Partisan Opinions, but Common Language: Similarities in Topic Use by Appellate Judges

Anonymous ACL submission

Abstract

As the final word on thousands of legal matters each year, appellate courts make some of the most impactful decisions in modern society. Understanding partisan behavior by their judges is therefore critical for the rule of law. However, judicial language is technical, making partisanship challenging to objectively measure and creating a unique opportunity for natural language processing. Using fine-tuned language embeddings from transformer models, we leverage the random assignment of individual judges to three-judge panels, and of those panels to cases, to causally estimate how discussion of legal topics on U.S. appellate courts differs across partisan environments. We show that while Democratic judges write more dispersed opinions, judges of both parties agree on average about the important topics in each legal case. Further, we demonstrate that mandatory bipartisanship does not reduce the range of topics considered. Judicial partisanship is thus driven by disagreements within legal issues rather than disputes about which issues apply. These results provide a clearer understanding of the structure of judicial language and open new directions for natural language processing research and impact.

1 Introduction

Politically-motivated behavior by judges is a long-standing topic of contention which has become central to U.S. politics leading up to the 2024 presidential election. Key to the judicial system are the thirteen federal courts of appeals, which serve as the final word on about 40,000 cases each year on topics ranging from criminal sentencing to antitrust to government regulation. Appellate court opinions are among the most impactful decisions in today's society, but despite their influence, they are often highly technical and difficult for laypeople to parse. Heavy use of domain-specific language sets judicial opinions apart from recent areas where polarized text has been studied, including social media (Jiang et al., 2023), news (He et al., 2021), and Congressional speech (Gentzkow et al., 2019). Given the importance of a politically neutral judiciary, a systematic and setting-aware study of partisan language in appellate court opinions is both a new challenge for natural language processing (NLP) and a critical tool for understanding and maintaining judicial accountability.

In this work, we study U.S. appellate courts using random assignment of individual judges to three-judge panels, and of those panels to cases. While prior work has found that opinions are distinguishable by author party (Lu and Chen, 2024), the mechanisms for this difference have not yet been explored. We take a step towards this important question by causally estimating differences in legal topics across two sets of partisan environments. Legal opinions aim to explain a panel's decision, e.g., is a ruling on a worker's termination about First Amendment protections, or the nature of employment contracts? By quantitatively analyzing legal topics, we can understand whether partisan differences are driven by differences of opinion within shared legal approaches (both parties agree the case is a First Amendment dispute but disagree on the extent of speech protections) or by the lack of a common framework (a Democratic judge focuses on the First Amendment while a Republican judge emphasizes contract law). Separating *what* judges discuss from *how* they discuss it provides a clearer picture of partisanship on appellate courts.

We learn a topic-driven embedding of judicial opinions, and succeed at predicting the primary topic among eight standard options with 73.2% accuracy, validating our approach and proving sufficient for the tasks at hand. Using predicted topics, we first compare all-Democratic panels to all-Republican panels to understand how partisan judges approach legal issues. We show that both parties cover highly similar topics, with Demo-

cratic judges showing a slight tendency towards opinions in broader areas of law and exhibiting more intra-party dispersion. Next, we explore the effect of mandatory bipartisanship by comparing bipartisan panels to party-unanimous ones. We demonstrate that bipartisan opinions include the same number of major topics, show no topic-distributional differences, and are just as dispersed as party-unanimous opinions. Encouragingly, these results suggest that the counter-partisans do not harm the richness or breadth of appellate courts' discussion. Even when rulings are influenced by partisan beliefs, we show that the terms and language of the discussion remain shared.

2 Setting and Data

The U.S. federal court system is divided into 94 districts, each covering a particular geographic area. Cases from those courts can be appealed to twelve geographically based circuit courts or to the Federal Circuit.[1] The circuit courts' total caseload is about 40,000 cases per year; about 7,500 of these are filed for further appeal with the Supreme Court, but less than 200 are actually reviewed. Thus, for ≈99.5% of cases, a circuit court is the final word.

The vast majority of cases on circuit courts are heard by panels of three judges. Those panels are randomly generated at the beginning of each term (subject to administrative constraints—see, e.g., Levy and Chilton 2015) and have been the subject of decades of academic study (Sunstein et al., 2006; Ash et al., 2024). Panels rule by majority vote on whether the district court's decision should be upheld or reversed. For cases where the decision creates substantive legal precedent, the panel issues a lead opinion describing their reasoning; those opinions are the focus of this work.

