynamics of the One Big Beautiful Bill Act illustrate that parties may ultimately vote for their president's signature policy even if they rhetorically disagree about the underlying substance.

The study most similar to ours is Gaynor (2025), which leverages text reuse methods to identify similarity in paired leader/rank-and-file press releases. As leaders often encourage the rank-and-file to adopt partisan talking points, this method can help us understand which members adopt party messages. However, unlike studies of legislative hitchhiking (Casas, Denny and Wilkerson 2020) or policy diffusion (Desmarais, Harden and Boehmke 2015; Hertel-Fernandez 2019) where direct re-use is of substantive interest, in the context of party messaging, direct re-use is a rough proxy for a more complicated quantity of interest. Lawmakers' speeches or social media posts often express similar ideas using different words. That is, they can be semantically similar even if they are not lexically similar. Conversely, speeches can share a set of pat phrases (scoring high on a text reuse measure) even if the remaining context is not semantically similar. An ideal measure would account for the fact that partisans may adopt the *spirit* of someone else's talking points even if not directly plagiarizing them.

As an illustration of these critiques, we turn to excerpts from three speeches given by Republicans in opposition to the Affordable Care Act on November 3, 2009 in Table 1. The first two excerpts in Panel A, labeled Q1 and Q2, frame opposition in terms of the Medicare Advantage Program. Both representatives reference Democratic Majority Leader Pelosi, name the Medicare Advantage Program, and lament that millions of seniors will lose coverage. A valid measure of rhetorical cohesion should assign this pair

Table 1: Comparison of ACA Message Cohesion: Three Republican quotes on the ACA and their measured similarity

Panel A. ACA Quotes

Q1	**Rep. Joseph Pitts (R-PA):** "The Pelosi health care bill we will consider later this week effectively eliminates the popular Medicare Advantage health plans that millions of seniors rely on for medical, vision, and dental care."
Q2	**Rep. John Fleming (R-LA):** "The CBO estimates in PelosiCare that it will cut over $150 billion from Medicare Advantage... That will knock about 6 to 11 million seniors off of Medicare Advantage."
Q3	**Rep. John Boozman (R-AR):** "The newly created public option will be authorized to fund elective abortions. The Pelosi health care bill does not include the pro-life language...As the bill is written, Federal funds will pay for elective abortions."

Panel B. Similarity Measures

Quote Pair	True Similarity	Topic Model	Text Reuse
Q1 & Q2	High	High	Low
Q1 & Q3	Low	High	Moderate
Q2 & Q3	Low	High	Low

a high score, given their substantive and semantic overlap. They make nearly identical claims and arguments. These two excerpts contrast with Q3, which also names Pelosi but focuses on the issue of abortion funding. It does not discuss the Medicare Advantage Plan or seniors' healthcare coverage, and it substantively differs from the other two speeches.

How do existing measures of party cohesion fare? First, roll-calls cannot provide much insight. When the bill came to a vote four days later, all three members voted nay (as did all Republicans, save one). However, as these quotes make clear, Boozman (Q3) may have voted (or claimed to have voted) against the bill for a very different reason than his colleagues Pitts (Q1) and Fleming (Q2).[1] Next, none of these excerpts reference the pres-

[1]We recognize that these explanations are likely post-hoc justifications for partisan opposition that would have materialized no matter the bill's content, but that does not mean these statements are cheap talk. How lawmakers frame a debate can influence constituents' attitudes toward the policy and their representatives (Broockman and Butler 2017; Grose, Malhotra and Parks Van Houweling 2015; Hassell, Heseltine and Reuning N.d.; Hopkins 2018).

ident. Given that all three Republicans voted against the bill, one of President Obama's key legislative priorities, we could infer that they were appropriately aligned in opposing an out-party president. However, and as suggested previously, that opposition is not correlated with the cohesiveness of these explanations. Next, a standard topic model fit on the corpus of floor speeches from 1995–2020 would likely place all three speeches within a healthcare or ACA topic.[2] Finally, we turn to text reuse. Although Q1 and Q2 are semantically similar, they share no common 5-grams (a standard approach, as in Casas, Denny and Wilkerson 2020; Gaynor 2025) and have a similarity score of 0. Meanwhile, Q2 and Q3 share the 5-gram "The Pelosi health care bill," thereby achieving a modest similarity score. While this method accurately detects the shared use of this (likely) leader-driven phrase (an important topic of study in and of itself), it gets the underlying cohesion backwards. The discussion in Table 1 highlights the need for a better method of identifying rhetorical cohesion, which we introduce in the next section.

Before describing our measure, we highlight three literatures that are related but distinct. First, Lin (2025) introduces cross-encoders to measure the semantic similarity of short text pairs. Given token limitations, this method cannot be used when the quantity of interest is longer (like floor speeches), necessitating a different method. Second, our goals differ from scholars who measure partisanship and polarization (e.g., Gentzkow, Shapiro and Taddy 2019; Green et al. 2024; Peterson and Spirling 2018). Our interest is within-group similarities rather than between-group differences. Although these two quantities may correlate, they need not if there are e.g., two distinct sub-factions within a party that nonetheless differ from the opposition. Third, technological innovations have led to a set of new methods for ideologically scaling lawmakers using text data (e.g., Porter and Case N.d.; Rheault and Cochrane 2020; Slapin and Proksch 2008). In particular, Gaynor et al. (2025) is most similar in assessing the spread of ideal points from roll-call votes,

[2]While it's possible that a model with few documents and many topics would identify these framing differences (see Hopkins 2018), the standard approach in the discipline is the opposite: a small number of topics fit on a large corpus of documents.

floor speeches, and social media posts in the 115th–116th congresses. While our measure of rhetorical cohesion is correlated with ideology (as shown below), it is conceptually distinct. Ideology is a single dimension, and ideologically dissimilar speech pairs may still be similar in other ways if they e.g., make use of emotional arguments, focus on constituency concerns rather than national policy, etc.

As a final note, our method does not impose any left-right (or any unidimensional) scale on the outputs. We do not (nor can we) rank-order speeches in a meaningful way beyond noting which pairs of speeches or lawmakers are most or least cohesive. Although many research questions necessitate rank-ordered, unidimensional scales, we have a different goal in mind and are solving a different problem. We seek to understand how parties cohere across many dimensions at once, without having to specify those dimensions. This approach is useful when we want to understand general rhetorical party cohesion as opposed to e.g., ideological cohesion or presidential support. If one's goal is to isolate a specific dimension of cohesion, we refer readers to the excellent papers we have cited.

A General Measure of Party Cohesion

We draw on the legislative politics literature which defines party cohesion as "the extent to which, in a given situation, group members can be observed to work together for the group's goal in one and the same way" (quoted in Hazan 2003, 3), in particular, as it relates to policy agreement (Sinclair 2003). We extend this idea to rhetoric, seeking to capture the extent to which ingroup members speak in a cohesive manner—such as expressing similar positions, ideologies, justifications, frames, etc—on a specific policy issue.

To help operationalize our definition, imagine a fictional "maximally cohesive party." In the single issue setting, we define this party as one where every single member gives an identical speech on the policy issue. Of course, parties talk about many issues at once,

but maximal cohesion in our framework is issue-specific. That is, we do not expect parties to have the same message on healthcare and civil rights policy. These are different issue domains that require the use of different language and different arguments. A maximally cohesive party in a multi-issue space is one that gives identical statements *within* issues, even if statements differ across issues. There may be overlap, for example, if one party consistently evokes themes of limited government across issues, but there need not be.

Parties do not achieve this maximal level of cohesion on any issue in reality. The English language is rich, and co-partisans often express similar sentiments using slightly different language. Strategically, lawmakers' messages deviate from their peers due to personal idiosyncracies, ideological preferences, constituency pressure, variation in audience, representational style, and poor leadership coordination. So while we recognize maximal cohesion is generally infeasible (and perhaps, not even desirable), it is helpful methodologically when considering pairwise similarity and party cohesion. Here, our aim is to construct a measure that would reach its maximum were party members to achieve maximal cohesion on a given issue and decrease as paired rhetoric differs. In practice, the more two paired documents approach perfect overlap, the higher our measure. The more any two documents differ—whether due to idiosyncratic differences in phrasing or larger differences in framing or positioning—the lower our measure. Then, by aggregating—over party members, over topics, and/or over time—we can produce a summary statistic of rhetorical cohesion at any unit of analysis.

To construct our measure of rhetorical cohesion, we proceed in two stages. First, we identify documents that cover comparable policy issues using a supervised topic modeling framework. Our goal in this stage is to organize documents into a set of policy issues of interest. This step allows us to pair documents that discuss the same policy, which is essential for operationalizing our definition of party cohesion. Second, we capture message variation within issues by measuring the semantic similarity between co-partisans' policy statements using text embeddings from a large language model (LLM). This ap-

proach provides a fine-grained measure of rhetorical policy agreement, i.e., speaking "in one and the same way," which differs from existing measures of party unity relying on roll-calls (e.g., Lebo, McGlynn and Koger 2007; Crespin, Rohde and Wielen 2013).

A key limitation of our approach is that rhetorical cohesion is a black box. The scalar summary of this multi-dimensional concept does not provide any insight into *which* dimensions are responsible for any observed (dis-)similarity. As such, we do not seek to displace single-dimension studies. Rather, as existing literature has made clear, there is interest in a generalized measure of rhetorical party cohesion that is agnostic to any particular dimension (e.g., Gaynor 2025; Groeling 2010; Hughes and Koger 2022; Sellers 2009). The latter is what we develop and showcase here.

