A BERT-based Analysis of Evidence versus Intuition Rhetoric in U.S. House of Representatives and Constituent Education Levels

Avery Lee and Patrick Ruan and Sanjali Roy
University of California, Berkeley
{avery_lee, patrick.ruan, sanjaliroy}@berkeley.edu

Abstract

Differences in evidence-based versus intuition-based language have been found in politics across time, party, and individual politicians in previous literature. We explore how constituent education level affects House of Representatives' use of evidence-based versus intuition-based language from 2012 to 2023 in house hearings, using dictionary-based SemAxis methods and fine-tuning a BERT regression model. Although we find no meaningful correlation between constituent education level and representative rhetoric in the constituents' absence to observe unbiased speaking style, we discovered with mixed-effects regression that variation in evidence-based and intuition-based language is largely explained by differences between individual representatives, rather than their constituent education level or the party they belong to.

1 Introduction

The relationship between government representatives and their constituents is fundamental to modern-day governance. Previous research in political science reveals that the education level of a politician can influence the efficiency with which they govern (Sørensen, 2023), and constituents prefer to see representatives in power that have the same education level as them (Mayne and Peters, 2023). Additionally, social class affects language complexity, with children from working-class families using simpler language and sentences, while children from educated families use more complex grammar and reasoning (Bernstein, 2003). Individuals with higher cognitive ability are also more likely to engage in and require substantive evidence-based arguments for persuasion (Petty and Cacioppo, 1986). Studies in computational political science show that representatives across the globe use both evidence-based and intuition-based language in their argumentation (Aroyehun et al., 2025; Lasser et al., 2023; Carrella et al., 2025; Jordan et al., 2019), and much of this field of work examines public-facing speech (Carrella et al., 2025; Lasser et al., 2023; Jordan et al., 2019). An underexplored question at the nexus of these observations is whether constituent education levels affect representatives' rhetoric (i.e. use of evidence vs intuition-based language) when constituents are not present. We hypothesize that districts with higher education levels will be associated with greater use of evidence-based language by representatives.

2 Related Work

Related work includes studies investigating party-level rhetoric differences, linguistic complexity of politicians, and politicians' use of evidence-based versus intuition-based language in a variety of contexts, mostly when constituents are present.

2.1 Party-Level Rhetoric Factors

Furnas et al. (2025) analyzed a dataset tracking which types of papers policy documents in committee hearings cite. They found that if it was a Democrat-controlled committee, the documents were 1.8 times more likely to cite science than if it was from a Republican committee. However, this paper was not analyzing whether the language of the document was scientific, just whether it cited science. Simonsen and Widmann (2025) investigated how moral language is used in political discussions on immigration using moral dictionaries (i.e. dictionaries containing words which indicate a moral argument) and validation through crowdworkers. They found that party-level factors do not explain moral language use, but increasing levels of polarization does.

2.2 Politics & Linguistic Complexity

Some studies focus on how complex politicians' language is. Jiménez-Preciado et al. (2024) studied different presidential candidates' rhetoric in the US

presidential debates of 2024. They extracted a bag of words for each candidate and ran BERT (Devlin et al., 2019) for sentiment analysis. Democratic candidates Harris and Biden had higher lexical diversity, whereas the Republican candidate Trump evoked more emotion with his words. Similarly, Schoonvelde et al. (2019) analyzed speeches from 10 parliaments in Europe and found that liberal parties tend to use more complex language than conservative parties. In the US, Democratic (i.e. more liberal) voters are relatively more educated than Republican voters (Green and Van, 2023), and each district has a Democrat or Republican representative, so we were motivated to investigate whether voter education level affects the language of representatives.