Our main data is provided by the Data Science Justice Collaboratory (details in Ash et al. 2024). Text is from Bloomberg Law and contains the universe of published U.S. appellate court opinions from 1890 to 2013. Besides text, we have the names of the three judges on the panel, the opinion author, and a hand-coded primary legal topic from Bloomberg Law in one of eight categories: criminal, civil rights, First Amendment, due process, privacy, labor relations, economic activity/regulation, miscellaneous. There are ≈250,000 lead opinions with full metadata and at least one page of text. Judge characteristics originally come from the Songer Project, the Federal Judicial Center, and data collection by Chen and Yeh (2020). They include age, gender, and self-reported political party. There are ≈2,500 judges in the sample. In the Appendix, Table 2 summarizes opinion text; Table 3 and Figure 3 summarize judge data.

[1] See Figure 2 in the Appendix for a visualization.

3 Methods

We access two standard transformer models (DistilBERT, 66m param., from Sanh et al. 2019; RoBERTa, 110m param., from Liu et al. 2019) and one subject-specific model (LegalBERT, 110m param., from Chalkidis et al. 2020) via HuggingFace and fine-tune in native PyTorch to predict primary legal topic using cross-entropy loss. Before fine-tuning, we remove names of places and people using Presidio to ensure results are not driven by, e.g., learning that cases from Washington, D.C. are more likely to be about government regulation. We divide opinions exceeding the maximum input length into 500-token chunks with 25 tokens of overlap, treat each chunk as a separate document for fine-tuning and prediction, and average predictions and embeddings to obtain a single output per opinion. 25% of the opinion chunks (≈350,000 documents) are used for fine-tuning with a 70:30 train:test split. We follow hyperparameter search recommendations in the relevant works. Each fine-tuning run takes ≈15 hours on an NVIDIA Tesla V100 GPU with 32GB of RAM. Reported results use only the held-out 75% of opinion chunks.[2]

4 Results

4.1 Prediction of Legal Topics

Despite the unbalanced dataset (dominated by criminal, due process, and economic activity/regulation), our fine-tuned model is highly successful in predicting legal topic. The best-performing model is LegalBERT, with overall accuracy of 73.2%. The F1 scores for two of the three dominant topics are strong[3] and the weighted average F1 score across all topics is 0.722. Furthermore, most classification errors are intuitive and suggest overlap between topics—e.g., many incorrectly classified labor cases are predicted as economic activity/regulation, but almost none are

[2] We obtain results for almost all opinions since only opinions where all chunks were used for fine-tuning are omitted.

[3] Those are criminal (0.884) and economic activity/regulation (0.798). Under-performance for due process (0.581) is unsurprising since due process overlaps with other topics (e.g., due process violations can be criminal or civil).

predicted as criminal cases. A confusion matrix (Table 4) and per-topic metrics (Table 5) are provided in the Appendix. These results are strong for our setting, validate the approach, and form the basis for our comparisons of legal topic use across partisan environments.

4.2 All-Democratic vs. All-Republican Panels

Three-judge panels are randomly assigned to cases, preventing a judge from choosing which types of cases they hear. However, this institutional feature is insufficient for causal analysis of authorship, which is assigned at the discretion of the panel's senior judge and may be affected by administrative reasons, subject-matter expertise, or politics (Farhang et al., 2015). For example, a higher proportion of the economic activity/regulation topic among Republican-authored opinions may reflect both an increased tendency by Republicans to raise economic issues (our target estimand) and a greater likelihood that Republicans are chosen to write opinions on economically-driven cases (a confounder). We therefore compare all-Democratic panels to all-Republican panels to ensure that author party is independent from "facts of the case."

For each true label, representing the main legal topic of a case, we follow prior work on text partisanship (Gentzkow et al., 2019) and represent the prevalence of topics by their predicted probabilities. We compute three metrics for each opinion: per-topic probabilities, number of topics above a uniform threshold, and number of topics with higher predicted probability than their prior probability.[4] We then estimate a fixed-effects regression model to account for random assignment within each circuit-by-year and control for author age and gender.[5] E.g., for the probability of using the "criminal" topic we index by opinion o and estimate,

$$\begin{aligned}\mathbb{P}(crim.)_o = &\alpha\, party_o + \beta\, age_o + \gamma\, gender_o \\ &+ circuit_o \times year_o + \varepsilon_o.\end{aligned} \tag{1}$$

