Legislative Studies Quarterly
SPECIAL ISSUE The Politics of Congressional Reports

ORIGINAL ARTICLE

Partisan Knowledge Claims in Congressional Oversight

Kenneth Lowande[1] | Mark A. Weiss[2]

[1]Department of Political Science, Rice University, Houston, Texas, USA | [2]Department of Political Science, Ford School of Public Policy, University of Michigan, Ann Arbor, Michigan, USA

Correspondence: Kenneth Lowande (lowande@rice.edu)

Received: 14 June 2025 | **Revised:** 5 May 2026 | **Accepted:** 9 June 2026

Keywords: committees | congress | epistemology | information | oversight

ABSTRACT

Congress negotiates while uncertain about the effects of proposed policies, and often relies on committees to conduct oversight that reports about the past and projects the future. Little is known about their conclusions, their accuracy, and how politics seeps into what they choose to report. We apply a typology of knowledge claims contained in congressional oversight, which classifies them based on whether they are prospective or retrospective, and whether they contain an inference. We use a large language model (LLM) to extract findings, evaluations, predictions, and recommendations from a new dataset of committee reports issued in the 103rd–118th Congresses. We argue that partisan investigation teams reveal more raw findings and make fewer recommendations, relative to bipartisan teams, which is consistent with their need to demonstrate they have acquired expertise. Our study pilots a general approach to studying the acquisition and sharing of information in legislative oversight.

Oversight of public policy is hard. Members of Congress are uncertain about the effects of policies, both those enacted in the past and proposed for the future. To reduce that uncertainty, they rely on committees and their staff, who specialize and are charged with gathering information. In practice, gathering that information is messy. Most of the time, committees divide their voice, rather than form a bipartisan team. Moreover, the information committees produce is complex. Committees gather many facts, and much like social scientists, make causal inferences and out-of-sample predictions in order to arrive at actionable recommendations.

In this study, we argue that the knowledge claims of committees are conditioned by politics. We ask why members and staff would exert costly effort to compose an oversight report and reveal information to the rest of Congress. In our view, committees want their work to inform policy change, either in the legislature, the private sector, or in the Executive Branch. Partisan teams know their work is less credible to the floor both because of their failure to put together a coalition and their more limited staff resources. So they tend to publish more underlying findings of fact and make fewer explicit recommendations.

To evaluate this argument, we analyze a new dataset of House and Senate committee reports published in the 103rd to 118th Congresses. Many studies examine the composition of committees (e.g., Groseclose 1994; Provins et al. 2022), the frequency of hearings and members' conduct in them (e.g., Kriner and Schickler 2014; Bussing and Pomirchy 2022; Park 2021; Eldes et al. 2024), the amendment process (e.g., Curry 2015, 2019), and more. Yet, committee reports, which are supposed to transmit the summary views of committee members, have been largely ignored. We also propose and implement a typology of knowledge claims. Using a large language model, we extract findings, evaluations, predictions, and recommendations from committee reports. Findings of fact are simple

Previous versions were presented at the 2025 annual meeting of the Midwest Political Science Association, Chicago, IL and at the CORD Workshop in Detroit, MI. We thank Matt Dull, Christian Fong, Jeff Jenkins, Dave Rapallo, and Sean Theriault for helpful comments, as well as Alexis Castellanos for research support. We also thank Abbie Paegert, Saloni Singh, and Alan Andrade for helpful research assistance. Finally, we thank Elise Bean. This study is one of many inspired by her lifelong dedication to congressional oversight.

conclusions about the past, whereas evaluations are inferences about the past. Predictions involve inferences about the future, whereas recommendations are simpler, normative statements about what should be done. This novel measure indexes the types of claims committee authors are willing to make public.

Our analysis shows that politics is an important determinant of the kind of information published by Congress' committees. In general, committees ground their reports in findings of fact, and less often make inferences or recommendations for the future. However, bipartisan teams tend to be more prospective, containing, on average, an additional recommendation relative to partisan teams. Despite their access to less staff resources, partisan teams working alone make 1.5 additional descriptive findings public relative to their bipartisan counterparts. These results persist under different modeling assumptions, after accounting for political conditions that mark inter-branch conflict, like divided government.

Our results reveal an important dynamic in how Congress informs itself: the nature of the knowledge claims communicated by teams of lawmakers depends on internal disagreement. This opens important questions about how Congress processes and uses information. While legislative studies have largely focused on procedural nodes that terminate in a vote, our framework presents an opportunity to study how legislators come to consensus on particular knowledge claims, along with whether and how those claims are supported by evidence. Most importantly, with time, subsequent work might examine the reliability of these claims. Predictions that are falsifiable might later be proven wrong, and recommendations may or may not be adopted by their intended target. By making knowledge claims the object of analysis, and taking advantage of contemporary developments in LLMs, this research opens new opportunities in the study of legislatures—and of political institutions, writ large.

1 | Information and Policy Evaluation in Congress

Uninformed policies are typically bad polices. Congress needs information to pass laws and oversee their implementation. To acquire and use this information, Congress typically relies on committees with specialized jurisdictions and scope. These committees exert costly effort to get informed and allow their conclusions to be leveraged by the rest of Congress. A lot of scholarship investigates the strategic considerations that arise from this kind of organization.

Most importantly, knowing more than the rest of Congress might be tempting to exploit. Committees might misrepresent the truth to get the policies they want (Fenno 1973; Kiewiet and McCubbins 1991). One solution is to ensure committees are made up of members with diverse preferences, and to the extent that Congress tolerates extremists, it should do so only when it will benefit from their superior expertise (Gilligan and Krehbiel 1987; Krehbiel 1991). This way of doing things, however, requires diverse committees to find consensus. When committees draft and mark up bills, this consensus is arrived at mechanically. Their work product is a bill advanced to the floor.

Oversight, on the other hand, is less straightforward. Oversight of the executive branch and the private sector takes many forms, is often conflictual, and typically does not terminate in the drafting of a bill. Committee conflict is often on display, for example, in open hearings. It is the most obvious evidence that coming to consensus involves fundamental disagreements about the facts, how to frame them, and their implications. Eldes et al. (2024) show, for example, that members of the opposition party are more confrontational in their questioning of witnesses. Out-party legislators tend to make fewer falsifiable statements and engage in more rhetorical bluster (Park 2021). The likely goal is press that may aid their re-election prospects (Park 2023). By making a public display or "going viral," they might attract positive press that aids their own re-election. Indeed, alongside the nuts and bolts of obtaining information from the target of an investigation, training in oversight for legislative staff tends to emphasize ways to cultivate good press (Fong et al. 2024).