Identifying Topics: Comparable Policy Issues Across Time and Corpora

We apply our measure to House floor speeches from the 104th to 116th (1995-2020) Congress. Although our workflow is agnostic to the corpus, we begin with floor speeches for several reasons. Methodologically, the use of floor speeches ensures our estimates are comparable to other recent research exploring party cohesion (Gaynor et al. 2025; Gentzkow, Shapiro and Taddy 2019). Floor speeches are also readily available across our time series, allowing us to assess changes in cohesion across several presidential administrations and changes in majority control. Substantively, we believe floor speeches also constitute the best source of data to explore the relationship between rhetorical cohesion and roll-call cohesion *because* floor speeches are subject to the same agenda-setting considerations as roll-calls and, in the House, must be germane. We should observe the strongest relationship between rhetorical and roll-call cohesion here. However, floor speeches may not be representative of general party cohesion for these same reasons. Therefore, we later apply our measure to a corpus of e-newsletters, where we expect leader influence to be weaker (Green et al. 2024). By comparing across these two data sources, we are able to leverage cross-sectional variation in leader influence across venues to explore how proce-

dural control may affect rhetorical cohesion.

We begin by identifying the most prevalent topic of each speech using a keyword assisted topic model (keyATM, Eshima, Imai and Sasaki 2023). Unlike unsupervised LDA, keyATM allows researchers to nudge the topic model toward a set of topics using lists of pre-defined keywords. Our keyword lists come from the Comparative Agendas Project (CAP) master codebook (Jones et al. 2023), a project that categorizes political content into a set of 21 major topics such as macroeconomy, health, and foreign affairs. This model allows us to focus on a set of broad, stable policy issues that we expect to persist across our time series, facilitating over-time comparisons. To generate keywords, we downloaded Democratic and Republican Party Platforms that have been hand-coded at the quasi-sentence level by Wolbrecht et al. (2023) according to the CAP codebook. We calculated the tf-idf score of each stemmed word in this corpus within each topic and chose the top 15 words in each category as our topic-specific keywords (see Table A1 for the full keyword list). We also manually created two additional non-CAP topics specific to congressional speeches: *parliamentary language* (e.g., quorum, yield) and *other*, which included uninformative and common words (e.g., people, think). We apply standard pre-processing to all speech transcripts (see Grimmer, Roberts and Stewart 2022, and Appendix A.2) and fit a separate keyATM model to all speeches delivered within each two-year Congress. This approach allows vocabulary and policy content to vary over time, while the use of keyATM stabilizes the set of topics. After fitting, we assign each speech a single label according to its most prevalent topic and drop speeches labeled *parliamentary* or *other*, given our focus on *policy* cohesion.[3] Below, we follow Ying, Montgomery and Stewart (2022) to validate these topic labels.

[3]Speeches may cover multiple topics, and we could have used this mixed-membership to iteratively re-pair speeches across several topics. Doing so creates a computational scalability problem that makes the analysis infeasible and our measure impractical for use in standard political science research.

Beyond Topics: Capturing Party Cohesion with Semantic Embeddings

Topics are inherently coarse and do not capture the nuances of how lawmakers frame, justify, or take positions on policy issues. Further, floor speech topic selection is directly set by procedural rules and the agenda-setting power of party leaders. Thus, we move beyond topic-level analysis to measure the substance of political communication by leveraging OpenAI's `text-embedding-3-small` model. This model converts each speech to a high-dimensional numeric vector where speeches with more similar values are positioned closer to one another in vector space.[4] There are three advantages to using these LLM-based embeddings. First, embeddings are trained, in part, by predicting words in context, giving them an understanding of how words and phrases relate. As a result, they can capture similarity between documents based on meaning, not just exact word matches. Second, `text-embedding-3-small` is a contextual embedding model, meaning that the resulting vectors account for how meaning changes based on the context. For example, "alien" will have a different vector representation depending on whether it is used in the context of immigration or UAP sightings. These embeddings improve upon static models like word2vec (Mikolov et al. 2013) and GloVe (Pennington and Manning 2014), which assign a fixed, static vector to each word in a vocabulary.[5] Third, due to scaling model size and the massive training corpus of LLMs (Minaee et al. 2024), the `text-embedding-3-small` model is capable of handling longer sequences (8,191 tokens per input) as compared to, e.g., BERT (Devlin et al. 2018, 512 tokens) and other smaller, pre-trained language models (PLM).[6] In addition, PLMs require task-specific fine-tuning with domain data to achieve better model performance (Wang 2023), but extant literature

[4]Details about the model architecture and training processes of OpenAI's embedding models are not fully disclosed. To the best of our knowledge, OpenAI's embedding models are built upon transformer architectures closely associated with the GPT (Generative Pre-trained Transformer) family of LLMs (Yenduri et al. 2024).

[5]For a comprehensive discussion of the disadvantages of word embedding approaches in measuring text similarity, see Lin (2025).

[6]With a window length of 8,191 tokens, the model can encode nearly every speech in our corpus. We drop the vanishingly small number of speeches that exceed this maximum length.

has demonstrated that LLMs (especially GPT) are effective zero- and few-shot learners (Rathje et al. 2024), possessing the ability to solve unseen tasks based on instructions and prompt engineering (Yang et al. 2023; Zhang et al. 2023). OpenAI's embedding models are versatile and well-suited to our task. Below, we validate these embedding-based similarity scores with a set of human-labeled comparisons.

With these speech-level embedding vectors, we measure rhetorical cohesion by computing the distance between text pairs using cosine similarity, as recommended in OpenAI's documentation.[7] Here, we pair speeches at the day-party-topic level. That is, we pair each speech given by a lawmaker on a topic with each other speech given by a co-partisan member on the same topic on the same day.[8] By restricting pairs to the same period of time, we maximize the likelihood that speeches on the same topic cover the same underlying sub-issue. However, we note that researchers can choose their unit of analysis as befits their study. For example, a scholar interested in party factions could aggregate at the caucus level; someone interested in differences between foreign and domestic policy could aggregate across coarser topics.

The output of our measure is a scalar quantity that can theoretically range from -1 to 1, which we interpret as the level of rhetorical cohesion expressed by two co-partisans on the same policy topic. Higher values indicate that two statements are more related to each other. Negative values would indicate dissimilar or unrelated inputs (however, we do not observe negative values in our data given that we pre-pair within-topic and party). Our final dataset contains 1,825,739 speech pairs, and our rhetorical cohesion score ranges from 0.04 to 1.00 with an inter-quartile range of 0.56 to 0.76.

[7] See: https://platform.openai.com/docs/guides/embeddings/.

[8] We exclude independents and party-switchers in the Congress they switch. Throughout, all covariates come from VoteView (`https://voteview.com/`) and The Center for Effective Lawmaking (`https://thelawmakers.org/`). We exclude a small subset of lawmakers for whom we do not have key covariates.

Validating our Measures

In this section, we assess the construct validity of both elements that constitute our measure of rhetorical cohesion: the topic labels and the embedding-based similarity scores. Then, we turn to the convergent validity of our measure.

Construct Validity: Topic Validation

To ensure our topic model and subsequent top-1 labels are accurate, we follow Ying, Montgomery and Stewart (2022) in conducting an Optimal Labeling (OL) task with two trained coders. One of the authors randomly sampled 105 speeches from the corpus (five per policy topic) and presented the coders (the other author and a trained RA, both blind to the model labeling) with each full speech and four potential labels. One of these was the top label assigned by the model, while the other three were selected from among the remaining possible topics.[9] Coders were asked "What is the main policy category for this speech?" and selected one of the options. The model matched the author and RA in 80 and 81% of cases respectively—accuracy similar to that of Ying, Montgomery and Stewart (2022). Inter-rater reliability between the two coders was substantial, achieving a Cohen's-κ of 0.77. We provide more information on topic validation in Appendix A.3.

Construct Validity: GPT-Similarity Validation

To determine whether our GPT-embedding based similarity measure accurately captures broad-based similarity, we worked with a trained research assistant to conduct a novel pairwise validation exercise (cf. Carlson and Montgomery 2017). We presented the RA with a *focal speech* sampled from our corpus along with two *comparison speeches* that were paired with the focal speech in the underlying data.[10] The RA was asked to com-

[9] We exclude the next two most probable topics given that several CAP topics overlap (e.g., macroeconomy and domestic commerce) and could cause confusion.

[10] We worked with a single trained RA given the demanding and complex nature of the task. Speeches in our sample are 372 words on average, and these comparisons require careful reading, sustained attention,

plete a series of *comparison tasks*. For each, they carefully read the focal speech and both comparison speeches, then they selected the comparison speech that was more similar to the focal speech, where similarity was broadly defined.[11] When sampling documents and constructing pairs, we ensured we sampled a mix of within- and across-focal speech pairs, to validate within-pair and across-pair comparisons. Document sampling information, complete instructions, an example, and additional detail on this validation task can be found in Appendix A.4.

