

C-QUERI: Congressional Questions, Exchanges, and Responses in Institutions Dataset

Manjari Rudra School of Computing Binghamton University mrudra1@binghamton.edu	Daniel Magleby Department of Political Science Binghamton University dmagleby@binghamton.edu	Sujoy Sikdar* School of Computing Binghamton University ssikdar@binghamton.edu
--	--	--

Abstract

We introduce C-QUERI, a dataset of question-answer exchanges from U.S. congressional committee hearings spanning the 108th–117th Congresses, with the pipeline used to construct it. C-QUERI transforms congressional hearing records into structured question-answer exchanges, enabling researchers to study political questioning and witness responses systematically and at scale. Congressional hearings provide a particularly valuable setting for such research: interactions are structured by formal rules, members of different parties are guaranteed opportunities to question witnesses, and witnesses are expected to respond. As an initial application demonstrating the resource’s utility, we examine whether questioner party affiliation can be predicted from question text alone. Question text contains a modest aggregate partisan signal, with substantially greater distinguishability in some committees, hearing types, and configurations of partisan control than others. C-QUERI provides new infrastructure for studying congressional politics and a general framework for analyzing question-answer interactions in other interview-like settings.

1 Introduction

Congressional *hearings* are among the most visible and consequential activities of the U.S. Congress (Ban et al., 2024; Ban and Hill, 2025). Congressional hearings are especially valuable for studying political communication because they generate structured exchanges between legislators and witnesses. Existing research demonstrates that the character of these exchanges, including their informational and confrontational content, varies with political and institutional incentives (Eldes et al., 2024). Yet the scale and structure of hearing transcripts have made such interactions difficult to study systematically across Congresses, committees, and political contexts. In these forums,

elected representatives question witnesses including cabinet officials, agency heads, policy experts, and industry leaders. Hearings shape the legislative agenda, provide oversight of executive agencies, and serve as a stage for partisan conflict and coalition-building. These exchanges offer the public a rare window into the complex, and often contested, processes of deliberation, lawmaking, investigation, and accountability at the highest levels of government.

Yet, despite their centrality to democratic accountability, the language of questions and responses in hearings has been difficult to study systematically and at scale. While transcripts are public, they rarely appear in a form that supports utterance and exchange level analysis of who asked what, how questions were framed, and how witnesses responded. Researchers interested in these interactions must therefore transform lengthy, inconsistently structured transcripts into analyzable exchanges before they can address substantive questions about congressional questioning and witness response. The substantial effort required to structure these data limits both the scale of existing research and the range of questions that researchers can readily investigate.

This paper introduces C-QUERI, a new dataset of question-and-answer exchanges constructed from 16,126 congressional hearings spanning the 108th to the 117th Congresses, together with a computational pipeline for constructing it. We segment hearing transcripts into *more than 3 million* utterances, identify question-answer exchanges using fine-tuned transformer-based classifiers, and pair questions with the responses they elicit. We validate key pipeline steps using human annotation and provide a rich set of linguistic features and metadata to support downstream analysis and natural language processing tasks.

By structuring congressional hearings at the level of individual utterances and question-answer

* Corresponding author.

exchanges, C-QUERI makes it possible to investigate aspects of congressional behavior that are difficult to observe using more conventional sources of legislative data. Much existing research on congressional partisanship has focused on floor votes and roll-call data, which capture only issues that reach the floor and obscure upstream processes in which committees decide what gets debated and what gets a vote (Mayhew, 1974; Cox and McCubbins, 1993, 2005). Hearings offer a complementary window: members with different affiliations are guaranteed opportunities to ask questions, witnesses are obliged to respond, and questioning alternates across majority and minority members, enabling direct comparisons across legislators, parties, and time.

To demonstrate one type of substantive analysis that C-QUERI enables, we use the resource to examine whether the party affiliation of members of Congress can be predicted from the language of their questions alone. We hypothesize that political agendas and strategic goals are reflected in the linguistic structures, tone, and style of congressional questioning. In this paper, we ask: **Do party affiliation and majority/minority status affect the form and content of questions and the answers they elicit in congressional hearings?** Using BERT-based classifiers, classical machine-learning models based on linguistic features, and zero-shot prompting with large language models, we find detectable partisan signals in congressional questioning and substantial variation across committees, hearing types, and configurations of partisan control. Using C-QUERI data, we characterize systematic variation in congressional questioning while also identifying questions for future substantive research.

1.1 Our Contributions

Our contributions are three-fold. First, we introduce a novel dataset constructed from transcripts of 16,126 committee hearings in the U.S. House and Senate spanning the 108th to 117th Congresses, from which we extract over 3 million utterances. The dataset identifies question and answer utterances, includes human annotations to validate classification, and provides linguistic features for each exchange, opening new avenues for studying congressional behavior, polarization, and question-answer dynamics at scale.

Second, we develop a pipeline for processing unstructured hearing transcripts into analyzable

question-answer units. The pipeline combines BERT-based and heuristic named entity recognition for speaker identification, fine-tuned BERT classifiers for question-answer labeling, and a comprehensive linguistic feature extraction step. Beyond congressional hearings, this framework generalizes to other structured interview-like settings including media interviews and judicial proceedings, enabling comparative studies of political questioning across institutions.

Third, we provide an empirical assessment of partisan signals in hearing dialogue using BERT-based classifiers, classical machine learning on human-engineered linguistic features, and zero-shot prompting with large language models. Across all hearings, we predict questioner party affiliation with 59% accuracy which is 8% above the majority-class baseline, showing that partisan signals are detectable in deliberative and oversight functions of congressional committees. Notably, aggregate accuracy masks important variation: partisan differences range from near-random in some committees to over 70% in others. We show that the ability to distinguish party affiliation varies across committees, hearing types, and configurations of partisan control. These findings speak to longstanding debates about whether congressional committees primarily serve informational or partisan ends (Krehbiel, 1991; Cox and McCubbins, 2005).

2 Related Work

Legislative and Political Discourse Datasets.

Several datasets provide access to transcripts of floor speeches and debates in the U.S. Congress (Gentzkow et al., 2018), European democracies (Rauh and Schwalbach, 2020; Fiva et al., 2025), and local council meetings (Hoang, 2025). Delano et al. (2025) archive congressional data including committee hearings and reports. The Congressional Committees Hearing dataset (CoCoHD) is the most similar dataset to ours, consisting of transcripts from the 105th to the 118th sessions (Hiray et al., 2024). Our dataset differs from CoCoHD in three ways: we segment transcripts into individual utterances, identify question and answer utterances within each hearing, and pair questions with their corresponding answers to enable downstream question-answer analysis. We also provide utterance-level human annotations and a comprehensive set of linguistic features. Birkenmaier et al. (2025) extract transcripts of political in-

interviews from news media, and [Maher et al. \(2020\)](#) analyze testimonies of social scientist witnesses from 1946 to 2016. None of these datasets identify question and answer utterances for large-scale computational analysis.

Question Analysis in NLP. The NLP question-answering literature focuses predominantly on information-seeking questions, evaluated on extracting correct responses from text or knowledge bases ([Zhang et al., 2023](#); [Biancofiore et al., 2024](#); [Yu et al., 2024](#)), conversational question answering ([Zaib et al., 2022](#)), and natural language understanding ([Kočiský et al., 2018](#)). Most closely related to our work, [Zhang et al. \(2017\)](#) introduce an unsupervised framework for identifying rhetorical question types in UK parliamentary discourse, revealing systematic differences between government and opposition questioners. [Ferracane et al. \(2021\)](#) study conversation acts and intents in 20 U.S. congressional hearings, examining whether witnesses answer the questions posed to them and how question form shapes responses. Our dataset extends this line of work to a corpus three orders of magnitude larger, pairing questions with answers across all major committees and congressional sessions, and providing the infrastructure for studying partisan signals in questioning strategies at scale. More broadly, NLP research has examined rhetorical question identification ([Bhattasali et al., 2015](#)) and dialogue act classification in naturalistic conversational settings ([Stolcke et al., 2000](#); [Raheja and Tetreault, 2019](#)). Questions in congressional hearings, however, are embedded in a distinct institutional context where they are frequently rhetorical, adversarial, or performative. A member asking “Isn’t it true that your agency failed to act?” is making an accusation in interrogative form, not seeking information. Our dataset provides infrastructure for studying such questions at scale, complementing existing question-answering paradigms.

Computational Analysis of Political Text. A substantial body of NLP work has analyzed political discourse, including ideology detection from text ([Thomas et al., 2006](#); [Iyyer et al., 2014](#)), framing and agenda-setting ([Card et al., 2015](#)), and partisan language in floor speeches ([Gentzkow et al., 2019](#); [Lauderdale and Herzog, 2016](#)). Text-as-data approaches have tracked polarization in congressional speech and linked linguistic features to roll-call voting patterns. Our work extends this line of research from floor speeches to committee hearings, where the structured question-answer format

enables finer-grained analysis of how partisanship is expressed in dialogue rather than monologue.

