The Predictable Court: Lessons from Algorithmic Forecasting

Edward H. Stiglitz*
April 27, 2026

ABSTRACT

To date, the best-performing legal prediction technology remains the oldest: the human mind. Human crowds consistently outperform algorithmic approaches to forecasting judicial outcomes. This Article challenges that arrangement. Using a fine-tuned transformer architecture trained on fifteen years of unstructured Supreme Court data, the Article introduces the "Experience Model," which learns the personalities, preferences, and jurisprudential commitments of the sitting justices. Tested against the most recent term (OT2024), the model achieves 74 percent vote-level accuracy and 82 percent case-level accuracy—outperforming existing benchmarks and, at the case level, human crowds in the test term.

Beyond mere forecasting, the model's calibrated predictions point to a novel methodology for addressing longstanding questions of legal theory and institutions, illustrated in four applications. First, legal theorists from Hart to Raz prize law's guidance function—or capacity to provide citizens a discernible pathway for their conduct—a concept tied to predictability. By this measure, law's guidance function operates more fully in some domains than others: for instance, criminal procedure cases are difficult to predict; judicial power cases are highly predictable. Second, the model provides a novel methodology for testing the Priest-Klein hypothesis that litigated cases tend to be uncertain propositions—using more theory-aligned but previously unobservable calibrated ex-ante win probabilities rather than observed win rates. Results largely support the hypothesis, as adjusted for docket selection through the certiorari process. Third, in the first justice-level study of the marginal predictive contribution of oral argument, an ablation study demonstrates that oral argument substantially boosts predictive accuracy, but does so heterogeneously, with large gains for some justices and negligible effects for others. Fourth, cases decided in June, often considered the term's most complex, turn out to be its most predictable. This suggests that, though controversial, end-of-term cases divide along ideological lines and other contours the model easily tracks rather than presenting truly uncertain legal or coalitional questions.

The Article concludes by examining the normative implications of accurate prediction. Some predictive capacity enhances law's guidance function, but over-reliance risks reducing innovation in legal doctrine and displacing authoritative judicial reasoning. Legal institutions, however, are not passive subjects of technological disruption. Courts control the data on which prediction models depend—including panel disclosure timing, authorship norms, and decision structure—and can titrate that information flow to foster a healthy legal system.

---

* Cornell University. [Acknowledgements.]

TABLE OF CONTENTS

I. Introduction

Observers have long been interested in predicting the outcomes of judicial decisions. A century ago, Holmes envisioned a "bad man" who was governed by predictions about legal sanctions, gleaned from the "sibylline leaves" of law books.[1] Efforts to use computational methods to study judicial votes stretch back to the 1940s and 1950s, with notable strides in the last twenty years.[2] However, to date the best performing legal prediction machine is the oldest available technology: the human mind.[3] Human crowds consistently outperform algorithmic approaches.[4]

This Article challenges that arrangement. It introduces a fine-tuned transformer-based prediction engine on Supreme Court data, tests its capacity against unseen data, and extracts empirical lessons for legal institutions based on its predictions. Using fifteen years of data from Supreme Court decisions, filings, and oral arguments, the model learns the personalities, preferences, and jurisprudential commitments of the sitting justices. It learns, for instance, that Justice Jackson may be sympathetic to claims involving criminal defendants, or that Justice Gorsuch may be sympathetic to claims involving states' rights. Or that, even if Justice Thomas says little during oral argument, he tends to vote to reverse when Justice Sotomayor is critical of the respondent during argument. Modern models can exploit these deep inter-dependencies between justices and case characteristics to predict vote and case outcomes.[5]

Using the last term completed as a test period, the model outperforms the existing state of the art in algorithmic legal prediction.[6] Perhaps more impressive, it also outperforms *human* predictions at the case level in the test term. The state of the art in algorithms produces vote-level and case-level accuracy in the low 70s.[7] Human crowds tend to do better: last term, the leading aggregator of human predictions reported vote-level accuracy of 80 percent and a case-level accuracy of 76 percent.[8] By comparison, for October Term 2024, the model introduced in this Article has a vote-level accuracy of 74 percent and a case-level accuracy of 82 percent, thus surpassing human performance at the case level.[9]

---

[1] Oliver Wendell Holmes, *The Path of the Law*, 110 HARV. L. REV. 991, 992 (1897) (imagining a person "wishing to avoid an encounter with public force").
[2] *See infra* Part II.A.
[3] *Id*.
[4] *Id*.
[5] John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014) (identifying prediction as one of five areas in which machines will disrupt the practice of law).
[6] *See infra* Part II.
[7] Katz et al., *infra* note 51, https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0174698 (reporting 70 percent case level accuracy and 72 percent vote level accuracy).
[8] Josh Blackman, *Congratulations to Kirill Muzyka, Chief Justice of FantasySCOTUS for OT2024*, REASON (July 16, 2025), https://reason.com/volokh/2025/07/16/congratulations-to-kirill-muzyka-chief-justice-of-fantasyscotus-for-ot-2024/
[9] As noted later, the sample size at the case level urges a degree of caution, yet performance for other models in the test term is consistent with their published performance levels in earlier terms, suggesting the present model represents a substantial step in predictive capacity.

A further virtue of the model is that it is well-calibrated, implying accurate confidence estimates, a capacity humans notoriously lack.[10] This means that if the model says it is seventy-five percent sure of outcome X rather than Y, then outcome X rather than Y will in fact occur about seventy-five percent of the time. It knows when it is confident of an outcome and when it is more uncertain. Calibration is important not only as an aspect of model performance; it also allows the model to be used as an instrument to study long-standing questions about legal institutions.

As shown in four applications, the model's calibrated predictions point to a novel methodology for addressing longstanding questions of legal theory and institutions. First, legal theorists broadly agree that for law to guide citizen conduct—a standard rule of law desideratum—it must be knowable or, in other words, predictable.[11] Yet the extent to which the legal system, or specific areas of law, satisfy that criterion is largely unstudied empirically. The model provides a novel methodology for studying this topic: legal areas that prove to be unpredictable ex ante may be areas governed by arbitrary discretion, rather than the regularity of law. Using this approach, the model finds that cases focused on criminal procedure are difficult to predict, suggesting an area of law governed by discretion; by comparison, cases involving questions of due process or judicial power tend to be relatively easy to predict, suggesting areas governed by empirically detectable regularities.[12] Normatively, areas that prove unpredictable may be candidates for refinement through legislative or judicial action.

The model, second, provides a novel, conceptually grounded methodology for testing perhaps the most discussed hypothesis from the theory of strategic litigation—Priest and Klein's hypothesis that litigated cases tend to be uncertain, fifty-fifty propositions.[13] For cases with lopsided odds of victory, they theorize, parties tend to be best off settling, splitting in some way the foregone costs of litigation. Numerous studies examine that hypothesis using observed win

---

[10] Justin Kruger & David Dunning, *Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*, 77 J. PERSONALITY & SOC. PSYCH. 1121 (1999) (showing the now-labeled Kruger-Dunning effect, whereby people over-estimate their knowledge or ability).

[11] H.L.A. HART, THE CONCEPT OF LAW 89 (1961) (referring to the "internal point of view" as central to the rule of law); Lon L. Fuller, *A Reply to Professors Cohen and Dworkin*, 10 VILL. L. REV. 655, 657 (1965) (regarding law "as a means for achieving a certain kind of order … a particular kind of control or power, that obtained through subjecting people's conduct to the guidance of general rules by which they may themselves orient their behavior"); JOSEPH RAZ, THE MORALITY OF FREEDOM 54 (1986); SCOTT J. SHAPIRO, LEGALITY 170-73 (2011) (arguing that "legal institutions are supposed to enable communities to overcome the complexity, contentiousness, and arbitrariness of communal life by resolving those social problems that cannot be solved, or solved as well, by non-legal means alone"); PAUL GOWDER, THE RULE OF LAW IN THE REAL WORLD (2016) (arguing that "the rule of law makes it possible for citizens to become committed to the legal order in which they live, and demands that commitment in order to permit the law to be used as a tool to coordinate their behavior in order to preserve their equal standing").

[12] As detailed in Part IV.A, several competing explanations must be considered, including that the litigation incentives differ across issue areas.

[13] George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL. STUD. 1, 5 (1984) (claiming that the individual maximizing decisions of the parties will create a strong bias toward a rate of success for plaintiffs at trial or appellants at appeal of 50 percent regardless of the substantive standard of law). This fifty percent hypothesis depended on certain conditions discussed later, such as symmetry in the stakes of litigation and predictive capacity. *Id*.

rates.[14] The main challenge with observed win rates, however, is that they do not necessarily inform us of ex-ante win probabilities, the more theory-relevant quantity. The observed win rate might be fifty percent, for example, but ex-ante half the cases may have been sure losers and the other half sure winners. The model allows analysis using calibrated ex ante win probabilities. Using this approach, forecasts of the Supreme Court's test-term largely support the Priest-Klein hypothesis—the mass of predicted reversal probabilities concentrates at a little above fifty percent, with few ex-ante sure-losers.[15]

A third lesson concerns oral argument. The conventional view among Supreme Court scholars is that oral argument is "mere window dressing" and matters little to outcomes or predicting votes.[16] That said, the justices themselves maintain that oral argument is vital,[17] and revisionist accounts by social scientists likewise point to its value.[18] The model provides a new method for studying the predictive capacity of oral argument: we can train sister models, one with argument data and one without, and compare their performance. Through this ablation study, we can produce the first estimates of the marginal contribution of oral argument to predictive capacity at the justice level. The basic result from this exercise aligns with Justice Roberts' statement that "oral argument is terribly, terribly important."[19] Argument data notably boosts predictive performance. However, it does so more for some justices than for others—Justices Jackson and Sotomayor a great deal, Justices Alito, Thomas, and Barrett hardly at all.

A final lesson relates to the timing of Supreme Court decisions. Observers have long noted that the biggest, most controversial, and complex cases tend to come down at the very end of the term, during the so-called June Crunch.[20] The present analysis adds a dimension to this

---

[14] Joel Waldfogel, *The Selection Hypothesis and the Relationship Between Trial and Plaintiff Victory*, 103 J. Pol. Econ. 229 (1995) (drawing inferences from ex post win rates); Theodore Eisenberg, *Testing the Selection Effect: A New Theoretical Framework with Empirical Tests*, 19 J. Legal Stud. 337 (1990) (same); Peter Siegelman & John J. Donohue III, *The Selection of Employment Discrimination Disputes for Litigation: Using Business Cycle Effects to Test the Priest-Klein Hypothesis*, 24 J. Legal Stud. 427 (1995) (same); Kevin M. Clermont & Theodore Eisenberg, *Do Case Outcomes Really Reveal Anything About the Legal System? Win Rates and Removal Jurisdiction*, 83 Cornell L. Rev. 581 (1998) (same); Theodore Eisenberg & Henry S. Farber, *The Litigious Plaintiff Hypothesis: Case Selection and Resolution*, 28 RAND J. Econ. S92 (1997) (same); Yun-chien Chang & William H.J. Hubbard, *New Empirical Tests for Classic Litigation Selection Models: Evidence from a Low Settlement Environment*, 23 Am. L. & Econ. Rev. 348 (2021) (same). Daniel Klerman & Yoon-Ho Alex Lee, *The Selection of Disputes at Forty*, 42 YALE J. ON REG. 1030 (2025) (reviewing the literature, and in part pointing out the challenge of relying on observed win rates for drawing inferences about the fifty-fifty thesis).

[15] Part IV.B.

[16] TIMOTHY R. JOHNSON, ORAL ARGUMENTS AND DECISION MAKING ON THE UNITED STATES SUPREME COURT 2-3 (2004) (noting the "dominant view" that argument has "little influence over case outcomes", referring to common view that argument is "window dressing"); DAVID W. ROHDE & HAROLD J. SPEATH, SUPREME COURT DECISION MAKING (1976); JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED 280 (2012) (observing that, though justices say argument matters, this "does not mean that oral argument regularly, or even infrequently, determines who wins and who loses," and noting that argument comes up rarely in the justices' papers).

[17] John G. Roberts, Jr., *Oral Advocacy and the Re-emergence of a Supreme Court Bar*, 30 J. SUP. CT. HIST. 68 (2005).

[18] JOHNSON, *supra* note 16 (studying Burger Court era papers and arguments to show, for example, connections between topics of oral argument and topics addressed in decisions).

[19] Roberts, *supra* note 17, at 69.

[20] Lee Epstein, William M. Landes & Richard A. Posner, *The Best for Last: The Timing of U.S. Supreme Court Decisions*, 64 DUKE L. J. 991 (2015) (noting the common thesis and testing it, finding support for the claim that end

strand of the literature—are the end of term cases the least predictable? If they represent the most complex, difficult-to-negotiate coalition-challenging cases, they might be the least predictable. It turns out that the opposite is true. By a fair margin, they represent the *most* predictable cases decided during the term. The model can predict fully 90 percent of June cases correctly. In this sense, though they may be many of the term's most important cases, the June cases do not present particularly "difficult" or "complex" legal or coalitional questions that defy forecasting. These cases instead tend to follow predictable lines of empirical regularity, ideological and otherwise, that the model easily tracks.