Hearings themselves may be tactical engagements in larger conflicts between the president and the opposition party (Kriner and Schickler 2014; Lowande and Peck 2017). But committee hearings are only the most dramatic part of a larger oversight effort. Earlier in that process, committees and individual members request information from bureaucratic agencies (Lowande 2018, 2019; Ritchie 2018), as well as pre-interview and decide on witness lists (Ban et al. 2021). All of the above involves observable behavior at some stage of the committee's oversight process. But it leaves out the last stage: packaging and transmitting the committee's conclusions.

Committees produce reports, sometimes hundreds of pages long, presenting their top-line conclusions, along with supporting reasoning and evidence. Committee staff might spend years on such reports, which are built on hearing transcripts, private inquiries, witness interviews, expert consultations, and more. These reports are to hearings and oversight as roll-call votes are to bill mark-up and legislating. Yet, there are important differences.

First and foremost, oversight reports are not themselves instruments of changing policy. They are not bills and will never be enacted into law. Moreover, nothing requires committees to compose a single oversight report to release to the rest of Congress, and they often choose not to. Committees are made up of members of both the majority and minority party. The final work products might not represent the median of the committee, but instead, the medians of each partisan team.

Though oversight reports are not explicitly legislative, they are required to serve a "legislative function" (Levin and Bean 2018). The act of oversight might garner members' publicity, but its core purpose is to help members influence the ultimate direction of policy.[1] This happens via several channels. Most obviously, oversight might identify problems and solutions which can be addressed in legislation at some later date. Less formally, oversight might lead other actors to voluntarily change their behavior without a legislative fix. Executive branch officials might agree to change patterns in enforcement, revise guidance documents, spend money differently, or be influenced in the way they write rules. Likewise,

in the private sector, businesses might discontinue practices members of Congress frame as predatory, like revising fee schedules for late credit card payments.

In summary, when committees conduct oversight, they might produce multiple work products, representing different views—and at best, their efforts only indirectly influence policy change. In this way, oversight reports complicate standard views of committee influence and behavior. If legislative oversight was entirely about the consumptive act of holding public hearings, it is not clear why members would go to the trouble of producing a report. On the other hand, if these reports were about transmitting information to the rest of Congress, the way they are produced seems to prevent them from becoming a clear, informative signal. This leads to the questions we address in the following sections. Why do oversight reports exist, and *how do they help committees sway the direction of policy?*

2 | Why Committees Report the Way They Do

We see two ways oversight reports might help members accomplish their goal of influencing policy. If oversight only influences policy indirectly, the key task is convincing either the rest of Congress or the target of the investigation that the committee's recommendations are well-informed and not merely self-serving. Reports do this, first and foremost, by their list of authors. Most obviously, bipartisan reports signed by the full committee demonstrate a minimum level of partisan consensus. They also show the oversight conducted by the committee has leveraged its full resources and expertise.

Of course, all reports are not bipartisan, which comes about for different reasons. One party may take issue with the selection of a particular target for investigation. They may decline, for example, to participate in an investigation of a co-partisan president occupying the White House. Moreover, during the course of an investigation, one party team may develop divergent views about how to interpret the facts and evidence. Either of these paths can lead to the oversight reports with a single party signing on. Not surprisingly, this should lead the opposition party to be more willing to sign on to investigations of the incumbent president's administration. Patterns in the partisanship of reports should mirror patterns seen in most of the other oversight activities that feed into them (e.g., holding hearings, deposing witnesses, etc.).

But these familiar patterns pose an important, unsettled issue. On the one hand, the reports might be similar to what has been supposed about oversight hearings and the questioning of witnesses. That is, they might be public-facing documents designed only to make a symbolic splash—just one more tool in the seemingly dominant "messaging" game played by both major parties (Lee 2016). If that is the case, partisan reports should be characterized by the same strategies seen in events like confrontational hearings—grandstanding, non-sequitur, and non-falsifiability.

On the other hand, reports could be useful for the internal operations of Congress; they might be meant to help members of the committee influence policy. But when the oversight team is not bipartisan, their audience, whose behavior will determine whether they are influential, does not know if the team has acquired expertise. Moreover, the issue cannot be resolved by the simple act of writing a report, because the other party may do the same.

Instead, we argue that members demonstrate their effort by the *way they choose to compose* their reports. Oversight reports differ in the kinds of knowledge claims they make. Suppose, for example, a committee chooses to conduct an investigation of the Strategic Petroleum Reserve (SPR). It requests and obtains information from the Department of Energy (DOE), interviews witnesses, consults outside experts, sends its staff down to West Hackleberry, Louisiana, where brine caverns have been converted to store barrels of oil in the tens of millions. The committee now has a mound of raw information and must distill its work for an audience who did not do the investigation or watch the committee do its work and cannot verify the quality of the report with its own information.

The report might contain headline conclusions that simply summarize facts obtained by the committee—for example, that the total reserve contains about 700 million barrels of crude oil. Alternatively, the committee might take a step further, and draw an inference about policy. It could say that current crude levels in the SPR are the result of the DOE's severe energy supply interruption policy (42 U.S.C. 77 §6202). It could take a more speculative tact, and say that if such policies continue, the levels will dip below tolerable thresholds. Finally, it could recommend a policy change. Each of these examples involves a choice about how to compose a report, and we argue, those choices reveal something about the work the reports' authors have (or have not) done.

To put it more generally, we argue that committees report one of four types of knowledge claims: findings, evaluations, predictions, and recommendations. A "finding" is the most basic conclusion a committee could contain in a report, mimicking findings of fact in the legal profession, due partly to the fact that many staff authors of committee reports are lawyers.[2] Findings, by our definition, are the equivalent of revealing some discrete piece of private information the committee uncovered (e.g., Callander 2008).

