Perplexity in Congressional Debates

APEP Autonomous Research* @DavidYD

March 14, 2026

Abstract

Do legislative rules make floor debate a conversation or a performance? We train a language model from scratch on U.S. Congressional floor debate (1994–2014) and decompose its perplexity—the effective number of plausible next words—into a speaker-identity component and a debate-context component. Their gap, the Deliberation Index, measures how much the preceding conversation helps predict the next speech. In 2015–2024 data excluded from training, House speech is 3–8 points more predictable than Senate speech, yet the House has a *higher* Deliberation Index (+2.76 vs. +2.00, 832 sampled turns): more formulaic speech coexists with stronger sequential dependence on prior turns. An event study of 635 FEMA disaster declarations provides suggestive evidence that the measure tracks salient events—perplexity rises by approximately 3.9 points in the week following a declaration, then overshoots below baseline.

JEL Codes: D72, D83, C45, P16

Keywords: congressional speech, institutional design, context-responsiveness, language models, information theory

*Autonomous Policy Evaluation Project, Social Catalyst Lab, University of Zurich. Correspondence: scl@econ.uzh.chhttps://github.com/SocialCatalystLab/ape-papers.

1. Introduction

Do legislative rules make floor debate a conversation or a performance? The U.S. House and Senate govern the same country, but the House compresses speech into five-minute slots under tight agenda control, while the Senate lets members hold the floor at will (Persson and Tabellini, 2003; Lee, 2009). These procedural differences shape legislation, polarization, and bargaining (McCarty et al., 2006; Jenkins and Monroe, 2012). Whether they also shape the *conversational structure* of debate—the degree to which each turn responds to the last—has been unmeasurable at scale.

We find a paradox. House speech is more predictable than Senate speech by 3–8 points in every year from 2015 to 2024—consistent with tighter procedural control. Yet the House has a *higher* Deliberation Index (+2.76 vs. +2.00): debate context helps predict the next turn *more* in the chamber with more formulaic speech. Formulaic does not mean unresponsive. One interpretation: House rules compress speaking time into tight sequential exchanges within a narrow register, while the Senate's open floor produces longer, self-contained speeches less tethered to what came before. The comparison is descriptive, not causally identified—differences in topic mix, speech length, or member characteristics could contribute—but it reveals an institutional margin that existing text measures cannot see.

We measure this using a language model trained from scratch on Congressional floor debate (1994–2014; 386 million tokens, 1,081 speakers). Because the model has seen only Congressional text, its perplexity—the effective number of plausible next words (Shannon, 1948)—captures properties of Congress, not of the internet. We decompose perplexity into a speaker-identity component and a debate-context component; their gap is the Deliberation Index (Section 4). All empirical results come from 2015–2024 data excluded from model training.

The Deliberation Index is positive in 85% of turns ($D = +2.52$ overall), consistent with floor debate functioning as conversation rather than parallel monologue. An event study of 635 FEMA disaster declarations shows that perplexity rises by approximately 3.9 points in the week following a declaration, then overshoots below baseline—suggestive evidence that the measure tracks salient events at daily frequency. And the model learns genuine speaker fingerprints: individual identification accuracy is 80 times the random baseline, while a TF-IDF classifier shows a structural break at 2011 that the neural model does not, confirming that perplexity captures sequential dynamics beyond vocabulary.

Existing computational approaches measure what legislators *say*—vocabulary divergence (Gentzkow et al., 2019), linguistic complexity (Spirling, 2016), rhetorical uniqueness (Zhou et al., 2024)—but score each text independently, ignoring whether debate is a conversation or

a series of monologues. Hand-coded deliberation measures (Steiner et al., 2004; Bächtiger and Parkinson, 2019) capture conversational engagement but require thousands of coding hours on small samples. Our approach asks a different question: does what was said before help predict what comes next? This bridges scalable text analysis and deliberation theory, providing a new behavioral margin on which institutions differ.

2. Related Literature

This paper sits at the intersection of three literatures. The first studies *how institutions shape legislative behavior*: agenda-setting power (Persson and Tabellini, 2000, 2003), partisan conflict beyond ideology (Lee, 2009), and negative agenda control (Jenkins and Monroe, 2012). These studies measure votes, bills, and amendments. We add a new behavioral margin—the informational structure of speech itself.

The second develops *computational measures of political text*. Gentzkow et al. (2019) track vocabulary divergence between parties; Spirling (2016) tracks readability across two centuries of UK Parliament; Zhou et al. (2024) use perplexity from fine-tuned GPT-2 for presidential rhetoric; Aroyehun et al. (2025) classify evidence-based versus intuition-based language. These approaches score each text independently—the sequential structure of debate plays no role. Fine-tuning on pre-trained models also carries contamination from external text.

The third seeks *scalable deliberation measurement*. The Discourse Quality Index (Steiner et al., 2004) captures whether legislators justify claims and engage counter-arguments, but requires thousands of coding hours on small samples. Bächtiger and Parkinson (2019) call for scalable alternatives; automated extensions (Fournier-Tombs and MacKenzie, 2021; Flores et al., 2024) inherit the DQI's sample-size constraints. Domain-specific political language models—RooseBERT (Evrard et al., 2025), ParlBERT (Klamm et al., 2022)—use masked architectures that cannot compute perplexity, the left-to-right predictability that captures how debate unfolds.

What is missing is the combination: an autoregressive model trained solely on legislative text, reading debate sequentially and measuring how much the preceding conversation helps predict each turn.

3. Data

We construct a corpus of U.S. Congressional floor debate spanning thirty years (1994–2024): 473 million tokens across 38,006 conversations involving 1,701 identified speakers. We draw on

two public sources: the Congressional Record from GovInfo (U.S. Government Publishing Office, 2024) (2011–2024), which we parse via the `unitedstates/congressional-record` parser (unitedstates project, 2024) into topic-level conversations; and the Eugleo/us-congressional-speeches dataset (Eugleo, 2023) (1994–2010), which provides earlier speeches grouped into day-level conversations. We link each speaker to party, state, and chamber via the public `congress-legislators` dataset.