The Article then turns to the normative tradeoffs of accurate prediction, likely to become more acute as technology improves and accurate prediction penetrates the lower courts. As noted, legal theorists as diverse as Hart, Fuller, and Raz see law's guidance function as critical to rule of law values.[21] Law must guide conduct, providing a pathway for citizens (and officials) to carry out their lives. For law to fulfill this function, it must be knowable and forecastable—in other words, it must be predictable.[22] This predictability speaks also to pragmatic normative concerns: encouraging ex-ante deterrence from unlawful behavior and promoting efficient settlement of legal disputes.[23] Yet predictive capacity is not costless. For law to innovate and adapt to changing social, moral, and economic conditions, the legal system must have opportunities to revisit earlier decisions.[24] And the very cases that the system needs to adapt represent those most likely to be un-litigated due to settlement—i.e., those cases that, under the current law, would be very likely to lose, but that possess the potential to reshape the law to emerging social, technological, or economic conditions.[25] Over-reliance on predictions,

---

of term cases may be decided then due to their "complexities and internal divisions that the Court encountered in resolving them"); see also Thomas E. Baker, *The Intelligible Constitution*, 10 CONST. COMMENT. 167, 173 (1993) (referring to the "June Crunch").

[21] H.L.A. HART, THE CONCEPT OF LAW 89 (2d ed. 1994) (identifying rules that permit members of society to "use[] them as guides to conduct"); LON L. FULLER, THE MORALITY OF LAW 110 (rev. ed. 1969) (starting with the "obvious truth that the citizen cannot orient his conduct by law if what is called law confronts him merely as a series of sporadic and patternless exercises of state power"); JOSEPH RAZ, THE AUTHORITY OF LAW 213 (2009) (identifying guidance as one of two aspects of rule of law, i.e., "that the law should be such that people will be able to be guided by it").

[22] FULLER, *supra* note 21, at 35 (using the allegory of a failed law-maker, Rex, whose citizens complained that "when they said they needed to know the rules, they meant they needed to know them in *advance* so they could act on them").

[23] E.g., Louis Kaplow, *Rules versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992) (connecting predictability with deterrence); Priest & Klein, *supra* note 13 (connecting predictability with settlement); Benjamin Minhao Chen & Anthony Niblett, Bargaining in the Shadow of the Algorithm (Typescript, Aug. 14 2025) (empirically connecting prediction with beliefs leading to settlement); Robert H. Mnookin & Lewis Kornhauser, *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L.J. 950 (1979).

[24] E.g., CALABRESI, *infra* note 160 (arguing that common law should remedy obsolescent statutes). Similarly, theories of common law "efficiency" rest on different mechanisms, but they all require a flow of cases that allow courts to reconsider inefficient arrangements. Richard A. Posner, *A Theory of Negligence*, 1 J. L. STUD. 29 (1972); Paul H., Rubin, *Why is the Common Law Efficient?*, 6 J. L. STUD. 51 (1977); George L. Priest, *The Common Law Process and the Selection of Efficient Rules*, 6 J. L. STUD. 65 (1977); Nicola Gennaioli & Andrei Shleifer, *The Evolution of Common Law*, 115 J. POL. ECON. 43 (2007). For a take on how these claims over inefficient arrangements might develop, *see* Janice Nadler, *Flouting the Law*, 83 TEX. L. REV. 1399 (2005) (examining how a "commonsense view of justice" can lead to non-compliance with contrary laws).

[25] E.g., Fiss, *infra* note 162, at 1085 (arguing that the job of the legal system is to "explicate and give force to the values embodied in authoritative texts … to interpret those values and to bring reality into accord with them … [a] settlement will [] deprive a court of the occasion, and perhaps even the ability, to render an interpretation.")

moreover, risks displacing authoritative judicial reasoning with statistical inference, degrading the legal system over time.[26] A healthy legal system needs not just dispositions, forecasted or realized, but also explanations of those decisions that can in turn condition the beliefs and conduct of lower courts and non-parties.[27] The optimal level of predictive capacity is in the interior.[28]

The Article finally examines judicial levers and devices available to regulate predictive technology by manipulating the flow of data available for model development. For example, the D.C. Circuit made panel composition information available earlier in the litigation cycle, seeking to amplify predictive capacity.[29] That same move could be reversed to reduce predictive capacity. Courts could also change practices of authorship, moving anywhere between seriatim opinions (maximum information), which was the historical practice of the early Supreme Court,[30] to a norm of unsigned or per curiam opinions (minimal information);[31] similarly, courts could suppress dissents, as they do in other jurisdictions, such as the European Court of Justice.[32] Legal institutions control the data needed to create predictive technology and they should work within judicial norms to titrate information to foster a healthy legal system.[33]

II. Predicting Judicial Outcomes: Historically and with Modern Architectures

A. From Holmes to Transformers

Judicial predictability has been core to the literature for at least a century. Attempting to differentiate his position from legal formalists, Holmes provocatively conceived of the law as simply consisting of prediction. Law was what a "bad man" who found no moral suasion in legality thought a court would do on a set of facts.[34] "The prophecies of what the courts will do

---

[26] *See* Part V.A.

[27] EDWARD H. STIGLITZ, THE REASONING STATE (2022) (arguing that the distinctive feature of legal and administrative systems is their ability to generate credible public reasons); Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633 (1995); FULLER, *supra* note 21, at 35 (using Rex to illustrate a failed legal system that renders dispositions without reasons).

[28] For a similar conclusion, see Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics*, 68 U. TORONTO L.J. 12, 20 (2018) (arguing that "the uncertainty that grounds legal certainty is not a drawback but, rather, an asset," observing that "justice" can tradeoff with certainty).

[29] *See* Jordan, *infra* note 150.

[30] *E.g.,* G. EDWARD WHITE, THE MARSHALL COURT AND CULTURAL CHANGE, 1815-1835 (1988).

[31] Ronen Avraham, et al., *Lifting the American Supreme Court Veil: Identifying Authorship in Unsigned Opinions*, 17 J. L. ANAL. 2 (2025) (using AI to find "likely" authors of unsigned opinions—however, this is only possible because most opinions are signed).

[32] *E.g.,* Vlad Perju, *Reason and Authority in the European Court of Justice*, 49 VA. J. INT'L L. 307, 324 (2009) (noting a common tradition of unitary decisions in many European jurisdictions).

[33] This point can be seen as an instance of an important theme in the social science literature—that full transparency often leads to degenerate outcomes. *E.g.,* Emir Kamenica & Matthew Gentzkow, *Bayesian Persuasion*, 101 AMER. ECON. REV. 2590 (2011); David Stasavage, *Open-door or closed-door? Transparency in domestic and international bargaining*, 58 INT'L ORG. 58 667 (2004); Aviv Caspi & Edward H. Stiglitz, *Observability and Reasoned Discourse: Evidence from the U.S. Senate*, Typescript (2023); Bauke Visser & Otto H. Swank, *On Committees of Experts*, 122 Q. J. ECON. 337 (2007).

[34] Holmes, *supra* note 1, at 992.

in fact," he wrote, "and nothing more pretentious, are what I mean by the law."[35] Prediction, for him, consisted of traditional legal research: the study of "a body of reports, of treatises, and of statutes, in this country and in England, extending back for six hundred years, and now increasing annually by hundreds."[36]

Though influential among legal realists, Hart and other theorists criticized that view almost since it was put to paper.[37] One simple point is that in Holmes' conception the meaning of the law is calibrated to our ability to predict what courts will do. If the law is merely what we predict courts will do, noisy predictions imply that the law has less meaning. Absent some ability to predict what courts will do in fact, the law is meaningless in fact. Predictive capacity is a necessary condition—a foundation—for his conception of the law and that of many other realists.

Social scientists represent the intellectual descendants of Holmes and the realists.[38] Less interested in developing a theory of law, they adopt an observational, externalist viewpoint of the law—seeking to understand how judges and courts behave and what drives their behavior. One prominent social science understanding of law, the attitudinalist school, posits that judicial behavior, and therefore law, can best be understood by studying judges' policy preferences and ideology.[39] Others react to that stark view of the law, seeing it as overly reductionist, but it remains relevant both as a pole in the debate and as a pillar even in more high-dimensional social scientific understandings of the law.[40]

From early in the history of quantitative social science, scholars have used computational methods to predict Supreme Court votes. One early social science study noted a trend in the scientific study of judicial process reaching back to the 1940s,[41] and deployed then-available computational methods to observe that regularities existed in justices' voting patterns over issues and party types.[42] Scholars in cognate computational and mathematical fields, too, early recognized the challenge and opportunity of judicial prediction, propelling strands of literature in those fields that were sometimes in conversation with the social scientists and at other times operated on parallel tracks.[43]

---

[35] *Id.* at 994.
[36] *Id.* at 991.
[37] HART, *supra* note 21, at 89 (identifying the "internal point of view" as applying to "a member of the social group which accepts and uses [the social group's rules of conduct] as guides to conduct").
[38] For the roots of this connection, *see* JOHN HENRY SCHLEGEL, AMERICAN LEGAL REALISM AND EMPIRICAL SOCIAL SCIENCE (1995).
[39] C. HERMAN PRITCHETT, THE ROOSEVELT COURT (1948); GLENDON SCHUBERT, JUDICIAL POLICY MAKING (1967); JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL (1993).
[40] FRANK B. CROSS, DECISION MAKING IN THE U.S. COURTS OF APPEALS (2007); LEE EPSTEIN & JACK KNIGHT, THE CHOICES JUSTICES MAKE (1998); Jeffrey R. Lax, *Constructing Legal Rules on Appellate Courts*, 101 AMER. POL. SCI. REV. 591 (2007); TOM S. CLARK, THE LIMITS OF JUDICIAL INDEPENDENCE (2010).
[41] S. Sidney Ulmer, *Quantitative Analysis of Judicial Processes: Some Practical and Theoretical Applications*, 28 L. & CONTEMP. PROB. 164, 164 (1963) (citing earlier work stretching to the 1940s).
[42] *Id.*
[43] R. Koewn, *Mathematical Models for Legal Prediction*, 2 COMPUTER L.J. 829 (1980). In computer science related fields, the task of predicting votes often falls under the rubric of "legal judgment prediction," which is inclusive of a broader range of predictive tasks than predicting votes. Junyun Cui et al., *A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges*, IEEE ACCESS (2023) (reviewing much of the literature); *see also*, Olga Alejandra Alcantara Francia et al., *Survey of Text Mining Techniques Applied to Judicial Decision Prediction*, 12

Ruger and colleagues developed one of the more remarkable externalist prediction projects, running a competition between an algorithm and human experts to predict the outcomes of an upcoming term.[44] For their algorithmic model, they adopted a workhorse machine learning model and included only six features: (1) circuit of origin; (2) issue area of the case; (3) type of petitioner (e.g., employer, etc); (4) type of respondent; (5) ideological orientation of the lower court decision; (6) whether the petitioner made a constitutional claim.[45] Against the predictions from that model, they arrayed the wisdom of the human crowds: the predictions of legal experts, seventy-one law professors and twelve practicing appellate lawyers.[46] Notably, they asked the experts only about the cases within their areas of expertise.[47] They surveyed the experts before oral argument, making the information available to the experts roughly on par with that available to the algorithm.[48]

The results from their experiment? Although the human and the algorithm both predicted votes at about the same level (67-68 percent),[49] the algorithm outperformed humans at the case-level. The algorithm predicted about 75 percent of cases correctly, and the expert human crowds predicted only 59 percent of cases correctly.[50]

Building on the effort of Ruger and team, about a decade later another interdisciplinary team used a more advanced workhorse model, included more features, and found arguably stronger results.[51] They correctly classified about 72 percent of votes—and about 70 percent of cases.[52] Their task was more challenging than that of Ruger's team, in that they considered nearly two centuries of decisions using a rolling random forest design, yet they also analyzed a far larger number of features than Ruger and colleagues, including information about whether and when oral argument occurred and when the decision was rendered. A third team used another more advanced workhorse model, and included richer features relevant to the content of the oral argument, and they boosted case-level performance to 74 percent, though employing a different validation strategy.[53]

APPL. SCI. 10200 (2022); Mohammed Alsayed et al., *JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction*, arXiv:2112.03414 (2021).

[44] Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn, *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-making*, 104 COLUM. L. REV. 1150 (2004).

[45] *Id.* at 1154.

[46] *Id.* at 1168.