Our quantity of interest is whether the RA selects the comparison speech that that has a higher cosine similarity to the focal speech. In total, the RA and embedding-based cosine similarity score matched in 81% tasks, which is strong given the complexity. As a further demonstration of performance, in Figure 1, we compare the accuracy of our measure to others that are popular in the literature. Our competitors include the negative absolute difference of a pair's text-based ideal point (TBIP) scores (Gaynor et al. 2025),[12] 5-gram text overlap (Casas, Denny and Wilkerson 2020; Gaynor 2025), and the cosine similarity of the pairs' averaged GloVe vectors from the Stanford pre-trained embeddings (Rodriguez and Spirling 2022). In the first two cases, we code a match as when the negative absolute difference is smaller for the RA-selected pair. In the GloVe case, we define a match as when the cosine similarity score is higher for the RA-selected pair.

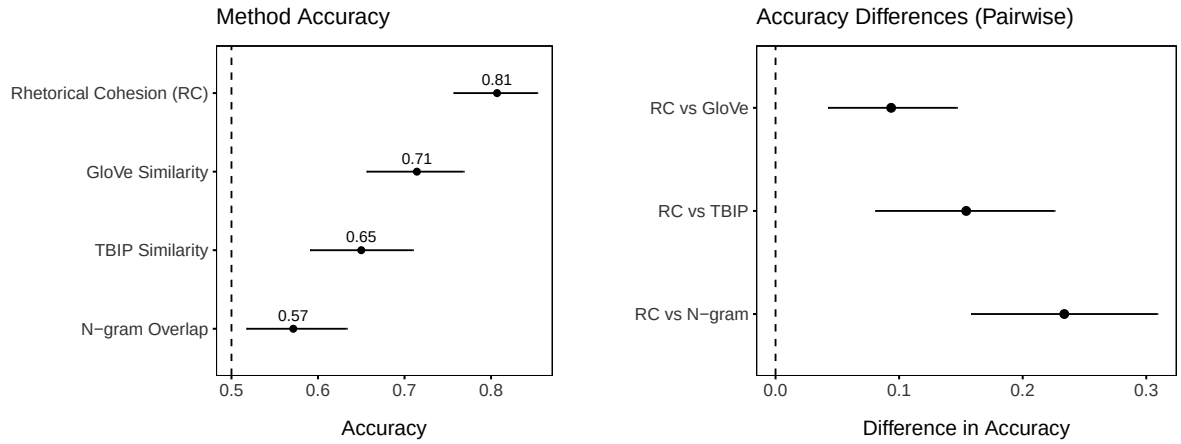
In the left panel of Figure 1, we present the accuracy of each method in recovering the RA's selections. Our measure of rhetorical cohesion beats all competitors. The closest competitor, GloVe, achieves only 0.71 accuracy—and these differences in accuracy are statistically distinguishable at the 95% level, as shown in the right panel, where we assess statistical differences via bootstrapping. Both 5-gram overlap and TBIP perform poorly

and thoughtful judgment over a span of weeks. Given cost and time constraints, our RA was able to complete 260 comparison tasks.

[11] To ensure the RA understood the assignment, one of the authors coded a small overlapping set (blind to the labels) which matched 90% of the RA's selections.

[12] Given the structure of our data, we construct Congress-topic pairs from our existing data and fit a Wordfish model on each of these subsets in our full dataset rather than computing a true TBIP score. We exclude speeches for which the model did not converge.

Figure 1: Accuracy of Rhetorical Cohesion and Comparison to Competitor Methods



Note: The left figure depicts accuracy of our rhetorical cohesion measure in recovering the human-selected more-similar speech. Our rhetorical cohesion measure outperforms other related measures. The 95% confidence intervals are generated via 1,000 bootstraps. The right figure depicts the pairwise accuracy differences with bootstrapped standard errors. The improved accuracy of our method is statistically distinguishable from competitors.

with accuracies of 0.65 and 0.57, respectively. Taken together, these results provide a strong indication that our measure accurately captures a general understanding of speech similarity and outperforms competing measures in the literature.

Convergent Validity

We have shown that our measure of rhetorical cohesion captures general similarity in text. However, we have also argued that our measure is multidimensional, capturing existing undimensional measures. If true, we expect our measure to positively correlate with existing measures of cohesion. To provide evidence to substantiate this claim, we compute the Pearson correlation of our measure and existing measures of pairwise similarity from the literature. Here, we focus on four major variables: NOMINATE scores, TBIP (Gaynor et al. 2025), presidential references (Groeling 2010), and 5-gram text overlap (Gaynor 2025; Casas, Denny and Wilkerson 2020). For our quantities of interest, we compute the negative absolute difference of the speakers' NOMINATE scores, the negative

Table 2: Correlation Between Rhetorical Cohesion and Relevant Proxies

	Rhetorical Cohesion	**NOMINATE Similarity**	**Both Ref. Pres.**	**Text Overlap**
NOMINATE Similarity	0.051			
Both Ref. President	0.170	0.019		
Text Overlap	0.126	0.018	-0.021	
TBIP Similarity	0.391	0.029	0.051	0.054

Note: All pairwise correlations are statistically significant at $p < 0.001$.

absolute difference of TBIP estimates of paired speeches, a binary indicator for whether both speeches reference the president,[13] and the set Jaccard similarity of 5-grams in both speeches, respectively. We expect each of these metrics to be positively correlated with our measure of rhetorical cohesion, however, we expect these correlations to be modest or low given that they are four of many components that comprise cohesion.

We present pairwise correlations in Table 2. First, we see that NOMINATE similarity has a positive, but small, correlation with rhetorical cohesion of 0.051. More ideologically similar members given more similar speeches, but the relationship is weak. Granted, this result is to be expected given that NOMINATE scores are computed across roll-calls (rather than text) and across all topics (rather than per-topic). Nonetheless, it is a useful benchmark for understanding the degree to which text-based measures can provide different kinds of information about party cohesion. Next, we turn to text based measures of similarity and find stronger, but still weak to moderate, correlations. Presidential references are correlated at 0.170, and text overlap is correlated at 0.126. The latter is particularly interesting given that, mechanically, our measure of rhetorical cohesion will increase the more overlapping 5-grams are present in a speech. If two speeches were identical, rhetorical cohesion would be at 1. This result suggests that, to the extent two speeches are similar, much of it is being driven by shared meaning rather than shared words or phrases. Finally, TBIP shows the highest correlation with our measure, 0.391. This cor-

[13]We follow Noble (2024) in coding a reference as when the president's last name or "the president" is used.

relation is modest, but note that it explains only about 15% of the variance ($R^2 = 0.15$) of our measure. They are related, but our measure is distinct from topic and ideology based measures of cohesion. The pairwise correlations between alternative measures are also of interest. Although each of these measures is positively correlated with ours, as expected, the correlations between these alternatives tend to be much weaker—and in one case, negative. The coefficients in Table 2 provide evidence that our measure of rhetorical cohesion captures similarity in a multi-dimensional sense and highlights the risk of using existing measures as proxies for a general measure of rhetorical cohesion.

Applications

Having validated our measure of rhetorical cohesion, next, we demonstrate how our measure can be used to uncover new insights and directions to explore in the congressional politics literature.

Does Rhetorical Cohesion Predict Roll-Call Cohesion?

To what extent does rhetorical cohesion predict roll-call cohesion? Answering this question can yield important insights about both legislative behavior and voter perceptions. If these two items are well correlated, we may conclude that roll-call based measures of party cohesion are sufficient for our purposes as both political scientists and constituents. However, a finding that these two types of cohesion are not tightly linked (or not linked at all) would raise several intriguing possibilities that merit further investigation. First, such a finding would suggest that lawmakers themselves might view these two types of cohesion as serving different purposes. Second, such a finding would have implications for how voters *perceive* party unity. Although parties have never been more cohesive when it comes to roll-call voting, voters observe both roll-calls *and* rhetoric. If what parties say is more diffuse than how they vote, voters may not think the parties are