Congressional Hearings in Political Science. Congressional hearings are a central but understudied arena of legislative politics. Prior research on oversight has relied on aggregate measures such as hearing counts ([Kriner and Schwartz, 2008](#); [Kriner and Schickler, 2016](#)), and classic work emphasizes that congressional monitoring of the executive is difficult to observe and episodic at best ([McCubbins and Schwartz, 1984](#); [Aberbach, 1990](#)). Recent work has begun to examine how hearings shape bureaucratic behavior and how partisan incentives determine from whom Congress seeks information ([Ban et al., 2024](#); [Ban and Hill, 2025](#)). Informational theories of Congress ([Krehbiel, 1991](#); [Gailmard and Patty, 2013](#)) and work on descriptive representation ([Mansbridge, 1999](#); [Swers, 2002, 2020](#); [Kanthak and Krause, 2013](#); [Kathlene, 1994](#); [Minta, 2011](#)) have lacked direct measures of the exchanges they theorize, and concerns have been raised about the underrepresentation of women among congressional witnesses ([Coil et al., 2024](#)). Studies of polarization have focused on floor votes and speeches ([McCarty et al., 2006](#); [Lee, 2016](#)), leaving oversight hearings largely unexamined, while theories of deliberative democracy ([Gutmann and Thompson, 1996](#); [Bessette, 1994](#); [Mansbridge and Parkinson, 2012](#)) lack direct empirical grounding in legislative practice. By structuring hearings into discrete question-answer exchanges, our dataset makes legislative dialogue observable at a fine-grained level, enabling systematic tests of when hearings generate genuine policy deliberation and when they serve partisan signaling functions.

3 Constructing the C-QUERI Dataset

Public Congressional hearings are made available in digital format at <https://www.govinfo.gov/app/collection/chrg> between 2 months to 2 years after the hearing is held, and as Fig. 1 illustrates, are published as a dialogue style transcript in plain text. Our dataset¹ is constructed from all available transcripts as of April 2022, covering 20 years spanning the 108th to the 117th congressional sessions with transcripts from 16,126 hearings by 76 committees involving 1,774 committee members and 69,149 witnesses in total.

Fig. 2 and 3 in Appendix A illustrates the distribution of hearings, witnesses, utterances and word-

¹<https://doi.org/10.5281/zenodo.15470512>

counts across congressional sessions and committees, revealing trends that track major domestic and global events.

Mr. BROOKS. Thank you, Mr. Chairman. I yield back the balance of my time.
Mr. LANGEVIN. Thank you, Mr. Brooks.
Ms. Slotkin is now recognized for 5 minutes.
Ms. SLOTKIN. Great. Thanks for being here and for doing this. I think what you're hearing from us, we're sort of—both sides of the aisle, frankly, asking different versions of the same question, which is, we have all heard testimony in the last Congress.

Figure 1: Excerpt of a congressional hearing transcript

We note that some committees have been renamed over time although their roles and responsibilities typically have not changed. For example, the House Committee on Foreign Affairs operated as the Committee on International Relations from 1975 to 1978 and 1995 to 2007. To avoid confusion, we maintain the name under which the committee operated at the time the hearing was conducted and only consider activity across committees when appropriate for further analysis.

Transcripts. Congressional hearings follow a common format that is reflected in the transcripts which are published as a single, unstructured text document. A hearing typically starts with a roll call of members present, followed by an address by the committee chair, and statements from committee members and witnesses. This is followed by exchanges where committee members ask questions which are answered by the witnesses. Some hearings have more structure with fixed time slots allocated to each member. Metadata published in the Extensible Markup Language (XML) identifies the hearing, committee(s), members in attendance, and witnesses present at the hearing along with their affiliations. We observe that there can be significant delays in the publication of the official transcripts and the metadata may sometimes fail to identify committee members and witnesses in attendance. Curating our dataset consisting of individual utterances together with linguistic features and identifying questions and answers from transcripts in plain text involves the following tasks.

3.1 Task 1: Identify Utterances

We identify individual utterances in a transcript using the following pipeline.

Preprocessing. We identify the beginning and end of proceedings using rules developed through careful observation of the structure of a transcript, allowing us to remove parts of the transcript that convey redundant metadata. Detailed documentation of these rules accompanies the dataset.

Extracting utterances. We use the word “utterance” to mean the consecutive words spoken by a single member or witness. As Figure 1 illustrates, the proceedings are transcribed so that the name of the speaker (member or witness) precedes their utterance, a structure that we exploit using named entity recognition. A key challenge we encounter is that standards for formatting these markers vary by committee and the time at which the transcript was produced. For example, honorifics, full name or last name alone, case and capitalization, and speaker’s affiliation are used inconsistently across the raw transcripts.

Named Entity Recognition. We develop a hybrid Named Entity Recognition (NER) method to identify these markers by combining large language model (LLM) based and heuristic techniques. We use a BERT-based NER following Devlin et al. (2019) to identify and classify entities, such as names of people, organizations, locations, dates, and other specific items within a text. To augment this, we develop heuristic NER rules by systematic examination of transcript and metadata formatting patterns which we implement using regular expressions. We tag named entities appearing at the beginning of utterances as either a member of the committee or a witness using data obtained from the metadata associated with the hearing, and through the Propublica Congress API (<https://projects.propublica.org/api-docs/congress-api/members/>) which provides details about members their roles and affiliations.

Code to reproduce results in this submission can be found at <https://github.com/QUERI-Project/CongressHearings.git> and data can be found at <https://doi.org/10.5281/zenodo.15470512>.

3.2 Task 2: Identify Questions and Answers

Next, we identify question and answer utterances and focus particularly on the question and answer sessions that follow the statements by the witnesses. We take a supervised transfer learning approach to fine-tune a variant of the BERT model obtained from Huggingface² coupled with a dropout layer and a linear layer with ReLU activation in order to classify the utterances as questions and answers (Figure 9 in Appendix F).

²<https://huggingface.co/google-bert/bert-base-cased>

Since transcripts of Congressional committee hearings are not annotated, we construct a training dataset from the Reddit Ask Me Anything (r/AMA) community consisting of 60756 Questions and 60756 Answers, and the U.K. Parliamentary dataset (Zhang et al., 2017) with 2344 questions and 2344 answers. The r/AMA community lets individuals field questions from Reddit users, with top-ranked comments serving as questions and the poster’s responses as answers; we construct our training set from the highest-ranked exchanges. The U.K. Parliamentary dataset consists of written questions by members of the U.K. Parliament directed at cabinet ministers who are typically required to respond within a set timeframe. Although these training sources differ in register from congressional hearings, our classifier achieves 87% accuracy on held-out congressional utterances, validating the transfer. Details are in Appendix B.

3.3 Dataset Curation Results

3.3.1 Task 1: Identify Utterances.

Our dataset consists of 3,319,386 utterances. We evaluate the efficacy of our utterance identification using an annotated evaluation dataset constructed by first sampling 50 hearings from each congressional session uniformly at random. From each of these hearings, we sample 10 utterances from the output of our utterance identification pipeline, for a total of 500 utterances from each congressional session. We then manually verify whether these utterances were correctly identified.

Session	# Incorrect	# Clubbed	# Broken
108	32	12	20
109	16	8	8
110	20	15	5
111	12	8	4
112	4	2	2
113	39	36	3
114	27	18	9
115	23	18	5
116	52	40	12
117	77	69	8
Total	302	226	76

Table 1: Performance on utterance identification (Task 1). We sample 50 hearings per session, from each of which we sample 10 utterances and manually verify them. Over 10 sessions and 5,000 verified utterances, 302 were incorrectly identified, of which 226 were clubbed and 76 were broken.

Overall, 93.96% of the utterances identified by our pipeline were correct in our manual verification. Incorrectly split utterances occur due to one of two

kinds of error; utterances that are *clubbed* together or a single utterance that is *broken* and split in two or more utterances. Table 1 summarizes our findings for the task of identifying utterances.

3.3.2 Task 2: Identify Questions and Answers.

To evaluate our question and answer annotations, we sampled 800 utterances uniformly at random from among all utterances that occurred between the 114th through to the 117th congressional sessions. These 800 utterances were then hand-labeled, yielding 379 questions and 421 answers. We achieved an accuracy of 87.14% at the question-answer labeling task using our BERT-based classifier. Table 2 shows the confusion matrix for this task. Overall, 291 and 380 out of the 379 and 421 question and answer utterances respectively were classified correctly.

Session		114	115	116	117	114–117
Questions	True	20	105	101	65	291
	False	6	25	29	28	88
Answer	True	26	127	134	93	380
	False	0	3	3	5	11
Accuracy		0.88	0.89	0.87	0.83	0.87

Table 2: Performance on identifying question and answer utterances (Task 2), against 800 human labeled utterances. For each session indicated by a column, a row entry provides the number of utterances either correctly (True) or incorrectly (False) identified as a question or answer. Over sessions 114–117, 291 out of 379 utterances identified as questions by our classifier were truly question utterances.

4 Methodology for Identifying Party Affiliation and Standing

The task is to predict the questioner’s *party affiliation*, either Democrat (D), Republican (R), or Independent (I), or *party standing*, i.e., whether their party holds a majority (M) or minority (m) in their respective chamber, from the text of a question utterance, an answer utterance, or both. Each data instance is a single utterance labeled with the speaker’s affiliation or standing. We take a three-pronged approach, with each method offering distinct advantages: classical machine learning provides interpretable, human-readable feature importance; fine-tuned BERT captures deep contextual semantics; and zero-shot LLM prompting establishes whether partisan signals are detectable without task-specific training.

Approach 1: Classical ML with linguistic fea-

tures. We train random forest classifiers on a comprehensive suite of human-engineered features from the News Landscape (NELA) toolkit (Horne et al., 2019), including **Style** (POS tags, punctuation usage), **Complexity** (lexical diversity, reading difficulty, sentence length), **Affect** (sentiment), and **Bias** (hedge words, factives, assertatives). Random forests achieved the best performance among classical methods including SVMs and logistic regression variants. We use an 80/20 train/test split with 5-fold cross-validation for hyperparameter tuning (Appendix E), and use feature importance scores to identify the most significant features and study class differences using Kolmogorov-Smirnov tests (Appendix D).