As we lay out in Figure 1, other categories of conclusions come from the answers to two distinct questions. First, is the conclusion about the *past or future*? Second, does the conclusion *make an inference*? Simple findings, by our definition, refer to the past and do not make inferences. To return to our hypothetical, "the SPR contains 700 million barrels of oil" is a finding. But the committee could also state that this amount was observed because of a policy of the Biden administration. The "because" marks an inference about a particular policy and implies a counterfactual that had this policy not been adopted, the level of oil would be different. We term this an "evaluation." It is a retrospective assessment of a policy, a causal inference about the past.

Both findings and evaluations deal with the past, but there are analogous claims about the future. The committee could project forward, arguing that if some policy is changed (or remains in place), we can expect to see the level of oil reach some new, different level. This is a "prediction." Again, it involves an inference because it implies a counterfactual projection in which the policy is not in effect. Finally, the committee can draw a conclusion

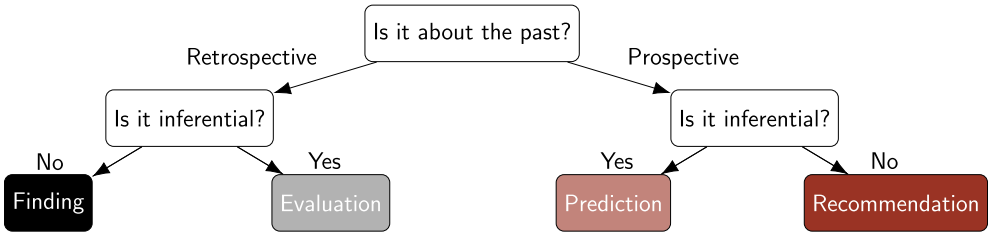

FIGURE 1 | Knowledge claims and information transfer.

about the future that does not involve an inference. It can make a "recommendation" or a normative statement about an action or policy some external body—in the executive branch, private sector, or the rest of Congress—should adopt. Reports differ in length, topic, author, and more, but all should feature knowledge claims of one kind or another.

The next question is, then, how should we expect the kinds of claims to differ, given the strategic problem for both author and audience? Again, if the reports themselves are meant to be informative, we argue there should be a few basic patterns in their claims.

Reports, overall, should tend to be more retrospective than prospective. This falls from their audience's expectations. What do they expect? That the team has conducted oversight, which almost by definition is concerned with the past. Deviating too far toward prospective predictions and recommendations, then, might give the audience the impression that they have strayed from the work of an oversight committee or investigation.

Second, audiences know that reliable knowledge claims tend to be built on each other. "Do not attempt to install a new plumbing fixture in your upstairs bathroom" is a simple recommendation. But it is based on the prediction that if you do, you might flood your first-floor dining room. This prediction is supported by an evaluation: a previous flood was caused by an amateur deciding to try their hand at plumbing installation. Moreover, our recommendation, prediction, and evaluation are collectively based on the observation (or "finding") that neither of you is a plumber, and that gravitational forces pull water toward Earth.

If you trust the source of the recommendation, you might ask no further questions. But if you are uncertain, and you want to know whether the original recommendation is something you should listen to, there is a good chance you want to know what the recommendation is based on. Report authors know that audiences will expect this general structure. Revealing the findings on which some recommendation is based tends to enhance their credibility to an audience. It also reveals that the team has actually conducted the investigatory work connected to their suggestions and acquired the right expertise. Thus, if committee authors want their work to be used, they will tend to include more claims about the past than the future.

This dynamic, then, naturally extends to the strategic problem for partisan authors. Bipartisan reports can rely on the identity of their authors to demonstrate to their audience that the committee is informed. Partisan teams cannot. But, in the absence of that consensus, oversight teams can demonstrate their effort by the kinds of claims they reveal in their reports. In particular, they can include more retrospective claims, rather than prospective, relative to bipartisan teams. Similarly, the partisan team whose views diverge most significantly from the audience should be the more retrospective, relative to the team who is in relative agreement with the audience. (The former is generally the minority party, while the latter is the majority.) Even if the reports are "solo-authored," audiences are predisposed to accept recommendations from oversight teams that they know they agree with.

The above expectations assume some kind of limitation of the allocation of effort.[3] Otherwise, findings would not serve the purpose of demonstrating the team has acquired expertise. It could be that collecting and including findings in a report is costly. These costs must be sufficient enough that bipartisan teams would prefer to trade on the identity of their authors rather than display the same efforts. Alternatively, this same budget constraint might be imposed by beliefs about the ideal length of a report. There is no formal limitation on the length of a report. But, in practice, authors emphasize the importance of executive summaries and the first pages for a reason: these are the pages most likely to be read. That is, bipartisan teams, knowing their authorship establishes their expertise, might prefer to allocate more prominent space in their reports to recommendations that might influence the behavior of targets of investigations or future legislation.

In summary, if oversight reports are meant to establish the expertise of overseers, with the aim of influencing the ultimate direction of policy, it should be reflected in their knowledge claims. All should tend to be more retrospective than prospective. Bipartisan teams should be more prospective than retrospective, relative to partisan teams. And partisans with preferences that diverge from the rest of Congress should reveal more findings in their reports, relative to partisans who are aligned with the floor.

3 | Data and Measurement

To evaluate the nature of messages transmitted by committees, we studied variation in the kinds of knowledge claims they chose to include in official reports. We did this by mining a new dataset of Congressional oversight reports from the 103rd to 118th Congress, compiled by researchers at the Levin Center for Oversight and Democracy.[4] As of this writing, these include 896 total reports produced by 25 committees in the House and Senate. These reports serve as the final step in official oversight activities and are distinct from filings made in connection with legislation. Because they all involve oversight, many reports were produced by the two committees with jurisdiction, namely,

the Senate Committee on Homeland Security and Governmental Affairs (HSGAC) and the House Committee on Oversight and Accountability (COA), along with their subcommittees and individual members. Beyond the content of each report, the data also contain information on the report's author, publication date, and author partisanship.

These data offer several advantages. Most importantly, they allow us to examine the conclusions of investigators, rather than the investigative activity itself. The data are also extensive, spanning 1994–2024, meaning that there are reports published during presidential administrations of both major parties as well as during periods of unified and divided government.