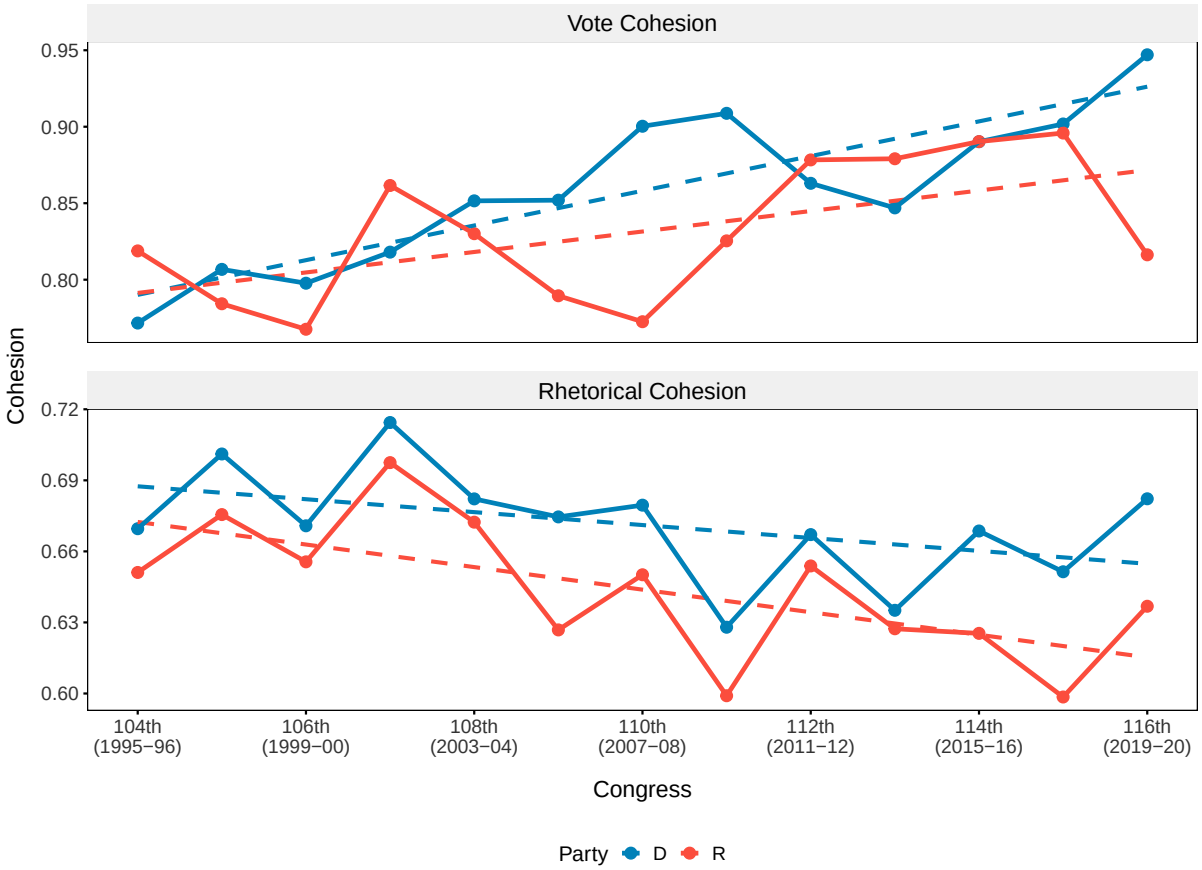
as unified as our roll-call based measures suggest.

At baseline, one should expect some positive correlation between these two measures. Lawmakers who promote similar messages on the floor should vote in the same way on roll-calls related to that issue. For example, two lawmakers who argue that "health care is a human right" should both vote for bills expanding healthcare access and against those limiting access. While we do not dispute this general proposition, we expect the relationship to be relatively weak given that rhetorical and roll-call cohesion may serve different purposes—especially as elite-level polarization and party unity have increased.

For rank-and-file lawmakers, roll-calls are highly salient. Defection can be costly to the member, who may be punished by the party (e.g., Hasecke and Mycoff 2007; Leighton and Lopez 2002) or primary opponents (Anderson, Butler and Harbridge-Yong 2020). Defection will also damage the party's collective reputation on which all co-partisans rely (Cox and McCubbins 2005; Lee 2009). In general, lawmakers have already made up their mind before a vote, often choosing to vote with their party, and rhetoric can provide cover for that decision (Edelman 1985). Party leaders care more about how members vote than what they say (Proksch and Slapin 2012), so speech provides a release valve for members to service constituents. And as Grose, Malhotra and Parks Van Houweling (2015) show, rhetorical presentation can increase legislator support, offsetting penalties that may materialize for voting against constituency preferences (Canes-Wrone, Brady and Cogan 2002) or in an overtly partisan fashion (Carson et al. 2010). Rhetoric also allows members to highlight ideological positions (Hassell, Heseltine and Reuning N.d.; Mayhew 1974), claim credit (Grimmer, Westwood and Messing 2014), and present themselves as a certain type of representative (Bernhard and Sulkin 2018; Grimmer 2013; Hill and Hurley 2002). Thus, what legislators say need not be a straightforward account of their voting behavior, but rather, serve as a justification for a decision they were going to make anyway.

To assess the relationship between these two types of cohesion, we construct a measure of roll-call voting cohesion at the pair-topic level for each Congress in our time series.

Figure 2: Rhetorical and Roll-Call Cohesion Over Time



Here, we use data from the Comparative Agenda's Project, which labels each vote on an HR bill according to the same 21-topic scheme as our keyATM model. We combine this data with information from VoteView, resulting in a dataset of 11,030 roll calls on 3,699 HR bills. For each vote, we assess whether two members of the same party voted in the same direction. We average this score across each topic and Congress in our data, resulting in a score for each pair of lawmakers ranging from 0 (always voting in opposite directions) to 1 (always voting in the same direction).

We visualize these two variables, roll-call and rhetorical cohesion, over time for each party in Figure 2. In the top panel, we see that each party's level of voting cohesion has generally increased between 1995 and 2020. This pattern is more consistent for Democrats

than Republicans, but even so, the linear trend is positive for both parties. However, this pattern contrasts markedly with that for rhetorical cohesion. Despite some modest increases and decreases over time, the general trend is flat—or possibly, even negative—for both parties. This relationship is especially surprising given that we use House floor speeches, which must be germane and are subject to many of the same agenda-setting and gate-keeping forces as the roll-calls used for the top panel. What's more, *Democrats* are consistently more rhetorically cohesive than Republicans. Although this comports with suggestions in Lee (2018), it is at odds with related text-as-data research using unidimensional measures that claim Republicans are more cohesive (Gaynor 2025; Russell 2018). Descriptively, these trends do not suggest a strong correlation between rhetorical and roll-call cohesion.

To more formally test the relationship between these two variables, we present two ordinary least squares models in Table 3. We begin, in column 1, with a model regressing pairwise roll-call cohesion on a series of key controls and fixed effects: the standardized negative absolute distance of a pair's DIME scores,[14] a pair's party, whether the pair is in the majority, whether the pair are presidential co-partisans, and both topic and Congress fixed effects. We cluster standard errors at the pair-level. In this first model, we exclude our measure of rhetorical cohesion to establish a baseline. Although this regression is not of particular interest in and of itself, it is interesting in comparison to the model in column 2, which includes our standardized measure of pairwise rhetorical cohesion as well as a control for the number of times the pair speaks on that topic in that Congress. This variable accounts for any potential relationship between frequency of speech and downstream rhetorical and roll-call cohesion. Here, we see that the coefficient on standardized rhetorical cohesion is positive and statistically significant, but the magnitude is quite small. A one standard deviation shift in rhetorical cohesion is associated with a 0.8 percentage point increase in vote cohesion. Standardized ideological similarity is five

[14]We use DIME rather than NOMINATE as we do not want to contaminate our analysis of roll-call votes using a measure of ideology created, in part, from these same roll-calls.

Table 3: Relationship Between Pairwise Rhetorical Cohesion and Pairwise Roll-Call Agreement

	(1)	(2)
Rhetorical Cohesion		0.008***
		(0.000)
DIME Similarity	0.041***	0.040***
	(0.000)	(0.000)
Republican	−0.040***	−0.039***
	(0.000)	(0.000)
Majority	0.059***	0.058***
	(0.000)	(0.000)
Presidential Co-Partisan	−0.005***	−0.005***
	(0.000)	(0.000)
Num. Speech Pairs		0.001***
		(0.000)
Fixed Effects		
Topic	✓	✓
Congress	✓	✓
Num.Obs.	835789	835789
R2 Adj.	0.209	0.212
R2 Within Adj.	0.117	0.120

Note: The dependent variable is the proportion of votes on a given topic in a given congress on which a pair of co-partisan lawmakers vote in the same direction. Coefficients come from ordinary least squares models with standard errors clustered at the pair-level.

times more predictive of roll-call cohesion. A standard deviation increase in DIME similarity is associated with a 4 percentage point increase in vote agreement. We also note that the adjusted R^2 increases by only 0.003 when we add rhetorical cohesion to our model.

The results in Table 3 are consistent with the idea that rhetorical and roll-call cohesion are weak complements. They are positively correlated, but the magnitude of rhetorical cohesion pales in comparison to known drivers of roll-call voting behavior, such as shared ideology or majority party status. Why is this the case? Although we cannot definitively answer the question here, we propose a few hypotheses that could be tested in future research. First, there may not be that much variation in the underlying roll-call record to exploit. Our time-series is defined by growing congressional polarization and the in-

creasing use of "unorthodox" lawmaking procedures (Sinclair 2017). As such, partisans have less discretion in how they vote than they did during the middle of the twentieth century. Within our time-series, our roll-call agreement score has a mean of 0.84 and an inter-quartile range of 0.79 to 0.94. Most of the time, co-partisans vote together. At an extreme, if all co-partisans vote the same way on every bill, then a pair of lawmakers who speak similarly will have the same pairwise roll-call agreement score as a pair of lawmakers who speak differently. Thus, we encounter a ceiling effect. If we were to analyze similar data during the middle of the twentieth century, we would expect the correlation between rhetoric and roll-calls to be higher. Second, lawmakers may not be striving for a high correlation between these two measures. That is, rhetoric and roll-calls may serve different purposes. Even if lawmakers have agreed upon a legislative course of action, they must persuade their constituents and justify their decisions—and different constituencies may require different explanations (cf. Grimmer 2013). A lawmaker in a rural district may generate support for legislation by highlighting the benefits to farmers while a lawmaker in an urban district may foreground its benefits to business interests. Perhaps we should expect divergence as lawmakers strategically tailor explanations to suit their constituencies.