2.3 Evidence-based versus belief-based political speech

Many studies apply NLP techniques to measure how evidence-based versus intuition-based politicians' speech is. Jordan et al. (2019) examines speeches by world leaders using the Linguistic Inquiry and Word Count and found there has been a general decline in analytical thinking in political speeches over time. Lasser et al. (2023) uses tweets by US Congress members to analyze how belief-based versus evidence-based language affects quality of speech. They measure quality using the NewsGuard information database which rates texts on trustworthiness. They found that belief-based speech lowers quality for Republican Congress members and fact-based speech increases quality for both parties, demonstrating the importance of rhetoric in affecting the trustworthiness of political speech. Similarly, Carrella et al. (2025) questioned whether replies to US Congress members posts on Twitter (now X) reflect the Congress members' original tweet's rhetoric, finding that the rhetoric (i.e. evidence-based versus belief-based language) used in the replies to a tweet align with those of the original tweet, emphasizing the influence of a politician's rhetoric. These three papers look at politicians' speech directed towards the public.

After synthesizing insights from previous research, our study builds primarily on Aroyehun et al. (2025). They analyzed evidence versus intuition-based language in congressional speeches and discovered that evidence-based language has declined over time. They made dictionaries of evidence versus intuition-based language, trained a Word2Vec embeddings model on congressional speeches, and calculated a score that is the cosine similarity of the text being analyzed and the dictionaries they made.

2.4 Our approach

Motivated by the distinction between evidence and intuition-based rhetoric, our study analyzes legislator speech when constituents are not present to observe their unbiased speaking style and correlates it with constituent education level. On top of this, we fine-tune BERT (Devlin et al., 2019), a contextual language model, because we think it could potentially improve performance of measuring evidence versus intuition-based language in dialogue, compared to previous literature that only uses dictionary-based methods.

3 Data

The dataset merges three data sources spanning Congress 112 to 118 (2011-2024). Congressional house hearing transcripts are obtained from GovInfo (U.S. Government Publishing Office, 2024). House committees discuss legislation proposals, conduct investigations, or evaluate government activities during these public congressional hearings. From these transcripts we extract a dataset that is on a sentence level, with each sentence annotated with the representative's metadata. The House of Representatives dataset contains representative state, district, congressional session, and party (Lewis et al., 2025). The US Census education dataset provides district-level population counts and the number of people who obtained bachelor's, master's, professional school, and doctorate degrees (U.S. Census Bureau, 2023). The mean percentage of higher education by each district for a congressional session is determined by the proportion of the population with a bachelor's degree or higher.

The final merged dataset comprises approximately 5.4 million sentences across 2,764 unique aggregated (state/district/congress) combinations, containing first name, last name, state, party, congress, district, dialogue, and bachelor's or higher percentages, enabling analysis of speech patterns as a function of the constituents' education levels.

4 Methods

4.1 Tokenization and Word2Vec Training

Tokenization is performed using SpaCy's english model (en-core-web-sm). A Word2Vec model is trained on the corpus using skip-gram architecture (Mikolov et al., 2013), 300-dimensional vectors, context window of 10 tokens, minimum word frequency of 5, and 20 training epochs. This captures domain-specific terminology and speech structure of the hearing corpus.

4.2 Semantic Axis Construction with SemAxis

SemAxis (An et al., 2018) is used to construct a bipolar axis between intuition versus evidence language. The initial seed words for both poles are adapted from Aroyehun et al. (2025). Our final seed dictionary consists of 48 evidence seed words and 32 intuition seed words; see Appendix A for the full dictionary. The seeds are expanded using cosine similarity. We added words with cosine similarity greater than 0.75 to the pole seeds and less than 0.35 to any opposing pole seed. This preserves the semantic separation between the words and reduces noise from ambiguous terms. If a word can either be evidence or intuition-based depending on different contexts, it should not be added to the dictionary. Lastly, we manually remove words with the same lemma and add words deemed relevant; since the original intuition dictionary predominantly contained negatively connoted terms, we also include intuition-based terms with positive connotations. Each word is assigned a SemAxis score along the axis via cosine similarity, with positive values indicating evidence-based language and negative values indicating intuition-based language.