One immediate question is whether data that comes primarily from two committees is indicative of information transfer in Congress, in general. We think they are indicative of a particular kind of information gathering and transfer. Since these reports concern oversight, the dataset excludes all reports tied to a specific piece of legislation. Most obviously, when committees author reports that concern future legislation, it is reasonable to assume that their tendency will be to focus on prospective, rather than retrospective information. Thus, the baseline descriptive figures we present later would likely differ significantly.

It is also possible that consensus-building in the very public process of passing legislation is different from consensus-building in the very private process of report-drafting. Our results therefore only apply to the theoretical questions we have posed about the oversight of policy. The data also pose a basic research challenge, which we address in the following section: many reports are hundreds of pages long, which makes them difficult for humans to parse with speed, accuracy, and reproducibility.

4 | Identifying Knowledge Claims

Our basic measurement challenge is to extract and classify informational conclusions contained in oversight reports. To accomplish this, we turned to a large language model (LLM), specifically, Open AI's Chat Generative Pre-Trained Transformer 4o (Chat GPT 4o). A rapidly growing body of work has shown that LLMs are adept at performing repetitive tasks often delegated to research assistants. Most importantly, they can quickly and accurately parse, analyze, and annotate large amounts of text data (Gilardi et al. 2023; Gatt and Krahmer 2018; Bail 2024; Ziems et al. 2024; Heseltine and von Hohenberg 2024). Social scientists have uncovered many applications of LLMs to the study of politics, including ideal point estimation (Wu et al. 2023; Napolio 2025) and hate speech detection (Hu and Collier 2024).

Beyond their substantive use, LLMs offer practical benefits. They allow us to complete time-consuming and expensive tasks quickly and affordably. Our sample data contain over 72,000 pages of text and figures, much of it written in legalese. Several oversight reports themselves contain hundreds of pages. While perhaps a triumph for good governance, this thoroughness means that coding the data manually would require a nontrivial investment in training and employing a team of research assistants (Goehring 2024). LLMs, by contrast, are so-called "few-shot learners," meaning they require little training data to comprehend the task assigned to them (Brown et al. 2020). There is also evidence that LLMs are particularly suited to our specific task. Several studies have found that LLMs are able to identify and extract from journal article abstracts evidence in support of/against certain claims (Koneru et al. 2024; Wadden et al. 2020). Employing Chat GPT thus allowed us to test our hypotheses on a large scale without sacrificing quality.

Of course, LLMs have drawbacks. As with any pre-trained model, we do not know the corpora on which Open AI trained GPT—accordingly, we are unaware of and unable to account for any biases they introduced in the process (Kato et al. 2024). Some noise in the data is inevitable. LLMs' accuracy wanes over time (Barrie et al., n.d.), and can "hallucinate" (produce misinformation) or amplify harmful stereotypes if improperly trained (Chang et al. 2024). Some decisions simply had to be left to the models, even if those decisions are important, and the way in which the model makes classification decisions is not always immediately obvious. The difference between one knowledge claim spanning multiple sentences and multiple distinct knowledge claims is subtle, for example. Training a model to recognize the difference is challenging, though, and it becomes increasingly difficult in the context of all the other decisions we force the models to make. This is somewhat concerning given that LLMs' reproducibility decreases as the complexity of their task increases (Wang and Wang 2025).

We began by creating a test model that we used to analyze small samples of oversight reports. We wrote specific guidelines on how to follow our typology, then instructed the model to read each document and extract all passages that qualified as a finding, evaluation, prediction, or recommendation (see Supporting Information 1–4). Each author hand-coded the output for accuracy.

After identifying recurring errors, we implemented a number of literature-based revisions.[5] Importantly, we split the single task of classifying all statements from one report into two tasks performed by separate models. For each report, one model extracts only retrospective statements and classifies them as either findings or evaluations. Then, a separate model extracts only prospective statements and classifies them as either predictions or recommendations. Loosely based on the idea of "chain of thought" prompt engineering (Zhang et al. 2022), our approach encourages the model to make classification decisions sequentially rather than simultaneously.

We obtained the knowledge claims from each report by feeding the document to Chat GPT and asking it to extract and classify for us the relevant statements from the first 20 pages of content (i.e., excluding the very first pages, which are typically cover letters or pages). We chose the first 20 pages because, setting aside cost and computational concerns, these pages typically contain the executive summary, which outlines the key arguments and evidence from the report, including knowledge claims from each section. This is also the part of the report that the audience is most likely to read, and thus may be more indicative of message committees choose to transmit. It is therefore unlikely that we over-sampled facts found toward the beginning of each investigation.

In total, Chat GPT extracted and classified 15,307 knowledge claims. We investigated both the extraction of knowledge claims

and the classification of particular statements into each of our four categories. We report the results of these validation exercises in Supporting Information 4–8. In general, we found that our final model has strong internal validity for classification, but is less consistent at the extraction stage. Moreover, in general, human and machine tend to agree on what constitutes an explicit recommendation and tend to disagree most on what counts as a simple, descriptive finding. We return to the implications of these stylized facts in the following sections.

5 | Results

Through our analysis of congressional oversight reports, we find evidence that partisan conflict changes the kind of information transmitted to the rest of Congress. We first examined whether simple empirical patterns are consistent with our argument and the typology it rests on. In particular, we argued that findings, evaluations, predictions, and recommendations should be nested and reliant on one another. All else equal, we would thus expect findings to outnumber evaluations, evaluations to outnumber predictions, and predictions to outnumber recommendations. This is mostly what we find. In particular, of our 15,307 conclusions contained in 876 parsible reports, 54% (8256) are findings, while 23% (3591) are evaluations.[6] However, recommendations did slightly outnumber predictions, at 14% (2096) to 9% (1364). Overall, this confirms the basic premises of our typology. More complex conclusions rest on a foundation of more basic, uncontroversial ones.

Next, we examine whether there is evidence that report authorship impacts the kind of information communicated. As a first cut, we classify reports by the staff who wrote them and the members who approved them. Some reports are co-authored by the majority and minority; some are solo-authored. This gives us both the status and partisanship of authorship teams, which we report in Figure 2. Specifically, we report the raw frequency distributions for each type of conclusion based on report author. A few stylized facts are worth noting. First, regardless of author, reports tend to be more retrospective than prospective. However, there is non-trivial variation in information type. Consistent with expectations, bipartisan teams tend to report fewer findings relative to partisan teams of either the majority or minority party.