Finally, we speculate that voters may not think parties are as cohesive as our usual methods suggest. We have shown that rhetorical cohesion on the floor has been flat or declining over time, even as pairwise roll-call agreement has increased. And this trend, as we show in the next section, is not a unique feature of the floor. Even if voters are not carefully attuned to floor debate (or even what lawmakers say on social media or in e-newsletters), media organizations are. Journalists prize conflict (Groeling 2010), and there is no shortage of headlines proclaiming that Democrats are in disarray or that Republicans are undergoing a Civil War. These types of stories are often prevalent before votes on major legislation like the Affordable Care Act or the One Big Beautiful Bill. These long periods of intra-party debate are heavily covered, even though they often precede unified

partisan votes. It seems unlikely that a single moment of partisan voting can outweigh months of coverage of intra-party debate and dissensus.

Exploring the Potential Effects of Procedure and Discipline

To this point, we have investigated rhetorical cohesion in floor speeches. We made this choice intentionally given our expectations that this venue would be the most constrained and the most policy-relevant. There, leaders set the agenda, and thus, the topics. They also have power to select who will speak, which may incentivize members to either toe the party line or stay silent. As Lee (2018, 1470) writes, "If Congress had no choice about the issues on which it took votes, the Republican Party would not look so cohesive." The same could be said for speech. To the extent that parties value rhetorical cohesion, and if leaders use procedural tools alongside carrots and sticks to enforce the rhetorical party line, rhetorical cohesion should be strongest on the House floor.

In other venues, leaders have weaker procedural tools and fewer opportunities to constrain member speech. For example, e-newsletters "are structured around home state and district concerns and are primarily meant for an audience of in-district constituents who have opted in to receive information from that specific member" (Green et al. 2024, 656). Here, any member can send a newsletter at any time on any topic. Leaders have less power (and perhaps, weaker incentives) to limit who speaks or what they say. The audience—supportive constituents (rather than elites who are targeted on the floor)—may also incentivize members to take on different voices (e.g., Fenno 1978; Grimmer 2013). Ultimately, we expect rhetorical cohesion to be weaker off the floor than on it given this loss of leadership control and audience diversity.

We apply our measure of rhetorical cohesion to a corpus of e-newsletters (Cormack 2023). Doing so gives us some leverage on the effects of procedural rules and leader-led discipline on rhetorical cohesion. If parties evince weaker rhetorical cohesion off the floor, that would suggest that discipline and procedure are doing some work in promot-

ing rhetorical cohesion. If on- and off-floor rhetorical cohesion are similar, that suggests little leader influence in constraining member speech. This comparative analysis also allows us to further interrogate our finding that Democrats are more rhetorically cohesive than Republicans. If Democrats are also more rhetorically cohesive in e-newsletters, that would suggest that previous measures have missed something. However, if Republicans are more cohesive than Democrats in e-newsletters, that would suggest Democrats are better at wielding procedural power and discipline, but that Republicans exhibit stronger preference overlap.

To compare rhetorical cohesion across these two venues, we download the text of all e-newsletters sent by House members from DC Inbox between 2010—2023 (Cormack 2023) and follow the same procedure and method described previously to measure party cohesion in this dataset. We plot average rhetorical cohesion at the party-Congress level in both newsletters and floor speeches for the overlapping period (2010–2020) in Figure 3. Consistent with our expectations, we see that rhetorical cohesion is higher in floor speeches (solid lines) than in e-newsletters (dashed lines) across this short time series for both parties. Although we do not provide causal evidence of discipline, this pattern is consistent with the idea that procedural tools may allow leaders to discipline members and promote a more cohesive party message than they otherwise would.

When we look specifically at e-newsletters, we see that Republicans are more cohesive than Democrats—the opposite of what we saw in floor speeches. Further, the cohesion gap between the floor and e-newsletters is large for Democrats and small for Republicans. Together, these patterns suggest that Republicans are resistant to the procedural power and discipline of leaders, but in general, the party has a stable level of alignment. Democrats are the opposite. Although they follow leaders in Congress, off the floor, they promote a much more diverse set of messages. These patterns are consistent with both conventional wisdom and academic accounts. For example, the 2020s have seen historic votes in the House Republican caucus over the Speaker, and major legislative initiatives

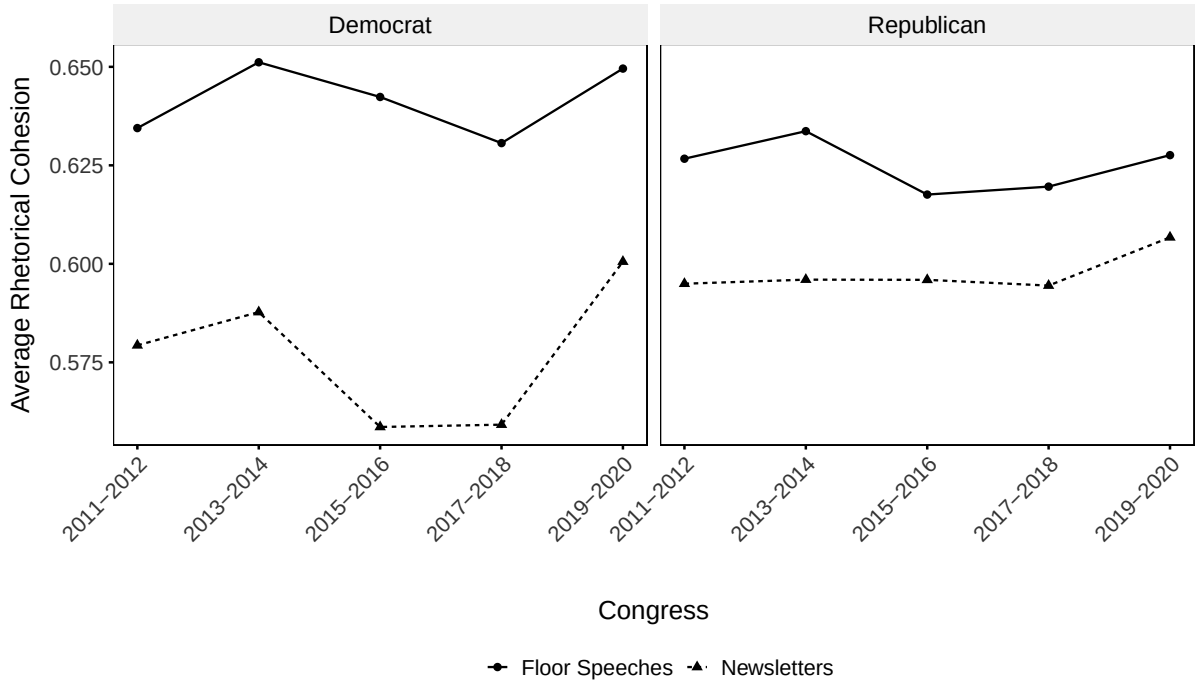


Figure 3: Comparison of party cohesion across data sources. The left panel reports results for Democrats and the right panel for Republicans. The x-axis indicates the congressional term. The y-axis shows the average cosine similarity of text pairs produced by legislators within the same party. Solid lines represent cohesion measured from floor speeches, while dashed lines represent cohesion measured from newsletters.

are often beset by rebellion and party infighting. The same cannot be said for Democrats, who tend to vote the party line and follow their leaders. This pattern also aligns with Lee (2018), who argues that House Republicans are much less cohesive than the roll-call record suggets. However, the off-floor patterns are more consistent with arguments about the party coalitions—with Republicans being much more cohesive and ideological as opposed to more group-oriented Democrats (Freeman 1986; Grossman and Hopkins 2016). Previous research has also found that Republicans are much better than Democrats at promoting leader talking points (Gaynor 2025) and partisan messaging (Russell 2018). Although these are suggestions provided by simple exploratory analyses, we believe this measure can be used in subsequent studies to further investigate the effects of discipline

and party coalitions on rhetorical cohesion.

Conclusion

How cohesive are the congressional parties? The answer depends on where you look—and how you measure it. Here, we have developed a novel measure of rhetorical cohesion using topic models and contextual text embeddings from large language models. Unlike existing measures, this approach accounts for the multi-dimensionality inherent in conventional understandings of similarity and cohesion. In the context of congressional rhetoric, we show that our measure both outperforms and incorporates other measures of rhetorical cohesion prevalent in the literature. We then use this measure to reveal a surprisingly weak correlation between rhetorical and roll-call cohesion, suggesting that the parties view these types of cohesion differently. We also apply this measure to a corpus of e-newsletters and show that while rhetorical cohesion is consistently weaker online, between-venue cohesion differs considerably within and between parties—suggesting some effect of party discipline.

Our research contributes to perennial debates about party cohesion in Congress with a novel measure of textual similarity. The unique value of our approach is that it can capture similarity across an arbitrary set of dimensions. This multidimensionality is advantageous in that it permits us to summarize the myriad ways that parties can cohere or diverge, without needing to choose a specific dimension or assume that a single dimension (e.g., ideology, text reuse, or presidential references) can reliably proxy for general party cohesion. As a result, our applications bring nuance to related arguments about the degree to which the roll-call record masks intra-party heterogeneity (Lee 2018) and arguments about lawmakers' presentational styles or types (Crosson and Kaslovsky 2025; Grimmer 2013; Bernhard and Sulkin 2018). Our exploratory analyses raise several questions about party differences, voter perceptions, and questions of discipline, which future

scholars should explore.