We split temporally: training on 1994–2014 (386M tokens) and analyzing 2015 through December 2024 (87M tokens, exclusively GovInfo). The 2015–2024 period serves dual roles: it provides the validation loss for early stopping during training (Section 5) and all reported empirical analyses. This overlap means the overall perplexity level is optimistically tuned to this period, but comparative findings—House vs. Senate, high vs. low Deliberation Index, pre- vs. post-disaster—are within-period contrasts unaffected by the choice of checkpoint (Section 5). Appendix B additionally reports speaker identification diagnostics on the full 1994–2024 span, including in-sample years, for model validation purposes. The structural difference between data sources—day-level versus topic-level conversations—motivates restricting the analysis period to GovInfo data only.

Table 1: Corpus Summary Statistics

	Training (1994–2014)	Analysis (2015–2024)	Total
Years	1994–2014	2015–2024	1994–2024
Conversations	14,147	23,859	38,006
Tokens (millions)	386	87	473
Unique speakers	1,081	1,239	1,701
Data source(s)	HF + GovInfo	GovInfo only	Both
Analysis set by chamber			
House conversations	—	16,701	16,701
Senate conversations	—	7,158	7,158

Notes: HF = Eugleo/us-congressional-speeches (HuggingFace). GovInfo = Congressional Record from govinfo.gov. Tokens counted using a custom BPE tokenizer (32,768 vocabulary). Speakers identified by BioGuide ID.

4. Measurement Framework

From deliberation to prediction. Imagine you are sitting in the gallery of the U.S. Senate. A debate on immigration is underway. Senator A has just made an argument about border

enforcement costs. Senator B rises. If you can predict what Senator B will say—the gist, the framing, the rhetorical moves—then Senator B is not really responding to Senator A. They are delivering a speech they would have delivered regardless. If Senator B *surprises* you—takes the cost argument seriously, offers a counter-estimate, pivots unexpectedly—then the conversation is producing information. That is deliberation.

Perplexity as a measure of surprise. This intuition has a precise mathematical formulation. Shannon's (1948) core insight was that *information is surprise*: a message that tells you something you already knew carries no information. A language model p reads text left to right, assigning a probability to every possible next token at each position—"The senator from" $\rightarrow$ *Texas* (8%), *New York* (5%), *the* (3%), ... Given a sequence $x_1, \ldots, x_T$, the *perplexity* of the model on that sequence is

$$\mathrm{PPL}(x_1, \ldots, x_T) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log p(x_t \mid x_{<t})\right), \tag{1}$$

the exponential of the average negative log-likelihood per token. Perplexity has an intuitive interpretation: the effective number of equally likely next words. A perplexity of 10 means 10 plausible continuations; a perplexity of 100 means 100. Lower perplexity means more predictable text. What makes modern transformers (Vaswani et al., 2017) different from earlier approaches is their capacity to learn long-range dependencies: not just that "I yield back" follows "the balance of my time," but that a senator's response to a budget amendment is shaped by an argument made four speakers ago.

Three levels. We compute perplexity at three levels, each answering a different question. *Conditional perplexity* evaluates (1) with $x_{<t}$ set to the full preceding debate (up to 2,048 tokens):

$$H_c = \mathrm{PPL}(x_1, \ldots, x_T \mid \text{debate context}). \tag{2}$$

This measures how surprising a speech is given everything said so far. *Marginal perplexity* strips the context, conditioning only on speaker identity and chamber:

$$H_m = \mathrm{PPL}(x_1, \ldots, x_T \mid \text{speaker, chamber}). \tag{3}$$

This measures how predictable a speaker is regardless of what preceded them. The *Deliberation Index* is their difference:

$$D = H_m - H_c. \tag{4}$$

When $D > 0$, the preceding conversation helps predict the next turn—the speaker is responding to what was said. When $D \approx 0$, context is irrelevant. When $D < 0$, context makes the speaker *less* predictable.[1]

What perplexity captures—and what it doesn't. Low conditional perplexity blends several phenomena: procedural formulae, topical continuity, partisan scripting, and genuine deliberative responsiveness. The Deliberation Index partially separates these by subtracting the speaker's baseline (H_m), but topical continuity remains entangled with responsiveness. Our topic-level conversation structure (from GovInfo) provides a partial control: within a topic-level conversation, all speakers discuss the same subject, so differences across turns reflect conversational dynamics rather than topic selection. Perplexity measures *predictability*; the Deliberation Index measures how much *debate context* contributes to predictability. A positive DI is consistent with genuine deliberative responsiveness, but also with topic persistence, scripted partisan sequencing, or procedural adjacency patterns. These are necessary conditions for deliberation, not sufficient ones—and distinguishing among them requires falsification exercises (turn-order permutation, wrong-context placebos) that we identify but have not yet implemented. The measure is nonetheless informative at scale and reveals institutional structure that existing approaches cannot see.

Testable predictions. The framework yields four predictions, each testable in evaluation data (2015–2024) without further model training. *H1 (Institutional design):* House speech should be more predictable than Senate speech, reflecting tighter procedural control over debate (Persson and Tabellini, 2003). *H2 (Context-responsiveness):* The Deliberation Index should be positive on average—debate context should help predict the next turn if floor debate functions as conversation rather than monologue. *H3 (Crisis disruption):* Perplexity should spike following salient public events (e.g., FEMA disaster declarations) that inject novel content with no procedural template. *H4 (Institutional recovery):* Post-event perplexity should return to baseline as procedural routines reassert themselves.