Raw frequencies, of course, cannot tell the whole story. They are partly a function of circumstances and events beyond the committees' control, as well as the partisan makeup of Congress at the time. To account for this, we model counts of each conclusion type separately as a function of political circumstances. Specifically, we predict the number of each type of knowledge claim as a function of the ideological distance between the author and the chamber median, as well as the disagreement within committee, and the total number of claims within a given report. We use the count of claims for several reasons. Estimating separate count models imposes the fewest modeling assumptions on the relationship between the counts. If there is a substitution effect between categories of information, it will emerge from the data summary, not be assumed. Finally, we can estimate predicted counts, which tend to be more interpretable than models for unordered factors. There are downsides to this measure. In particular, it says nothing about the quality of individual conclusions, treating each as equivalent within category.[7]

Our primary covariate of interest is the ideological distance between the relevant committee leader(s) and the chamber floor. Chairs and ranking members are the closest politicians to report authorship, since they hire and direct the staff who conduct investigations and write reports. Our measure is the absolute distance in DW-NOMINATE score between the report author(s) and the chamber median (voteview.com). For each report, we first identify the authors. We assign either the chair or ranking member as majority and minority author, respectively. If the report is bipartisan, we take the midpoint between the leaders as the author score. We then take the absolute distance between the author score and the chamber median in that Congress.

Our committee leader data come from Stewart III and Woon (2005), as updated by Eldes et al. (2024). There were 55 reports authored by staff in more than one committee. For these

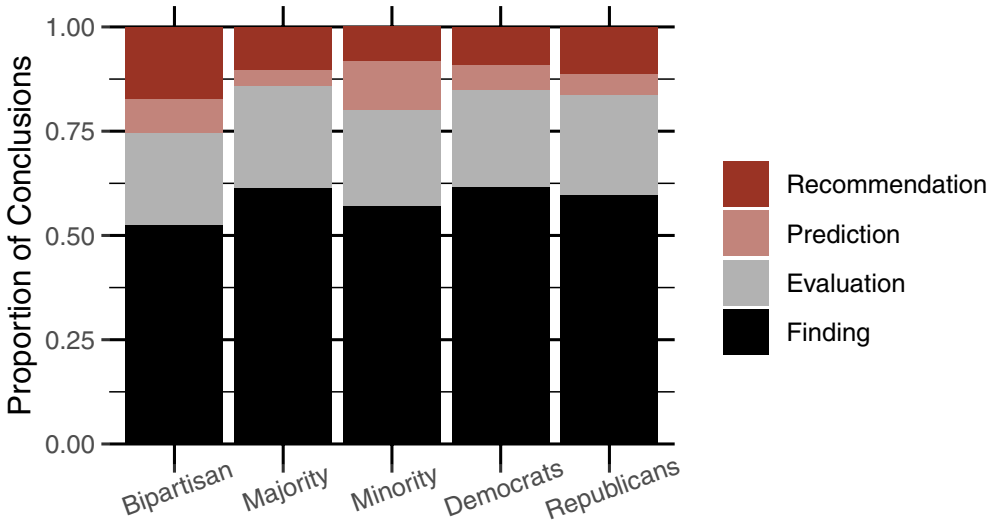

FIGURE 2 | Investigations are built on retrospective claims, but there is variation depending upon the author. Reports the proportion of knowledge claims classified as either a finding, evaluation, prediction, or recommendation, by the author of the report. Excludes 64 reports which are either committee activity reports or full committee reports without a designated author. All reports, regardless of size, receive equal weight.

reports, we again take the midpoint between any multi-author team, then the absolute distance between that hypothetical median and the floor median. Ultimately, this variable amounts to a continuous specification of the categorical authorship on display in Figure 2. Bipartisan reports tend to be close to the chamber median, with a mean NOMINATE distance of 0.22, whereas one-party reports are more than double that distance, at 0.48 on average. Minority reports, naturally, show the most disagreement, with a 0.63 average distance to the floor (Figure 3).

We also take into account the ideological distance between the chair(s) and ranking member(s) who could have, in theory, authored the report together. Greater disagreement may influence the distribution of the distance between the author and the floor by changing the set of authors. If the committee is divided, it is probable that we will see less cooperation.

We report the test of our main hypothesis in Figure 4, which plots estimated counts and standard errors for each conclusion type at three levels of distance to the floor: bipartisan author (0.23), partisan author (0.48), and minority party author (0.63). These estimates are simulated using an observed case approach from Poisson regressions which predict each dependent variable as a function of the five covariates we describe above, with

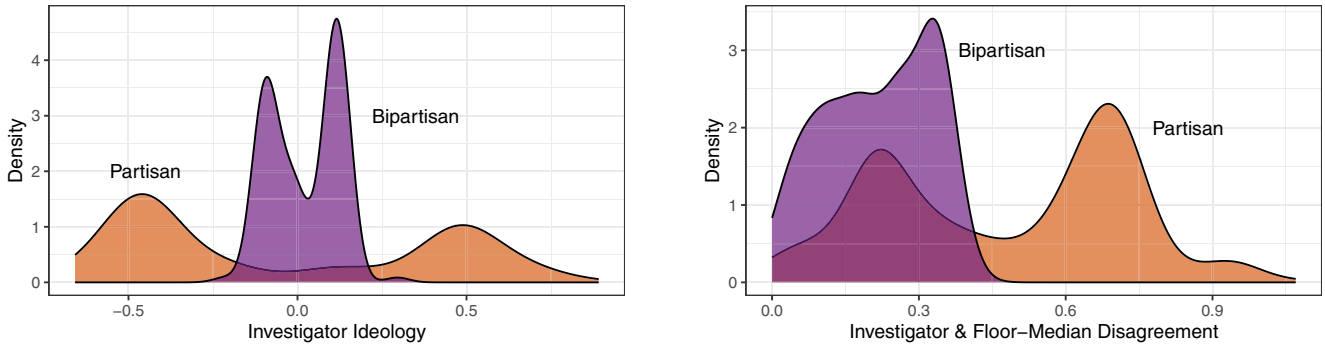

FIGURE 3 | Consensus within committee corresponds to consensus without. Distribution of common space DW-NOMINATE authors, the chair (majority) or ranking member (minority) in the committee (left), and distances between those authors and the chamber floor. Bipartisan reports are assumed to be the median distance between members, whereas inter-committee reports are assumed to be the median between committee authors.