[47] *Id.*

[48] *Id.*

[49] *Id.* at 1173

[50] *Id.* at 1171. This 59 percent figure is based on the individual predictions of the humans. Allowing each expert to vote, and aggregating over humans within cases, improves human performance to 66 percent in their sample. *Id.*

[51] Daniel Martin Katz, Michael J. Bommarito II, & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, 12 PLOS ONE 1 (2017). For a more recent effort to improve on Katz et al., *see* Ashkon Farhangi & Ajay Sohmshetty, *SCOTUS Outcome Prediction: A New Machine Learning Approach*, *in* RESEARCH HANDBOOK ON BIG DATA LAW (ed. Roland Vogl) (2021).

[52] Katz et al., *supra* note 51, at 8-9.

[53] Aaron Russell Kaufman et al., *Improving Supreme Court Forecasting Using Boosted Decision Trees*, 27 POL. ANAL. 381 (2019) (analyzing case outcomes but not votes). They use standard cross validation, wherein the sample is randomly divided into training and test sets. That validation strategy differs from Katz et al., who validate on a look-forward term, so that all test data temporally follows all training data. There is a risk of leakage across

That is about where algorithmic performance stands—when validated against temporally cordoned data, the performance tends to be in the low 70s in terms of accuracy at the vote level; the case level is similar, though approaches the mid-70s in some studies.[54]

To date, the most successful technology in fact remains that of Holmes—human study and intuition. Though Ruger et al. find vote-level accuracy of about 67 percent and case-level accuracy of about 59 percent in their sample, more recent efforts that allow access to oral-argument and near-term information consistently produce more optimistic assessments of human capability. Last term, a website that aggregates human predictions of Supreme Court votes reported vote-level accuracy of about 80 percent and case-level accuracy of about 76 percent.[55] The previous term, human performance was similarly strong.[56] The wisdom of the crowds approach easily outclasses the existing state of the art for algorithms.

The question now before us, though, is whether the new generation of transformer-based models can surpass human performance. These models, powering GPT and similar large language model products, enable context awareness and sensitivity to deep interactions among case and justice features.[57] The import of a specific term that appears in a brief, such as "jurisdictional," will depend not only on the context around that word, but perhaps also on the nature of the lower court decision, the identity of the plaintiff or defendant, or the procedural posture under which the case reaches the Supreme Court, in addition to patterns the model learns about specific justices. In this sense, the model may be able to approximate human judgments, holistically assessing the complex and high-dimensional decision-space of Supreme Court judgments.

B. A Modern Predicting Machine

1. The Problem and Task

Though the transformer-based models are more powerful than earlier models, note at the outset that this remains a difficult task. The reasons that previous efforts rarely exceeded performance of low-70s in terms of accuracy remain in play.

---

votes/cases within terms in standard cross-validation, which would inflate apparent performance. In this analysis, I likewise temporally cordon training and test data.

[54] A contemporaneous effort that uses a prompt-based strategy on various foundation models reports case-level performance for OT2023 that ranged from 76 to 82 percent, slightly under-performing relative to human crowds that term; in absolute terms, their best performing model approximates the performance of the Experience Model at the case level. It is not clear how the approaches compare in terms of F1 or calibration metrics, but prompt based strategies on foundation models represent another branch of AI-based predictors that should be studied. Amit Haim et al., LLMs as Court SuperForecasters? Benchmarking LLM Supreme Court Predictions (typescript, 2026).

[55] Josh Blackman, *Congratulations to Kirill Muzyka, Chief Justice of FantasySCOTUS for OT2024*, THE VOLOKH CONSPIRACY (Jul. 16, 2025), https://reason.com/volokh/2025/07/16/congratulations-to-kirill-muzyka-chief-justice-of-fantasyscotus-for-ot-2024/.

[56] Josh Blackman, *Congratulations to Brady Kelly, Back-to-Back Chief Justice of FantasySCOTUS for OT 2022 and 2023*, THE VOLOKH CONSPIRACY (Jul. 31, 2024), https://reason.com/volokh/2024/07/31/congratulations-to-brady-kelly-back-to-back-chief-justice-of-fantasyscotus-for-ot-2022-and-2023/.

[57] *See infra* Part II.B.

Those reasons are several.[58] First, law is exceptionally high dimensional, and it is challenging to capture the decision-relevant features without also capturing features that introduce noise and degrade model performance. Second, as the Court is at the top of the judicial hierarchy: the cases that it hears tend to be the most difficult, often pitting one valid signal against other valid signals (e.g., a conflict between precedents or legal provisions). The resolution of those clashing signals is the Court's job, and it will often be unclear a priori what the resolution will be. Third, it is difficult to account for temporal drift. Case composition, internal Court dynamics, and the larger political-economic decision-making environment evolve from year to year, meaning that the weights that optimize earlier cases may not optimize over future cases. Fourth, the Court is a collegial body, and a justice's vote may depend not only on complex interactions among case characteristics, but also on complex social interactions among the justices themselves. Justice X may be inclined to vote to affirm, for example, but then learns that Justice Y wishes to reverse and they re-evaluate their vote to maintain long-run bloc cohesion. Fifth, there is relatively little data on which to train a model. Most recent terms include between fifty and sixty cases on the merits docket, and every new case is unseen in the earlier data. Finally, it must be remembered that the Supreme Court is, in a meaningful sense, creating rather than following law when it decides cases, so understanding the law provides only a partial sense of where votes might fall. All these factors make it difficult for humans to predict judicial behavior—and the same is true of algorithms.

In addition to those difficulties, the present analysis introduces self-imposed challenges. First, the prediction model should be able to operate with unstructured data. Most previous efforts employ the Supreme Court Database,[59] a structured and highly curated dataset that requires a team months to construct for each term. In part because we want to be able to make live predictions,[60] the model should be able to start with raw, unstructured data, such as the lower court decision, briefs, and oral argument transcripts. Second, the prediction model should not draw on human commentary—the inputs should be the same as the hypothetical first human observers to study the case. This means that the model cannot use as input features the wisdom of the crowds, op-eds, or any other human-created content outside that contained in the filings or oral argument. This restriction reflects, in part, a commitment to exploring the capability of machines to understand the case, rather than to understand what humans say about the case—but also again, in part, because the model should be able to make live predictions, and not all cases will offer much, if any, commentary. Third, the final prediction should be rendered as oral argument completes—that is when the case closes for the model. This means that the model must exclude information about when the cases were decided, a feature that previous efforts include. Depending on prediction objectives, that information can be seen as a kind of leakage, where the outcome becomes part of the training data. When the Court is internally divided, for example, it may take longer to decide the case—but we want to learn that through case characteristics, rather than the timing of the decision.

---

[58] This paragraph borrows from my post at my research log, reasonablemachines.io. Edward H. Stiglitz, *Using AI to Predict Justices' Votes*, reasonablemachines.io (Sept. 11, 2025), https://reasonablemachines.io/posts/post-using-ai-to-predict-justices-votes.

[59] The Supreme Court Database, http://scdb.wustl.edu/ (Jan. 10, 2026).

[60] The model is deployed for live prediction at a research log associated with this project. reasonablemachines.io. This deployment informed several of the design constraints described in this section (for instance, the requirement that the model operate on unstructured data and exclude post-argument information).

Beyond those data restrictions, there are two important evaluation criteria absent from the literature. Previous efforts focus on measures of classification success to evaluate their models. Most typically, this means they examine the percent of correctly predicted votes or cases—that is, accuracy. However, other important classification metrics include F1 scores, which represent the (harmonic) mean of precision (percent of instances predicted as true that are true) and recall (the percent of true instances predicted as true). Ranging from zero to one, the F1 score provides a scalar representation of the prevalence of false positives and false negatives and is the dominant metric of classification success.[61]

Moreover, existing efforts do not consider calibration, which relates to the confidence of those classifications. If the model says there is a 60 percent chance that a justice votes to reverse, will that justice in fact reverse 60 percent of the time? This is an important evaluation criterion because we may be interested not just in predictions that a justice will vote to reverse, but also in predictions about which justices will be sure-reverse votes and which justices may be likely to vote for reverse but sit on the fence. Calibration can be important for properly assessing case-level outcomes, too.[62]

2. Data: Building a Supreme Court Corpus

This exercise draws on merits docket cases in terms between 2010 and 2025. The series starts with the term Justice Kagan took her seat and continues through the end of last term. In total, that amounts to about one thousand decisions, and about nine thousand votes. It would be possible to construct a larger dataset by extending deeper into history, but more recent cases carry the most information about what the Court is likely to do in the future. Including older cases risks distorting or diluting that more relevant information.

For each case, I attempt to collect a standard set of documents that would be part of the docket: the lower court decision, party and amicus briefs, and oral argument transcripts. Using that unstructured data, I experiment with various approaches to extracting and compiling signals, drawing on the existing literature where possible to identify plausible features.[63] However, I depart from the literature by constructing from the filings richer information about the legal and

---

[61] *See* Appendix; CHRISTOPHER M. BISHOP, PATTERN RECOGNITION AND MACHINE LEARNING (2016).

[62] Another overlooked evaluation criterion that I examine elsewhere is uncertainty. To analogize, calibration refers to an assessment of whether a coin is weighted. It says the coin will turn up heads x percent of the time. Uncertainty refers to how certain we are that the coin is weighted to turn up heads x percent of the time—maybe it is actually x-y percent of the time, or maybe x+y percent of the time. How sure can we be of the weight on the coin? This is also an important aspect of a model that helps to contextualize model predictions and informs us of where the model might perform well and where it might not. A prediction that a justice is likely to vote to reverse with 99 percent probability means a lot less if the model acknowledges that, though that is its best guess, there is a reasonable chance the probability to reverse could also be 1 percent. The model is just unsure—uncertain. A goal is to develop a prediction algorithm that acknowledges and estimates this form of uncertainty.

[63] A main challenge in this task is to include the optimal level of information. Simply dumping the lower court decision, briefs, and oral argument transcript into the model will include many relevant features. But it will also many irrelevant features. Just as humans can be confused by information overload, so too can algorithms. They "learn" patterns in the data that, in fact, represent peculiarities of the training data rather than transferable signals of case outcomes. And they miss features that would be durable signals of case outcomes. The raw, unstructured data, therefore, is refined as above to focus on the doctrinal and docket information that is most likely to inform judicial outcomes.

doctrinal stakes. Earlier attempts rely on relatively coarse legal categorizations available in curated datasets.[64] The present data includes features capturing legal and doctrinal stakes that are absent from existing structured datasets, such as the institutional consequences of adopting one rule over another—concerns long noted in the legal-process tradition.[65] Oral-argument transcripts were processed to identify the speaker for each utterance. For each justice, the model extracted features such as frequency of questioning, the polarity of questions toward each party, and the sequence of interruptions. Earlier work has shown the direction of questioning to be predictive of voting behavior,[66] but the present dataset captures these interactions and connects utterances with doctrinal elements of the case. This richer information provides the model with a textured set of considerations on which to draw for training and inference. After extracting signals from the docket filings, they are assembled into a structured JSON file that can in turn be fed into a model.

For most terms, voting information derives from the widely used Supreme Court Database (SCDB). For the most recent 2024 term, voting information derives from the Court's opinions themselves, as the SCDB had not yet processed the term at the time of analysis.[67] These justice-vote elements then become part of the JSON file. Individual justice votes serve as the outcomes that the model is trained to learn. As explained below, it is trained to understand the connection between the various case characteristics—noted earlier—and the outcome of interest, a justice's vote. Figure 1 illustrates the described data preparation pipeline.

[64] *E.g.,* Katz et al., *supra* note 51.
[65] Features were extracted through a pipeline of prompts to frontier LLMs, the outputs of which were then cleaned and refined through standard parsing techniques.
[66] *See* Kaufman et al., *supra* note 53.
[67] All distinct merits matters argued in OT2024 that reached decision included in the sample—this excludes, for example, the emergency docket and improvident grants.

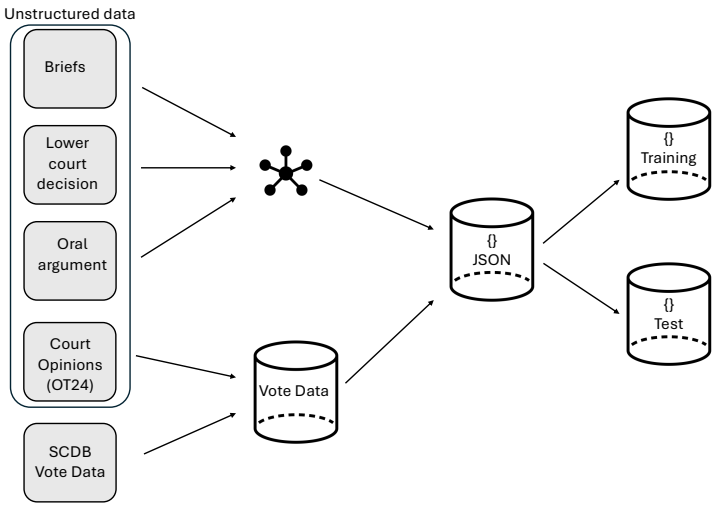

Figure 1: Data Preparation

3. Training, Validation, Calibration

The task of the model is to predict the vote of a justice conditional on case characteristics and questions asked during oral argument. Following the literature, the "vote" in a case is simplified to "reverse" or "affirm."[68] This allows the model to learn justice-specific tendencies, such as Justice Gorsuch is sympathetic to criminal defendants, or Justice Sotomayor is unlikely to vote to affirm if Justice Alito asks many questions of the respondent. As explained later, these predictions then aggregate to produce case-level predictions.

a. Experimental Design and Limitations

A key feature of the analysis is that the training and test data are completely separated in time. This departs from standard cross-validation, where cases are randomly shuffled into training and test sets. In many empirical settings cross-validation is appropriate, but in this setting it would allow the model to train on cases from the same term as the cases it is asked to predict. Cases within a term often share common shocks—such as the posture of the executive toward the Court, public mood, or collegial dynamics among the justices—that influence outcomes across multiple cases. Randomly mixing such cases into both training and test sets would therefore leak term-specific information into the training process. The model would, in effect, learn the social and institutional mood of the 2024 term from other 2024 cases, undermining the goal of evaluating how well it predicts that term before those dynamics can be known.