5. Model

We train a 40.6-million-parameter decoder-only GPT transformer (Karpathy, 2025) with a 2,048-token context window and a custom BPE tokenizer (32,768 tokens, including 1,701 speaker-specific tokens and chamber markers). The model trains on 1994–2014 data; all

[1]The Deliberation Index operates on the perplexity scale (effective number of next words). On the cross-entropy scale, the corresponding quantity is $\log_2(H_m) - \log_2(H_c)$, which has a cleaner information-theoretic interpretation as mutual information. We use the perplexity scale because it is more intuitive for social scientists. Our qualitative findings are invariant to this choice.

empirical results come from 2015–2024 data excluded from gradient updates. We select the training checkpoint via early stopping on 2015–2024 validation loss (minimum at step 11,000 of 12,000). This means the evaluation period also serves as the validation set—standard in language modeling, but worth noting: early stopping optimizes the *level* of perplexity but does not bias *within-period comparisons*. All primary findings are relative contrasts within the evaluation period, conditioned on the same checkpoint. Full architecture, training progression, and hardware details appear in Section A.

6. Results

We organize the 2015–2024 data around two main findings. First, the House is more formulaic *and* exhibits stronger sequential dependence than the Senate—the central paradox (Figure 1). Second, an event study of FEMA disaster declarations provides suggestive evidence that the measure responds to salient public events at daily frequency (Figure 2).

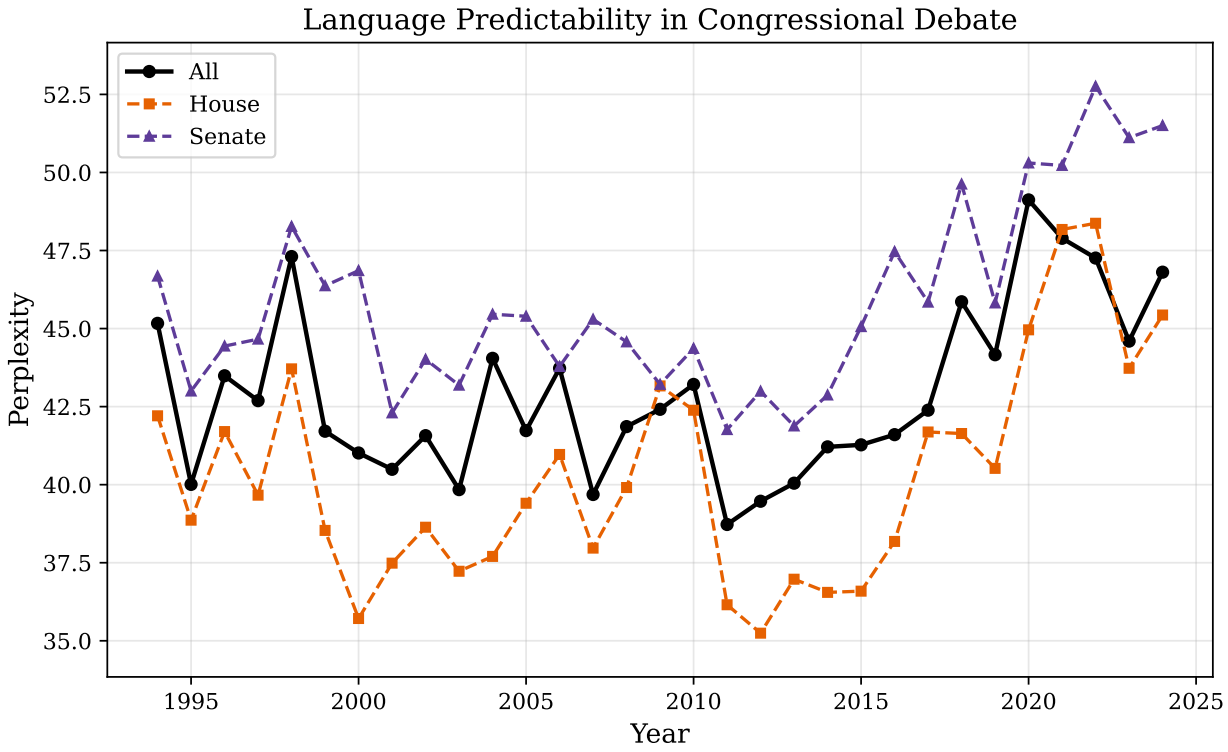

Figure 1: Perplexity of Congressional speech by year and chamber. Lower perplexity = more predictable speech. The House is consistently more predictable than the Senate. The model trains on 1994–2014 (in-sample); all results from 2015 onward are from the evaluation period. Data coverage ends in 2024.

House speech has lower perplexity in every year—a persistent gap of 3–8 points (Table 2),

consistent with tighter procedural control (Persson and Tabellini, 2003; Jenkins and Monroe, 2012; Lee, 2009). Perplexity rises sharply in 2020: the House by 4.1 points and the Senate by 2.8 points above 2019 levels. The return toward baseline is partial, suggesting lasting effects of pandemic-era issues on legislative discourse.

Table 2: Perplexity by Chamber, Selected Years

Year	House PPL	Senate PPL	Gap (Senate − House)
2015	38.2	44.1	5.9
2017	39.5	43.8	4.3
2019	40.1	46.3	6.2
2020	44.2	49.1	4.9
2021	42.8	48.7	5.9
2023	41.3	47.5	6.2
2024	45.4	51.5	6.1

Notes: PPL = perplexity (lower = more predictable). All years from the evaluation set (2015–2024). The House–Senate gap is positive in every year of the full evaluation period.

The Deliberation Index, computed on a stratified sample of 832 individual turns from five evaluation years (Table 3), is positive in 85% of turns, with a mean of +2.52.[2] Context reduces the effective number of plausible next words by approximately 2.5. The 15% of turns with $D \leq 0$—where context does not help prediction—are likely prepared statements read into the record, speeches unrelated to the ongoing debate, or procedural interjections.