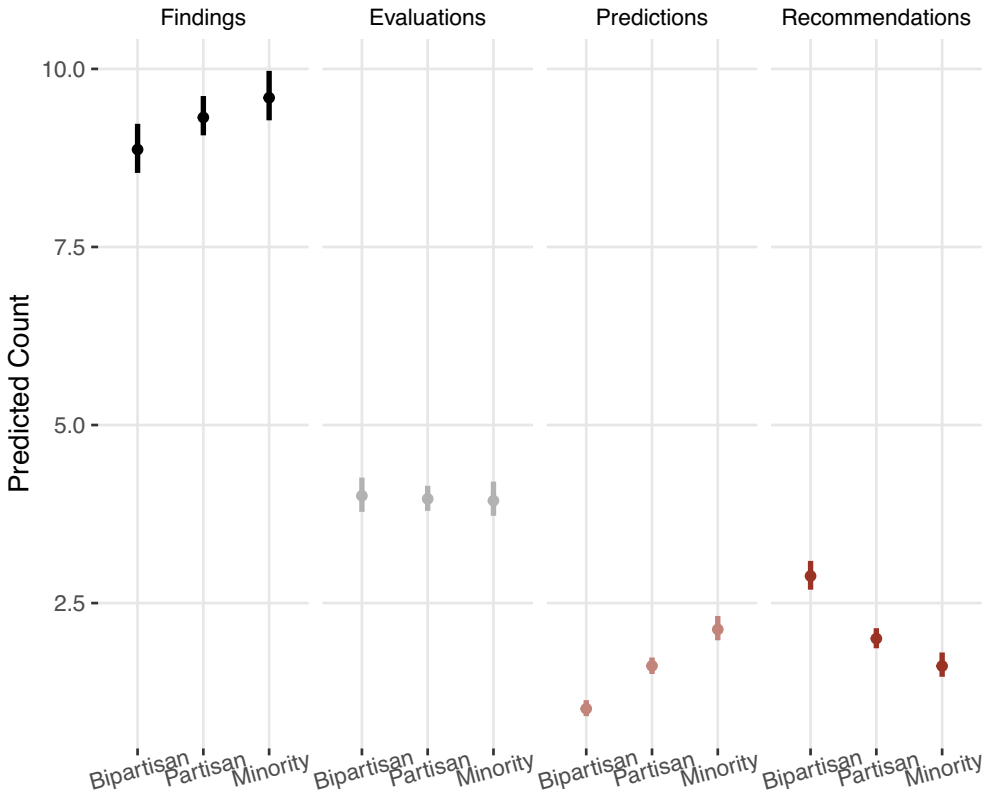

FIGURE 4 | Bipartisan teams tend to make more recommendations, while partisan teams reveal more findings. Reports estimated counts of findings, evaluations, predictions, and recommendations, simulated from Poisson regressions that predict the count of each conclusion type with NOMINATE distance to the floor, along with binary indicators for opposition author, copartisan author, House committee, and divided government. Bipartisan, partisan, and minority estimates constructed by setting the value of NOMINATE distance to the chamber floor to the mean observed value for each type of staff team. Full model results reported in Table B.1.

standard errors clustered at the Congress level. Several important patterns emerge from these estimates.

First and foremost, as expected, bipartisan reports tend to feature more recommendations. The typical bipartisan report contains about nine findings, whereas partisan reports contain around one additional finding and one fewer recommendation. The difference is starkest for reports authored by the minority party. The pattern is the reverse for predictions, with minority reports tending to make an additional 1.5 predictions, relative to bipartisan reports. This pattern is not replicated for evaluations. Reports, regardless of author, tend to contain about 3.5 of them. It is worth noting that, if anything, these differences underestimate the relative effort invested in each kind of informative signal. After all, partisan teams work with half the staff resources as their bipartisan counterparts. We therefore might expect them to have fewer conclusions as a function of their capacity. Yet, they reveal *more* findings, underscoring how much of their attention is focused on supporting the recommendations they do make.

Because our dependent variable is generated by an LLM, the standard errors are likely deflated, and worse, may systematically bias our coefficients. To investigate this, we implemented Egami et al.'s (2023) de-biasing algorithm, which constructs an estimate of measurement error based on the differences in hand-coded and LLM-coded responses. We used the algorithm to adjust each of our dependent variables—counts of each kind of knowledge claim in the report—and reran our primary analysis. Reassuringly, the sign of the key coefficients is mostly unchanged, with the exception of the association between recommendations and distance to the floor (see full results reported in Table B.3). The effects are most consistent for predictions, which both sets of results suggest are positively associated with partisan and minority status. However, the uncertainty estimates are, in general, greater and most effects are no longer distinguishable from zero. Based on the validation exercises we report in the SI, we think this reflects underlying ambiguity in the concepts and language itself, rather than errors in implementation on the part of the LLM. Put most simply, disagreement between human and machine does not always mean the machine was incorrect. This implies that future studies that take up this framework could refine both the computational implementation of the coding, along with the conceptual framework for what knowledge claims are and how they are made.

On the whole, we take these results as evidence consistent with our argument. Bipartisan teams tend to reveal fewer underlying findings to support more recommendations—perhaps leveraging the credible signal that coalition formation provides. Partisan ones, and especially minority partisans, driven by the need to make their oversight more influential, reveal more of their underlying work.

6 | Inter-Branch Conflict

We also examined the relationship between the committee author and the executive branch. Though some of these reports target other entities in Congress or the private sector, often the target is an organization ostensibly under the control of the sitting president. For example, the opposition party might focus intently on evaluating the policies of the incumbent, hoping to score political points while ignoring constructive claims like recommendations (e.g., Kriner and Schickler 2014; Eldes et al. 2024). This is pertinent to our argument, because it suggests the intended audience for reports is not the floor. Instead, in this reading, reports would be about the broader political combat between parties in government. If they predict the content of reports' conclusions, our argument about committee-floor relationships may simply miss the mark.