Our estimated coefficient $\hat{\alpha}$ then captures the causal effect on the share of text covering the "criminal" topic from switching an all-Republican panel to an all-Democratic one. The first column of Table 1 shows the coefficients for each metric.[6] Overall, there is little difference between the topics each party favors. Democratic judges are somewhat more likely to discuss First Amendment or "miscellaneous" issues in their opinions, while Republican judges are slightly more likely to discuss economic and regulatory issues. In absolute terms, these differences represent only about a percentage point change in the average share of text devoted to those topics. All-Democratic panels are about 2.4% more likely to discuss an additional major topic (defining "major" as posterior $>$ prior).[7] This result lines up with higher use of the wide-reaching miscellaneous topic, and may suggest that Republican judges have narrower discussions, e.g., because they have more similar legal approaches. We address within-party homogeneity using embeddings in Section 4.4.

Metric	All-D. vs. All-R.	Split vs. Unan.	Base Rate
$\mathbb{P}$(Crim.)	-0.011	-0.005	0.246
	(0.024)	(0.012)	
$\mathbb{P}$(Civ. Rights)	0.034	0.003	0.060
	(0.030)	(0.013)	
$\mathbb{P}$(1st Am.)	0.196*	-0.050	0.005
	(0.115)	(0.041)	
$\mathbb{P}$(Due Process)	-0.002	0.003	0.264
	(0.013)	(0.007)	
$\mathbb{P}$(Priv.)	-0.009	0.014	0.001
	(0.149)	(0.060)	
$\mathbb{P}$(Labor)	0.031	-0.001	0.052
	(0.043)	(0.015)	
$\mathbb{P}$(Econ./Reg.)	-0.036**	0.006	0.326
	(0.016)	(0.008)	
$\mathbb{P}$(Misc.)	0.132***	-0.013	0.047
	(0.042)	(0.016)	
# w/ $\mathbb{P} > 0.125$	-0.002	0.003	1.702
	(0.006)	(0.002)	
# w/ post. $>$ prior	0.014**	0.003	1.605
	(0.007)	(0.002)	
Observations	75,214	246,554	46,005

Table 1: Causal change in topic metrics. A coefficient of 0.1 represents ≈10% increase over the base rate. * = sig. at 10%, ** = sig. at 5%, *** = sig. at 1%.

4.3 Bipartisan vs. Party-Unanimous Panels

Another key dimension of partisanship is whether judges behave differently when collaborating across party lines. Given the prevalence of bipartisan institutions (e.g., Congressional committees, state redistricting commissions), systematically different discourse in the presence of counterpartisans may have major implications for accurate decision-making. We again leverage random assignment by comparing bipartisan panels to party-unanimous panels using the same three metrics—topic probabilities, above-uniform topic counts, and Bayes-more-likely topic counts.[8] We follow the estimation model of Equation (1) but replace the party variable with an indicator for bi-

[4] Tables 6 and 7 in the Appendix show that the latter two comparisons are robust to scaling those thresholds.

[5] For topic probabilities, we use a binomial specification; for number of topics we use a Poisson specification.

[6] See Figure 4 in the Appendix for full distributions.

[7] Uniform thresholds below 0.125 produce a similar effect.

[8] See Figure 5 in the Appendix for full distributions.

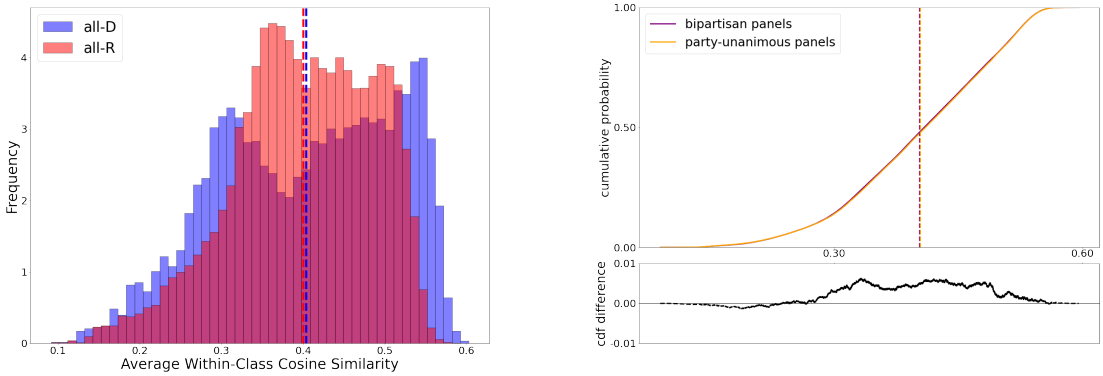

Figure 1: Within-class global similarity, comparing parties (left) and panel splits (right).

partisan panels and an additional control for author party to mitigate unknown author assignment. Results are in the second column of Table 1.