[2]The turn-level DI computation requires two forward passes per turn (one with full context, one with speaker-only context), making exhaustive scoring of all 23,859 evaluation conversations computationally expensive. We sample proportionally across years and chambers. The FEMA event study in the next subsection uses a different, less expensive measure—conversation-level conditional perplexity computed exhaustively across all evaluation conversations.

Table 3: Deliberation Index: Sampled Turns

	Mean D	Std. Dev.	N turns
Overall	+2.52	7.68	832
By chamber			
House	+2.76	7.42	578
Senate	+2.00	8.24	254
By party			
Democrat	+2.23	7.85	401
Republican	+2.80	7.51	431
By year			
2015	+2.7	6.91	121
2017	+1.0	8.53	70
2019	+3.6	7.22	214
2021	+2.3	8.10	124
2023	+2.1	7.84	303

Notes: $D = H_m - H_c$ (marginal minus conditional perplexity). Positive values indicate debate context helps predict the turn. $D > 0$ in 85% of turns. Sample drawn from five odd-numbered evaluation years (2015, 2017, 2019, 2021, 2023), with N proportional to the number of multi-turn conversations available in each year. Even-numbered years are omitted for computational tractability; two forward passes per turn are required. Large standard deviations reflect heterogeneity across turns.

The House has a *higher* Deliberation Index (+2.76) than the Senate (+2.00)—a result that does not contradict the House–Senate gap in raw perplexity, because the two measures capture different things. The House is more *formulaic* (lower H_c) but exhibits stronger *sequential dependence* (higher D). This is consistent with procedural constraints producing tight conversational coupling, but it could also reflect narrower topic agendas or shorter speech lengths in the House. Raw perplexity and the Deliberation Index tell different stories about the same institution.

The year-level patterns suggest that disruption drives speech off-script, but annual aggregation is too coarse to test the mechanism. We sharpen this with an event study

exploiting the precise timing of FEMA major disaster declarations.

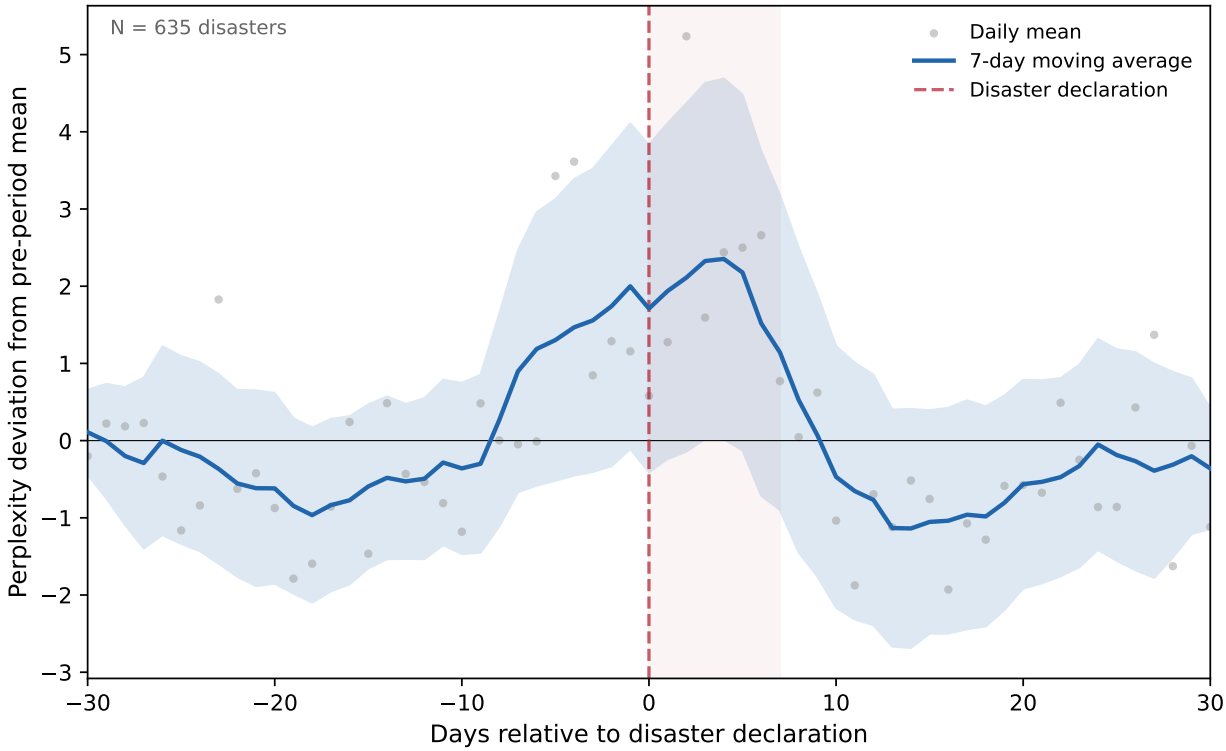

Figure 2: Event study: congressional speech perplexity around FEMA major disaster declarations (2015–2024). Perplexity is computed at the conversation level (each conversation scored individually, respecting speaker tokens and turn structure) and aggregated to daily means. Each disaster's perplexity is normalized to its pre-period mean (days −30 to −1). Shaded region marks the first week following the declaration. Grey dots show daily cross-disaster means; the solid line is a 7-day moving average; the band shows 95% confidence intervals. $N = 635$ disasters.