If members of Congress are simply position-taking when they write reports, partisan teams should care less about their report's credibility. They do not necessarily need to convince the audience that their findings are evidence-based, reducing the number of predicted findings/evaluations. And if credibility is less of a concern, partisan teams may feel empowered to make more costly claims (predictions, recommendations) without as much regard for professional consequences. This would increase the expected count of recommendations and predictions in each report relative to our above results. To ensure that executive-legislative conflict did not drive our primary results, we repeated the analysis while controlling for the presence of divided government in a given year, coded as zero when the president and majority in Congress share a party, one otherwise. We report the results in the Supporting Information (Table B.2 and Figure B.1). The patterns from the previous section hold.

Controlling for divided government did not change the direction or significance of the coefficients from the earlier Poisson regressions, nor did it alter the distribution of predicted counts of each kind of knowledge claim. Contrary to expectations under the 'combat' reading of congressional oversight reports, divided government leads to more findings and fewer recommendations transmitted, on average.

This is unsurprising, given what we have argued about the process that generates reports. We think the decision to write a report is itself associated with the sincere desire to influence policy. Under divided government, the oversight efforts tend to be more politically motivated and message-oriented. But those efforts do not require the composition of a lengthy report—members can grandstand more cheaply in hearings and public statements. Suppose, instead, all members were *required* to author a report when they conducted oversight of any kind. Then, indeed, we would expect divided government to change the distribution of knowledge claims, because the additional reports would tend to be less factual and more oriented toward political messaging. For the most part, the present institutional arrangements censor this kind of document.[8]

There is also not strong evidence that the type of information is informed by inter-branch conflict. Authors made up of only copartisans of the president write reports that include about as many findings, evaluations, and predictions as their bipartisan and even opposition party counterparts (Table B.1). Reports issued when the party controlling the chamber differs from the presidency do tend to contain fewer conclusions, in general. One potentially important difference, however, is worth noting and investigating further. We find that the opposition party, especially during divided government, is less likely to include recommendations in their solo-authored reports. Specifically, their reports contain about 1.5 fewer recommendations relative to bipartisan reports.[9]

7 | Discussion

In this study, we present a typology for understanding how oversight committees present their work. We provide evidence that partisan considerations influence the kinds of knowledge claims members make. Oversight committees are aware of the political environment in which they exist and tailor their messages accordingly. Bipartisan teams can focus their attention on projections and actionable steps for the target audience. Partisan teams, on the other hand, anticipate resistance to the recommendations they propose and so may spend more time revealing facts to show they have exerted the necessary effort to acquire relevant expertise. We have also shown that these patterns are not driven by intergovernmental conflict.[10]

Naturally, this evidence and argument has its limits. Our data are committee reports that deal with oversight investigations. So it is possible that those dealing more directly with positive legislative programs may be more prospective and contain fewer findings and evaluations of past policy. Moreover, impressive as it is, the CORD database remains a convenience sample of all oversight reports, the denominator of which is difficult to determine.[11] There is no systematic list of congressional reports in existence. In addition, this study should be regarded as a pilot attempt to scale coding of oversight reports using LLMs, and any future work should carefully review the validation reports contained in the SI and referenced throughout.

This study, however, suggests several important implications for the study of oversight and the use of information in Congress. While scholars generally focus on the work products of institutions, like bills, reports, rules, and orders, our study offers a way of studying a more fundamental unit. Within these work documents are knowledge claims, which travel from document to document, across institutions, and may or may not correspond to changes in the behavior of governments and private actors. In short, by taking advantage of contemporary developments in LLMs, we can measure these associations more directly and test theories which have, to date, been validated by second-order implications.

This naturally leads to other opportunities for future research. First and foremost, we have said nothing about the actual content of the knowledge claims. Some of them may be right, others wrong. It might be that partisan teams, weighed down by their reduced capacity, or blinded by their one-sidedness, make knowledge claims that are systematically lower quality. Likewise, the conventional wisdom suggests that bipartisan teams should have more of their recommendations adopted. But the target of these recommendations varies—it might be the executive branch, the rest of Congress, or private industry. Each of these audiences might be more or less responsive and likely to take up a committee's recommendation. Our study is unable to distinguish between those recommendations that were ignored and those that were followed, and by whom, and says nothing about which teams produce more accurate claims. This lays out obvious avenues, to us, that might lead to a better understanding of congressional oversight, and the influence of legislatures, more generally.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Endnotes

[1] In this assumption, we follow other authors in this issue like Theriault (Forthcoming) and Leavitt (Forthcoming).

[2] Note, in practice, committees sometimes label *everything* a finding, regardless of whether it is a fact, projection, or policy recommendation. Our study distinguishes between them because we believe they are generated for different reasons. In fact, the act of labeling a recommendation or speculation a "finding" itself is an indication that committees know facts or findings are more credible to the audience of the report.

[3] Another potential limitation on the ability to share findings is the willingness of targets of investigations to share messages. Committees cannot share information they are unable to collect. However, we do not think this possibility changes the expectations laid out in this section. The kinds of knowledge claims made in a report are not limited by the responsiveness of targets. Indeed, committees will report findings that state a target was unwilling to provide or did not have the information the committee requested. This is still a "finding" that demonstrates to an audience that the committee conducted an investigation, which is the key consideration for the audience.

[4] These can be found at: https://cord-levin-center.org/home. For a fuller description of their methodology for gathering and including reports, please see their codebook.

[5] A fuller description of our model development and validation process can be found in the Supporting Information.

[6] Note, these figures include committee activity reports, which we later exclude for the purposes of analysis. The basic pattern is the same.

[7] In earlier versions of this paper, we constructed a measure of sentence specificity using Chat GPT, but it failed validity checks and was abandoned.

[8] There are, of course, exceptions. Several dozen reports are under 10 pages and look more like brochures than reports supervised by legal counsel. We cannot test the hypothesis that fewer reports are written when earlier stages of oversight are oriented toward political messaging and grandstanding. Our sample of committee reports is neither comprehensive nor a random subsample of all reports. That makes it unsuitable for testing claims about associations between political covariates and the frequency of reports.

[9] This estimate is not statistically distinguishable, by convention, from reports authored only by copartisans, however.

[10] Our findings related to differences in minority party behavior are consistent with Drolc and Butcher (Forthcoming), who examine the frequency and length of oversight reports within this special issue.