Similarities are even more pronounced than in Section 4.2; there are no significant differences in any individual topic share. Given the lack of major differences between all-Democratic and all-Republican panels, these results are largely unsurprising—if judges of each party approach the law from similar angles, a bipartisan panel will likely have a qualitatively similar discussion to a party-unanimous one. Further, bipartisan panels discuss the same number of major topics. This result suggests, encouragingly, that mandatory bipartisanship does not curtail discussion.

4.4 Comparisons in Embedding Space

Our prior comparisons focused on reduced-form predictions. To take advantage of more fine-grained semantic information, we also compute global similarity metrics using the full embedding vectors.[9] For a given reference opinion u, we compute the average similarity to opinions v_p produced by panels p in the class P, which we denote $\bar{s}_P$:

$$\bar{s}_u^P = \sum_{p \in P} \frac{u \cdot v_p}{|u|\,|v|}. \tag{2}$$

Letting D denote all-Democratic panels, R denote all-Republican panels, B denote bipartisan panels, and U denote party-unanimous panels, we compute $\bar{s}_u^P$ for all u and all $P \in \{B, U, D, R\}$. The left panel of Figure 1 shows that $\bar{s}_u^D$ has a bimodal distribution, with the one peak below the single mode of $\bar{s}_u^R$ and one peak above. The significant mass of Democratic-authored opinions which are more dispersed than the typical Republican-authored opinion corroborates the results of Section 4.2, while the second peak suggests a cluster of Democratic-authored opinions using similar language. We leave characterization of this cluster for future work.

Since the histograms for $\bar{s}_u^B$ and $\bar{s}_v^P$ are less qualitatively different, we compare their cdfs in the right panel of Figure 1. The difference in distributions is small in magnitude, but the cdf of $\bar{s}_u^B$ lies above that of $\bar{s}_v^P$ for over 90% of the probability mass of opinions: at the vast majority of quantiles, average within-class similarity is smaller for opinions from bipartisan panels. This result again supports the reduced-form analysis—bipartisan panels do not entrench discussion along simplistic party lines—using the full information in our learned embedding representation.

5 Conclusion

We used NLP to take a setting-specific approach to text polarization: assessing causal differences in legal topics discussed by U.S. appellate courts, one of the most impactful decision-making bodies in modern society. Our fine-tuned transformers learn the distribution of legal topics over a large range of different legal cases (criminal, individual rights, regulatory, etc.) and time periods (1890-2013). Our model shows no major cross-party differences in topic use but suggests slightly more dispersion among Democratic-authored opinions. We confirm this dispersion using embedding similarity, and additionally show that mandatory bipartisanship does not reduce the breadth of legal discussion. This research agenda is ongoing; as a key next step we plan to analyze the complexity and emotional valence of opinion text. Further unpacking the differences not only between partisan judges, but between bipartisan and party-unanimous panels, will help ensure informed and comprehensive oversight of this critical branch of government.

[9] See figures 6 and 7 in the Appendix for visualizations.

Limitations

While our topic labels cover the entire dataset, they are coarse and may not capture finer differences that speak more clearly to partisan interpretations of each area of law. An alternative approach is to obtain finer-grained labels from legal experts on a small subset of opinions and use few-shot learning approaches with large language models (LLMs) to learn a finer topic distribution for the entire data set, then merge the hand-labeled and LLM-labeled samples using the approach in Egami et al. (2023). Doing so would provide a middle ground between the low-dimensional, reduced-form approach of sections 4.2-4.3 and the high-dimensional embedding comparisons in Section 4.4. We can also use this LLM approach with our existing broad labels, though the strong performance of the transformer model and existing comparisons of LLMs with domain-specific transformers for legal tasks in Chalkidis (2023) suggest that any gains would be marginal.