[68] *E.g.,* Katz et al., *supra* note 51, at 4.

Temporal cordoning avoids that risk. The model therefore is trained on cases from the 2010 term through the 2023 term. The test term, the 2024 term, is completely held out from training. The model's performance will then be evaluated against the vote and case outcomes from the 2024 term. The performance in this test term provides a reasonable sense of how successful the model would be generally if asked to predict future case outcomes based on historical cases, which is the standard setup for a Holmes-like consequentialist individual.

Though this experimental design provides a clean window into predictive capacity, there are inherent limits in evaluating model performance. The performance in the test term is not assured to be the performance in future terms. The most significant risk involves drift over time. The Court might be more ideological and partisan one term, more collegial the next. Or, as politics shifts outside the Court, the meaning of legal issues could shift, too. Administrative law, for example, might go from a relatively technocratic area of law to the core of partisan disputes;[69] or a relatively settled area, such as reproductive rights, may become unsettled. Or a new member might join the Court, leading to different social dynamics on the bench—a justice may vote differently in a new coalition, or the meaning of their questions during argument may change. All these changes mean that the patterns learned on earlier terms may carry more limited information about future terms than would be expected based on this exercise. Another inherent limit relates to the sparsity of Supreme Court cases and training data. Performance may be especially unstable on relatively rare issues, such as Indian law, where the model does not see many cases to learn textured patterns in voting behavior. Finally, though holding out the 2024 term as the test data is correct given concerns about within-term, cross-case leakage, the Court decided fewer than sixty cases that term. The small number of cases means that estimates of performance will be inherently noisy.

b. Modeling Approach

The model is a modern transformer-based classifier fine-tuned to predict individual justices' votes from the materials available prior to decision.[70] The architecture is designed to integrate the heterogeneous set of signals that experienced court-watchers routinely use, such as oral argument, while also learning the stable tendencies and interaction patterns among the justices.

The base of the system is a high parameter open-weight transformer model, similar in architecture to those powering contemporary large language models, such as GPT and Gemini.[71] The key advantage of transformers is that they represent meaning contextually. A phrase like "major questions" does not carry a fixed meaning across cases—its meaning depends, for example, on the nature of the dispute, the history in the lower court, and the justice under consideration. The model will learn to be sensitive to this context.

---

[69] *E.g.,* Gillian Metzger, *The Roberts Court and Administrative Law*, SUP. CT. REV. 1, 6 (2019) (observing of the Roberts Court, "its categorical and uncompromising stance, commitment to limited government and aggressive judicial review, this approach has the potential to radically transform American governance.").

[70] For a review, *see* Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens, *Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement*, *In* PROCEEDINGS OF THE 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 9252 (2023).

[71] The current analysis is based on Llama 3.1 70b model.

To fully understand context, the model must be trained, or fine-tuned, to align its weights with the legal domain and vote-prediction task.[72] During training, the model develops internal representations that capture latent doctrine, institutional features, rhetorical cues that correlate with justices' eventual votes.[73] That fine-tuned model "understands" the issues confronted by justices, and the justices' personalities and social dynamics, in just that sense—it learns the correlation between those case representations and justice votes. High parameter transformer models, moreover, are adept at something like the analogistic reasoning that is taught in law schools. They appreciate that, though a new case is unseen, it is analogous to an earlier case it does know in some respects, and that provides a foundation for a vote prediction.[74] The technical appendix details strategic choices around model training.[75]

Once trained, the model can be deployed to generate predictions of each justice's vote in each case in the 2024 term. The predictions generated by the model consist of the probability that the justice votes to reverse.[76] Those probabilities can then be aggregated to generate a case-level outcome probability using Monte Carlo simulations, also detailed in the appendix.[77] For labeling purposes, the trained, calibrated, and ensembled version of this model will be called the "Experience Model," in a nod to Holmes' dictum that the life of the law is found not in logic but in experience.[78]

III. Results from a Modern Predicting Machine

A. Main Performance Results

To test the model, we conduct inference on OT2024, which was held out entirely from model training. How well does it predict vote-level outcomes? Case-level outcomes? And are the estimates of the model well-calibrated?

1. Vote-Level Results

Consider first vote-level outcomes. The overall accuracy of the Experience Model in the October 2024 term is 74 percent. As reported in table 1, evaluated for that same term, other

---

[72] Thalken et al., *supra* note 70.
[73] *Id.*; Jason Wei et al., *Finetuned Language Models are Zero-Shot Learners*, *arXiv preprint arXiv:2109.01652* (2021).
[74] Vaswani, Ashish, et al., *Attention is All You Need*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (2017) (the foundational transformer paper); Tom Brown, et al., *Language Models Are Few-shot Learners*, 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1877 (2020) (suggesting that models can learn in analogy-like ways); Taylor Webb, Keith J. Holyoak, & Hongjing Lu, *Emergent Analogical Reasoning in Large Language Models*, 7 NATURE HUM. BEHAV. 1526 (2023).
[75] *See* Appendix.
[76] The reported probabilities are not Platt scaled, though I experimented with that approach. John C. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, in ADVANCES IN LARGE MARGIN CLASSIFIERS 61 (Alexander J. Smola et al. eds., 1999).
[77] *See* Appendix.
[78] OLIVER WENDELL HOLMES. JR., THE COMMON LAW 1 (1881) (writing with characteristic epigraphic flair that "[t]he life of the law has not been logic: it has been experience").

algorithmic efforts return with performance of about 68 percent.[79] Earlier work on this prediction task reports accuracy, which is informative and serves as a useful top-line benchmark. However, voting is not balanced between reverse and affirm—at least for the modern Court, reverse votes are more likely. This follows from the Court's discretionary docket. More often than not, if the Court views the lower court's decision as correct, they will not grant cert; the cases they grant cert to disproportionately represent cases that justices think may have been decided incorrectly.[80] As a result, accuracy can be misleading as a performance measure. Naively predicting the majority class (reverse) would yield predictive accuracy of 65 percent, as reported in table 1. Some existing models perform only slightly better than the naïve always-reverse model.

Table 1. Predictive Accuracy

		Period	Accuracy	
			Votes	Cases
Algorithm	Experience Model	OT2024	**74**	**82**
	Katz et al.	1816-2015	72	70
	Katz et al.*	OT2024	68	75
	Katz et al.**	OT2024	68	73
	Ruger et al. (algorithm)	OT2002	67	75
	Ruger et al. (algorithm)*	OT2024	68	72
	Kaufman et al.	2005-2015	--	74
	Naïve Reverse Benchmark	OT2024	65	73
Human	Ruger et al. (humans)	OT2002	68	59
	FantasyScotus (humans)	OT2024	**81**	**76**

Note: best performing model within the algorithmic and human categories bolded. * denotes replication and extension of analysis to new term. ** denotes replication and extension of analysis to new term, with features restricted to those known at time of argument.

The F1 score provides a more informative assessment of model performance.[81] The F1 score consists of the (harmonic) mean of precision and recall. Precision is the correctly predicted

[79] To determine the performance of other models against OT2024, we needed to replicate and extend their analyses. Katz et al. provide a replication repository, which allowed for fairly direct replication and extension. The only notable departure from their procedure is that we limited attention to the "modern" SCDB, which starts in 1946, which allowed for training on approximately 80,000 votes. Ruger et al. clearly describe their procedure in the text of their article and replication and extension was straightforward but complicated by the fact that they analyze a natural court with a longer series of terms. The natural court of interest includes only two terms before OT2024. I examine their approach with a short natural court and find that it performs relatively poorly (vote accuracy of 66 percent, and case accuracy of 70 percent). So for the results reported below in table 1, I adjust their analysis slightly and break the natural court construction, instead training on data since Kagan was appointed (i.e., the same sample as used in the Experience Model). Departing from the natural court set up, further, required simplifying their model slightly by dropping inter-justice features (they use predictions from some justices as predictors for other justices, which is messy if court composition is unstable). Outside of OT2024, the closest existing effort is Katz et al.'s random forest model, which returned with a reported vote level accuracy of 72 percent.

[80] PERRY, *supra* note 124 (noting a reverse bias in the cert process).

[81] Bishop, *supra* note 61; *see* Appendix.

instances divided by the number of predicted instances of a class, informing us of the false positive rate (higher is better); recall is the number of correctly predicted instances over the number of true instances of a class, informing us of the false negative rate (higher is better). The F1 score ranges from zero to one, with higher figures indicating a more successful model. The measure is helpful for assessing performance when classes are unbalanced because it penalizes for false positives and false negatives.

Table 2: Experience Model Performance (Votes)

	Precision	Recall	F1	Support
Affirm	0.63	0.61	0.63	173
Reverse	0.8	0.81	0.81	328
Macro (unweighted)			0.72	
Macro (weighted)			0.74	

Note: see text for definitions of performance metrics. Support refers to the number of observations in a class. The weighted macro F1 score weights class scores by the support for that class; unweighted averages the scores for each class with equal weight.

As reported in table 2, the Experience Model performs well overall. The macro F1 score weighted by class frequency is 0.74, or 0.72 when unweighted by class frequency. On this difficult task, that indicates a model with high predictive capacity. For example, Katz et al. report a lower unweighted macro F1 score of 0.62 over the affirm and reverse classes of voting.[82] The naïve reverse model provides another benchmark: it would return with an unweighted F1 score of 0.40, or a weighted F1 score of 0.52. Though more predictive than those benchmarks, the Experience Model has more trouble with the affirm class than the reverse class, as indicated in the table. Precision is respectable for the affirm class, but recall is low. This means that, when the model predicts affirm, we can be reasonably sure that the vote will be to affirm, but that there are many affirm votes that the model does not identify. With respect to the reverse class, precision and recall are both strong.

In sum, in the test term, the Experience Model outperforms existing algorithmic models at the vote level. Yet as suggested by table 1, the strongest vote-level model remains the wisdom of the human crowds: humans can predict individual justice votes with about 80 percent accuracy. That performance is not strictly comparable, as the Experience Model is forced to make its final prediction based only on docket materials and immediately following oral argument, whereas the humans can draw on commentary, non-docket information about the

[82] Katz et al., *supra* note 51, at 8. They also report and try to classify a third class of votes, "Other," which includes recusals. I do not attempt to model that class of "votes" and drop it any recusal from the analysis. To compare results with their paper, I likewise remove the Other class from their results where possible. As another benchmark on a challenge classification task, a team attempting to predict ideological direction from case text report and F1 score of 0.65 and refer to it as "highly predictive." Carina I. Hausladen, Marcel H. Schubert, and Elliott Ash, *Text Classification of Ideological Direction in Judicial Opinions*, 62 INT'L REV. L. ECON. 1 (2020).

political climate, and the timing of the decision, and can wait until just before the decision is made to make their predictions.

Part of the reason that the model performs well at the case level is that it produces relatively well-calibrated estimates of the probability that a justice will vote to affirm. Calibration, again, refers to whether a model's predicted probabilities correspond to the actual frequencies of outcomes. A model is well-calibrated if, among all cases to which it assigns a 70 percent probability of reversal, roughly 70 percent are in fact reversed. This is an important measure of model performance because it allows us to aggregate votes accurately for case level classification through Monte Carlo trials. Moreover, we will often want to know, not simply whether a justice is predicted to vote affirm or reverse, but whether they sit on the fence or represent a firm vote. To my knowledge, previous efforts to predict Supreme Court votes do not calculate or report calibration metrics.

The two most common calibration-relevant metrics are the Brier score and the expected calibration (ECE).[83] The Brier score is the squared difference between vote outcomes and the predicted probability of vote outcomes.[84] A lower score is better. For instance, if a vote outcome is assigned a value of "1", a predicted probability of one would score 0; a predicted probability of zero would score one. The Experience Model's Brier score is 0.17. As a benchmark, a naïve reverse model that predicted reverse with probability one for all votes would yield a Brier score of 0.35, essentially double the Experience Model's score.