Although we have focused on party cohesion in Congress, our framework is incredibly general. This same workflow can be adapted to measure cohesion at different levels of aggregation—such as committees, factions or even within-member—or in different text sources, such as bills themselves, to understand hitchhiking and diffusion in cases where policy is adopted in spirit rather than verbatim. We also see this method as extending naturally to the study of inter-branch communication, administrative rulemaking, and the diffusion of court precedent. Finally, while we have focused on a single topic model and one set of pre-trained embeddings, these choices are customizable, allowing scholars to take advantage of the ever-increasing array of closed and open-source models, as befits their research question and goals. As language models continue to evolve and become increasingly prevalent in political science research (e.g., Lin 2025), we look forward to seeing how researchers apply these methods to shed new insight on longstanding questions in the discipline.

References

Adams, John. 1780. "From John Adams to Jonathan Jackson, 2 October 1780.". **URL:** *http://founders.archives.gov/documents/Adams/06-10-02-0113*

Ainsley, Caitlin, Clifford J Carrubba, Brian F Crisp, Betul Demirkaya, Matthew J Gabel and Dino Hadzic. 2020. "Roll-call vote selection: implications for the study of legislative politics." *American Political Science Review* 114(3):691–706.

Aldrich, John H. 1995. *Why Parties? The Origin and Transformation of Political Parties in America*. Chicago: University of Chicago Press.

Aldritch, John H., Mark M. Berger and David Rohde. 2002. *The Historical Variability in Conditional Party Government, 1877-1994*. Stanford: Stanford University Press p. 17–35.

Anderson, Sarah E., Daniel M. Butler and Laurel Harbridge-Yong. 2020. *Rejecting Compromise: Legislators' Fear of Primary Voters : Legislators' Fear of Primary Voters*. Cambridge: Cambridge University Press.

APSA. 1950. "Toward a More Responsible Two-Party System: A Report of the Committee on Political Parties." *The American Political Science Review* 44(3).

Arnold, R. Douglas. 1990. *The Logic of Congressional Action*. Yale University Press.

Ban, Pamela and Jaclyn Kaslovsky. 2024. "Local orientation in the U.S. House of Representatives." *American Journal of Political Science* .

Bernhard, William and Tracy Sulkin. 2018. *Legislative Style*. Chicago, IL: University of Chicago Press.

Broockman, David E. and Daniel M. Butler. 2017. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* 61(1):208–221.

Canes-Wrone, Brandice, David W. Brady and John F. Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members' Voting." *American Political Science Review* 96(01):127–140.

Carlson, David 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.

Carson, Jamie L., Gregory Koger, Matthew J. Lebo and Everett Young. 2010. "The Electoral Costs of Party Loyalty in Congress." *American Journal of Political Science* 54(3):598–616.

Casas, Andreu, Matthew J. Denny and John Wilkerson. 2020. "More Effective Than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process." *American Journal of Political Science* 64(1):5–18.

Clinton, Joshua D. 2007. "Lawmaking and Roll Calls." *The Journal of Politics* 69(2):457–469.

Cormack, Lindsey. 2023. "DCinbox.".
URL: *https://www.dcinbox.com/*

Cox, Gary W. and Matthew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge: Cambridge University Press.

Crespin, Michael H, David W Rohde and Ryan J Vander Wielen. 2013. "Measuring variations in party unity voting: An assessment of agenda effects." *Party Politics* 19(3):432–457.

Crosson, Jesse and Jaclyn Kaslovsky. 2025. "Do Local Roots Impact Washington Behaviors? District Connections and Representation in the U.S. Congress." *American Political Science Review* 119(2):887–904.

Curry, James M. and Frances E. Lee. 2020. *The Limits of Party: Congress and Lawmaking in a Polarized Era*. Chicago: University of Chicago Press.

Desmarais, Bruce A., Jeffrey J. Harden and Frederick J. Boehmke. 2015. "Persistent Policy Pathways: Inferring Diffusion Networks in the American States." *American Political Science Review* 109(2):392–406.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805* .

Duck-Mayr, JBrandon and Jacob Montgomery. 2023. "Ends Against the Middle: Measuring Latent Traits when Opposites Respond the Same Way for Antithetical Reasons." *Political Analysis* 31(4):606–625.

Edelman, Murray Jacob. 1985. *The Symbolic Uses of Politics*. University of Illinois Press.

Eshima, Shusei, Kosuke Imai and Tomoya Sasaki. 2023. "Keyword-Assisted Topic Models." *American Journal of Political Science* n/a(n/a).
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12779*

Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. Little, Brown.

Fiorina, Morris P. 1980. "The Decline of Collective Responsibility in American Politics." *Daedalus* 109(3):25–45.

Freeman, Jo. 1986. "The Political Culture of the Democratic and Republican Parties." *Political Science Quarterly* 101(3):327–356.

Gaynor, SoRelle, Kristina Miler, Pranav Goel, Alexander M. Hoyle and Philip Resnik. 2025. "Express Yourself (Ideologically): Legislators' Ideal Points Across Audiences." *The Journal of Politics* .

Gaynor, SoRelle Wyckoff. 2025. "Following the leaders: Asymmetric party messaging in the U.S. Congress." *Legislative Studies Quarterly* 50(1):85–106.

Gennaro, Gloria and Elliott Ash. 2022. "Emotion and Reason in Political Language." *The Economic Journal* 132(643):1037–1059.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78(1):35–71.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87(4):1307–1340.

Green, Jon, Kelsey Shoub, Rachel Blum and Lindsey Cormack. 2024. "Cross-Platform Partisan Positioning in Congressional Speech." *Political Research Quarterly* 77(3):653–668.

Grimmer, Justin. 2013. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press.

Grimmer, Justin, Margaret E. Roberts and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grimmer, Justin, Sean J. Westwood and Solomon Messing. 2014. *The Impression of Influence: Legislator Communication, Representation, and Democratic Accountability*. Princeton: Princeton University Press.

Groeling, Tim. 2010. *When Politicians Attack: Party Cohesion in the Media*. Cambridge University Press.

Grose, Christian R., Neil Malhotra and Robert Parks Van Houweling. 2015. "Explaining Explanations: How Legislators Explain their Policy Positions and How Citizens React." *American Journal of Political Science* 59(3):724–743.

Grossman, Matt and David A. Hopkins. 2016. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. Oxford University Press.

Hasecke, Edward B. and Jason D. Mycoff. 2007. "Party Loyalty and Legislative Success: Are Loyal Majority Party Members More Successful in the U.S. House of Representatives?" *Political Research Quarterly* 60(4):607–617.

Hassell, Hans J G, Michael Heseltine and Kevin Reuning. N.d. "Believing What Politicians Communicate: Ideological Presentation of Self and Voters' Perceptions of Politician Ideology." . Forthcoming.

Hazan, Reuven Y. 2003. "Does Cohesion Equal Discipline? Towards a Conceptual Delineation." *Journal of Legislative Studies* 9(4):1–11.

Hertel-Fernandez, Alexander. 2019. *State Capture: How Conservative Activists, Big Businesses, and Wealthy Donors Reshaped the American States–and the Nation*. Oxford University Press.

Hill, Kim Quaile and Patricia A. Hurley. 2002. "Symbolic Speeches in the U.S. Senate and Their Representational Implications." *The Journal of Politics* 64(1):219–231.

Hopkins, Daniel J. 2018. "The Exaggerated Life of Death Panels? The Limited but Real Influence of Elite Rhetoric in the 2009–2010 Health Care Debate." *Political Behavior* 40(3):681–709.

Hughes, Tyler and Gregory Koger. 2022. "Party Messaging in the U.S. House of Representatives." *Political Research Quarterly* 75(3):829–845.

Jones, Bryan D., Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Cheyenne Lee and Miranda E. Sullivan. 2023. "Policy Agendas Project: Codebook.".

Kaslovsky, Jaclyn and Michael R. Kistner. 2025. "Responsive rhetoric: Evidence from congressional redistricting." *Legislative Studies Quarterly* 50(3):e12473.

Krehbiel, Keith. 1993. "Where's the Party?" *British Journal of Political Science* 23(2):235–266.

Lauderdale, Benjamin E. and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24(3):374–394.

Lebo, Matthew J, Adam J McGlynn and Gregory Koger. 2007. "Strategic party government: Party influence in Congress, 1789–2000." *American Journal of Political Science* 51(3):464–481.

Lee, Frances E. 2009. *Beyond Ideology: Politics, Principles, and Partisanship in the U. S. Senate*. Chicago: University of Chicago Press.

Lee, Frances E. 2016. *Insecure Majorities: Congress and the Perpetual Campaign*. Chicago: University of Chicago Press.

Lee, Frances E. 2018. "The 115th Congress and Questions of Party Unity in a Polarized Era." *The Journal of Politics* 80(4):1464–1473.

Leighton, Wayne A and Edward J Lopez. 2002. "Committee Assignments and the Cost of Party Loyalty." *Political Research Quarterly* 55(1):53–90.

Lin, Gechun. 2025. "Using cross-encoders to measure the similarity of short texts in political science." *American Journal of Political Science* .

Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven, CT: Yale University Press.

McCarty, Nolan, Keith T. Poole and Howard Rosenthal. 2001. "The Hunt for Party Discipline in Congress." *American Political Science Review* 95(3):673–687.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781* .

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain and Jianfeng Gao. 2024. "Large language models: A survey." *arXiv preprint arXiv:2402.06196* .