Topics are only one dimension of judge partisanship, and an important question is whether the *quality* of discourse differs across the categories considered here even if the *subject* of discourse does not. In ongoing work, we are deploying tools from sentiment analysis to test whether opinions become more emotionally charged on all-partisan panels and using measures of linguistic complexity to see whether partisanship leads to simpler (and perhaps less informative) text. These rich NLP approaches, as well as others like argument mining, can be used to verify whether the encouraging conclusions of this work about common ground among partisan judges remain true when looking at stylistic features rather than content-based ones.

Finally, there are some setting-specific considerations that can guide further research. This work focuses on lead opinions, which communicate the binding ruling of the panel. However, an important margin of expression are dissenting and concurring opinions, which are non-binding but express the preferences of an individual judge on the panel. Dissenting behavior can be subtle and is the subject of ongoing study in the judicial politics literature (e.g., Chen et al. 2023). Examining dissenting opinion language using the full range of NLP tools discussed here could shed light on an important and high-profile margin for judicial action. Also of interest are unpublished opinions, which make up a substantial share of appellate court cases (Cohen, 2024). There may be an opportunity to use generative tools to simulate the distribution of unobserved opinions given extensive judge- and case-level metadata.

Ethical Considerations

All data on appellate judges is publicly accessible, primarily via the Federal Judicial Center. Appellate court decisions and opinion text are also publicly available, e.g., via CourtListener. There is no personally identifying used in this project that is not available to the general public. The contributions of the researchers cited in Section 2 are in collating and merging these data sources rather than adding previously unavailable information. Even so, to respect the privacy of judges and litigants, we work with judge identification numbers rather than names and additionally remove the names of judges and litigating parties from the opinion text used in this work.

Our analysis and conclusions are intended to contribute to a clearer understanding of appellate court function. To this end, code for the analysis in this work is available upon request. Data subsamples are available upon request and full data will be made public as soon as possible around the time of publication.

This work represents our own conclusions and does not reflect the opinions of any appellate courts or parties involved in appellate litigation.

References

Elliott Ash, Daniel L. Chen, and Ariana Ornaghi. 2024. Gender attitudes in the judiciary: Evidence from u.s. circuit courts. *American Economic Journal: Applied Economics*, vol. 16 (1):pp. 314–350.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *Preprint*, arXiv:2304.12202.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages pp. 2898–2904. Virtual conference.

Daniel L. Chen, Moti Michaeli, and Daniel Spiro. 2023. Non-confrontational extremists. *European Economic Review*, 157:104521.

Daniel L. Chen and Susan Yeh. 2020. Growth under the shadow of expropriation? the economic impacts of eminent domain. *Working paper*, this version Mar. 2020.

Alma Cohen. 2024. The pervasive influence of ideology at the federal circuit courts. *Working paper*, this version Feb. 2024.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601.

Sean Farhang, Jonathan P. Kastellec, and Gregory J. Wawro. 2015. The politics of opinion assignment and authorship on the us court of appeals: Evidence from sexual harassment cases. *The Journal of Legal Studies*, 44(S1):S59–S85.

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.

Zihao He, Negar Mokhberian, António Câmara, Andres Abeliuk, and Kristina Lerman. 2021. Detecting polarized topics using partisanship-aware contextualized topic embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2102–2118, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2023. Retweet-bert: Political leaning detection using language features and information diffusion on social networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):459–469.

Marin K. Levy and Adam S. Chilton. 2015. Challenging the randomness of panel assignment in the federal courts of appeals. *Cornell Law Review*, vol. 101 (1):pp. 1–56.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wei Lu and Daniel L. Chen. 2024. Motivated reasoning in the field: Polarization in precedent, prose, vote, and retirement in u.s. circuit courts, 1800-2013. *Working paper*, this version Feb. 2024.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Cass R. Sunstein, David Schkade, Lisa M. Ellman, and Andres Sawicki. 2006. *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Brookings Institution Press.

Appendix: Supplementary Figures and Tables

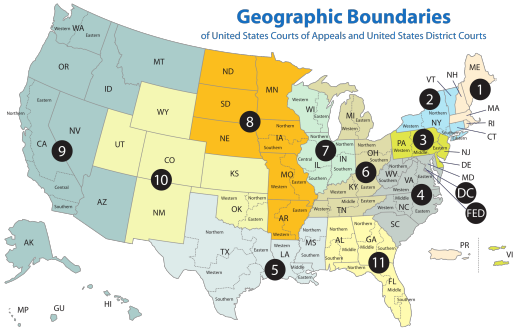

Figure 2: The thirteen circuit courts of appeals and the district courts they cover. Source: US Federal Government.