The ECE is even more intuitive.[85] To compute it, we first divide the model's predicted probabilities into ten equal-width bins: one bin contains predictions from 0 to 0.1, the next from 0.1 to 0.2, and so on. Within each bin, we then calculate the empirical frequency of votes to reverse. A well-calibrated model will produce predicted probabilities that match these empirical frequencies: for example, in the 0.9–1.0 bin, roughly 90 percent of the actual votes should be reversals. The ECE is simply a weighted average—weighted by the number of observations in each bin—of the absolute deviations between these observed frequencies and the predicted probabilities. Lower values indicate better calibration.

The model is relatively well-calibrated by this measure, as shown visually in figure 2. There, we plot the proportion of reverse votes in each bin and include a dashed line of perfect correspondence running at a 45-degree angle, and the size of the dot reflecting the number of observations in the bin. If the model were perfectly calibrated, the dots would appear exactly on the line. They follow the line closely but depart from it more or less randomly. The weighted average of those deviations constitutes the ECE score and can be interpreted as approximately the average error in the predicted probability. The ECE score for the model is 0.058. Earlier efforts do not report ECE scores, so it is difficult to benchmark against them, but the Experience

[83] Tilmann Gneiting & Adrian E. Raftery, *Strictly Proper Scoring Rules, Prediction, and Estimation*, 102 J. AM. STAT. ASS'N 359, 363–65 (2007) (showing relevance of Brier scores); Chuan Guo, Geoff Pleiss, Yu Sun, & Kilian Q. Weinberger, *On Calibration of Modern Neural Networks*, *In* INT'L CONF. ON MACH. LEARNING 1321 (2017); Mahdi Pakdaman Naeini, Gregory F. Cooper & Milos Hauskrecht, *Obtaining Well Calibrated Probabilities Using Bayesian Binning into Quantiles*, 29 PROC. AAAI CONF. ON A.I. 2901 (2015).
[84] Glenn, W. Brier, *Verification of Forecasts Expressed in Terms of Probability*, 78 MONTHLY WEATHER REV. 1 (1950).
[85] Guo et al., *supra* note 83.

Model's ECE score is in the range of what the literature regards as a well-calibrated model.[86] As shown later, this calibration represents a key advantage of the Experience Model over its predecessors—the model can be used as a powerful research tool rather than as a mere predictive classifier of votes.

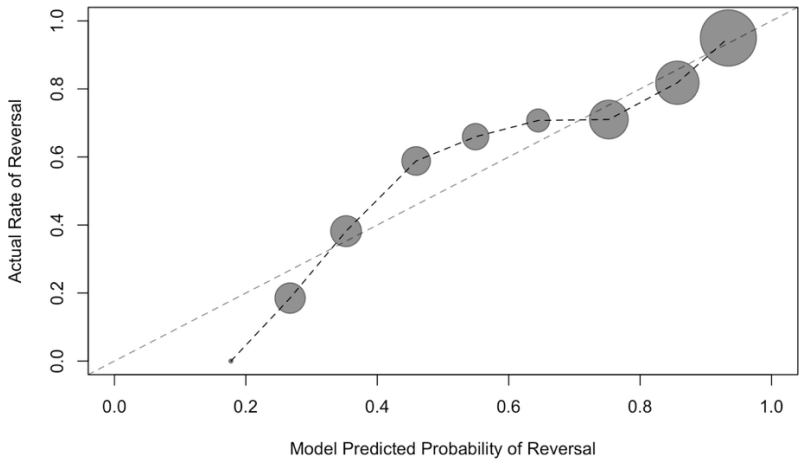

Figure 2: Vote-level Calibration

2. Case-Level Results

Unlike most existing models—including the human crowd source model—the Experience Model performs better at the case level than the individual vote level. Accuracy at the case level for OT2024 is 82 percent. That outperforms existing algorithmic efforts in the test term. It also outperforms humans. Last term, humans predicted case outcomes with a 76 percent accuracy, six points lower than the accuracy of the Experience Model.

The more detailed case-level performance measures appear in table 3. Both the unweighted (0.73) and weighted (0.81) F1 scores improve over the vote level results. For comparison, though they do not use precisely the same classes for case outcomes, at the case level Katz et al. report an unweighted (weighted) F1 score of 0.67 (0.69). A naïve always-reverse prediction would return with an unweighted (weighted) F1 score of 0.42 (0.62). The model continues to show more difficulty with the affirm class—though precision is strong, recall remains a challenge. This again means that, if the model predicts an affirm outcome, it is quite likely that the outcome will be to affirm—but that it misses a large fraction of the affirming cases, mistakenly predicting they will reverse. It is difficult for the model to spot the affirm cases.

[86] Chuan Guo et al., *On Calibration of Modern Neural Networks*, PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (2017) (reporting ECE results for calibrated models on text-based tasks of between 0.02 and 0.09).

Table 3: Experience Model Performance (Cases)

	Precision	Recall	F1	Support
Affirm	0.78	0.47	0.58	15
Reverse	0.83	0.95	0.89	41
Macro (unweighted)			0.73	
Macro (weighted)			0.81	

Note: see text for definitions of performance metrics. Support refers to the number of observations in a class. The weighted macro F1 score weights class scores by the support for that class; unweighted averages the scores for each class with equal weight.

In total, therefore, in this test term the Experience Model outperforms existing algorithms at both the vote and case level. Perhaps even more impressive, it outperforms even humans at the case level. The case level cell size is small in the test term, of course, calling for on-going assessment. As part of this research project, a version of the Experience Model is being used to live-predict the current (OT2025) term, posted publicly at a research log.[87]

Calibration at the case level is roughly the same as it is at the vote level. The Brier score decreases slightly to 0.16, suggesting a somewhat closer relationship between observed outcomes and the model's estimated probabilities. With respect to ECE, the number of case observations is about nine times smaller than the number of vote observations, meaning that the cells become very thin and noisy with the ten bins we used for votes. Given the small number of case observations, and the fact that the ECE is sensitive to the number of bins, Brier may be a better comparison; on that measure, the vote and case level differ only marginally.[88]

B. Disaggregated Results

The main results establish the Experience Model as outperforming earlier benchmarks and human performance in the test term. Beyond predictive performance, however, the promise of a modern model is that it can teach us lessons about individual justices, judicial behavior, and areas of law. This section starts that analysis by disaggregating the main results in three ways.

1. Justice-Level Predictability

Which justices are most predictable? A large literature posits a strong relationship between ideology and judicial voting.[89] A conjecture that follows from that perspective is that the votes of

[87] *See* Edward H. Stiglitz, Reasonable Machines, reasonablemachines.io (Sept. 2025–).

[88] If we aim for an average of about 20 observations per bin we can analyze ECE with three bins at most. Using that binning strategy, the ECE is 0.09.

[89] E.g., SEGAL & SPAETH, *supra* note 39; Peter K. Enns & Patrick C. Wohlfarth, *The Swing Justice*, 75 J. POL. 1089 (2013); for context-dependence on voting, *see* Emerson H Tiller & Frank B. Cross, *Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals*, 107 YALE L.J. 2155 (1998); Max M. Schanzenbach & Emerson H. Tiller, *Reviewing the Sentencing Guidelines*, 75 U. Chi. L. Rev. 715 (2008); Adam Chilton, Christopher Cotropia, Kyle Rozema & Schwartz, *Political Ideology and Judicial Administration*, 41 J.L.

the most ideologically extreme justices ought to be easier to predict.[90] The extreme justices tend to move clearly in one direction or the other in cases—but the moderates tend to be conflicted and often in a pivotal voter position, runs the reasoning, torn between two alternative visions of the law. That conflicted nature may mean that, in principle, they could support either of the visions, with their vote turning for instance on the narrowness of the decision or other difficult-to-predict aspects of the case subject to inter-justice bargaining.

An analysis of model performance at the justice level partially supports this conjecture, as reported in table 4. The votes of Justices Sotomayor and Jackson, widely recognized as the most liberal justices on the Court,[91] tended to be the easiest to predict. The model correctly predicted their votes 86 percent of the time in OT2024. However, contrary to the conjecture, the most conservative justices were among the most *difficult* to predict. Justices Alito and Thomas, the two most conservative justices according to the Martin-Quinn scores,[92] had the lowest and third-lowest predicted accuracies, respectively. Their accuracies were closer to 70 percent. Moreover, the justices often considered to be swing, such as Roberts or Kavanaugh, returned with vote accuracies near the highest in the set.

This pattern resists at least a reductionist ideological account of voting and judicial behavior. The attitudinal model, most fully developed by Segal and Spaeth,[93] posits that Supreme Court justices decide cases primarily based on their ideological preferences. A strong form of that thesis predicts that the most ideologically committed justices should be the most predictable: their preferences point consistently in one direction, making their votes easier to forecast. The results here offer partial, asymmetric support. The liberal justices conform, but the conservative justices complicate the prediction.

It may be that Justices Alito and Thomas produce surprise votes more often, indeed, precisely because of their (typically) strident ideology. To take an example from the test term, in *Thompson v. United States*, the Court reversed a lower court's conviction for making false statements to the FDIC on the grounds that, unlike other statutes, the relevant statute criminalizes only "false" statements and not statements that are "misleading" but not "false."[94] The model predicted that Alito and Thomas would vote to affirm the conviction—likely keying off a general skepticism of criminal defendants. Yet in this case they voted to reverse, possibly because their general skepticism of criminal defendants was overridden by their commitment to textualism. This might be an instance, therefore, of the model over-learning their ideology and under-playing jurisprudence.

Or to take another example, in *Free Speech Coalition v. Paxton*, the Court upheld a Texas law requiring websites hosting sexually explicit material to verify ages of their visitors. Thomas

---

Econ. & Org. 91 (2025). Cf. Michael S. Kang & Joanna M. Shepherd, *The Partisan Price of Justice*, 86 N.Y.U. L. Rev. 69 (2011).

[90] SEGAL & SPAETH, *supra* note 39.

[91] According to the widely used Martin-Quinn scores, the most liberal justices in OT2024 were Sotomayor and then Jackson.

[92] *See* https://mqscores.wustl.edu/measures.php; Andrew D. Martin & Kevin M. Quinn, *Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999*, 10 POL. ANAL. 134 (2002).

[93] SEGAL & SPAETH, *supra* note 39

[94] Thompson v. United States, 604 U.S. 408 (2025).

wrote the opinion for a 6-3 majority, which Alito joined. Yet the model had predicted they would vote to reverse, thereby upholding the law. Here, it is possible that the model over-keyed on the tendency of the justices to use the First Amendment to reject government regulation—in recent years, a general posture that favors conservative outcomes.[95] Yet in this case, the conservatives favored the regulation and looked past qualms about the First Amendment. This case might be understood as one in which the model over-learned the jurisprudential layer as a window into ideology.

By comparison, the predictability of the liberal wing may well be due to their status as structural minorities in the Court. One consequence of being in the minority is that they have less control over the docket and direction of the Court. Though the Court will grant cert on consensus conservative priorities, many of the hard, contested cases they take will feature divisions *within* the conservative coalition. For instance, in *NRC v. Texas* the Court reversed a lower court decision that had vacated an agency action.[96] In an opinion written by Justice Kavanaugh, they held that the party challenging the agency action was not entitled to judicial review under the relevant statute, the Hobbs Act. Justices Alito and Thomas dissented. The dissenters would have looked past the question of reviewability and reached the merits, affirming the lower court and vacating the agency action. The conservatives, therefore, split: some went one way based on procedure; others went the other way on the merits.

Table 4: Experience Model Performance (Justice-Level)

Justice	N	Accuracy	Precision	Recall	F1
SAAlito	55	0.67	0.78	0.58	0.67
ACBarrett	56	0.7	0.75	0.81	0.78
CThomas	56	0.7	0.77	0.7	0.73
NMGorsuch	53	0.7	0.66	0.93	0.77
EKagan	56	0.71	0.78	0.78	0.78
BMKavanaugh	57	0.74	0.83	0.81	0.82
JGRoberts	56	0.77	0.82	0.88	0.85
KBJackson	56	0.86	0.92	0.88	0.9
SSotomayor	56	0.86	0.88	0.92	0.9

Note: precision and recall statistics provided for the reverse class.

These conclusions must be read against the caution regarding power noted earlier in the case-level analysis. Once we disaggregate votes by justice, the number of observations becomes relatively small and differences between justices may be due to sampling variability. The

[95] Amanda Shanor, *The New Lochner*, 2016 WIS. L. REV. 133 (2016) (documenting the use of the First Amendment as a deregulatory tool); see also Janus v. AFSCME, Council 31, 585 U.S. 878 (2018) (Kagan, J., dissenting) ("The First Amendment was meant for better things. It was meant not to undermine but to protect democratic governance — including the role of public-sector unions.")