Approach 2: Fine-tuned BERT Classifier. We fine-tune a BERT-based classifier (Devlin et al., 2019) consisting of the BERT encoder followed by a dropout layer and a linear layer with LeakyReLU activation using the cross-entropy loss function and the Adam optimizer. We use a 20/5/75 train/validation/test split with 5-fold cross-validation to select hyperparameters including batch size, learning rate, and number of epochs. We train separate classifiers for each combination of congressional session and committee to capture context-specific partisan signals. Architecture details are in Appendix F.

Approach 3: Zero-shot LLM Prompting. We evaluate the ability of a wide range of popular pre-trained LLMs to distinguish party affiliation and standing using zero-shot prompting techniques. Details are presented in Appendix G.

5 Results

We examine whether the party affiliation and party standing (majority/minority status in their respective chamber) of members of Congress can be distinguished using linguistic features and semantic content of the questions they ask and the answers they receive. We first report these patterns across the full corpus, summarized in Table 3, before exploring how they vary across committees, hearing types, and configurations of partisan control.

Question-based prediction. After removing names of committee members and witnesses, which appear at the beginning of utterances in the transcript, we fine-tune the BERT model. We find that party affiliation can be predicted with an accuracy of 59%, i.e., 8% higher than the baseline of predicting the majority class in the test set (see

Model	Affiliation		Standing	
	Q	A	Q	A
NELA	53.75	52.22	61.05	60.97
NELA Baseline	51.52	51.53	60.97	60.97
BERT w/o speaker	59.04	54.70	58.90	57.39
BERT with speaker	92.40	58.00	75.00	63.00
BERT Baseline	51.66	51.62	61.45	61.45

Table 3: Results for identifying the questioner’s party affiliation (Republican vs Democrat) and standing (majority vs minority) when the input is a question (Q) or the answer (A) in response to a question. Speaker names were removed from utterances unless otherwise specified. The Baseline is the fraction of test examples with the majority class.

Table 3). Thus, question text contains detectable information about the party affiliation of the questioner, although the aggregate partisan signal is modest. As we show below, this overall result also masks substantial variation across institutional contexts: partisan affiliation is considerably more distinguishable from question text in some committees, hearing types, and political configurations than in others.

As a methodological check informing our choice of classifier, we also evaluated zero-shot prompting with five open-weight LLMs on a random sample of 10,000 question-answer pairs. The fine-tuned BERT classifier exceeded its majority-class baseline by a larger margin than any of the LLMs with zero-shot prompting on their sample (Table 17, Appendix I), so we use it in the pipeline.

Notably, when speaker names are retained in the question text, party affiliation prediction accuracy jumps to 92% (See Table 3). This suggests that embeddings of committee members’ names capture systematic differences in how politicians from different parties are referenced in text corpora, reflecting partisan patterns in media coverage, political discourse, and public communication, but are unable to resolve linguistic or semantic differences when trained on all committee hearings. This underscores the importance of name removal in isolating genuine differences in questioning style. Therefore, all remaining results use utterances with speaker names removed.

Answer-based prediction. Identifying the party affiliation or standing of the questioner based on answer utterances appears to be more challenging as we achieve accuracy that is only 1% and 0% above their respective baselines.

Party standing. In contrast to party affiliation, we

BERT	BERT Baseline	NELA	NELA Baseline	Committee
64.48	54.75(R)	58.73	54.71(R)	(H) Energy and Commerce
65.39	56.40(R)	57.85	56.38(R)	(H) Financial Services
62.54	55.19(D)	57.10	54.87(D)	(H+S) Judiciary
72.48	70.21(R)	71.07	70.42(R)	(H) Oversight and Govt. Reform
64.69	66.47(R)	68.76	66.91(R)	(H) Ways and Means
67.13	63.99(D)	63.66	63.69(D)	(S) Homeland Security and Govt. Affairs
60.95	60.54(D)	62.41	61.47(D)	(S) Comm, Sci, and Transportation
59.27	51.20(D)	54.81	51.35(D)	(H) Armed Services
61.04	50.32(D)	54.88	50.26(D)	(H+S) Veterans' Affairs
60.00	50.80(R)	57.72	51.79(R)	(H) Foreign Affairs

Table 4: Results for identifying the questioner’s party affiliation (Republican vs Democrat) on question utterances across the 10 largest committees by number of committee hearings using BERT and Random Forest Classifier. An (R) or (D) indicates that Republican or Democrat speakers have more utterances, and H = House, S = Senate.

find little evidence that majority/minority status is distinguishable from either question or answer text in the full corpus. Neither the BERT nor NELA models meaningfully improves on the majority-class baseline. This contrast suggests that party affiliation produces a more readily detectable linguistic signal in congressional questioning than majority/minority status.

Linguistic features. Classical linguistic features from the NELA toolkit alone are insufficient to predict either party affiliation or standing (Table 3).

To interpret these results, we conduct statistical testing on the NELA features using two-sample Kolmogorov-Smirnov tests (Appendix D), which show that question utterances by Republicans contain significantly more assertive and hedge words, while those by Democrats contain more positive sentiment and implicative phrasing.

5.1 Differences Occur in Certain Committees and Hearing Types

The modest aggregate partisan signal may conceal important variation across institutional contexts. We therefore examine whether party affiliation is more distinguishable from question text under particular conditions. We examine this question along three dimensions: (1) *individual committees* and Congressional sessions, (2) *types of hearings* and legislative functions, and (3) across times of *unified* or *divided* government. Here, government is unified during a given Congressional session if the same party controlled the presidency as well as both the House of Representatives and the Senate.

Differences across Committees. Table 4 (and Figure 10 in Appendix I) shows that improvement over the majority-class baseline varies considerably across committees, with committees falling roughly into three tiers based on the magnitude of

this improvement.

The greatest improvement over baseline ($\geq +8.0$) (See Table 4 rows 1, 9 and 10) emerge in Veterans’ Affairs, Energy and Commerce, Foreign Affairs, Financial Services, and Armed Services. One possibility, which C-QUERI could be used to investigate more directly in future work, is that both parties often share broad headline goals (e.g., supporting veterans or national security) but differ on how they frame issues, have different priorities, invoke different values, and attribute problems to different causes, creating linguistically distinct questioning even when substantive positions overlap. This is consistent with theories of issue ownership (Egan, 2013), issue frame ownership (Arbour, 2014), and party reputation and branding that highlight collective incentives for coordinated messaging (Cox and McCubbins, 1993, 2005).

We observe moderate improvement over baseline (< 8.0 and ≥ 3.0) (See Table 4 rows 3 and 6) in Judiciary, and Homeland Security and Governmental Affairs, which handle issues where partisan positions are well-established, but policy complexity and legal constraints may impose some convergence in questioning style.

We observe little or no improvement over baseline in Oversight and Government Reform, Ways and Means, and Commerce, Science, and Transportation (See Table 4 rows 4, 5 and 7). This is initially surprising, because Oversight and Ways and Means are among the most visibly partisan arenas in Congress. A possible explanation is that in these high-stakes settings, both parties adopt similar adversarial or technically focused questioning strategies. For example, Oversight hearings often feature aggressive questioning and demands for accountability regardless of party, and the partisan difference may lie in whom members target rather than

Committee on	Government	BERT	Baseline
(H)Energy and Commerce	Unified	60.33	52.22(D)
	Divided	65.53	56.57(R)
(H) Financial Services	Unified	61.20	54.05(R)
	Divided	64.88	59.15(R)
(H+S) Judiciary	Unified	58.93	56.23(D)
	Divided	60.66	55.72(D)
(H) Oversight and Govt. Reform	Unified	66.50	68.17(R)
	Divided	71.20	69.34(R)
(S) Homeland Security and Govt. Affairs	Unified	63.16	60.15(D)
	Divided	67.08	66.72(D)
(S) Comm, Sci, and Transportation	Unified	59.64	58.28(D)
	Divided	52.42	65.24(D)
(H) Armed Services	Unified	59.66	54.96(D)
	Divided	50.60	50.61(R)
(H+S) Veterans' Affairs	Unified	55.06	52.20(D)
	Divided	60.34	52.74(R)
(H) Foreign Affairs	Unified	48.31	58.79(D)
	Divided	60.87	57.31(R)

Table 5: Predictions of party affiliations from question utterances across committees, comparing unified and divided Congress settings. An annotation of (R) or (D) indicates that Republican or Democrat speakers have the most utterances.

how they question. Ways and Means deals with tax policy where questioning may be constrained by established institutional norms (Fenno Jr., 1973).

Taken together, these results show that party affiliation is considerably more distinguishable from question text in some committees than in others, and that this variation does not map cleanly onto conventional expectations about which committees are most partisan. This conditionality is consistent with arguments that the strength and expression of partisan organization depend on political and institutional context (Aldrich and Rohde, 2000) and that legislators have incentives to differentiate their parties and cultivate distinct party reputations (Lee, 2009). Although our results do not directly test these theories, C-QUERI provides a resource for investigating whether variation in issue framing, committee function, electoral context, or other institutional features accounts for these differences.