4.3 Sentence-Level Scoring with TF-IDF

Sentence-level scores are computed by aggregating the word-level SemAxis scores weighted by term frequency-inverse document frequency (TF-IDF), treating each sentence as a document. TF-IDF weighting accounts for words with different semantic significance. For example, function words like "the" or "is" provide little context on rhetoric, whereas content-rich words like "believe" and "data" provide informative context for classification.

4.4 BERT Fine-tuning

An English BERT-based-uncased regression model (Devlin et al., 2019) is fine-tuned on normalized SemAxis sentence scores as training labels. The min-max normalization to $[0, 1]$ is implemented for better interpretation and alignment with educational attainment percentages.

The dataset is stratified into five bins to address high class imbalance, as there are more intuition-based sentences and neutral sentences than evidence-based sentences, and partitioned into training (70%), validation (15%), and test (15%). The five bins are very intuition $[0, 0.2)$, intuition $[0.2, 0.4)$, mixed $[0.4, 0.6)$, evidence $[0.6, 0.8)$, and very evidence $[0.8, 1.0]$. We use the BERT uncased WordPiece tokenizer with a maximum length of 128 tokens. The model is trained for three epochs with batch sizes of 64 for training and 128 for evaluation, using mean absolute error (MAE) as the evaluation metric. Mean absolute error is chosen as the evaluation metric because it is interpretable and robust against outliers. The BERT predictions are then min-max normalized to a $[0, 1]$ scale to maintain consistency with the SemAxis scores and education attainment percentage.

4.5 Validating Model Performance

The SemAxis and BERT models are validated against human-annotated "ground truth" labels. A stratified sample of 200 sentences (40 per bin) is selected from the test set and independently labeled by three annotators on a 0-5 scale (0 as entirely intuition, 1 as mostly intuition, 2 as some intuition, 3 as some evidence, 4 as mostly evidence, and 5 as entirely evidence). Intuition is defined as emotional or moral language and evidence as data-driven reasoning. For each sentence, human-labeled scores are computed as the mean of the three annotators' ratings, then transformed into a [0, 1] scale by dividing by 5.

We compare the SemAxis and BERT scores to the human labels using Pearson correlation and majority class baseline, getting the baseline MAE by using the mean human label for every sentence. A paired t-test and paired Cohen's D is performed to determine if the improvement is significant.

4.6 Representative Rhetoric vs Constituent Education Levels

The mean sentence scores are calculated for each representative in a congressional session. To examine the relationship between rhetoric and district education levels, the Pearson correlation coefficient and coefficient of determination (R^2) is computed between district education level and mean scores.

Since the data includes multiple sentences from the same representative, across 2 parties and 7 congressional sessions, the observations are not independent and form a hierarchical clustering structure. This means each observed data point is not independent from each other, so we implement a mixed-effects regression model to account for the clustering structure, which models both the fixed effects of the predictor variables and the random effects of these nested clusters. We group at the party, then individual representative level for our variation analysis.

5 Analysis

5.1 Sentence-Level Scores with SemAxis and TF-IDF

Using the seed dictionary, we construct a bipolar SemAxis representing a continuum from intuition-based (negative score) to evidence-based language (positive score), where each word is assigned a score along this axis. The scores range from -0.446 to 0.318. Examples of highly intuition-based words using SemAxis include "victimhood," "arrogance," "fear," and "immoral," while highly evidence-based words include "documenting," "assessments," "feasibility," and "study."

These scores weighted with TF-IDF give sentence-level scores that range from -0.27 to 0.19, with a mean of -0.03. We apply min-max normalization to scale these scores to a [0, 1] interval. These scores are then put into 5 bins, resulting in 1,329 very intuition, 316,469 intuition, 4,553,866 mixed, 547,910 evidence, and 864 very evidence sentences.

5.2 BERT Regression

We use the pseudo-labeled SemAxis scores to fine-tune a BERT regression model. We stop at 3 epochs because the training loss hits 0.000000 with a steady decline in validation loss and MAE.