[96] Nuclear Regulatory Commission v. Texas, 605 U.S. 665 (2025).

accuracy for Justice Sotomayor, for instance, is statistically distinguishable from that of Justices Alito,[97] Barrett,[98] Thomas,[99] and Gorsuch[100]—but not from Justice Kagan.[101]

2. Issue-Level Predictability

As discussed later, the extent to which judicial determinations can be predicted informs rule of law values.[102] An area of law that is predictable is likely to be well-defined and help guide citizens in their conduct. An area of law that cannot be easily predicted, on the other hand, may involve substantial, idiosyncratic judicial discretion, falling short of rule of law ideals regarding foreseeability and failing to give citizens a pathway in their lives. In this way, predictability is a virtue.

In principle, it is possible to analyze predictability at a fine level of granularity. We could examine predictability by individual statute, for instance, or by constitutional provision. Doing so, however, requires many observations, and accordingly this illustrative exercise examines predictability by main issue type.[103] Especially for the issues with few cases, it bears repeating the caution about statistical power and cell size.

With that caution noted, as reported in table 5, there is significant heterogeneity in predictability across issue type.[104] The range between the most and least predictable area is substantial: the most predictable area can be correctly predicted almost 90 percent of the time; the least predictable area, less than 70 percent of the time. Focusing on issues for which there exist at least 20 votes, Due Process, Judicial Power, and even First Amendment and Civil Rights cases tend to be relatively predictable, all with vote-level accuracies at or above 75 percent. The least predictable cases, by comparison, tend to involve Economic Activity or Criminal Procedure as their main topic.

Table 5: Predictable Issues (Vote Level)

Issue Area	N	Accuracy	Precision	Recall	F1
Criminal Procedure	89	0.69	0.71	0.68	0.7
Economic Activity	143	0.71	0.79	0.8	0.79
Civil Rights	72	0.75	0.84	0.77	0.8
First Amendment	36	0.75	0.67	1	0.8
Judicial Power	108	0.77	0.8	0.88	0.84
Due Process	26	0.88	0.95	0.9	0.92

[97] $p = 0.02$.
[98] $p = 0.04$.
[99] $p = 0.04$.
[100] $p = 0.05$.
[101] $p = 0.07$.
[102] *Infra* Part IV.A.
[103] Issue area for this analysis derives from the Supreme Court Database.
[104] The table reports all issue areas for which three or more cases were decided in the term. This excludes the SCDB main issue areas of "Attorneys," "Federal Taxation," and "Unions."

Note: precision and recall statistics provided for the reverse class. Issue area coded according to the Supreme Court Database.

Though the cell sizes naturally decrease at the case level, the pattern of predictability is even more stark at that level, as reported in table 6.[105] There it is evident that criminal procedure is, by a substantial margin, the least predictable area of law.[106] Indeed, excluding criminal procedure cases from the analysis would boost predictability fully five percentage points over the full-sample results, to 87 percent at the case level.[107]

Part IV interprets what these differences may imply for rule of law values, including whether predictability might signal legal indeterminacy.

Table 6: Predictable Issues (Case Level)

Issue Area	N	Accuracy	Precision	Recall	F1
Criminal Procedure	10	0.6	0.67	0.67	0.67
Economic Activity	16	0.75	0.75	1	0.86
First Amendment	4	0.75	0.67	1	0.8
Judicial Power	12	0.92	0.9	1	0.95
Civil Rights	8	1	1	1	1
Due Process	3	1	1	1	1

Note: precision and recall statistics provided for the reverse class. Issue area coded according to the Supreme Court Database.

3. Vote Margin

Are unanimous or near-unanimous votes easier to predict than 5-4 or 6-3 votes? One might speculate that the greatest predictive uncertainty is in the "close" cases—that the 9-0 or 8-1 cases can be easily predicted. However, that conjecture would be only partially correct based on results from the test term.

Though performance appears somewhat stronger for lopsided decisions than close decisions, the difference is not overwhelming. Table 7 decomposes model performance at the vote and case levels by observed vote margin. The top row of the table shows, for instance, that on unanimous decisions the model predicted votes with 75 percent accuracy and cases outcomes with 87 percent accuracy. That level of performance is, indeed, stronger than for 5-4 cases, where the model predicts votes with 68 percent accuracy and case outcomes with 75 percent accuracy. This suggests that 5-4 decisions likely contain more contradictory signals or cross-cutting doctrinal

[105] Statistics reported for issue areas with more than one case in the test term.
[106] Cf. Joshua B. Fischman & Max M. Schanzenbach, *Do Standards of Review Matter? The Case of Federal Criminal Sentencing*, 40 J. LEGAL STUD. 405 (2011).
[107] At the vote level, predictability would boost to 76 percent from 74 percent.

lines that make prediction difficult. The difference in performance between these two sets of cases, however, is of degree and not kind.[108]

Moreover, the weakest performance was for cases that ended up decided by 7-2 votes. There, the model correctly predicted votes with only 66 percent accuracy, and case outcomes with only sixty percent accuracy. It is not as though performance monotonically increases with vote margin. Instead, the pattern is more irregular—even quite lopsided cases can be difficult to predict. And the strongest-performing vote margin was for 6-3 votes.

Table 7: Accuracy by Vote Margin

Margin	N Votes	Vote Accuracy	N Cases	Case Accuracy	Case Accuracy (Humans)
9-0	207	0.75	23	0.87	0.87
8-1	28	0.75	3	1	0.67
7-2	89	0.66	10	0.6	0.5
6-3	81	0.86	9	0.89	0.89
6-2	8	0.75	1	1	1
5-4	72	0.68	8	0.75	0.75

Note: Margin refers to observed majority-dissent vote counts, as reported in the SCDB. In OT2024 there was a single 6-2 case. Human accuracy based on FantasyScotus.

Lopsided votes may be difficult to predict because they represent relatively technical legal questions on which the justices enter with weak priors and therefore can be persuaded.[109] Or they might represent decisions in which the justices settle and agree on a dimension not salient in the lower court—for instance, jurisdictional grounds, such as standing or mootness.[110] Decided on the merits, the case goes one way, decided on procedural ways, the case goes unanimously the other way. Given the high dimension nature of Supreme Court decisions, it is difficult for models—or humans—to isolate the dimension of conflict that will govern the case.

Closer votes, on the other hand, may sometimes be easier to predict. The 6-3 vote margin especially stands out. Many of the 6-3 decisions represent cases in which the justices voted along their conventional ideological lines.[111] On high profile cases with salient ideological contours, the vote count is likely to be 6-3—"close" in a sense, but also highly predictable.

[108] Notably, human performance, reported in the final column, is similar to model performance along the various vote margins.
[109] E.g., Kousisis v. United States, 604 U.S. ___, 145 S. Ct. 1047 (2025) (narrow statutory question); Thompson v. United States, 604 U.S. ___, 145 S. Ct. 807 (2025) (narrow statutory question).
[110] Royal Canin U.S.A., Inc. v. Wullschleger, 604 U.S. ___, 145 S. Ct. 41 (2025) (potentially politicized merits question avoided by deciding on jurisdictional grounds); Bouarfa v. Mayorkas, 601 U.S. ___, 145 S. Ct. 9 (2024) (immigration case decided on jurisdictional grounds).
[111] E.g., Mahmoud v. Taylor, 145 S. Ct. 1567 (2025) (religious rights and opting out of LGBT curriculum).

Investigating the errors, moreover, reveals a consistent pattern of the model over-predicting reverse. Of six unanimous affirmances, for example, the model was able to predict only one correctly;[112] of the two 5-4 affirm cases, the model predicted zero correctly.[113] What likely occurs in these cases is that the model parses only a weak signal regarding the direction of the outcome—perhaps related to the relatively technocratic nature of the cases—and therefore leans into the fact that the Court reverses most lower court decisions. That is, the model essentially defaults to the majority class (reverse) if the idiosyncratic case signal is weak, which is a sensible optimization strategy.

C. Summary

Predictive technology is not yet an oracle, but the new generation of models allow for predictions that, in the test term, exceed existing algorithmic state of the art and surpass human capabilities on many measures. At the case level, the Experience Model predicted last term's cases with 82 percent accuracy, a stronger performance than wisdom of the human crowds. It is reasonable to expect performance to continue to improve as models improve and scholars refine the craft of data curation and structuring.

Disaggregated results further indicate that some justices, issue areas, and types of decisions can be predicted with relative ease. Justices Jackson, Sotomayor, and Roberts tend to be relatively predictable; Justices Alito, Barrett, and Thomas tend to be less predictable. Cases involving criminal procedure or economic activity tend to be less predictable; cases involving judicial power or due process tend to be more predictable. Predictability does not move monotonically with vote count—more lopsided cases are not necessarily easier to predict. The easiest cases to predict appear to be those that track with well-worn ideological patterns, with the conservatives splitting from the liberals 6-3.

IV. Empirical Implications for Legal Institutions

A highly predictive model such as the Experience Model is valuable for more than merely forecasting votes—it provides a method for learning empirical lessons about legal institutions. This Part illustrates that point through four important questions of legal theory. First, we use the model to assess the extent to which the legal system might guide citizen conduct; second, the ex-ante calibrated model predictions can be used to test theories of strategic litigation; third, the model allows insight into the role and utility of oral argument; finally, the model is used to study the late-term cases decided in June, which tend to attract the most media attention. Are these complex, difficult-to-predict cases? Or merely more controversial cases? The aim throughout is less to resolve these questions definitively than to point to a novel methodology that can be used to address them.

[112] It correctly predicted Rivers v. Guerrero, 605 U.S. 443 (2025), and incorrectly predicted Wisconsin Bell, Inc. v. United States ex rel. Heath, 604 U.S. 115 (2025); Bouarfa v. Mayorkas, 604 U.S. 6 (2024); Royal Canin U.S.A., Inc. v. Wullschleger, 604 U.S. 22 (2025); Kousisis v. United States, 605 U.S. 114 (2025); TikTok, Inc. v. Garland, 604 U.S. 56 (2025).

[113] Perttu v. Richards, 605 U.S. 460 (2025); Medical Marijuana, Inc. v. Horn, 604 U.S. 593 (2025).

A. Assessing Guidance

Despite disagreements on many points, prominent legal theorists converge on the centrality of law's guidance function—its capacity to provide citizens with a pragmatic or moral pathway for conduct.[114] Yet for law to guide conduct, it must be predictable: it must be relatively foreseeable, knowable, and stable.[115] The empirical findings reported in Part III bear directly on this shared commitment: they measure, at the case and vote levels, how predictable the Court's authoritative determinations actually are.

The model, that is, provides an empirical methodology for examining the rule of law characteristics of the legal system. The extent to which a legal system, or area of law within the legal system, is algorithmically predictable supplies a measure of the extent to which it might guide citizens and other government actors. Highly predictable legal systems, or areas of law, likely provide a clear pathway for conduct; those that cannot be easily predicted likely do not. Algorithmic predictability provides a window into the rule of law virtues of the legal system and particular pockets of law.

Overall, by this approach, the Court is quite predictable, suggesting that the legal system more broadly plausibly satisfies the guidance function criteria. Recall again that, viewed over the entire litigation pipeline, these likely represent the *least* predictable cases in the legal system; nevertheless, it is possible to correctly predict case outcomes over 80 percent of time.

Some areas of law, moreover, appear to be more predictable than others. As reported in table 6, the Court is more predictable when it comes to cases involving civil rights, due process, and judicial power. Accordingly, this suggests that the law might be broadly regarded as able to guide in these areas, with regulated parties able to adjust their conduct to reliable pathways. On the other hand, predictability is relatively low in the areas of criminal procedure, economic activity, and the first amendment, with only 60-75 percent of cases being correctly forecasted. These might be considered areas of law that would fit Hart's definition of "open texture" in law, where the law runs out and outcomes become indeterminate.[116]

---

[114] HART, *supra* note 11 (referring to the "internal point of view" as central to the rule of law); Fuller, *supra* note 11, at 657 (regarding law "as a means for achieving a certain kind of order … a particular kind of control or power, that obtained through subjecting people's conduct to the guidance of general rules by which they may themselves orient their behavior"); RAZ, *supra* note 11; SHAPIRO, *supra* note 11, at 170-73 (arguing that "legal institutions are supposed to enable communities to overcome the complexity, contentiousness, and arbitrariness of communal life by resolving those social problems that cannot be solved, or solved as well, by non-legal means alone"); GOWDER, *supra* note 11 (arguing that "the rule of law makes it possible for citizens to become committed to the legal order in which they live, and demands that commitment in order to permit the law to be used as a tool to coordinate their behavior in order to preserve their equal standing").

[115] *E.g.,* FULLER, *supra* note 21, at 39 (pointing to critical features of legal systems as "publiciz[ing]" rules, making them "prospective," "understandable," ensuring they are stable and not subject to "frequent changes" such that "the subject cannot orient his action by them."); RAZ, *supra* note 11, at 214-16 (similarly pointing to the importance of "prosective, open, and clear" rules, that are "relatively stable"); HART, *supra* note 11, at 144 (pointing to the importance of law having a "core of settled meaning").