Differences across Government Type. The relationship between partisan distinguishability and unified versus divided government appears to vary by committee. The findings reported in Table 5, and Table 20 in Appendix I, obtained by fine-tuning a BERT-based classifier and zero-shot prompting respectively, provide insights that connect ongoing debates in congressional scholarship. First, partisan identity shows the strongest classifier improvement over baseline in committees such as Energy and Commerce and Financial Services (See Table 5 rows 1 and 2), where regulatory debates often divide along party lines, consistent with theories

of partisan agenda control (Cox and McCubbins, 2005; Lee, 2009). Judiciary (Table 5 row 3) also shows consistent improvement across both government configurations. Second, classifiers perform inconsistently in committees associated with distributive politics or bipartisan traditions: Armed Services and Veterans' Affairs show stronger signal under unified and divided governments, respectively (See Table 5 rows 7 and 8). Third, while many committees show stronger partisan signals under divided government, consistent with findings on the dynamics of contested institutional control (Mayhew, 2005), there are notable exceptions. These results therefore do not support a simple account in which divided government uniformly strengthens partisan linguistic differences. Instead, they suggest that the relationship between partisan questioning and institutional control is itself conditional on committee context.

Differences across Hearing Types. Table 6 summarizes how predictive performance varies by hearing type and political context when we use question utterances to infer the speaker's party affiliation. These patterns reinforce the broader finding that partisan information in questioning is conditional rather than uniformly expressed across hearing settings. Determining whether this variation reflects differences in visibility, institutional function, issue composition, or other features of hearings is an important question for future research using C-QUERI.

Overall, these results are consistent with the idea

Hearing Type	Government	BERT	Baseline
General	All	58.42	51.51(R)
	Unified	51.72	50.46(D)
	Divided	59.57	52.14(R)
Field	All	57.21	54.90(R)
	Unified	53.34	53.30 (R)
	Divided	55.90	50.47(R)
Oversight	All	61.36	58.15(R)
	Unified	59.36	56.98(R)
	Divided	67.25	70.45(D)
Authorization	All	51.9	51.81(R)
	Unified	53.65	51.16(D)
	Divided	52.07	50.15(R)
Nomination	All	58.90	58.16(D)
	Unified	61.58	57.49(D)
	Divided	54.14	56.25(D)

Table 6: Predictions of party affiliation from question utterances across hearing types, comparing unified and divided Congress settings. The majority class is our baseline. An annotation of (R) or (D) indicates that Republican or Democrat speakers have more utterances.

that hearings with high visibility or explicit accountability functions create incentives for partisan messaging and framing that differ by party (Lee, 2009), and with accounts emphasizing strategic use of hearings for oversight and political positioning (McCubbins and Schwartz, 1984). By contrast, authorization hearings exhibit weaker affiliation signal, suggesting that when hearings focus on technical or distributive policy work, partisan language may be less distinctive (Krehbiel, 1991; Gilligan and Krehbiel, 1990). Results on predicting party standing across hearing types are provided in Appendix I.

6 Discussion and Summary

Our results suggest that questions in congressional hearings are not purely neutral instruments of oversight since partisan and institutional signals are detectable in how members question witnesses. More importantly, this signal is highly conditional. Partisan signals vary substantially across committees, hearing types, and government configurations, suggesting that language-based measures can detect where partisan conflict leaves strong imprints in questioning style and where it does not. The asymmetry between question-based and answer-based prediction underscores the distinct rhetorical roles of questioners and witnesses in congressional hearings. Whether partisan patterns in institutional dialogue differ qualitatively from those in the political texts on which LLMs are typically evaluated is a question our dataset can support in future work.

Our dataset and pipeline generalize beyond con-

gressional hearings to structured interview-like settings including media interviews, judicial proceedings, and expert testimony, providing a framework for comparative study of questioning practices across institutions. Future work will extend the dataset to other democratic institutions, link linguistic patterns to substantive outcomes such as policy changes and media coverage, and more systematically analyze how unified vs. divided government shapes rhetorical strategies.

Limitations

Data Coverage and Lag. Official transcripts published by the Government Publishing Office (GPO) are typically available only 2 months to 2 years after a hearing is held, meaning the dataset will always lag behind current congressional activity; expansion to more recent sessions is planned as transcripts become available. Many earlier congressional hearings are available only as scanned documents requiring OCR, introducing potential transcription errors and limiting extension to archival sessions; addressing this is a direction for future work.

Metadata Quality. Hearing-type metadata contains known classification errors, such as treaties occasionally misclassified as general hearings, which may affect analyses that condition on hearing type.

Interpretability. While our BERT-based classifier achieves strong predictive performance, interpreting what linguistic features drive predictions remains challenging; our NELA-based analysis provides a partial remedy by identifying interpretable linguistic differences between parties. While we demonstrate that party affiliation is detectable from questioning style, our analysis does not fully characterize the specific strategies that distinguish partisan questioning; a deeper qualitative and quantitative analysis of how members from different parties structure questions, deploy rhetorical devices, and frame issues remains an important direction for future work.

LLM Evaluation. Our evaluation of zero-shot LLM prompting reflects the models available at the time of the study; given the rapid pace of development in large language models, an extensive evaluation of LLMs at detecting party affiliation, a systematic comparison against lay and expert

human annotators, and the development of a benchmark for this task, deserve a dedicated future study.

Scope and Generalizability. Our substantive findings are specific to the U.S. congressional context and may not generalize directly to legislative questioning in other institutional settings, though our pipeline is designed to be extensible to such settings. The dataset covers only public hearings and does not include congressional meetings, where members deliberate without witnesses; question-answer dynamics in such settings may differ substantially.

Pipeline Robustness. Our hybrid NER pipeline relies on the formatting conventions of GPO hearing transcripts and may require adaptation if transcript structure changes, such as those introduced by remote hearings or evolving publication standards; however, the framework is flexible enough to accommodate domain-specific NER implementations for other settings.

Party Representation. The affiliation task treats Congress as effectively two-party, excluding Independent members; this simplification may become less accurate if third-party representation grows.

Open Questions. Our findings raise broader questions including whether partisan linguistic patterns in hearings have intensified over time in tandem with polarization, whether questioning strategies vary by member seniority or electoral vulnerability, and whether the linguistic features identified here predict downstream outcomes such as media coverage or policy change. Connecting committee hearing proceedings to actual legislative outcomes and lawmaking remains a broader challenge, as the causal pathways between hearing language and policy are complex and difficult to trace. Although we do not provide methods or results to address these questions in this work, we expect that our dataset will enable such future investigations.

Ethical Considerations

Data and Privacy. The C-QUERI dataset is constructed exclusively from public congressional hearing transcripts published by the Government Publishing Office. All statements made in congressional hearings are part of the public record, and members of Congress and witnesses appear in their official capacities. No personally identifiable information beyond what is already in the public

record is collected, stored, or released as part of this dataset. We do not include any data about private individuals.

Potential for Misuse. Our dataset and methods are designed to advance scholarly understanding of legislative behavior and democratic accountability. However, we acknowledge that tools capable of detecting partisan signals in language could potentially be misused, for example to profile individuals based on their communication style or to generate targeted political messaging. We encourage users of the dataset and methods to consider the ethical implications of downstream applications and to adhere to responsible use guidelines. We also note a dual-use tension inherent in our work: while the ability to identify effective questioning strategies may benefit legislators and researchers seeking to improve the quality of oversight, the same methods could be used defensively by witnesses or their advisors to anticipate and prepare for partisan lines of questioning. We believe the transparency afforded by open publication of our methods and data is preferable to leaving such analysis to well-resourced private actors alone.

Bias and Representation. Our analysis focuses on the two major U.S. political parties, which reflects the structure of Congress but excludes Independent members and third-party voices. Users of the dataset should be aware that models trained on this data may reflect and reinforce existing partisan categories. Additionally, our dataset spans a period of increasing political polarization, which may mean that models trained on it encode the specific rhetorical patterns of this era rather than more general properties of political discourse.

Broader Impact. By making the dataset and pipeline publicly available, we aim to lower the barrier to entry for interdisciplinary research at the intersection of NLP and political science. We hope that this resource will support research on democratic accountability, legislative representation, and the health of democratic institutions, and that the framework we provide will be extended to study legislative questioning in other democratic contexts worldwide.

References

Joel D. Aberbach. 1990. *Keeping a Watchful Eye: The Politics of Congressional Oversight*. Brookings Institution.