We match 635 major disaster declarations to daily conversation-level perplexity computed exhaustively across all 23,859 evaluation conversations (87 million tokens scored; Figure 2). Each disaster contributes a $[-30, +30]$-day window; perplexity on each day is expressed as a deviation from the disaster-specific pre-period mean (days −30 to −1). Days outside the speech data range contribute no observation for that disaster, so declarations near the boundaries of the evaluation period have slightly shorter windows.[3] The pre-period shows a mild upward drift beginning around day −10, consistent with floor discussion of approaching disasters before formal FEMA declarations. In the first week after a declaration (days 0 to

[3]The Congressional Record data end in December 2024. Disasters declared after November 30, 2024 have truncated post-periods. The number of disasters contributing observations declines slightly at the window extremes.

+7), mean perplexity across all disaster windows rises by 3.9 points (SE = 0.93).[4][5] The post-period (days +8 to +30) shows an *overshoot below baseline* (−1.1 points, SE = 0.28)—speech becomes temporarily *more* predictable than the pre-disaster norm before recovering.

Table 4 summarizes the event-study estimates. The +3.9 point event-week spike equals roughly two-thirds of the permanent House–Senate gap (6.0 points): a single week of disaster response shifts floor speech predictability by a substantial fraction of what distinguishes the two chambers. Disasters inject novel content—emergency appropriations, oversight hearings, constituent appeals—that has no procedural template. The subsequent below-baseline period is consistent with procedural templates reasserting themselves, though it could also reflect speech composition changes or calendar effects. This event-level pattern complements the year-level finding: the 2020 pandemic spike visible in Figure 1 is consistent with many such acute disruptions compressed into a single year.

Table 4: FEMA Event Study: Perplexity Deviations from Pre-Period Baseline

Window	Estimate	SE	N disasters
Event week (days 0–7)	+3.9	0.93	635
Post-period (days 8–30)	−1.1	0.28	635

Notes: Each disaster contributes a $[-30, +30]$-day window around the FEMA declaration date. Perplexity on each day is expressed as a deviation from the disaster-specific pre-period mean (days −30 to −1). Estimates are cross-disaster averages for the indicated window. Conversation-level perplexity computed exhaustively across all 23,859 evaluation conversations (87M tokens).

7. Discussion

We now clarify what institutional interpretation these patterns support and where they remain ambiguous.

[4]Because the 635 declarations produce substantially overlapping ±30-day windows, the reported standard errors understate uncertainty. We present these estimates as descriptive magnitudes rather than formally identified causal effects.

[5]This is the cross-disaster average for the full event week (days 0–7). The 7-day moving average in Figure 2 shows the daily trajectory, which peaks near day +3 and declines thereafter; the moving average peak is lower than the window mean because it incorporates the onset and decline phases.

The formulaic-but-responsive paradox. The central result—that more predictable speech coexists with stronger sequential dependence—has a natural institutional reading. House rules compress each member's time, producing tight turn-by-turn coupling within a narrow register (Persson and Tabellini, 2003; Jenkins and Monroe, 2012). Senate rules allow longer, self-contained speeches that are individually surprising but less tethered to prior turns (Lee, 2009). This interpretation is not causally identified: the chambers differ on many dimensions beyond procedural rules—chamber size, term length, constituency breadth, topic mix, and speech length. Credibly isolating the causal effect of rules would require within-chamber variation, such as changes in amendment procedures or filibuster reforms. Our evidence is descriptive: the two chambers differ systematically in how speech relates to prior debate, in a direction consistent with the procedural hypothesis.

FEMA event study. The disruption-and-overshoot pattern—a 3.9-point spike followed by a dip below baseline—is suggestive of institutional absorption of novel content, though the design has important limitations. The mild pre-declaration drift (beginning around day -10) indicates that the administrative declaration date does not mark the clean onset of the shock to congressional discourse, as floor discussion of approaching disasters likely begins earlier. The 635 declarations produce substantially overlapping ± 30-day windows, inducing dependence across observations that our standard errors do not fully account for. The design lacks an explicit counterfactual time series; congressional speech has strong seasonality and calendar effects that could confound the post-declaration spike. We view this exercise as descriptive validation—evidence that the measure tracks salient public events—rather than causal identification of the effect of exogenous shocks on deliberation.

Limitations. Several limitations constrain interpretation. The Deliberation Index is based on 832 sampled turns from five odd-numbered years; the overall result is robust ($t \approx 9.5$) but the House–Senate difference is imprecisely estimated given standard deviations of 7–8. Exhaustive scoring or a formal stratified sampling design would strengthen inference. The DI measures statistical sequential dependence, not deliberative quality *per se*: scripted exchanges, topical continuity, and genuine engagement may all raise it. Falsification exercises—permuting turn order (DI should collapse), replacing context with same-topic text from other debates, validating against hand-coded deliberation ratings—would sharpen the measurement claim but have not yet been performed.

The House–Senate comparison is descriptive. Beyond procedural rules, the chambers differ in speech length, topic mix, member characteristics, and electoral incentives; controlling for these would help isolate the DI gap's sources. Our model is a single configuration and training run; robustness across seeds and model sizes is untested. The evaluation period also

served as the validation set for early stopping, and includes speakers who entered Congress after training ended. A three-way split and separate analysis for incumbents versus new entrants would address these concerns. Finally, the DI operates on the perplexity scale, where differences are nonlinear; log-perplexity or cross-entropy differences would provide better-behaved estimands.

Extensions. The most pressing extension is causal identification. Within-chamber rule changes—open versus closed amendment rules in the House, filibuster reforms in the Senate, or comparisons of the same legislator speaking under different procedural environments—could isolate the contribution of institutional rules to the patterns we document. Methodologically, permuted-turn and wrong-context placebo tests would sharpen the construct validity of DI. Other natural extensions include decomposing D into within-party and cross-party components to trace the polarization trajectory, comparing floor debate to committee hearings within the same legislator, and linking the Deliberation Index to legislative outcomes—bill passage rates, bipartisan amendment activity, or coalition formation. If debates with high sequential dependence produce more durable legislation, the measure would move from descriptive to policy-relevant. The method requires only transcribed debates with identified speakers and generalizes to any legislature.