Epoch	Training Loss	Validation Loss	MAE
1	0.000100	0.000029	0.003092
2	0.000100	0.000018	0.002613
3	0.000000	0.000012	0.002002

Figure 1: BERT Training Loss, Validation Loss, and MAE on Each Epoch

Predictions on the validation dataset achieve 0.000012 validation loss and 0.002 MAE, and the test dataset achieves 0.000012 test loss and 0.00199 MAE, indicating high performance with minimal overfitting. Using this BERT model, we predict the scores for the entire dataset, then apply min-max normalization to fit the [0, 1] interval.

5.3 Performance of SemAxis and BERT Compared to Human-Annotation

First, we manually extract a few examples to ensure both models predict as expected. Figure 2 shows a sample of intuition and evidence-based sentences.

Sentence	SemAxis Score	BERT Score	Rhetoric
The GAO has written a thorough and detailed report evaluating the CBM Program.	0.864602	0.870584	Evidence
The Air Force has completed 114 of the 189 Installations identified for Site Inspections.	0.828187	0.830321	Evidence
The Federal Investigative Standards outline the required elements of the investigation.	0.827293	0.835115	Evidence
It's hyperbole and lies.	0.191850	0.170180	Intuition
This is crazy.	0.190595	0.184200	Intuition
I just despise it.	0.096519	0.083544	Intuition

Figure 2: Sample of Intuition and Evidence-based Sentences

As shown in Figure 3, the distribution of scores for both models are nearly symmetric and bell-shaped, with extremely thin tails and most values close to the center (neutral rhetoric).

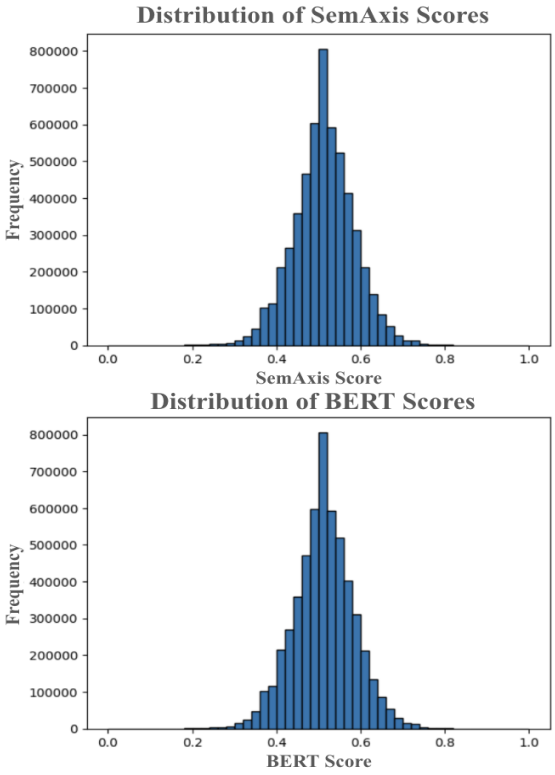

Figure 3: Distribution of Rhetoric Scores

In order to assess if the SemAxis and BERT predictions are accurate, we use human-labeled scores as ground truth. With 3 annotators, we get a Krippendorff's alpha of 0.802, indicating high inter-annotator agreement. Comparing the SemAxis and BERT scores with the human labels, we get a Pearson correlation value (r) of 0.866 for SemAxis and 0.864 for BERT, and MAE of 0.1105 for SemAxis and 0.1108 for BERT. At first glance, this seems to indicate that SemAxis performs slightly better than BERT; we use Steiger's Z-test to confirm if this is actually better performance or random chance. The Pearson correlation between SemAxis and BERT is 0.9996, meaning both models predict scores extremely similarly. Since the models perform nearly identically, the Steiger's Z value is high at 68.8443 with p-value of 0.0000, which makes the effect size (Δr = 0.002) and difference in performance negligible.