Müller, Stefan and Sven-Oliver Proksch. 2023. "Nostalgia in European Party Politics: A Text-Based Measurement Approach." *British Journal of Political Science* p. 1–13.

Nicholson, Stephen P. 2012. "Polarizing Cues." *American Journal of Political Science* 56(1):52–66.

Noble, Benjamin S. 2024. "Presidential Cues and the Nationalization of Congressional Rhetoric, 1973–2016." *American Journal of Political Science* 68(4):1386–1402.

Osnabrügge, Moritz, Sara B. Hobolt and Toni Rodon. 2021. "Playing to the Gallery: Emotive Rhetoric in Parliaments." *American Political Science Review* 115(3):885–899.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014. Glove: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP pp. 1532–43.

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1):120–128.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A Political-economic History of Roll Call Voting*. Oxford University Press.

Porter, Rachel and Colin Case. N.d. "Measuring Policy Positioning in U.S. Congressional Elections." . Forthcoming.
URL: *https://osf.io/zhrmv*$_v$2

Proksch, Sven-Oliver and Jonathan B. Slapin. 2012. "Institutional Foundations of Legislative Speech." *American Journal of Political Science* 56(3):520–537.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1):209–228.

Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson and Jay J Van Bavel. 2024. "GPT is an effective tool for multilingual psychological text analysis." *Proceedings of the National Academy of Sciences* 121(34):e2308950121.

Reynolds, Molly E. and Naomi Maehr. 2024. "Vital Statistics on Congress.".
URL: *https://www.brookings.edu/articles/vital-statistics-on-congress/*

Rheault, Ludovic and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28(1):112–133.

Rodriguez, Pedro L. and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84(1):101–115.

Rohde, David W. 1991. *Parties and Leaders in the Postreform House*. Chicago, IL: University of Chicago Press.

Russell, Annelise. 2018. "U.S. Senators on Twitter: Asymmetric Party Rhetoric in 140 Characters." *American Politics Research* 46(4):695–723.

Schattschneider, E. 1942. *Party Government: American Government in Action*. New York: Routledge.

Schwarz, Daniel, Denise Traber and Kenneth Benoit. 2017. "Estimating Intra-Party Preferences: Comparing Speeches to Votes." *Political Science Research and Methods* 5(2):379–396.

Sellers, Patrick. 2009. *Cycles of Spin: Strategic Communication in the U.S. Congress*. Cambridge University Press.

Sinclair, Barbara. 2002. *Do Parties Matter?* Stanford: Stanford University Press p. 36–63.

Sinclair, Barbara. 2003. "Legislative cohesion and presidential policy success." *The Journal of Legislative Studies* 9(4):41–56.

Sinclair, Barbara. 2017. *Unorthodox Lawmaking: New Legislative Processes in the U.S. Congress*. CQ Press.

Slapin, Jonathan B. 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.

Snyder, James M. and Tim Groseclose. 2000. "Estimating Party Influence in Congressional Roll-Call Voting." *American Journal of Political Science* 44(2):193–211.

Wang, Yu. 2023. "Topic Classification for Political Texts with Pretrained Language Models." *Political Analysis* 31(4):662–668.

Witko, Christopher, Jana Morgan, Nathan J. Kelly and Peter K. Enns. 2021. *Hijacking the Agenda: Economic Power and Political Influence*. Russell Sage Foundation.

Wolbrecht, Christina, Brooke Shannon, E.J. Fagan, Bryan D. Jones, Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Cheyenne Lee and Miranda E. Sullivan. 2023. "Policy Agendas Project: Democratic and Republican Party Platforms.".

Yang, Kailai, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie and Sophia Ananiadou. 2023. "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis." *arXiv preprint arXiv:2304.03347* 4.

Yenduri, Gokul, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang et al. 2024. "Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions." *IEEE Access* .

Ying, Luwei, Jacob M Montgomery and Brandon M Stewart. 2022. "Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures." *Political Analysis* 30(4):570–589.

Zhang, Bowen, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li and Liwen Jing. 2023. "Investigating chain-of-thought with chatgpt for stance detection on social media." *arXiv preprint arXiv:2304.03087* .

Online Supporting Information:
A More Coherent Measure of Party Cohesion

A Measurement of Message Discipline

A.1 Keywords for keyATM Model

In Table A1, we present the set of topics and keywords used to fit our keyATM models. These keywords are the top 15 keywords ranked by tf-idf within the party platforms, treating each topic as a single document. The *parliamentary* and *other* topic keywords were generated by the authors.

Table A1: Topics and keywords used to fit the keyATM model

Category	**Keywords**
Agriculture	farm, rancher, farmer, ranch, export, commod, agricultur, fiber, grain, embargo, pariti, crop, livestock, food, wheat
Civil Rights	abort, disabl, gender, religi, sex, discrimin, religion, ballot, desegreg, vote, equal, segreg, reproduct, marriag, racial
Culture	art, artist, endow, film, museum, danc, leisur, opera, orchestra, theatr, scholar, heritag, writer, scholarship, music, cultur
Defense	nato, nuclear, missil, weapon, ballist, veteran, iraq, soviet, treati, troop, korea, allianc, deploy, vietnam, arm
Domestic Commerce	antitrust, merger, mortgag, gambl, dodd, lend, patent, sba, theft, conglomer, ftc, frank, consum, small, loan
Education	student, classroom, teacher, math, tuition, parent, read, academ, graduat, teach, elementari, english, childhood, bilingu, secondari
Energy	oil, gas, coal, solar, energi, nuclear, electr, petroleum, atom, geotherm, opec, decontrol, wind, fossil, ethanol
Environment	speci, pollut, emiss, wetland, superfund, toxic, air, carbon, greenhous, esa, soil, brownfield, wildlif, fish, habitat
Foreign Trade	export, trade, tariff, currenc, negoti, textil, reciproc, monetari, bilater, nafta, china, agreement, protectionist, gatt, foreign
Government Operations	postal, district, columbia, lobbi, census, elector, mail, servant, statehood, ballot, incumb, branch, candid, vote, sunset, usp

Health	medicar, medicaid, patient, hiv, healthcar, drug, coverag, nurs, diabet, mental, cancer, medic, prescript, diseas, health
Housing	homeownership, slum, mortgag, fha, rent, rental, urban, tenant, homeless, rural, fanni, freddi, mac, mae, neighborhood
Immigration	immigr, refuge, undocu, deport, visa, alien, reunif, english, amnesti, newcom, flee, asylum, citizenship, illeg, admiss
International Affairs	israel, africa, soviet, taiwan, palestinian, east, arab, cuba, peac, korea, terrorist, ireland, asia, afghanistan, cuban
Labor	overtim, hartley, taft, pension, bargain, picket, employe, bacon, davi, collect, worker, arbitr, autom, osha, union
Law and Crime	gun, crime, crimin, drug, sentenc, offend, firearm, juvenil, polic, prison, victim, narcot, pornographi, traffick, marijuana
Macroeconomics	deficit, inflat, monetari, bracket, spend, debt, incom, wealthi, wealthiest, recess, taxat, loophol, inflationari, estat, code
Public Lands	puerto, indian, rico, guam, forest, nativ, hawaiian, tribal, virgin, samoa, mariana, tribe, miner, park, wilder
Social Welfare	welfar, parent, needi, nutrit, stamp, social, elder, child, recipi, disabl, lunch, older, charit, mother, poverti
Technology	space, nasa, broadband, internet, broadcast, scientif, telecommun, orbit, saturn, spacecraft, satellit, scienc, cyber, entertain, media
Transportation	highway, railroad, merchant, passeng, rail, freight, airport, transport, mode, maritim, congest, traffic, amtrak, marin, truck
Parliamentary	yield, gentleman, consent, amend, time, minut, senat, hous, bill, order, thank, committe, move, vote, quorum, motion, tabl
Other	peopl, go, get, got, laughter, know, thing, want, say, think, thank

A.2 Pre-Processing

To pre-process our speeches, we first subset to speeches with more than thirty words (following Noble 2024). We also tokenize to unigrams, remove non-text characters, lowercase words, remove a set of stop words (those listed as stop words in the `quanteda` package) and those with fewer than three characters, stem words, and remove words that appear fewer than 20 times or across fewer than 15 documents in each corpus.

A.3 Topic Validation

In Table A2, we provide additional summary statistics on our Optimal Label topic model validation task.

Table A2: Human Coder Validation Statistics

Statistic	**RA**	**Author**
Total Speeches Assigned	105	105
Correct Classifications	85	84
Overall Accuracy (%)	81	80
Inter-Rater Reliability		
Percent Agreement (%)	78.1	
Cohen's Kappa	0.77	
Kappa p-value	< 0.001	

Notes: Validation based on 105 manually coded speeches compared against the keyATM top-1 topic labels. Percent agreement calculated as proportion of speeches where both coders provided identical classifications.

In Table A3, we break down our OL accuracy by topic category.