- John H. Aldrich and David W. Rohde. 2000. The Logic of Conditional Party Government: Revisiting the Electoral Connection. In Lawrence C. Dodd and Bruce I. Oppenheimer, editors, *Congress Reconsidered*, 7th edition, pages 269–292. CQ Press.
- Brian Arbour. 2014. *Issue Frame Ownership: The Partisan Roots of Campaign Rhetoric*. *Political Communication*, 31(4):604–627.
- Pamela Ban and Seth J. Hill. 2025. *Efficacy of Congressional Oversight*. *American Political Science Review*, 119(4):1742–1760.
- Pamela Ban, Ju Yeon Park, and Hye Young You. 2024. *Hearings on the Hill: The Politics of Informing Congress*. Political Economy of Institutions and Decisions. Cambridge University Press.
- Joseph M. Bessette. 1994. *The Mild Voice of Reason: Deliberative Democracy and American National Government*. University of Chicago Press.
- Shohini Bhattachali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749.
- Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. 2024. *Interactive question answering systems: Literature review*. *ACM Computing Surveys*, 56(9):Article 239, 38 pages.
- Sarah A. Binder and Forrest Maltzman. 2009. *Advice and Dissent: The Struggle to Shape the Federal Judiciary*. Brookings Institution Press, Washington, D.C.
- Lukas Birkenmaier, Laureen Sieber, and Felix Bergstein. 2025. *PolInterviews – A Dataset of German Politician Public Broadcast Interviews*. Preprint, arXiv:2501.04484.
- Dallas Card, Amber Boydston, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The Media Frames Corpus: Annotations of Frames Across Issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.
- Collin Coil, Caroline Bruckner, Natalie Williamson, Karen O’Connor, and Jeff Gill. 2024. *Still under-represented? Gender Representation of Witnesses at House and Senate Committee hearings*. *Journal of Women, Politics & Policy*, 45(4):429–445.
- Gary W. Cox and Mathew D. McCubbins. 1993. *Legislative Leviathan: Party Government in the House*. University of California Press.
- Gary W. Cox and Mathew D. McCubbins. 2005. *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge University Press.
- Ryan Delano, Aaron Rudkin, and In Song Kim. 2025. *Bulk Ingestion of Congressional Actions and Materials Dataset*. *Scientific Data*, 12(1):1567.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Patrick J. Egan. 2013. *Partisan Priorities: How Issue Ownership Drives and Distorts American Politics*. Cambridge University Press, New York.
- Ayse Eldes, Christian Fong, and Kenneth Lowande. 2024. *Information and Confrontation in Legislative Oversight*. *Legislative Studies Quarterly*, 49(2):227–256.
- Richard F Fenno Jr. 1973. *Congressmen in Committees*. Little, Brown and Company.
- Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Kartrik Erk. 2021. *Did they answer? Subjective acts and intents in conversational discourse*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.
- Jon H. Fiva, Oda Nedregård, and Henning Øien. 2025. *The Norwegian Parliamentary Debates Dataset*. *Scientific Data*, 12(1):4.
- Sean Gailmard and John W. Patty. 2013. *Learning While Governing: Expertise and Accountability in the Executive Branch*. University of Chicago Press.
- Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2018. *Congressional Record for the 43rd–114th Congresses: Parsed Speeches and Phrase Counts*. Stanford Libraries. Accessed: 2026-08-21.
- Matthew Gentzkow, 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.
- Thomas W. Gilligan and Keith Krehbiel. 1990. Organization of Informative Committees by a Rational Legislature. *American Journal of Political Science*, 34(2):531–564.
- Amy Gutmann and Dennis Thompson. 1996. *Democracy and Disagreement*. Belknap Press of Harvard University Press, Cambridge, MA.

- Arnav Hiray, Yunsong Liu, Mingxiao Song, Agam Shah, and Sudheer Chava. 2024. [CoCoHD: Congress Committee Hearing Dataset](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15529–15542, Miami, Florida, USA. Association for Computational Linguistics.
- Bai Linh Hoang. 2025. “Are You Too Busy to Listen Up?” Legislative (Dis)engagement from Constituents in Local Public Meetings. *Political Communication*, 42(4):715–734.
- Benjamin D. Horne, Jeppe Nørregaard, and Sibel Adali. 2019. [Robust Fake News Detection Over Time and Attack](#). *ACM Transactions on Intelligent Systems and Technology*, 11(1):Article 7, 23 pages.
- Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1113–1122.
- Kristin Kanthak and George A. Krause. 2013. *The Diversity Paradox: Political Parties, Legislatures, and the Organizational Foundations of Representation in America*. Oxford University Press.
- Lyn Kathlene. 1994. Power and Influence in State Legislative Policymaking: The Interaction of Gender and Position in Committee Hearing Debates. *American Political Science Review*, 88(3):560–576.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Keith Krehbiel. 1991. *Information and Legislative Organization*. University of Michigan Press.
- Douglas L. Kriner and Eric Schickler. 2016. *Investigating the President: Congressional Checks on Presidential Power*. Princeton University Press.
- Douglas L. Kriner and Liam Schwartz. 2008. [Divided Government and Congressional Investigations](#). *Legislative Studies Quarterly*, 33(2):295–321.
- Benjamin E. Lauderdale and Alexander Herzog. 2016. [Measuring Political Positions from Legislative Speech](#). *Political Analysis*, 24(3):374–394.
- Frances E. Lee. 2009. *Beyond Ideology: Politics, Principles, and Partisanship in the U.S. Senate*. University of Chicago Press.
- Frances E. Lee. 2016. *Insecure Majorities: Congress and the Perpetual Campaign*. University of Chicago Press.
- Thomas V. Maher, Charles Seguin, Yongjun Zhang, and Andrew P. Davis. 2020. [Social scientists’ testimony before Congress in the United States between 1946-2016, trends from a new dataset](#). *PLoS ONE*, 15(3):e0230104.
- Jane Mansbridge. 1999. [Should Blacks Represent Blacks and Women Represent Women? A Contingent “Yes”](#). *The Journal of Politics*, 61(3):628–657.
- Jane Mansbridge and John Parkinson, editors. 2012. *Deliberative Systems: Deliberative Democracy at the Large Scale*. Cambridge University Press.
- David R. Mayhew. 1974. *Congress: The Electoral Connection*. Yale University Press. 2nd ed. 2004.
- David R. Mayhew. 2005. *Divided we govern: Party control, lawmaking and investigations, 1946-2002*. Yale University Press.
- Nolan McCarty, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.
- Mathew D. McCubbins and Thomas Schwartz. 1984. Congressional Oversight Overlooked: Police Patrols versus Fire Alarms. *American Journal of Political Science*, 28(1):165–179.
- Michael D. Minta. 2011. *Oversight: Representing Interests of Blacks and Latinos in Congress*. Princeton University Press.
- Vipul Raheja and Joel Tetreault. 2019. [Dialogue Act Classification with Context-Aware Self-Attention](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.
- Christian Rauh and Jan Schwalbach. 2020. [The Parl-Speech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies](#). Harvard Dataverse, V1. Accessed: 2026-08-21.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–374.
- Michele L. Swers. 2002. *The Difference Women Make: The Policy Impact of Women in Congress*. University of Chicago Press.
- Michele L. Swers. 2020. *Women in the Club: Gender and Policy Making in the Senate*. University of Chicago Press.
- Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335.

- Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. [Natural Language Reasoning, A Survey](#). *ACM Computing Surveys*, 56(12):Article 304, 39 pages.
- Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. [Conversational question answering: A survey](#). *Knowledge and Information Systems*, 64(12):3151–3195.
- Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. [Asking too much? The rhetorical role of questions in political discourse](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1572, Copenhagen, Denmark. Association for Computational Linguistics.
- Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. [A Survey for Efficient Open Domain Question Answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

A Supplementary Statistics for the Dataset

This section contains the supplementary statistics and related discussions.

Fig.2 (a) illustrates the distribution of hearings across congressional sessions and the top ten most active committees. We notice some clear trends of congressional committees responding to domestic and global events. The House Committee on Financial Services which oversees the financial services industry in the United States shows a clear increase in the number of hearings in response to the 2007–2008 global financial crisis between the 110th and 112th congressional sessions which span the years 2007 to 2013. We also see a clear increase in the average number of witnesses called to hearing within the same period in Fig.2 (b). A spike in the number of hearings by the Judicial committee in the 110th session coincides with growing concerns over domestic warrantless surveillance and wiretapping by law enforcement and national security agencies.

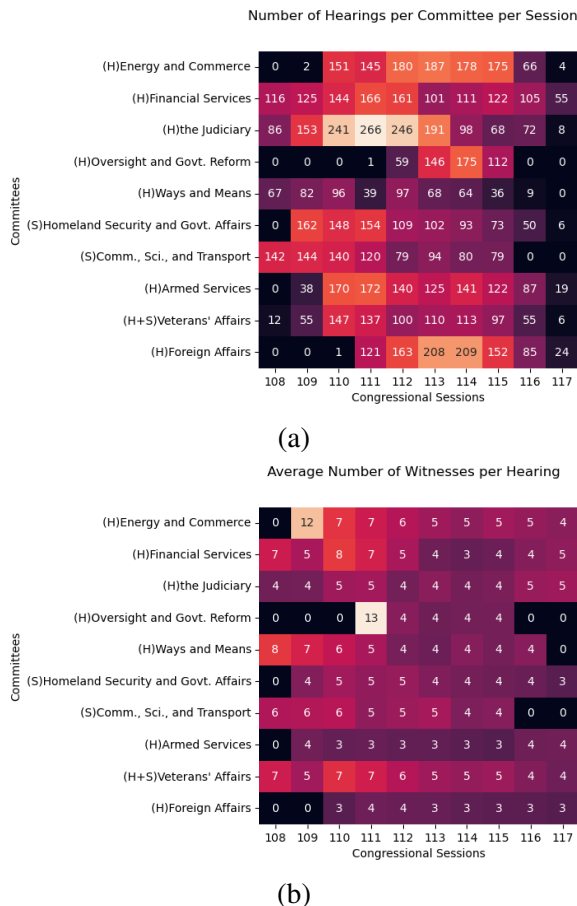


Figure 2: The (a) number of hearings and (b) average number of witnesses by committee and session

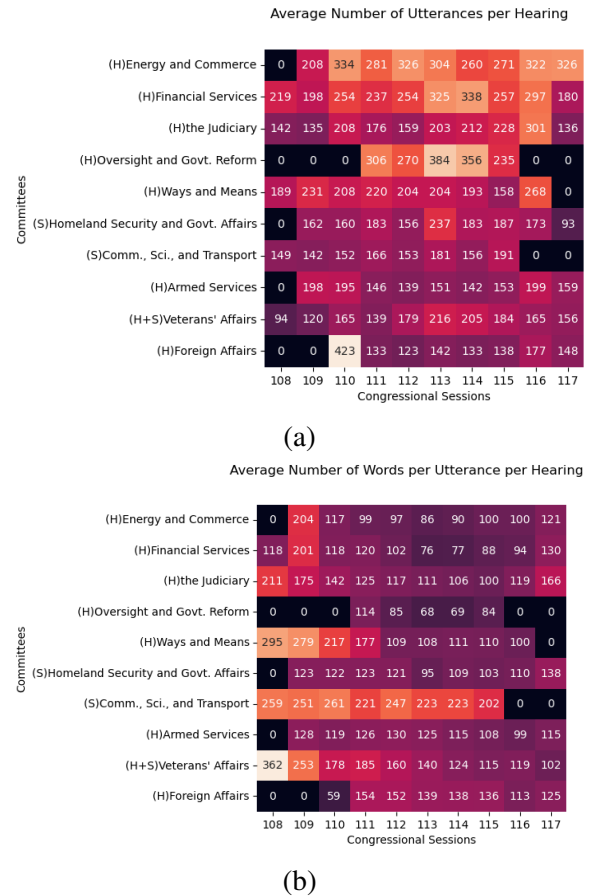


Figure 3: Heatmaps of (a) the number of utterances per hearing and (b) the average number of words per utterance, across committees and congressional sessions

B Identifying Questions and Answers

This section discusses, in further detail, the datasets used for training the BERT classifier for the Question and Answer identification task discussed in Section 3.2.