[11] This issue is further complicated by the fact that some knowledge claims are used more than once, cited in different reports on similar subjects (e.g., two reports may deal with the implementation of the same policy in different states).

References

Bail, C. A. 2024. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences of the United States of America* 121, no. 21: e2314021121.

Ban, P., J. Y. Park, and H. Y. You. 2021. "How Are Politicians Informed? Witnesses and Information Provision in Congress." *American Political Science Review* 117, no. 1: 122–139.

Barrie, C., A. Palmer, and A. Spirling. n.d. "Replication for Language Models." Forthcoming.

Brown, T., B. Mann, N. Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Bussing, A., and M. Pomirchy. 2022. "Congressional Oversight and Electoral Accountability." *Journal of Theoretical Politics* 34: 35–58.

Callander, S. 2008. "A Theory of Policy Expertise." *Quarterly Journal of Political Science* 3: 123–140.

Chang, Y., X. Wang, J. Wang, et al. 2024. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 15, no. 3: 39:1–39:45.

Curry, J. M. 2015. *Legislating in the Dark: Information and Power in the House of Representatives*. University of Chicago Press.

Curry, J. M. 2019. "Knowledge, Expertise, and Committee Power in the Contemporary Congress." *Legislative Studies Quarterly* 63: 203–237.

Drolc, C., and J. Butcher. Forthcoming. "Ulterior Oversight: Minority Party Investigations and the Politics of Exclusion." *Legislative Studies Quarterly*.

Egami, N., M. Hinck, B. Stewart, and H. Wei. 2023. "Using Imperfect Surrogates for Downstream Inference: Design-Based Supervised Learning for Social Science Applications of Large Language Models." *Advances in Neural Information Processing Systems* 36: 68589–68601.

Eldes, A., C. Fong, and K. Lowande. 2024. "Information and Confrontation in Legislative Oversight." *Legislative Studies Quarterly* 49: 227–256.

Fenno, R. F. 1973. *Congressmen in Committees*. Little, Brown and Co.

Fong, C., K. Lowande, and A. Rauh. 2024. "Expertise Acquisition in Congress." *American Journal of Political Science* 69: 1–14.

Gatt, A., and E. Krahmer. 2018. "Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation." *Journal of Artificial Intelligence Research* 61: 65–170.

Gilardi, F., M. Alizadeh, and M. Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 30: e2305016120.

Gilligan, T. W., and K. Krehbiel. 1987. "Collective Decisionmaking and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures." *Journal of Law, Economics, and Organization* 3, no. 2: 287–335.

Goehring, B. 2024. "Improving Content Analysis: Tools for Working With Undergraduate Research Assistants." *PS: Political Science & Politics* 57, no. 1: 57–63.

Groseclose, T. 1994. "Testing Committee Composition Hypotheses for the U.S. Congress." *Journal of Politics* 56: 440–458.

Heseltine, M., and B. C. von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11, no. 1: 20531680241236239.

Hu, T., and N. Collier. 2024. "Quantifying the Persona Effect in LLM Simulations." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kato, K., A. Purnomo, C. Cochrane, and R. Saqur. 2024. "L(u)PIN: LLM-Based Political Ideology Nowcasting." *arXiv Preprint* arXiv:2405.07320.

Kiewiet, D. R., and M. D. McCubbins. 1991. "American Politics and Political Economy Series." In *The Logic of Delegation*. University of Chicago Press.

Koneru, S., J. Wu, and S. Rajtmajer. 2024. "Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, 2787–2797. ELRA and ICCL.

Krehbiel, K. 1991. *Information and Legislative Organization*. University of Michigan Press.

Kriner, D. L., and E. Schickler. 2014. "Investigating the President: Committee Probes and Presidential Approval, 1953–2006." *Journal of Politics* 76, no. 2: 521–534.

Leavitt, C. Forthcoming. "The Purpose-Driven Congress: Measuring and Explaining Legislative Outcomes of House Investigations."

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

Levin, C., and E. J. Bean. 2018. "Defining Congressional Oversight and Measuring Its Effectiveness." *Wayne Law Review* 64: 1–22.

Lowande, K. 2018. "Who Polices the Administrative State?" *American Political Science Review* 112, no. 4: 874–890.

Lowande, K. 2019. "Politicization and Responsiveness in Executive Agencies." *Journal of Politics* 81, no. 1: 33–48.

Lowande, K., and J. Peck. 2017. "Congressional Investigations and the Electoral Connection." *Journal of Law, Economics, and Organization* 33: 1–27.

Napolio, N. G. 2025. "Measuring Executive Agency Ideology Using Large Language Models." *Journal of Political Institutions and Political Economy* 6: 161–180.

Park, J. Y. 2021. "When Do Politicians Grandstand? Measuring Message Politics in Committee Hearings." *Journal of Politics* 83: 214–228.

Park, J. Y. 2023. "Electoral Rewards for Political Grandstanding." *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 17: e2214697120.

Provins, T., N. W. Monroe, and D. Fortunato. 2022. "Allocating Costly Influence in Legislatures." *Journal of Politics* 84: 1697–1713.

Ritchie, M. N. 2018. "Back-Channel Representation: A Study of the Strategic Communication of Senators With the US Department of Labor." *Journal of Politics* 80, no. 1: 240–253.

Stewart, C., III, and J. Woon. 2005. "Congressional Committee Assignments, 103rd to 105th Congresses, 1993–1998: House and Senate." Technical Report Massachusetts Institute of Technology.

Theriault, S. M. Forthcoming. "The Legislative Effectiveness of Congressional Oversight Reports." Legislative Studies Quarterly.

Wadden, D., S. Lin, K. Lo, et al. 2020. "Fact or Fiction: Verifying Scientific Claims."

Wang, J. J., and V. X. Wang. 2025. "Assessing Consistency and Reproducibility in the Outputs of Large Language Models: Evidence Across Diverse Finance and Accounting Tasks."

Wu, P. Y., J. Nagler, J. A. Tucker, and S. Messing. 2023. "Large Language Models Can Be Used to Estimate the Latent Positions of Politicians."

Zhang, Z., A. Zhang, M. Li, and A. Smola. 2022. "Automatic Chain of Thought Prompting in Large Language Models."

Ziems, C., W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50, no. 1: 237–291.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** lsq70079-sup-0001-Supinfo.pdf.