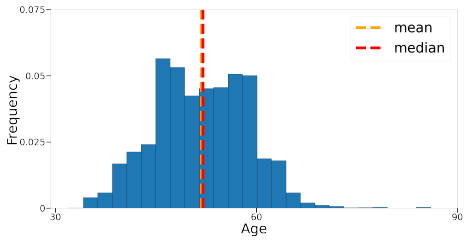

Figure 3: Histogram of judge ages; the distribution skews right, though there is a substantial mass of judges below the mean at 40-50 years old.

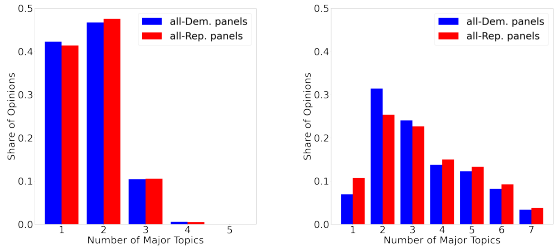

Figure 4: Comparing distributions of topics with greater-than-uniform probability (left panel) and posterior > prior (right panel) across all-Democratic and all-Republican panels.

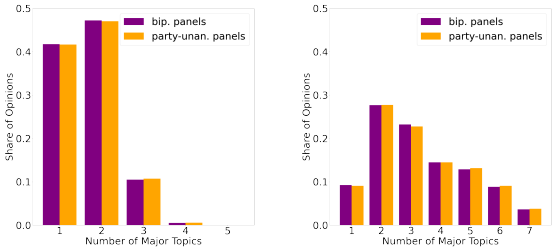

Figure 5: Comparing distributions of topics with greater-than-uniform probability (left panel) and posterior > prior (right panel) across bipartisan and party-unanimous panels.

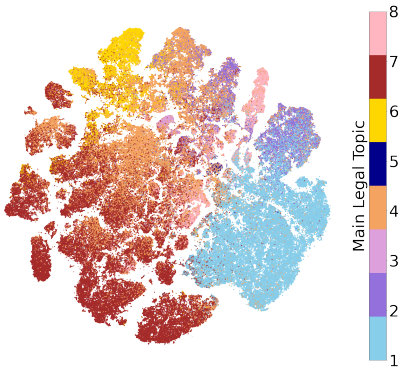

Figure 6: Opinion embeddings visualized using openTSNE in Python. Following package guidelines we use PCA to reduce the usual 768-dimensional embeddings to 50 dimensions, then apply t-SNE to represent those compressed vectors in two dimensions. We set t-SNE perplexity to 50. Colors represent main legal topic following the true Bloomberg labels:
1. criminal,
2. civil rights,
3. First Amendment,
4. due process,
5. privacy,
6. labor relations,
7. economic activity/regulation,
8. miscellaneous.

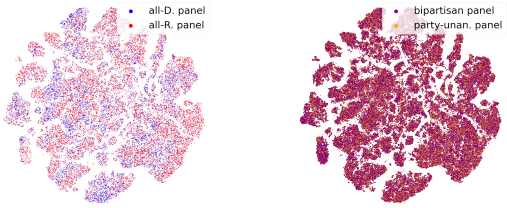

Figure 7: Opinion embeddings colored by party (left) and panel type (right), visualized using openTSNE in Python.

Type	Has Text	Has Author	$>$ 500w.	$>$ 1000w.	$>$ 5000w.
Lead/Majority	363,038	272,979	263,520	230,715	25,558
Dissenting	31,086	30,708	22,386	15,685	1,677
Concurring	15,641	15,641	8,119	5,011	491

Table 2: For opinions where the lead/majority has a known author and contains some text, 9.9% have a dissent with a known author; 99.4% of those dissents have some text, and 73.4% have at least one single-spaced page (500 words) of text. Again restricting to opinions where the lead/majority has a known author and contains some text, 6.8% have a concurrence with a known author; 74.9% of those concurrences have some text, and 38.7% have at least one single-spaced page of text.

Category	Total	Wrote 1+ Opinions	$>$ 5 Op.	$>$ 10 Op.	$>$ 100 Op.
All	2,684	2,467	1,636	1,283	619
Male	2,398	2,239	1,533	1,200	570
Female	286	228	103	83	49

Table 3: The sample contains many judges, most of whom have written nontrivial numbers of opinions. Female judges are much less common (10.7% of our sample) and slightly less prolific.