Table A3: Category-Specific Accuracy by Coder

Topic Category	**N**	**RA (%)**	**Author (%)**
Agriculture	5	100	100
Civil Rights	5	100	100
Culture	5	100	80
Education	5	100	100
Environment	5	100	100
Health	5	100	100
International Affairs	5	100	80
Social Welfare	5	100	80
Defense	5	80	80
Domestic Commerce	5	80	80
Energy	5	80	100
Foreign Trade	5	80	60
Government Operations	5	80	80
Labor	5	80	60
Law & Crime	5	80	100
Macroeconomics	5	80	80
Technology	5	80	100
Housing	5	60	40
Transportation	5	60	60
Immigration	5	40	60
Public Lands	5	20	40
Overall Accuracy	**105**	**81**	**80**

Notes: Category-specific accuracy rates for each of the 21 topic categories. Each category contains exactly 5 speeches in the validation sample. Accuracy calculated as the percentage of speeches correctly classified within each category.

A.4 Construct Validity: GPT-Similarity Validation

A.4.1 Comparsion Sampling

We sampled focal speeches and paired comparison texts across our time series subject to a few constraints. To validate within-focal speech similarity scores, not just across-focal speech similarity, we ensured that each focal speech appeared multiple times with different sets of comparison speeches. We also wanted to assess whether our RA could discriminate between speech pairs that were more or less similar. To that end, we calculated the standard deviation of GPT-similarity scores across our corpus (0.14) and then determined, for each focal speech, how large was the cosine-GPT difference between each of the comparison speeches. We grouped comparison tasks into buckets based on the size of this difference, such as below 0.25 of a standard deviation, 0.25–0.50 of a standard deviation, up to greater than 2 standard deviations. We sampled approximately 30 comparison tasks per bucket for a total of 260 tasks.

A.4.2 Example Task

Below are the instructions we provided to the RA in how to think about similarity when selecting speeches:

> Your task is to help validate an automated measure of similarity between political speeches. Below, you'll be shown a series of sets of two **comparison speeches**. For each pair of comparison speeches, choose the one that you believe is **more similar** to the **focal speech** above.
>
> **How to think about similarity:**
> Similarity in this context is **broad and multidimensional**. There is no single correct criterion. Instead, consider the following dimensions when making your judgment:
>
> - **Topic:** Are the speeches about the same issue or subject?
> - **Position:** Do the speakers take similar stances or make similar arguments?
> - **Language:** Do they use similar words or phrases?
> - **Tone/Sentiment:** Is the emotional tone or attitude similar (e.g., optimistic, angry, urgent)?
> - **Style:** Do they share rhetorical features like repetition, emphasis, or structure?

You do **not** need to agree with the speaker or analyze the truth of the content—just assess overall similarity in meaning, purpose, and presentation.

There will often be some overlap across both comparison speeches. Use your best judgment to select the one that feels closer **overall** to the focal speech. There are no trick questions—your intuitive assessment is valuable for this task.

Please note: sometimes, you will encounter a repeated comparison speech paired with a different comparison speech. That is expected. Just keep choosing the most similar one.

Then, the RA was presented with a series of tasks like the one below.

Parimary Speech (1120072647): mr. chair. i rise today in strong opposition to the socalled ”workforce democracy and fairness act” . the changes to union election procedures promoted in this bill are the exact opposite of the kind of fair and democratic policies that our working families need. instead of focusing on job creation and the revitalization of our middle class. the republicans in this chamber are once again promoting legislation that undermines the rights of american workers. this proposed legislation would limit the ability of the national labor relations board to interpret our nations labor laws and to protect workers right to unionize. for over 75 years. the national labor relations act has guaranteed the rights of employees to organize and bargain collectively. or to refrain from such activity if they choose. during the new deal. our predecessors in this body created the national labor relations board as an independent agency charged with the oversight and enforcement of these rights. h.r. 3094. which overturns the rulings of the nlrb. undermines its charge to maintain fair and democratic relationships between unions and employers. this legislation allows the problem of prolonged delays in union elections to continue unchecked by adding mandatory and arbitrary waiting periods. it seizes from workers the right to determine their own representative membership groups. which would allow unscrupulous businesses to suppress election drives and vote down union representation. it would also make it possible for irresponsible and frivolous litigation to endlessly delay the election process. effectively barring workers from their fundamental right to collective bargaining representation in the workplace. supporting and

protecting americas workers is an essential part of rebuilding our economy and ensuring that all families and communities share in our nations prosperity. our middle class was built on the rights and safeguards that labor unions fought to obtain. from the 40 hour workweek to ending child labor. union representation has helped to guarantee rights that many of us take for granted today. unions negotiate for safe working conditions. living wages. and basic benefits that impact all workers. efforts to decrease the power of collective bargaining in this country in recent decades have been accompanied by an erosion of workers benefits and greater income inequality. this year in wisconsin and ohio. we have seen voters reject recent attempts to strip away the rights of government workers. and we should likewise reject this attempt to limit access to these rights for those in the private workforce. this bill does nothing to protect and support working families. and i urge my colleagues to stand up for workers rights and oppose this bill.

Comparison A (1120072704): i yield myself such time as i may consume. what i say to my good friend from south carolina is that i have the greatest respect for employers. id like the gentleman to join me in passing the american jobs act to give them payroll tax relief and to give them tax credits for hiring new employees. but you have to ask the question: after this bills implementation. will workers view their workplaces more favorably? will their wages match the growth rates of the companies and economy? will workers feel like american employers. supported by government. provide meaningful safety for community survival? this legislation. frankly. undermines the american workers. can we all get along? can we find a way to address the concerns of making sure that we are fair to the employer but not have delay after delay after delay to deny someone his constitutional right of organizing freedom of expression? i think we can. the elimination of the provisions that i have spoken of is a dilatory upper hand of employers to get the better hand of our employees. i reserve the balance of my time. the acting chair. the gentlewoman from texas has 15 seconds remaining. and the gentleman from south carolina has 45 seconds remaining.

Comparsion B (1120072655): mr. chair. i rise today in opposition to h.r. 3094. the republican plan to crush workers rights and destroy any glimmer of hope our working families have at economic recovery. the republicans designed this bill to destroy 75 years of national labor review board case law in their attempt to

dismantle the middle class. collective bargaining and the right to organize helped build a strong american middle class. it doesnt cost the federal government one dime in real money. instead of taking steps to create jobs and strengthen working families. republicans are dismantling key worker protections. all workers should have the ability to negotiate with their employer about salary and benefits. whether theyre in a union or not. organized labor is great for business. thousands of companies across the country thrive with a unionized workforce. those businesses recognize that their employees deserve to have a safe workplace and fair wages and benefits. thats just good business. this bill encourages corporations to stall nlrb elections while they mount a onesided. antiunion campaign. at its core. this is an undemocratic bill that undermines our values. we have a long established process for workers to attempt to form a union and collectively bargain with employers. employers and employees should stay on equal ground in the process. there is no need to deny workers their right to a free and fair union election. many of my republican friends like to talk about the issue of tort reform. they like to tell us that we have to prevent frivolous lawsuitsthey cost taxpayers millions and millions of dollars and they drag down the economy. i have news for my republican friends: the election prevention act encourages frivolous litigation. this bill will mean mountains of litigation before union elections can be held. the result is a massive backlog. guess who picks up the tab? the american taxpayer! we have important issues facing our country and it boggles my mind that we are taking up yet another bill that does nothing to get our friends and neighbors back to work. we need to focus on lowering the unemployment rate and creating jobsnot taking away the rights of hardworking americans. i urge my colleagues to recognize this veiled attempt to destroy the rights of american working families.

Which comparison speech (A or B) is more similar to the primary speech?
A: 1120072704
B: 1120072655

Here, the RA and embedding-based choice was comparison B. As can be seen on a careful reading, comparison B is more similar to the focal speech for several reasons. First, although all speeches concern relationships between capital and labor, the focal speech and speech B explicitly discuss the NLRB whereas comparison A does not. Both the focal speech and comparison B also take a clear pro-labor position, whereas comparison A

takes a more balanced approach favoring both employers and labor. Both the focal speech and comparison B are emotional, making using of hyperbole and extreme claims such as "irresponsible and frivolous litigation to endlessly delay" (focal speech) and "destroy any glimmer of hope our working families have" (comparison B). Comparison A makes use of a different rhetorical strategy, questioning. For these reasons, it is clear why both the RA and model selected comparison B.

A.4.3 Bucket Accuracy

In Table A.4.3, we break down the accuracy of comparisons tasks by standard deviation bucket. For each focal-comparison pair, we note the GPT cosine similarity. Within a comparison task, we determine how close these two pairwise differences are as compared to the standard deviation of differences across our full dataset. For example, if the pairs within a comparison task have a cosine similarity difference of less than 0.035, this is a very small difference and they fall into the first "bucket."

This table demonstrates that as the difference between the two pairs grow, the RA and embedding-based cosine similarity are more likely to select the same comparison speech. When differences are small, accuracy decreases. These results suggest that humans may struggle to interpret small differences between speeches that our embedding method captures. This could result from either the embedding method detecting imperceptible differences or error in the machine measurement based on differences that are not relevant when humans process similarity.

Table A4: GPT-RA Agreement Rates by Standard Deviation Bucket (0.25 SD intervals)

Std. Dev. Bucket	Num. Responses	Agreement Rate (%)
0 to 0.25	32	65.6
0.25 to 0.50	30	73.3
0.50 to 0.75	25	68.0
0.75 to 1.00	25	72.0
1.00 to 1.25	31	83.9
1.25 to 1.50	25	84.0
1.50 to 1.75	29	86.2
1.75 to 2.00	31	90.3
over 2.00	32	100.0