[116] HART, *supra* note 21, at 124 (observing that rules, "however smoothly they work over the great mass of ordinary cases, will, at some point where their application is in question, prove indeterminate; they will have what has been termed an open texture").

The differential predictability documented above is consistent with the open-texture interpretation, but two alternative interpretations cannot be ruled out based on the present data. The first is model performance: an area may perform poorly due to model-specific deficits in that domain—for example, perhaps the model had less pre-training on certain types of disputes. To provide an example by analogy, AI models tend to be very good at coding in Javascript, Python, and many other languages—but they appear to struggle with Stata code, an historically popular but more recently disfavored statistical program.[117] This deficit does not mean that Stata code is more indeterminate or, in an underlying sense, more difficult to predict. It is simply likely that the model saw less Stata code in pre-training because the community posting coding solutions online is smaller, so the model is less able to generate code in that language. A similar kind of concern might apply for different areas of law. A concern of this nature is valid in principle, but for large, frontier models, it appears relatively unlikely as the models would be trained on virtually every area of law. Moreover, with larger datasets, scholars could examine calibration by issue area, which would provide a window into the model's understanding of the issue area apart from classification performance. If the model is well calibrated in the issue area, it is less likely that classification performance is due to a model-specific deficit in that area.

A second threat is based on litigant behavior: parties in some area may systematically pursue different types of cases. Case stakes fit into this category. Criminal defendants, for instance, may pursue legal claims that, translated to civil litigation analogs, would be abandoned or settled as unlikely to succeed. They might pursue these likely-to-lose claims because the stakes of losing, the loss of liberty, are so high that they will chase any possibility of relief. If this is right, the observed cases for criminal defendants may be more likely to dig into any slight ambiguity or contradiction in the law, whereas cases in other areas would be more likely to settle rather than explore these low-probability crannies of doctrine. In turn, this may mean that the more exploratory criminal law area *appears* more unpredictable—because it engages the slightest ambiguity. But the reality is that this lack of relative predictability reflects litigant behavior rather than the nature of the law in that area.

This concern is valid in principle, mitigated by a conceptual consideration: normatively, the level of predictability that we view as necessary may vary by issue. Some legal issues require more predictability, and others less. The predictability of who owns the mineral rights in Pennsylvania, for instance, should be strong. The predictability of who owns the mineral rights on Mars, on the other hand, does not need to be strong. Why? Because contestation over the mineral rights on Mars is unlikely—due to the infeasibility of transport—whereas contestation between mineral and surface rights is a staple of disputes in the United States—due to the accessibility of high value minerals.[118] To put it another way, the stakes of mineral rights in Pennsylvania are higher than the stakes of similar rights on Mars; so normatively, we should be more demanding about the law in Pennsylvania than on Mars. That same logic suggests that we ought to be more demanding of the law in criminal procedure than in other, lower stakes areas. Saying that predictability is lower in criminal procedure because litigants push harder on claims in that space only shows that this is a high-stakes area—not that, in the relevant normative sense,

[117] E.g., https://blogs.worldbank.org/en/impactevaluations/can-ai-write-your-stata-code.
[118] E.g., Pennsylvania Coal Co. v. Mahon, 260 U.S. 393 (1922); Amoco Production Co. v. Southern Ute Indian Tribe, 526 U.S. 865 (1999); Watt v. Western Nuclear, Inc., 462 U.S. 36 (1983).

criminal law is sufficiently predictable. It may be useful to think about predictability in terms of litigation-adjusted terms, wherein we ask—conditional on observed litigation—how predictable is the law? By that measure, criminal procedure is less predictable than other areas.

B. Litigating Losers? Exploiting Model Self-knowledge

An important feature of the model is that its predictions are well calibrated—cases the model predicts will reverse with about fifty percent probability, for example, tend to reverse about fifty percent of the time. The Brier score, a standard way to score predicted probabilities,[119] is about 0.17, close to human superforecaster benchmarks.[120] And in cases in which the model was at least seventy-five percent confident, it was correct one hundred percent of the time. This measure of self-knowledge, which most humans famously do not possess,[121] is useful for studying litigation and legal systems.

These calibrated scores provide a methodology to study the extent to which parties litigate sure losers. Priest and Klein's theory suggests that parties will settle cases in which one side or the other is likely to lose, and that most observed litigation should constitute relatively close calls.[122] At least with accurate predictions of outcomes, and symmetry in litigation stakes, most observed cases, their theory suggests, should be about fifty-fifty propositions.[123] That theory requires modification for the Supreme Court, as the justices themselves control their docket—they select among the cases seeking cert those that they wish to review. A common understanding is that the Court is more likely to grant cert to a case they wish to reverse,[124] thereby correcting a lower court understanding of the law—merely affirming is rarely worth the time or docket space—and this second layer of selection nudges the observed cases towards reversal.

Numerous empirical studies examine the fifty-fifty hypothesis.[125] They tend to partially support the claim, which is difficult to test cleanly for reasons discussed below. The predictions from a model such as the Experience Model offer two methodological advantages over existing studies. The first is that existing studies rely on observed win rates to test the hypothesis. They ask if the fraction of wins approximates fifty percent. However, the theoretically relevant quantity is the ex-ante win probability—that is, the win probability that parties experienced at the time they made decisions about settlement. It is possible, for example, that a fifty percent observed win rate derives from one-half ex-ante sure losers and one-half ex-ante sure winners; in this case, a fifty percent win rate would provide misleading support to the hypothesis. An

[119] Gneiting et al., *supra* note 83.
[120] PHILIP E. TETLOCK & DAN GARDNER, SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION (2015) (using Brier scores to evaluate predictors).
[121] Kruger & Dunning, *supra* note 10.
[122] Priest & Klein, *supra* note 13.
[123] Klerman & Lee, *supra* note 14 (observing the results hold only in the limit).
[124] H.W. PERRY, JR., DECIDING TO DECIDE: AGENDA SETTING IN THE UNITED STATES SUPREME COURT (1991).
[125] *E.g.*, Waldfogel, *supra* note 14; Eisenberg, *supra* note 14; Siegelman & Donohue, *supra* note 14; Clermont & Eisenberg, *supra* note 14; Eisenberg & Farber, *supra* note 14; Chang & Hubbard, *supra* note 14; Klerman & Lee, *supra* note 14 (critically reviewing the literature).

advantage of the Experience Model is that it generates calibrated ex-ante win probabilities, approximating the theoretically relevant quantity.

A second advantage of using ex-ante predictions is that, in principle, they can also be generated for cases that parties do not fully litigate. That is, they allow for more explicit comparison of the un-litigated and litigated cases. Does the distribution of ex-ante win probabilities differ for the settled and litigated cases? In the Supreme Court context, this exercise would require developing a model similar to the Experience Model, but using only data common to all appellate decisions—the distribution of ex-ante win probabilities for litigated Supreme Court cases can then be compared to the corresponding distribution for cases that rested at the appellate level.

Even under the conventional theories, win probabilities may diverge from fifty percent for various reasons.[126] To start, the theory only holds if parties can in fact predict outcomes reasonably well—noisy predictions will reduce the selection pressures that converge outcomes to fifty percent probability.[127] Moreover, asymmetric stakes of litigation affect the theory.[128] A party with more at stake will, in general, litigate lower probability claims. It would not be surprising, for instance, to find a lower predicted probability of reversal for cases involving criminal appeals, as the defendant may face loss of liberty whereas the prosecutor may face merely the loss of a case.

Examining the ex-ante probabilities for the test term, the probability of reversal approximately matches theoretical expectations, as shown in figure 3. There is a mass of probability around the theoretically relevant fifty percent mark. And the tails of the distribution become thin, particularly on the affirm side—very few cases can be said to be sure losers for either party. The distribution is also shifted to the right, plausibly reflecting the reverse-bias of the cert granting process, which layers an additional level of selection on top of the classic theory.[129] The average predicted probability of reversal is about 0.65. [130]

[126] Other theoretical approaches would produce additional reasons for departure from the fifty percent prediction. For example, information asymmetry, not in the Priest Klein framework, could produce departures. E.g., Lucian A. Bebchuk, *Litigation and Settlement Under Imperfect Information*, 1 RAND J. ECON. 404 (1984).
[127] Klerman & Lee, *supra* note 14 (critically reviewing the literature).
[128] *Id*.
[129] PERRY, *supra* note 124 (noting a reverse bias in the cert process).
[130] A Kolmogorov-Smirnov test easily rejects the hypothesis that the distribution is uniform ($p << .01$); on the other hand, we cannot reject the hypothesis that the distribution is normal with mean 0.65 and standard deviation 0.15 ($p = 0.56$).

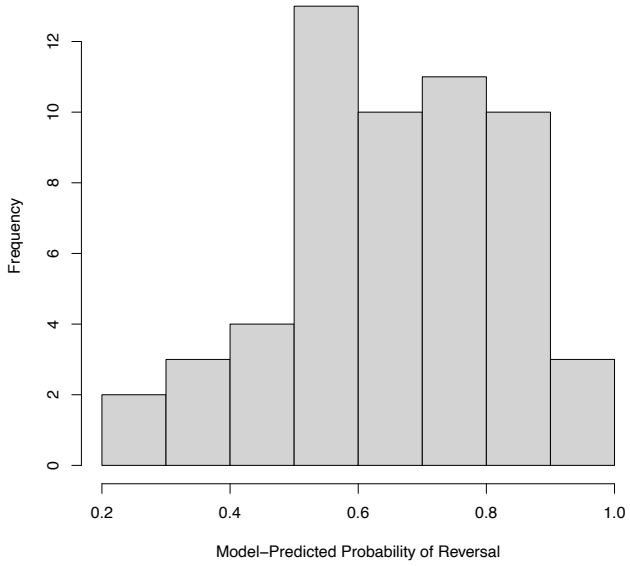

Figure 3: Few Sure-Losers Litigated

What is more, this pattern appears to hold once we disaggregate cases by issue area. Though the cells become very small at the case level when disaggregated, the average probability of reversal within areas ranges from 0.54 to .69.[131] Among these areas, the criminal cases show the lowest average ex-ante probability of reversal. That again may reflect the asymmetric stakes featured in criminal cases. The number of observations at the case level, particularly broken down by area, is too small to resolve the question, but the methodology in this section suggests a new strategy for addressing claims involving settlement and strategic litigation.

C. Understanding the Predictive Role of Oral Argument

The conventional wisdom is that oral argument matters little predicting vote or case outcomes—it is mostly a ceremony.[132] Yet the justices themselves maintain argument is key to decision-making,[133] and revisionist accounts by social scientists likewise point to its importance.[134] Shortly after joining the Court, for instance, Justice Roberts wrote an article studying predictability and Supreme Court oral argument.[135] Using a sample of twenty-eight

[131] Issue areas with more than five cases reported. Within this set, the lowest average probability of reversal was for criminal procedure (0.54) and the highest average probability of reversal was economic activity (0.69).
[132] See ROHDE & SPAETH, *supra* note 16; JOHNSON, *supra* note 16 (noting the "dominant view" that argument has "little influence over case outcomes"); SEGAL & SPAETH, , *supra* note 16 (observing that, though justices say argument matters, this "does not mean that oral argument regularly, or even infrequently, determines who wins and who loses," and noting that argument comes up rarely in the justices' papers).
[133] Roberts, *supra* note 17.
[134] JOHNSON, *supra* note 16.
[135] Roberts, *supra* note 17.

cases, he found that the side that receives more questions from the justices loses 86 percent of the time.[136]

One general concern with studying the role of argument is that much depends on the quality of the baseline model. That is, if the baseline model is very poor, adding argument is likely to matter—but not because argument is vital, instead because the baseline is poor. A strong baseline model, on the other hand, may show that the *marginal* contribution of argument is modest. Studies such as Justice Roberts', in effect, do not include a baseline model, so argument appears important to outcomes. Yet a more capable baseline model without argument may perform almost as well as a model with argument.

Another concern with existing studies is that they typically do not pull out the marginal effects of oral argument by justice.[137] Justice Thomas, for instance, famously asked few questions for much of his time on the bench. Is argument data uninformative for him? Or, instead, is the model able to learn patterns from the questioning of *other* justices about Justice Thomas's likely behavior? Some justices, likewise, may speak often, but mask their vote intentions. Others may speak little, but do so emotively, such that their vote intentions become clear with only a few words.

It is thus useful to identify two relevant dimensions of studies of oral argument. The first is whether they study the *marginal* contribution of argument, or instead its overall predictive power. The second is whether that effect is studied at the case or justice level. Existing studies examine the total effect of argument at the case level,[138] the total effect of argument at the justice level,[139] or the marginal contribution at the case level.[140] However, no existing study analyzes the marginal effects of argument at the justice level. As we seek to understand whether argument matters, when, and why—we must know for whom it matters and in what cases.

The Experience Model provides a natural inroad for this query. We can generate a powerful, modern baseline model trained with rich data—but without argument data—that promises a relatively high-performing baseline prediction. The full model, trained with oral argument data, then tells us of performance with argument, and the difference between these sister models informs us of the marginal contribution of argument data.