We use the majority class (mean) baseline to confirm if the MAE's are low enough to conclude high model performance. If we simply predict the mean human label of 0.5127 for every sentence, we get a baseline MAE of 0.2557. Comparing this baseline MAE to our SemAxis and BERT MAE's, we get 56.8% (SemAxis) and 56.7% (BERT) improvement. A paired t-test reveals statistically significant differences relative to the baseline, where the t-statistic measures how many standard errors away the effect is from the baseline. SemAxis (t-statistic = 13.901, p-value = 3.80e-31) and BERT (t-statistic = 13.657, p-value = 2.15e-30) both have high t-statistics, indicating significant improvement from the baseline. The paired Cohen's D indicates large effect size for both models (SemAxis = 0.983, BERT = 0.966), further confirming both models' high performance.

5.4 Relationship between Rhetoric and Educational Attainment

Calculating the mean of all sentence scores grouped by individual representatives for a congressional session, we find the Pearson correlation between rhetoric scores and their constituents' education levels during that session. Even though the p-value is less than 0.05 for both models (SemAxis = 0.0167, BERT = 0.0145), indicating statistical significance, due to the large sample size of 2,764, the relationship is negligible as the Pearson correlation (r) and coefficient of determination (R^2) are nearly 0 (SemAxis r = -0.046, SemAxis R^2 = 0.0021, BERT r= -0.047, BERT R^2 = 0.0022). There is no meaningful predictive relationship between constituent education levels with representative rhetoric in house hearings. The data points in Figure 4's scatterplot are generally spread out, which aligns with our finding that there is no meaningful relationship between these two variables.

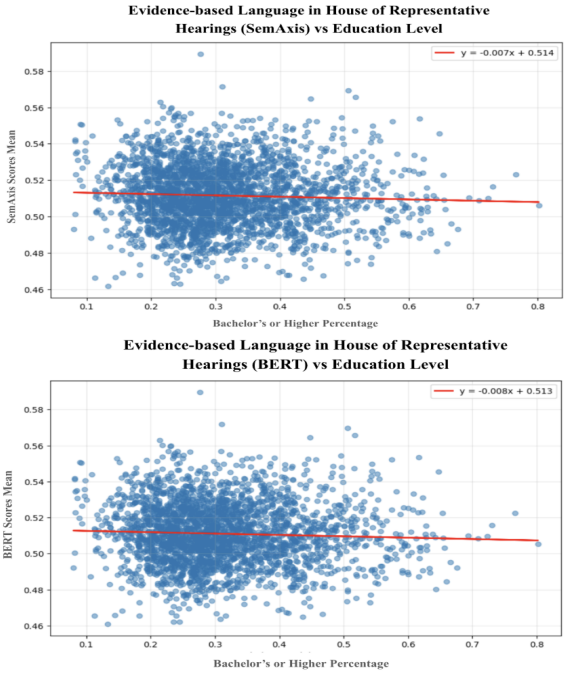

Figure 4: Relationship between Higher Education and House of Representative's Rhetoric in House Hearings

Each data point is a representative in a particular congressional session, so multiple data points can represent the same person across multiple sessions; the data points are not independent from each other. To account for this, we used random coefficients (not fixed slope) mixed-effects regression to find both fixed and random effects by grouping by different attributes. We first group by the two political parties in Congress: Democrat and Republican.