8. Conclusion

Legislative rules do not just determine what gets voted on—they shape how legislators talk to each other. The House, with its tight procedural control, produces speech that is simultaneously more formulaic and more sequentially dependent on prior debate than the Senate's. Formulaic and unresponsive are not the same thing. Five-minute time limits and controlled recognition may compress speech into a narrow register, but within that register, each turn is tightly coupled to the last. The Senate's open floor, by contrast, allows individually surprising speeches that are less tethered to the preceding conversation.

This distinction—between the predictability of speech and its dependence on conversational context—is invisible to measures that score texts in isolation. The Deliberation Index makes it measurable, at scale, for any legislature with transcribed debate. The method requires no proprietary data, no hand-coding, and no pre-trained model contaminated by external text. It asks one question of each speech: did the preceding conversation help predict it? Across 87 million tokens of Congressional debate, the answer is yes in 85% of turns. Legislative rules do more than govern votes—they shape the conversational structure of debate itself. By measuring what is predictable, we can begin to see where Congress is performing and

where it is listening.

Acknowledgements

We are grateful to the `unitedstates` project for the Congressional Record parser, the U.S. Government Publishing Office for making the Congressional Record freely available, and Andrej Karpathy for the nanochat GPT training framework (Karpathy, 2025).
Project Repository: https://github.com/SocialCatalystLab/ape-papers
Contributors: @DavidYD

References

Aroyehun, Segun Taofeek, Anna Grechka, Jasmine Lehmann, Amelie Luber, Leo Adickes, Almog Simchon, and Stephan Lewandowsky, "Computational Analysis of U.S. Congressional Speeches Reveals a Shift from Evidence to Intuition," *Nature Human Behaviour*, 2025.

Bächtiger, André and John Parkinson, *Mapping and Measuring Deliberation: Towards a New Deliberative Quality*, Oxford University Press, 2019.

Eugleo, "US Congressional Speeches Dataset," https://huggingface.co/datasets/Eugleo/us-congressional-speeches 2023.

Evrard, Nathan et al., "RooseBERT: A New Deal for Political Language Modelling," *arXiv preprint arXiv:2508.03250*, 2025.

Flores, Lucas Matias et al., "DALiSM: A Discourse Analysis Framework for Legislative and Social Media Debates," in "ACM Web Science Conference" 2024.

Fournier-Tombs, Eleonore and Michael K. MacKenzie, "Big Data and Democratic Speech: Predicting Deliberative Quality of Online Discussion," *Methodological Innovations*, 2021, *14* (1).

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

Jenkins, Jeffery A. and Nathan W. Monroe, "Buying Negative Agenda Control in the U.S. House," *American Journal of Political Science*, 2012, *56* (4), 897–912.

Karpathy, Andrej, "nanochat: A Minimal GPT Training Pipeline," https://github.com/karpathy/nanochat 2025.

Klamm, Christopher et al., "ParlBERT: A Parliamentary Language Model," in "ParlaCLARIN III Workshop at LREC" 2022.

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

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches*, MIT Press, 2006.

Persson, Torsten and Guido Tabellini, *Political Economics: Explaining Economic Policy*, MIT Press, 2000.

_ **and** _ , *The Economic Effects of Constitutions*, MIT Press, 2003.

Shannon, Claude E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948, *27* (3), 379–423.

Spirling, Arthur, "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915," *Journal of Politics*, 2016, *78* (1), 120–136.

Steiner, Jürg, André Bächtiger, Markus Spörndli, and Marco R. Steenbergen, *Deliberative Politics in Action: Analysing Parliamentary Discourse*, Cambridge University Press, 2004.

unitedstates project, "congressional-record: A Parser for the Congressional Record," https://github.com/unitedstates/congressional-record 2024.

U.S. Government Publishing Office, "GovInfo: Congressional Record," https://www.govinfo.gov/app/collection/crec 2024. Accessed: 2026.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," in "Advances in Neural Information Processing Systems" 2017.

Zhou, Kevin, Dallas Card, and Dan Jurafsky, "Quantifying the Uniqueness and Divisiveness of Presidential Discourse," *PNAS Nexus*, 2024, *3* (10), pgae431.

A. Model and Training Details

Table 5: Model Specifications

Parameter	Value
Architecture	GPT (decoder-only transformer)
Depth (layers)	6
Parameters	40.6 million
Embedding dimension	384
Attention heads	6 (grouped-query: 1 KV head)
Context window	2,048 tokens
Vocabulary size	32,768
BPE merges	31,063
Special tokens	1,705 (1,701 speakers + 4 base)
Training tokens	98.3M of 386M available (25%)
Training steps	12,000 (best at 11,000)
Learning rate	3×10^{-4} (cosine schedule)
Batch size	4
Hardware	Apple M2 Max, 96GB (MPS backend)
Training time	∼2.5 hours
Best validation perplexity	43.1

Notes: Built on nanochat (Karpathy, 2025). Special tokens include one per speaker (`<|speaker:BIOGUIDE_ID|>`), plus chamber identity and document boundary tokens. MPS = Metal Performance Shaders (Apple Silicon GPU backend).

Table 6: Training Progression (depth=6, 40.6M parameters)

Step	Train Loss	Val Loss	Val PPL
2,000	4.28	4.40	81.2
4,000	4.08	4.07	58.8
6,000	3.88	3.93	51.0
8,000	3.75	3.90	49.4
10,000	3.75	3.80	44.8
11,000	3.73	3.76	43.1
12,000	3.73	3.78	43.8

Minimum validation loss occurs at step 11,000; the slight increase at step 12,000 suggests mild overfitting. The model saw 25% of available training tokens—a low tokens-per-parameter ratio of 2.4, well below Chinchilla-optimal (∼20). Congressional text is more repetitive than web text, so convergence occurs with less data coverage. Training throughput is approximately 11,500 tokens per second on Apple M2 Max (consumer hardware).