	Predicted Label								
True Label	Crim.	Civ. Rights	1st Am.	Due Process	Priv.	Labor	Econ./Reg.	Misc.	Sum
Crim.	64,156	3,968	16	1,471	0	15	1,101	438	71,165
Civ. Rights	2,902	9,255	24	5,444	0	133	586	649	18,993
1st Am.	103	22	1,226	577	0	16	333	39	2,316
Due Process	4,196	3,615	321	41,535	0	2,805	22,828	2,126	77,426
Priv.	13	4	0	349	0	0	23	3	392
Labor	48	36	1	3,473	0	13,846	2,181	47	19,632
Econ./Reg.	1,418	437	67	8,869	0	1,289	87,041	413	99,534
Misc.	1,195	889	55	3,922	1	199	4,622	8,967	19,850
Sum	74,031	18,226	1,710	65,640	1	18,303	118,715	12,286	308,912

Table 4: Confusion matrix for predicting legal topic using the best-performing model (fine-tuned LegalBERT). Overall, 226,026 out of 308,912 opinions (73.2%) were classified correctly.

True Label	Accuracy	Precision	Recall	F1 Score
Crim.	0.945	0.867	0.902	0.884
Civ. Rights	0.940	0.508	0.487	0.497
1st Am.	0.995	0.717	0.529	0.609
Due Process	0.806	0.633	0.536	0.581
Priv.	0.999	0.000	0.000	0.000
Labor	0.967	0.756	0.705	0.730
Econ./Reg.	0.857	0.733	0.874	0.798
Misc.	0.953	0.707	0.452	0.551
Weighted avg.	0.884	0.724	0.731	0.722

Table 5: Performance metrics by topic for the best-performing model (fine-tuned LegalBERT). Of the three main topics—criminal, due process, and economics/regulation—the first and last show strong performance, and the weighted averages reflects an overall successful classifier.

Metric	All-D. vs. All-R.	Split vs. Unan.	Base Rate (All-R)
# topics w/ $\mathbb{P} > 0.25$	-0.003	0.002	1.344
	(0.005)	(0.002)	
# topics w/ $\mathbb{P} > 0.125$	-0.002	0.003	1.702
	(0.006)	(0.002)	
# topics w/ $\mathbb{P} > 0.1$	0.001	0.004*	1.825
	(0.006)	(0.002)	
# topics w/ $\mathbb{P} > 0.05$	0.011*	0.001	2.194
	(0.006)	(0.002)	
# topics w/ $\mathbb{P} > 0.01$	0.013***	0.001	3.079
	(0.005)	(0.002)	
Observations	75,214	246,554	46,005

Table 6: Robustness to alternative uniform thresholds for major topics. A threshold of 0.125 (reported in the main text, and again here) represents a uniform distribution over all 8 topics. The number of major topics on all-Democratic panels compared to all-Republican panels grows slightly as the threshold decreases, but remains small; the number of major topics is the same on bipartisan and party-unanimous panels.
The estimated model is binomial/Poisson, so a coefficient of 0.1 represents ≈10% increase over the base rate.
* = significant at 10% level, ** = significant at 5% level, *** = significant at 1% level.

Metric	All-D. vs. All-R.	Split vs. Unan.	Base Rate (All-R)
# topics w/ posterior $>$ prior	0.014**	0.003	1.605
	(0.007)	(0.002)	
# topics w/ posterior $> 1.25\times$prior	0.016**	0.001	1.445
	(0.007)	(0.002)	
# topics w/ posterior $> 1.5\times$prior	0.016**	0.002	1.293
	(0.007)	(0.002)	
# topics w/ posterior $> 2\times$prior	0.019**	0.000	1.027
	(0.008)	(0.001)	
# topics w/ posterior $> 3\times$prior	0.040***	-0.007	0.585
	(0.011)	(0.005)	
Observations	75,214	246,554	46,005

Table 7: Robustness to alternative Bayesian thresholds (posterior $> \alpha\times$ prior) for major topics. A value of $\alpha = 1$ (reported in the main text, and again here) represents those topics where the posterior distribution places more weight than the prior. Across thresholds, all-Democratic panels have slightly more major topics than all-Republican panels, and bipartisan panels have the same number of topics as party-unanimous panels.
The estimated model is binomial/Poisson, so a coefficient of 0.1 represents ≈10% increase over the base rate.
* = significant at 10% level, ** = significant at 5% level, *** = significant at 1% level.