B.1 Datasets used for transfer learning

U.K. Parliamentary Question Hour The U.K. Parliamentary written answered questions are an important aspect of the democratic process in the United Kingdom. Members of Parliament (M.P.s) can submit written questions to government ministers on any issue within their portfolio. These questions are typically answered in writing within a set timeframe and are publicly available online on the U.K. Parliament's official website. This provides an opportunity for M.P.s to hold ministers accountable for their actions, even when Parliament is not in session. This dataset also provides an excellent source for labeled examples of questions and answers. We have used 4688 such exam-

ples (2344 questions and 2344 answers) to test our question-answer prediction model. Figure 4 is an example of an answered question from the House of Commons.

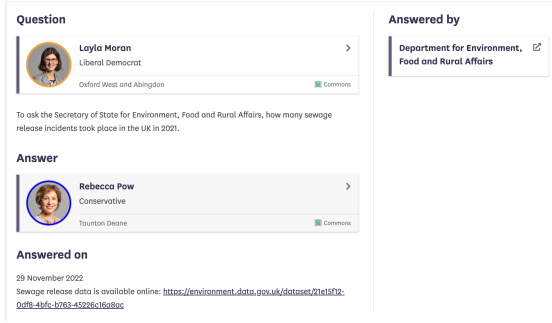
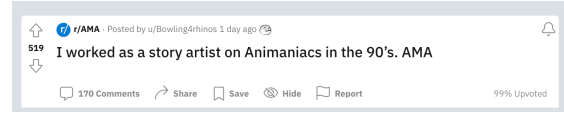


Figure 4: An example illustrating a question answered by a member of the U.K. Parliament in writing and the response from the concerned minister. Questions and answers are clearly annotated.

r/AskMeAnything The subreddit AskMeAnything, also known as AMA, is a popular online platform on Reddit that allows users to host a question-answering session that allows other users to ask the host any question. Hosts may include celebrities, politicians, scientists, athletes, and everyday individuals with unique experiences or perspectives to share. AMA sessions are typically hosted in a question-answer format, where the person answering questions, also known as the “host,” responds to questions posted by users in real time. This makes the posts a very rich source of labeled question-answer examples. We construct a dataset with 121512 examples, 60756 questions, and 60756 answers as follows to train our questions-answer prediction model. Each post is an introduction where the expert introduces themselves; Figure 5 (a) is an example of such a post and Figure 5 (b) is an example of comments on the same post. The first level of comments on these posts are typically questions, and comments by the post creator in the second level of comments answer the questions.



(a)



(b)

Figure 5: An Ask Me Anything (AMA) post in the subreddit r/AMA (a), and the subsequent comments on the post (b).

C Feature Descriptions from the NELA Toolkit

This section contains details about the features from the NELA toolkit (Horne et al., 2019) referred to in Section 4. Table 7 contains the computations of the Complexity features, Table 8 and Table 9 contain the descriptions of the Affect and Bias features respectively.

Feature	Description
vneg	Negative sentiment score using Vadar Sentiment
vneu	Neutral sentiment score using Vadar Sentiment
vpos	Positive sentiment score using Vadar Sentiment
wneg	Number of weak negative words
wpos	Number of weak positive words
wneu	Number of weak neutral words
sneg	Number of strong negative words
spos	Number of strong positive words
sneu	Number of strong neutral words

Table 8: Description of the **Affect** category of NELA Features

Feature	Description
ttr	Lexical Diversity also known as Type-Token Ratio
avgWlen	Average number of characters in a word
wCount	Number of words
FKGLvl	Flesch–Kincaid grade level: Standard readability measure computed by $0.39 * \frac{totalwords}{totalsentences} + 11.8 * \frac{totalsyllables}{totalwords} - 15.59$
SmgIn	Smog Index: Standard readability measure computed by $1.0430 * \sqrt{\#polysyllables * \frac{30}{\#sentences}} + 3.1291$
CLIn	Coleman–Liau index: Standard readability measure computed by $0.0588 * L - 0.296 * S - 15.8$ where L = avg # letters per 100 words and S = avg # sentences per 100 words
lix	LIX: Standard readability measure computed by $S + L$ where S = average sentence length and L = percentage of words with more than 6 letters. The scores usually range from 20 to 60.

Table 7: Description of the **Complexity category of NELA Features**

Feature	Description
bias	The number of bias words
assert	the number of assertive verbs
facts	The number of factive verbs
hedges	The number of hedge words
implctv	The number of implicatives
repVerb	Count of report verbs
poWords	number of positive pinion words
noWords	number of negative opinion words

Table 9: Description of the **Bias category of NELA Features**

D Statistical Testing with NELA Features

This appendix reports a detailed linguistic analysis based on two-sample Kolmogorov–Smirnov (K–S) tests of NELA features as discussed in Section 5. For each comparison, we sampled question utterances from members of different parties, with the null hypothesis that both samples were drawn from the same distribution.

Our analysis reveals several patterns; on average, both Democrats and Republicans use more complex language than Independents (Figure 6). Democrats also employ more positive sentiment words, whereas Republicans rely more heavily on neutral sentiment (Figure 8). Finally, Republicans make greater use of assertive and hedge words, while Democrats more often use implicative and positive opinion words (Figure 7).

The hues of each cell represent the difference

between the means of the two distributions under study. If the difference between the two distributions is not statistically significant, the cell has been hatched with the crossed pattern. Here, the R stands for Republican, D for Democrat, I for Independent, M for Majority party, m for Minority party, R.M. for Republican in Majority, and D.M. for Democrat in Majority.

E Hyperparameter Search

We discuss the hyperparameters that we have used for our gridsearches for the different models in Tables 10, 12, 11, and 13.

Hyperparameter	Search Grid
n_estimators	10, 20, ..., 100 (step of 10)
criterion	gini, entropy
max_depth	50, 100, ..., 500 (step of 50)
min_samples_split	0.0001, 0.005, 0.001, 0.05, 0.01, 0.5, 0.2, 0.1

Table 10: Random Forest Hyperparameters

Hyperparameter	Search Grid
C	0.001, 0.01, 0.1, 1, 10, 100
penalty	l1, l2
solver	liblinear, saga

Table 11: Hyperparameters for Logistic Regression

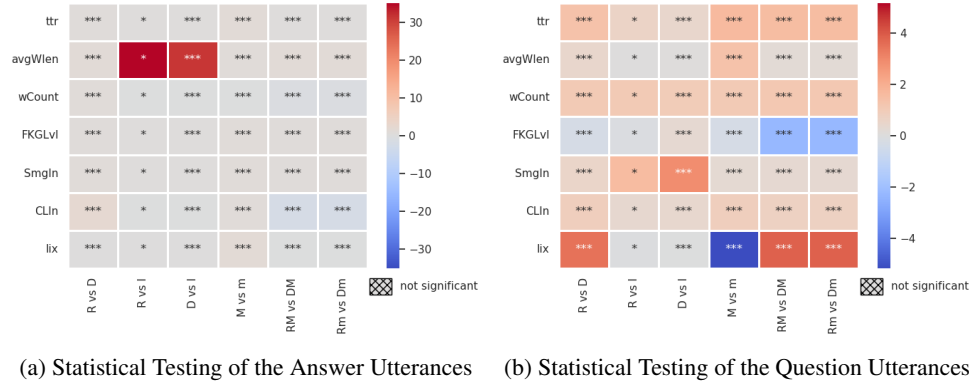


Figure 6: Statistical Testing of the Complexity NELA Features using two-sample Kolmogorov-Smirnov test. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks ‘*’ indicates the value of confidence for that statistical test. ‘***’ $\Rightarrow p \in [0, 0.001)$, ‘**’ $\Rightarrow p \in [0.001, 0.01)$, ‘*’ $\Rightarrow p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.



Figure 7: Two-sample Kolmogorov-Smirnov test on the Affect NELA Features using. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks ‘*’ indicates the value of confidence for that statistical test. ‘***’ $\Rightarrow p \in [0, 0.001)$, ‘**’ $\Rightarrow p \in [0.001, 0.01)$, ‘*’ $\Rightarrow p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.