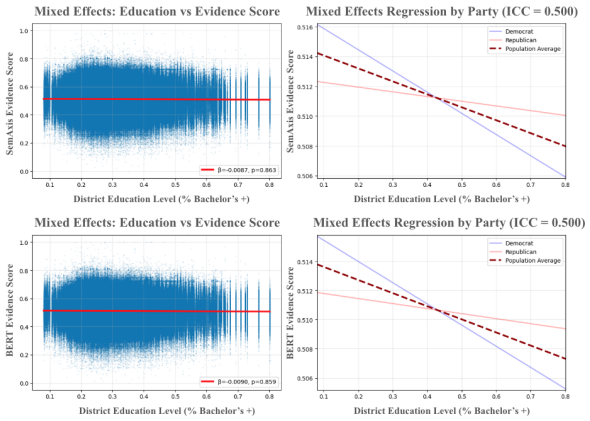

Figure 5: Mixed-Effects Regression by Political Party

Each data point on the left plot in Figure 5 is an individual sentence score. The slope is nearly flat (SemAxis = -0.0087, BERT = -0.0090) with high p-value (SemAxis = 0.863, BERT = 0.859). There is insufficient statistical evidence to reject the null hypothesis that the slope is 0, which aligns with our previous correlation observation. The Intraclass Correlation Coefficient (ICC) indicates that 50% of the variance in rhetoric scores is between parties, while the other 50% is within parties. There is poor reliability of measurements as half of the variance is due to true differences between parties and the other half is simply error variability.

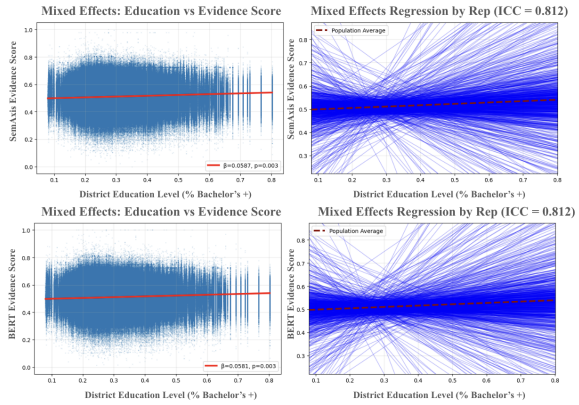

Figure 6: Mixed-Effects Regression by Political Party

Performing mixed-effects regression with a more granular attribute, we analyze data at the individual representative level. The slope is nearly flat (SemAxis = 0.0587, BERT = 0.0581) with low p-value (SemAxis = 0.003, BERT = 0.003), so even though there is strong statistical significance to reject the null hypothesis that the slope is non-zero, the magnitude of the effect is extremely small. Each line on the right graph in Figure 6 represents one representative. The ICC is 0.812, meaning 81.2% of the variance in rhetoric scores is between the representatives, and the other 18.8% is within an individual representative. Most of the variation in scores is due to true differences among representatives, and a small portion is from measurement error; rhetoric is largely consistent within a representative and there is high distinction in rhetoric between different representatives regardless of the party or district education level they represent. Representatives have distinct speaking patterns, but when rhetoric is aggregated by broader categories like party, or analyzed across the entire dataset, these individual differences average out, resulting in a near-zero correlation with constituents' education levels.

6 Conclusion

Even though there is no meaningful relationship between the constituents' education level of a congressional district and how much their elected House of Representative uses evidence versus intuition-based language in their absence, the mixed-effects regression reveals that there are significant individual differences between representatives in terms of their rhetoric. That is, which representative is speaking relates much more to their use of evidence-based vs intuition-based language than their constituent education level or which party they belong to. Additionally, previous studies in this area mostly rely on dictionary-based methods, and our study shows that BERT regression that utilizes context can be a viable method for rating rhetoric in a dialogue-based dataset.

There are some limitations of our study. District-level education census data was only available for 2012-2023, so effects may have been stronger if analysis was conducted over a longer period of time. Additionally, including state-level education data and Senators' rhetoric could have yielded different results.

There are several promising future research directions that could build off our work. One is investigating whether representatives' own education level affects their use of evidence versus intuition based language. Another is correlating other demographic traits of the constituents, such as immigration status, ethnicity, and socioeconomic status, with legislator rhetoric. We also studied representatives' rhetoric in their constituents' absence to observe their unbiased speaking style, but further research can use public speeches to directly study rhetoric shifts based on the audience's education level. Finally, one can examine whether representatives change their rhetoric over time. Future studies in this area could offer greater understanding into how legislator's rhetorical strategies relate to the constituents they serve.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

Segun T. Aroyehun, Almog Simchon, Fabio Carrella,

Jana Lasser, Stephan Lewandowsky, and David Garcia. 2025. Computational analysis of us congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, 9(6):1122–1133.