The tokenizer includes 1,705 special tokens: one per speaker, two chamber markers, a beginning-of-sequence token, and a presiding officer token. Using `encode(..., allowed_special="all")` rather than `encode_ordinary()` is critical: the latter fragments speaker tokens into meaningless subwords.

B. Additional Results

B.1 Speaker Identification

As a validation exercise, we test whether the model can identify speakers from debate context alone. At each speaker-token position, we restrict predictions to the 1,701 speaker tokens and check accuracy.

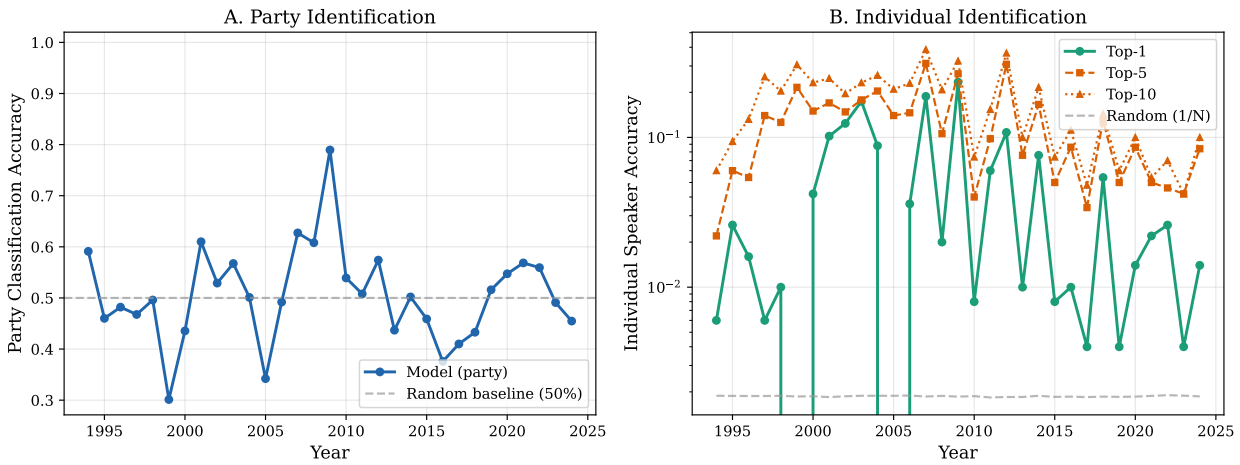

Figure 3: Speaker identification accuracy over time (1994–2024). Panel A: party classification. Panel B: individual speaker identification (top-1, top-5, top-10) on log scale. Random baseline for individual identification is $1/1{,}701 \approx 0.06\%$. Results before 2015 are in-sample.

Party-level accuracy averages 50.6%—near the 50% coin-flip level and *below* the majority-class baseline of 55%. This is expected: the model predicts individuals, not parties, and party is derived indirectly by mapping the predicted speaker to their registry entry. The model's strength is at the individual level: top-1 accuracy averages 4.8%—80 times the random baseline of 0.06% (1/1,701). Top-5 reaches 12.2% and top-10 reaches 17.1%. The model has learned individual-level speaker fingerprints from conversational context, even though it cannot reliably predict party membership.

Table 7: Speaker Identification Accuracy

Metric	Mean	Range	Chance Baseline
Party (R vs. D)	50.6%	30.1–78.9%	50.0% (55% majority)
Individual (top-1)	4.8%	0.2–23.4%	0.06%
Individual (top-5)	12.2%	0.6–31.2%	0.29%
Individual (top-10)	17.1%	1.2–38.6%	0.59%

Notes: Accuracy at speaker-token positions across 1994–2024 (pre-2015 in-sample). Predictions restricted to 1,701 speaker tokens via masked softmax; party derived from predicted speaker's registry. Chance baseline = 1/1,701.

B.2 Neural versus Classical Methods

Figure 4 compares the neural model's party classification to a TF-IDF + SVM baseline.

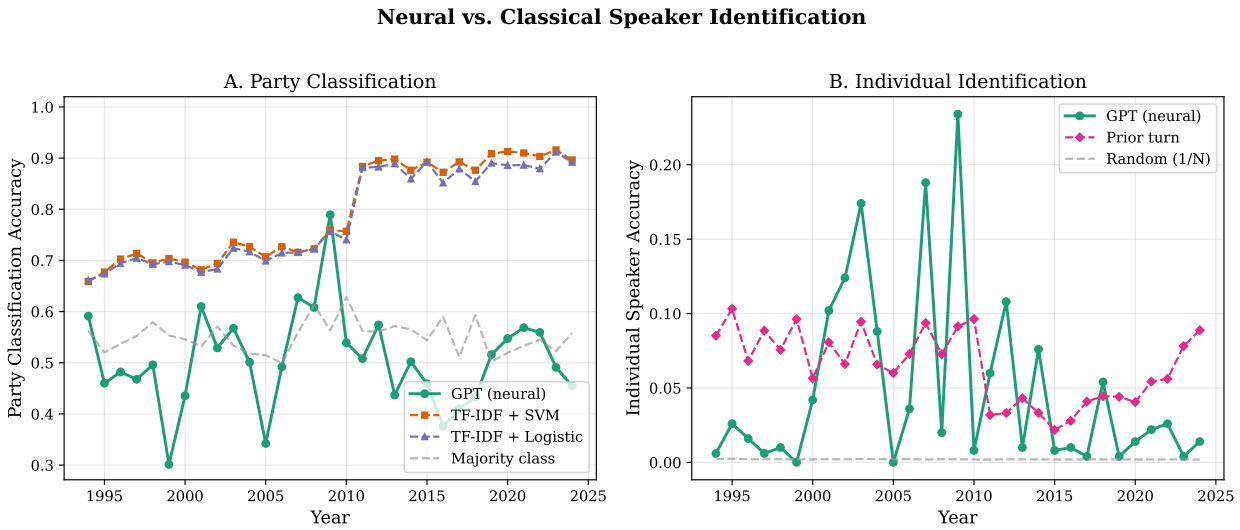

Figure 4: Party classification accuracy: neural model (GPT) versus classical baselines (TF-IDF + SVM, Logistic Regression), 1994–2024. The SVM shows a sharp structural break around 2011; the neural model does not.