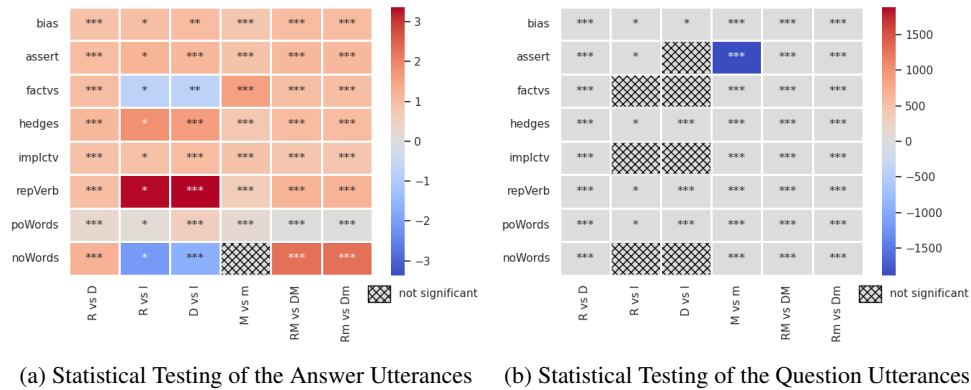


Figure 8: Two-sample Kolmogorov-Smirnov test on the Bias NELA Features using. The hue of each cell signifies the ratio of the means of the two samples being studied; red hue signifies that the left sample is larger where as the blue signifies that the right side sample has a larger mean. The number of asterisks ‘*’ indicates the value of confidence for that statistical test. ‘***’ $\Rightarrow p \in [0, 0.001)$, ‘**’ $\Rightarrow p \in [0.001, 0.01)$, ‘*’ $\Rightarrow p \in [0.01, 0.05)$. Differences that are not statistically significant are indicated with a cross-hatched pattern.

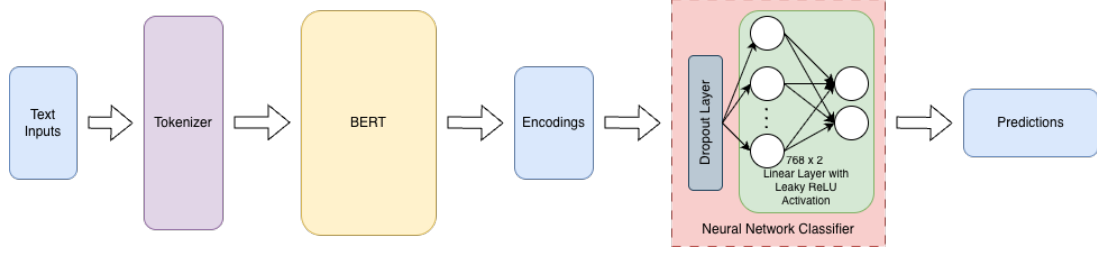


Figure 9: Architecture of the BERT classifier model.

Hyperparameter	Search Grid
C	0.1, 1, 10, 100
gamma	0.001, 0.01, 0.1, 1
kernel	rbf, sigmoid, linear

Table 12: Hyperparameters for Support Vector Classifier

F BERT Architecture and Fine-Tuning

This section discusses the BERT architecture and models used for the various tasks in the main paper.

Figure 9 shows the BERT architecture we have used in Section 3.1, Section 3.2, and Section 4 Approach 2. The model mentioned as "BERT" in Figure 9 used for Section 3.1 is <https://huggingface.co/dslim/bert-base-NER>, and the model used for Section 3.2 and Section 4 Approach 2 is <https://huggingface.co/google-bert/bert-base-cased>.

Hyperparameter	Search Grid
Initial Learning Rate	1e-5, 1e-7
# Epochs	5, 10, 15, 20
Batch Sizes	2, 32, 64, 128

Table 13: BERT Hyperparameters

G LLM Model Versions and Configurations

In this section we discuss the large language models (LLMs) that we have used in our experiments. The identifiers in parentheses reflect the version strings:

- LLaMA-3-70B (llama3:70b)
- DeepSeek-R1-70B (deepseek-r1:70b)
- Qwen-3-32B (qwen3:32b)
- Gemma-3-27B (gemma3:27b)
- Mistral (mistral:latest)

For each prediction task, we queried the large language models using zero-shot prompts designed to elicit the target label (party affiliation or party standing) directly from the text of the question utterance. The speaker’s name, typically present at the

start of each utterance, was removed to avoid trivial cues. Prompts were phrased in a standardized format across all models to ensure consistency, with variations only in the input text corresponding to each utterance. The exact wording of the prompts is provided in Table 14, for reproducibility.

H GPU and Environment Configurations

All LLM experiments were run on an NVIDIA RTX 6000 Ada Generation GPU with 49,140 MiB of memory. Inference was performed in a zero-shot setting without fine-tuning, using various batch-sizes across the models. Experiments were conducted in a Python 3.10 environment with PyTorch 2.2.1 and Hugging Face Transformers 4.41.2, running on CUDA 12.1 with cuDNN 8.9.

I Full Results (Supplementary)

This appendix provides supplementary tables containing additional results and related discussion that did not fit in Section 4 and Section 5. These include extended breakdowns by committee, hearing type, and linguistic feature, complementing the analyses presented in the paper.

Predicting standing. The results in Table 23, on predicting the party standing of the questioner, provide several insights for scholars interested in studying politics. First, we observe from the baselines that majority party members (M) often dominate question utterances, reflecting the procedural advantages of majority status. Committee chairs control agendas, allocate time, and structure witness panels (Cox and McCubbins, 2005).

Second, BERT outperforms the baseline in nomination hearings under unified government and Republican presidencies, where the Democrat minority generates the most questions. This indicates that minority questioners adopt linguistically distinctive strategies when challenging nominees, consistent with accounts of confirmation proceedings serving as an arena for opposition messaging and execu-

Prompt: “What follows is a {type_text} in a congressional hearing: {utterance_text} The question was asked by a person who is a member of a congressional committee, and whose party affiliation is either Democrat, Independent, or Republican. Based on the {type_text_2} above, what is the party affiliation of the person who asked the question? Answer with either D for Democrat, I for Independent, or R for Republican. Do not explain.”			
Variable name	Text according to utterance type		
	Question	Answer	Both
“type_text”	“question that has been asked”	“response to a question asked”	“question and its answer”
“utterance_text”	“Question:” followed by the text of the utterance	“ Answer:” followed by the text of the utterance	“Question:” followed by the text of the question and then “Answer:” followed by the text of the answer
“type_text_2”	“question”	“answer”	“question and answer”

Table 14: Prompts used for zero-shot predictions of party affiliation for different utterance types, standardized across all LLMs.

tive scrutiny (Binder and Maltzman, 2009). We note that the high overall accuracy in oversight and field hearings largely reflects class imbalance, suggesting that the content of questions may be less distinguishable despite different participation rates in monitoring contexts, perhaps because both parties engage in credit-claiming or position-taking with similar rhetorical templates (Mayhew, 1974).

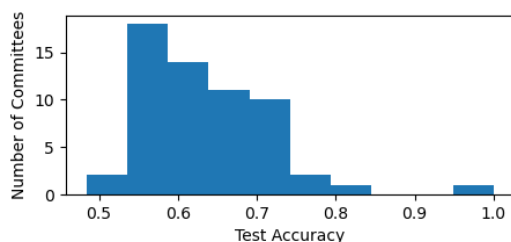


Figure 10: Histogram of test accuracies of the Random Forest Classifier on the NELA features across all committees for party affiliation.

J Generative AI Use

We used generative AI assistants (ChatGPT, Claude) to help write code for generating plots and visualizations, and to edit and, in some sections, draft prose from author-provided content and guidance grounded in the authors’ own analysis of the data and results. Generative AI was not used for research design, data collection, analysis, or interpretation of results. All AI-assisted text was reviewed, revised, and approved by the authors, and the authors take full responsibility for the content of this paper.

Hearing Type	#Hearings	#Utterances	#Questions
General	14774	2234567	579282
Field	403	46933	11962
Oversight	199	51581	14075
Authorization	439	58738	15381
Nomination	309	34360	8179
Treaty	1	88	22
Markup	8	2001	16

Table 15: The number of hearings, utterances and questions per hearing type.

NELA	Baseline	Session	Committee
100.00	51.67(R)	109	(H) Committee on Financial Services and Committee on Resources
73.78	61.39(D)	113	(S) Commerce, Science, and Transportation
76.00	55.17(D)	110	(S) Education and Labor and Committee on Health, Education, Labor, and Pensions
96.00	58.33(D)	113	(H) Committee on Armed Services Meeting Jointly with Subcommittee on Asia and the Pacific
76.36	50.00	117	(H) House Administration

Table 16: The top 5 combinations of Committees and Congressional session on which random forest classifiers using NELA features have the highest prediction accuracy in identifying party affiliations on question utterances. The Baseline column represents the majority-classifier.

Model	Question	Answer	Both
Llama	55.91	54.67	58.37
Gemma	55.63	52.81	58.50
Deepseek	55.62	52.84	60.78
Qwen	50.02	48.61	55.20
Mistral	54.01	50.24	56.14
LLM Baseline	52.12	52.12	52.12

Table 17: Results for identifying the questioner’s party affiliation when the input is the question, the answer in response to a question or both question and answer text using zero-shot prompting. The Baseline is the percentage of question-answer pairs with the majority class.