Basil Bernstein. 2003. *Theoretical Studies towards a Sociology of Language*, volume 1 of *Class, Codes and Control*. Routledge, London.

Fabio Carrella, Segun T. Aroyehun, Jana Lasser, Almog Simchon, David Garcia, and Stephan Lewandowsky. 2025. Different honesty conceptions align across us politicians' tweets and public replies. *Nature Communications*, 16(1):1409.

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.

Alexander C. Furnas, Timothy M. LaPira, and Dashun Wang. 2025. Partisan disparities in the use of science in policy. *Science*, 388(6745):362–367.

Andrew Daniller Scott Keeter Green, Hannah Hartig and Ted Van. 2023. Demographic profiles of republican and democratic voters.

Ana Lorena Jiménez-Preciado, José Álvarez García, Salvador Cruz-Aké, and Francisco Venegas-Martínez. 2024. The power of words from the 2024 united states presidential debates: A natural language processing approach. *Information*, 16(1):2.

Kayla N. Jordan, Joanna Sterling, James W. Pennebaker, and Ryan L. Boyd. 2019. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, 116(9):3476–3481.

Jana Lasser, Segun T. Aroyehun, Fabio Carrella, Almog Simchon, David Garcia, and Stephan Lewandowsky. 2023. From alternative conceptions of honesty to alternative facts in communications by us politicians. *Nature Human Behaviour*, 7(12):2140–2151.

Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2025. Voteview: Congressional roll-call votes database. Accessed: December 15, 2025.

Quinton Mayne and Yvette Peters. 2023. Where you sit is where you stand: education-based descriptive representation and perceptions of democratic quality. *West European Politics*, 46(3):526–549.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Richard E. Petty and John T. Cacioppo. 1986. *The Elaboration Likelihood Model of Persuasion*, volume 19, page 123–205. Elsevier.

Martijn Schoonvelde, Anna Brosius, Gijs Schumacher, and Bert N. Bakker. 2019. Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PLoS ONE*, 14(2):e0208450.

Kristina Bakkær Simonsen and Tobias Widmann. 2025. When do political parties moralize?: A cross-national study of the use of moral language in political communication on immigration. *British Journal of Political Science*, 55:e33.

Rune J. Sørensen. 2023. Educated politicians and government efficiency: Evidence from norwegian local government. *Journal of Economic Behavior Organization*, 210:163–179.

U.S. Census Bureau. 2023. Educational attainment: ACS 5-year estimates detailed tables. American Community Survey, Table B15003. Accessed: October 15, 2025.

U.S. Government Publishing Office. 2024. Congressional hearings collection. `https://www.govinfo.gov/app/collection/chrg/`. Accessed: 2024-11-15.

A Seed Words

Evidence Seed Words for SemAxis Dictionary: accurate, exact, intelligence, precise, search, analyze, examination, investigate, procedure, show, analysis, examine, process, statistics, correct, expert, knowledge, proof, study, correction, explore, question, trial, data, fact, learn, read, real, dossier, logic, reason, education, findings, logical, research, truth, evidence, information, method, science, pinpoint, intel, scientist, reasons, showing, discuss, examining, processes, investigations, educational, finding, methods, techniques, figures, numbers

Intuition Seed Words for SemAxis Dictionary: advice, doubt, mislead, suggestion, belief, fake, mistaken, suspicion, believe, mistrust, view, bogus, feeling, opinion, genuine, perspective, wrong, happy, optimistic, excited, deceive, guess, phony, deception, gut, viewpoint, think, dishonest, instinct, propaganda, trust, sense, intuition, honest, distrust, lie, suggest, recommend, standpoint