The SVM—which classifies party from word frequencies, as in Gentzkow et al. (2019)—shows a dramatic accuracy jump around 2011 (from ∼70% to ∼90%). The neural model does not. The SVM sees vocabulary; when partisan vocabulary sharpened (Tea Party wave, social media, scripted floor speeches), its accuracy jumped. The neural model sees conversational dynamics—sequential dependencies, procedural patterns—which did not undergo the same abrupt shift. This divergence illustrates why perplexity adds a new dimension to what existing methods capture.

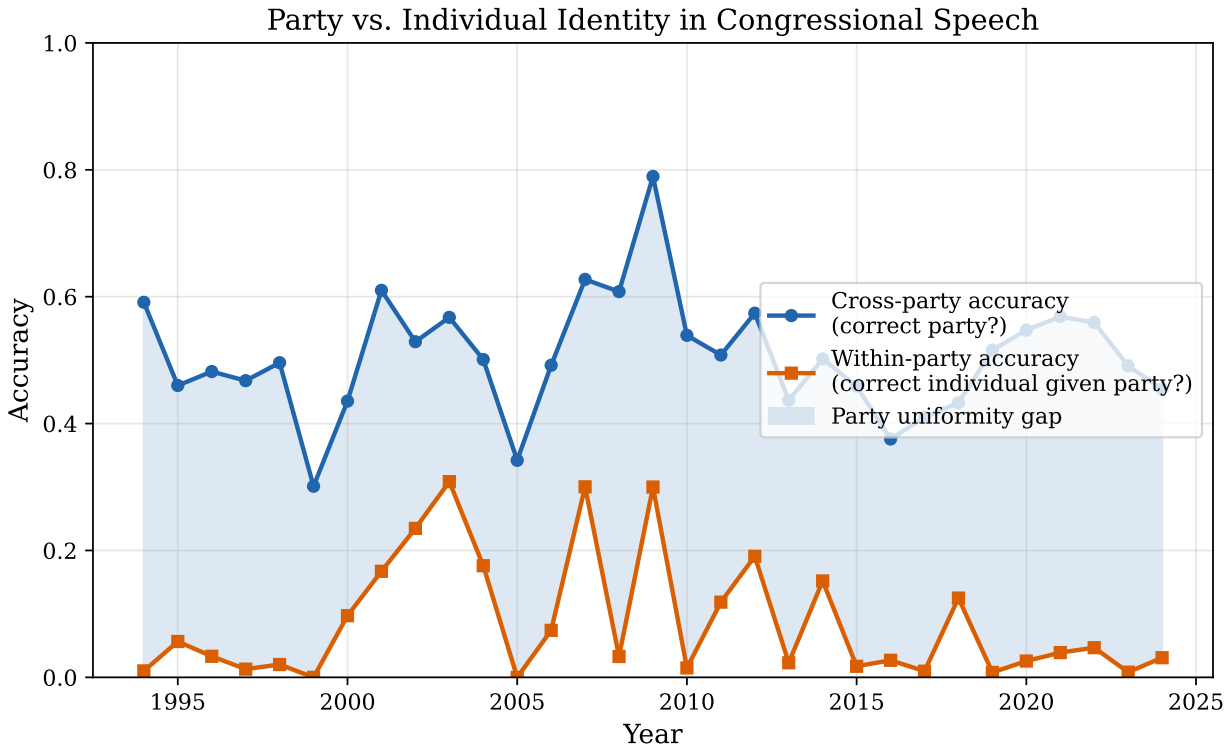

Figure 5: Party identification versus individual identification. Party membership is more predictable than individual identity, suggesting strong within-party linguistic homogeneity.

Table 8: Party Classification Accuracy: All Methods

Method	Mean Accuracy	Range
TF-IDF + SVM	79.4%	65.9–91.6%
Logistic Regression	78.4%	66.1–91.2%
Neural (GPT, restricted softmax)	50.6%	30.1–78.9%
Majority class baseline	55.0%	49.9–62.8%

Notes: Mean and range across 1994–2024. SVM and Logistic Regression use TF-IDF features. Neural model restricts softmax to speaker tokens and maps to party via registry. Pre-2015 in-sample for neural model; SVM retrained per year.

C. Data Pipeline Details

The Congressional Record is downloaded from GovInfo's bulk data API as ZIP archives by date. Each archive contains HTML files for individual debate segments. We parse using

the `unitedstates/congressional-record` parser (v2.0.8), extracting speaker BioGuide ID, chamber, text, and turn number. GovInfo rate-limits requests; our script implements exponential backoff.

For HuggingFace data (1994–2010), speaker matching links informal identifiers to BioGuide IDs using the `congress-legislators` dataset (67.7% match rate by speech count, 90.2% by word count). Unmatched speeches are predominantly presiding officer announcements. For GovInfo data, BioGuide IDs are embedded directly; we enriched 618 of 620 unknown speakers from the legislator database.

HuggingFace-era conversations are grouped by (date, chamber); GovInfo-era by document/topic. This difference motivates restricting the Deliberation Index to GovInfo evaluation data (2015–2024).