All		No Stop Words		No Speaker		Committee on
BERT	Base	BERT	Base	BERT	Base	
98.31	54.21	98.40	54.20	64.48	54.75	(H) Energy and Commerce
98.73	56.72	98.72	56.67	65.39	56.40	(H) Financial Services
97.88	54.88	97.87	55.05	62.54	55.19	(H+S) the Judiciary
98.80	69.48	98.84	69.42	72.48	70.21	(H) Oversight and Government Reform
95.06	54.69	94.91	54.92	64.69	66.47	(H) Ways and Means
98.43	63.71	98.18	63.59	67.13	63.99	(S) Homeland Security and Governmental Affairs
98.94	61.34	99.07	61.43	60.95	60.54	(S) Commerce, Science, and Transportation
95.82	51.75	95.68	51.62	59.27	51.20	(H) Armed Services
98.82	50.45	99.03	50.10	61.04	50.32	(H+S) Veterans’ Affairs
97.10	52.91	97.01	53.38	60.00	50.80	(H) Foreign Affairs

Table 18: Question utterance party affiliation prediction accuracy by session and committee obtained by training a BERT-based classifier

Llama	Gemma	Deepseek	Qwen	Mistral	Baseline	Committee on
56.38	58.60	56.82	50.51	56.04	54.36(R)	(H) Energy and Commerce
52.09	58.05	58.38	51.66	55.76	57.31(R)	(H) Financial Services
60.41	52.45	56.45	52.72	51.85	57.30(D)	(H+S) the Judiciary
52.01	66.85	59.08	44.53	63.18	70.85(R)	(H) Oversight and Govt. Reform
45.95	62.16	54.55	60.00	71.43	65.00(R)	(H) Ways and Means
54.53	44.12	47.90	43.23	36.89	60.47(D)	(S) Homeland Security and Govt. Affairs
56.92	51.36	54.69	51.26	52.55	57.64(D)	(S) Comm, Sci, and Transportation
59.28	51.89	56.79	50.54	51.49	50.77(R)	(H) Armed Services
50.00	47.40	52.63	45.97	49.18	51.18(D)	(H+S) Veterans' Affairs
55.25	56.91	54.31	49.58	53.65	56.00(R)	Foreign Affairs
0.86	0.63	5.21	3.24	2.72	-	Not Answered by LLM

Table 19: Zero-shot predictions of questioner party affiliation from question utterances (with the speaker's name removed), across the ten committees with the most utterances, using various LLMs. The Base column reports the majority-class baseline.

Committee on	Government	Llama	Gemma	Deepseek	Qwen	Mistral	Base
(H) Energy and Commerce	Unified	59.50	49.38	55.23	51.27	49.21	53.89(D)
	Divided	55.12	62.34	57.11	50.19	61.54	57.71(R)
(H) Financial Services	Unified	51.66	53.05	54.84	48.06	53.49	56.90(R)
	Divided	53.70	62.24	62.08	55.38	54.55	57.89(R)
(H+S) the Judiciary	Unified	57.73	54.11	56.58	52.82	53.82	56.51(D)
	Divided	62.34	51.46	56.74	52.46	68.44	58.03(D)
(H) Oversight and Gov Reform	Unified	47.52	63.37	57.61	48.48	70.41	71.03(R)
	Divided	53.02	67.63	61.34	43.62	33.33	70.82(R)
(S) Homel Sec and Gov Affairs	Unified	52.29	44.70	43.28	40.67	35.41	59.66(D)
	Divided	56.42	43.63	48.36	45.38	16.16	61.13(D)
(S) Comm, Sci, and Transportation	Unified	55.25	46.11	56.47	46.89	49.15	56.92(D)
	Divided	60.53	54.61	51.35	52.67	100.00	64.71(D)
(H) Armed Services	Unified	64.33	50.63	50.34	50.33	50.32	55.43(D)
	Divided	56.78	52.52	54.00	50.64	40.00	53.94(R)
(H+S) Veterans' Affairs	Unified	59.26	48.75	48.65	50.00	45.57	64.71(D)
	Divided	46.33	47.95	49.52	43.66	42.86	51.84(R)
(H) Foreign Affairs	Unified	55.06	46.07	54.02	51.69	48.86	53.00(D)
	Divided	55.31	60.44	53.38	48.86	50.00	59.00(R)

Table 20: Predictions of party affiliations (Republican, Democrat or Independent) from question utterances across hearing types, comparing unified and divided Congress settings. The Base column reports the majority-class baseline.(R) denotes that speakers from the Republican party have the majority number of utterances, (D) denotes Democratic speakers hold the majority class in the dataset.

Hearing Type		All		Democratic Presid		Republican Presidency	
		Majority	Minority	Majority	Minority	Majority	Minority
All	Unified	128189	85670	50065	28057	78124	57613
	Divided	224285	136869	197205	109798	27080	27071
General	Unified	116560	78621	46011	26064	70549	52557
	Divided	210931	129141	185003	103790	25928	25351
Field	Unified	2582	1187	802	240	1780	947
	Divided	4066	1433	3710	1059	356	374
Oversight	Unified	4008	1533	1455	612	2553	921
	Divided	2194	977	1729	806	465	171
Authorization	Unified	2630	2053	1035	785	1595	1268
	Divided	5827	3792	5496	3313	331	479
Nomination	Unified	2409	2276	762	356	1647	1920
	Divided	1252	1507	1252	811	0	696

Table 21: Number of question utterances from members of the party holding the majority/minority in the respective chamber under unified or divided governments from all sessions and sessions under Democratic and Republican presidencies.

Hearing Type		All		Democratic Presidency		Republican Presidency	
		Democrat	Republican	Democrat	Republican	Democrat	Republican
All	Unified	107678	106181	50065	28057	57613	78124
	Divided	173633	187521	141800	165203	31833	22318
General	Unified	98568	96613	46011	26064	52557	70549
	Divided	162430	177642	132269	156524	30161	21118
Field	Unified	1749	2020	802	240	947	1780
	Divided	2737	2762	2330	2439	407	323
Oversight	Unified	2376	3165	1455	612	921	2553
	Divided	2209	962	1729	806	480	156
Authorization	Unified	2303	2380	1035	785	1268	1595
	Divided	4684	4935	4219	4590	465	345
Nomination	Unified	2682	2003	762	356	1920	1647
	Divided	1548	1211	1228	835	320	376

Table 22: Number of question utterances from members of the Democratic or Republican parties under a unified or divided government over all sessions, and from sessions under Democratic and Republican presidencies.

Hearing Type	Government	All		Democrat President		Republican President	
		BERT	Base	BERT	Base	BERT	Base
General	Unified	58.30	59.43(M)	61.46	63.48(M)	59.80	57.02(M)
	Divided	58.65	61.27(M)	61.26	63.80(M)	62.20	50.66(M)
Field	Unified	69.08	69.00(M)	74.72	76.48(M)	63.63	64.49(M)
	Divided	72.63	73.72(M)	77.49	77.74(M)	50.91	53.20(m)
Oversight	Unified	72.28	73.00(M)	66.03	68.10(M)	73.61	73.75(M)
	Divided	67.72	68.30(M)	69.30	67.98(M)	70.60	71.13(M)
Authorization	Unified	55.72	55.63(M)	52.01	57.05(M)	56.62	55.40(M)
	Divided	59.25	59.96(M)	59.70	62.20(M)	59.30	59.10(M)
Nomination	Unified	62.55	50.52(m)	62.43	66.29(M)	62.08	54.38(m)
	Divided	57.00	55.87(m)	59.75	60.75(M)	100.00	100.00(m)

Table 23: Accuracy on predicting party standing (majority vs. minority) from question utterances across hearing types, comparing unified and divided governments. The Base column reports the majority-class in the test set, and are annotated either (M) or (m) to indicate that a majority of the utterances belonged to speakers from the chamber’s majority or minority party respectively. Results are grouped based on Democrat and Republican presidencies, while the ‘All’ column represents all governments.

Llama	Gemma	Deepseek	Qwen	Mistral	LLM Baseline	Hearing Type
55.89	58.60	55.62	50.31	54.15	51.89(R)	General
50.85	58.05	44.58	40.12	53.53	59.79(R)	Field
58.82	66.85	56.94	45.37	61.64	61.51(R)	Oversight
53.15	62.16	55.02	46.58	51.82	51.46(R)	Authorization
64.29	44.12	54.05	49.53	50.00	59.32(D)	Nomination
0.86	0.55	4.91	3.58	2.58	-	Not Answered by LLM

Table 24: Zero-shot predictions of questioner party affiliation from question utterances (with the speaker’s name removed), across hearing types, using various LLMs. The Baseline column reports the majority-class baseline.

Hearing Type	Government	Llama	Gemma	Deepseek	Qwen	Mistral	Base
General	Unified	55.08	53.17	54.50	49.70	53.70	50.16(D)
	Divided	57.46	57.65	57.50	51.13	56.77	52.48(R)
Field	Unified	46.30	38.89	45.10	32.08	0.00	58.33(R)
	Divided	54.12	59.14	48.72	45.68	66.67	57.30(R)
Oversight	Unified	61.90	56.19	55.56	46.15	66.67	60.18(R)
	Divided	69.44	50.00	54.29	52.94	100.00	58.54(D)
Authorization	Unified	55.70	45.00	46.05	42.86	80.00	50.59(D)
	Divided	53.78	50.41	66.17	50.83	40.00	51.54(R)
Nomination	Unified	63.64	48.48	56.06	50.00	50.00	61.19(D)
	Divided	61.76	50.00	50.00	42.42	50.00	56.41(D)

Table 25: Predictions of party affiliation, Democrat (D) vs. Republican (R) from question utterances across hearing types, comparing unified and divided Congress settings. The Base column reports the majority-class baseline and an annotation of (R) denotes that speakers from the Republican party have the majority number of utterances, (D) denotes Democratic speakers hold the majority class in the dataset.