---

[136] Roberts, *supra* note 17, at 75.

[137] Prior studies examine either the marginal predictive contribution of oral argument or justice-level variation but not both. For marginal contribution at the case level, see Aaron Russell Kaufman, Peter Kraft & Maya Sen, *Improving Supreme Court Forecasting Using Boosted Decision Trees*, 27 POL. ANAL. 381 (2019); Bryce J. Dietrich, Ryan D. Enos & Maya Sen, *Emotional Arousal Predicts Voting on the U.S. Supreme Court*, 27 POL. ANAL. 237 (2019). For justice-level analysis without marginal contribution, see Timothy J. Dickinson, *A Computational Analysis of Oral Argument in the Supreme Court*, 28 CORNELL J.L. & PUB. POL'Y 449 (2019); Lee Epstein, William M. Landes & Richard A. Posner, *Inferring the Winning Party in the Supreme Court from the Pattern of Questioning at Oral Argument*, 39 J. LEGAL STUD. 433 (2010); Timothy R. Johnson, Paul J. Wahlbeck & James F. Spriggs II, *The Influence of Oral Arguments on the U.S. Supreme Court*, 100 AM. POL. SCI. REV. 99 (2006).

[138] *E.g.*, Roberts, *supra* note 17.

[139] *E.g.*, Dickinson, *supra* note 137; Tonja Jacobi & Kyle Rozema, *Judicial Conflicts and Voting Agreement: Evidence from Interruptions at Oral Argument*, 59 J. LEGAL STUD. 22 (2020).

[140] *E.g.*, Kaufman et al., *supra* note 137.

Relatedly, we can break down the marginal contribution of argument by issue area, identifying the issues over which justices might be most revealing of their votes during argument. It may be that argument is revealing for some topics and not others—for example, for civil rights cases but not civil procedure cases—because some topics elicit more unvarnished views from the justices.

1. How Much Does Oral Argument Add? An Ablation Study

To study the impact of oral argument on predictions, we can create a new training and test dataset that does not contain oral argument information.[141] Then the model can be re-trained on this ablated data, so that a new model is produced that takes as much advantage of the non-argument case features as possible. Comparing the performance of the model trained without argument data to the model trained with argument data provides information about the value of oral argument to predictability.

Consistent with Justice Roberts' study, removing oral argument data substantially reduced model performance. Accuracy drops from 74 percent at the vote level to 70 percent. As reported in table 8, precision, recall, and the F1 scores, likewise, diminish. The most notable decrease in performance is in recall for the affirm class, which informs the extent to which the model can correctly include affirm votes as affirm predictions; the lower recall value indicates that the model is now under-inclusive with respect to affirm predictions. This is likely because, without the oral argument data, the model leans into its prior, which is that the "typical" vote is to reverse the lower court. Oral argument helps the model learn which votes likely run contrary to that prior.

Table 8: Experience Model Performance (No argument, Votes)

	Precision	Recall	F1	Support
Affirm	0.63	0.3	0.41	173
Reverse	0.71	0.91	0.8	328
Macro (unweighted)			0.6	
Macro (weighted)			0.66	

Note: see text for definitions of performance metrics. Support refers to the number of observations in a class. The weighted macro F1 score weights class scores by the support for that class; unweighted averages the scores for each class with equal weight.

[141] The training data used for this ablation exercise is identical in every respect to the data used for the main model discussed earlier, save for the fact that it removes oral argument features. The ablated model is trained for one fewer epoch than the full model, as the reduced feature set converges more quickly and is prone to overfitting.

The story is much the same at the case level. Overall, accuracy dropped markedly, from 82 percent with argument to 71 percent without argument. As at the vote level, most under performance comes from the affirm class, where both precision (among those identified as affirm, what fraction are actually affirm) and recall (among those that are actually affirm, what fraction does the model identify as affirm) about halving from the model trained with argument data. Here, too, with poorer information, the model is leaning into the class balance in the data, predicting that more cases reverse.

Table 9: Experience Model Performance (No argument, Cases)

	Precision	Recall	F1	Support
Affirm	0.43	0.2	0.27	15
Reverse	0.76	0.9	0.82	41
Macro (unweighted)			0.55	
Macro (weighted)			0.68	

Note: see text for definitions of performance metrics. Support refers to the number of observations in a class. The weighted macro F1 score weights class scores by the support for that class; unweighted averages the scores for each class with equal weight.

2. Justice-Level Ablation Results

The results so far tell us that oral argument, in general, tends to produce stronger vote and case level predictions. What is unclear, however, is whether this is true for all justices or merely a subset of them. The questions asked by some justices may be more revealing of their votes than questioning of other justices, either because they ask fewer questions or because they obscure their likely vote more successfully.

The justice level ablation results, reported graphically in figure 4, indicate that the effect of argument on accuracy is highly heterogeneous. Several justices exhibit substantial gains in accuracy, most notably Justices Sotomayor, Jackson, and Gorsuch, each of whom increase in accuracy by ten or more percentage points. On the other hand, including argument produces very small, or in a few instances, even slight negative effects for some justices: Justices Kavanaugh, Barrett, Thomas, and Alito. Justices Roberts and Kagan show modest improvements in performance by including argument.

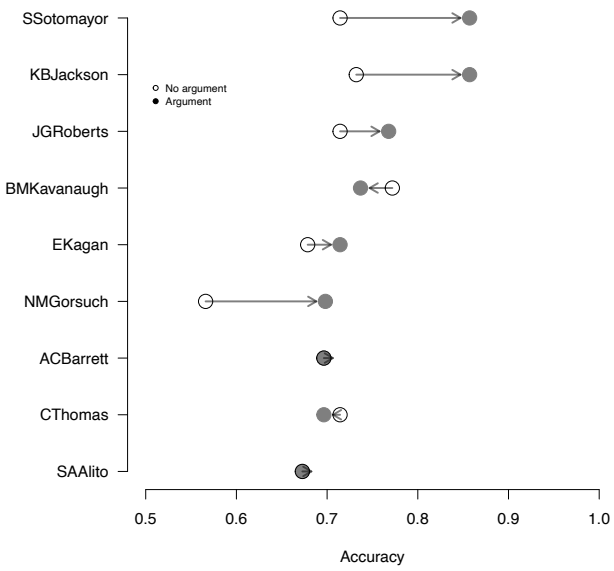

Figure 4: Justice-Level Performance with and without Argument

It is possible to imagine several accounts for why these justices exhibit the greatest performance gains from including oral argument data. As noted earlier, it may be that these justices tend to reveal the most about their positions during questioning—they lay their cards on the table. Other justices, on the other hand, may be less revealing. Such differences may arise from the simple volume of questions, or they may derive from the content of the questions themselves. Another possibility is that the justices with the most movement represent those with more finely articulated jurisprudence. The features in the baseline model are sufficient to predict many justices, but for these more elaborately articulated justices—whose votes appear to be more unpredictable—information from argument is helpful. To put it another way, there is more to learn about the jurisprudence of the justices with more movement than those with less movement; they have a higher dimensional method of evaluating judicial decisions. These two explanations rest on different foundations: the first on difference in the strength of signals produced during argument; the second on differences in the complexity of the underlying jurisprudence that the model is attempting to simulate.

In fact, figure 4 provides suggestive evidence that both explanations may be at play. Accuracy for Justice Gorsuch without argument information is very poor—the model correctly predicts his vote only about 55 percent of the time. This suggests that the baseline model without argument data provides a poor approximation of his jurisprudence; it is more complex or higher dimensional than the baseline case characteristics allow insight into. With argument information, however, the model closes the gap on his accuracy. The baseline accuracy for Justices Sotomayor and Jackson, on the other hand, resemble those of other justices. This suggests that the improved accuracy for these justices may primarily be due to being relatively transparent about their likely votes during argument.

3. Issue-Level Ablation Results

Just as performance variably improved across justices due to oral argument data, the same is true of issue areas. As shown in figure 5, argument is far more informative in some issue areas than others. It increases vote level accuracy in cases focused on criminal procedure, First Amendment, due process, and unions; but argument does not change vote level performance, or even decreases it, in cases involving economic activity, civil rights, attorneys, and federal taxation.

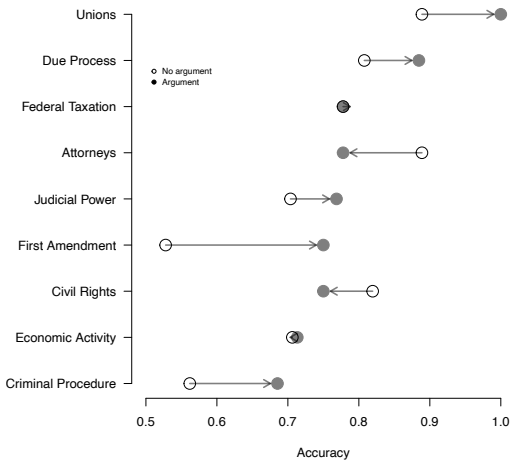

Figure 5: Issue-Level Performance with and without Argument

The fundamental categories of explanation for this pattern also remain the same as earlier. It is first possible that some issue areas represent a greater challenge for the model—they are more complex and therefore rely on argument to extract signals about likely votes. The most likely candidates for this category include First Amendment and criminal procedure. Without argument data, those classes of cases tend to markedly underperform other areas, suggesting that the baseline struggles to understand their jurisprudence and relevant signals. With argument data, however, those issue areas perform roughly in line with other issue areas.

A second possible interpretation is that oral argument in some areas is more revealing than in other areas. Some topics may, for instance, induce justices to reveal their preferences more starkly, whereas in other areas justices tend to obscure their intentions during argument. Unions and due process represent the most natural candidates for this possibility. Here, the baseline model performs relatively well—it is not as though the baseline model struggles to appreciate the jurisprudence in these areas (in a relative sense). However, the model still finds considerable information in argument, suggesting that justices may be especially transparent about their intended votes for these issues. The issues may, for example, elicit emotions or deeply held commitments, which the model detects during argument.[142]

[142] Ryan C. Black et al., *Emotions, Oral Arguments, and Supreme Court Decision Making*, 73 J. POL. 572 (2011); Bryce J. Dietrich, Ryan D. Enos & Maya Sen, *Emotional Arousal Predicts Voting on the U.S. Supreme Court*, 27 POL. ANALYSIS 237 (2019).

D. June Crunch Predictability

It is often thought that the Court releases the most "important and controversial" cases in June, late in the term, plausibly because they take the most time to negotiate and write opinions for.[143] Simpler, less controversial cases, on which there is broad consensus among the justices, on the other hand, may be issued earlier in the term.[144] If that conventional wisdom is right, does it also follow that these difficult end-of-term cases are the least predictable?

The end-of-term cases may be difficult because they combine high social or political salience with challenging legal questions. If so, they may be more difficult to predict. On the other hand, they may in fact be relatively simple legally, controversial only in the sense that they align with prominent ideological cleavages. If so, they may in fact be relatively easy to predict—votes and outcomes would likely run along standard ideological lines.

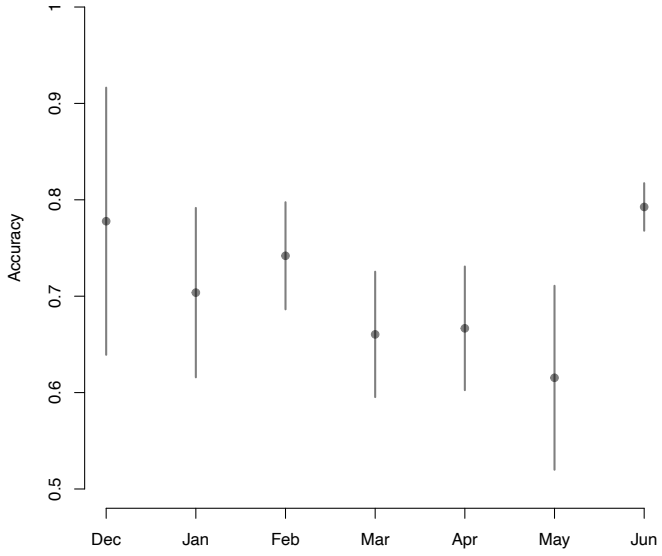

Figure 6: Vote-level Accuracy by Month of Decision

As shown in figure 6, vote level accuracies reveal a reasonably clear pattern: accuracy is declining for most of the term, from those cases decided in December until those decided in May; in June, however, accuracy improves markedly, reaching levels higher than in any other month.

[143] Lee Epstein, William M. Landes, and Richard A. Posner, *The Best for Last: The Timing of U.S. Supreme Court Decisions*, 64 DUKE L.J. 991, 1009 (2014).
[144] *Id*.

The end-of-term cases, this suggests, tend not to be the most complex ones likely to confuse the model—as a group, they instead represent the most predictable cases of the term. The more difficult cases—in the sense of complex and difficult to predict—tend to be decided in the months immediately prior. The June cases, instead, may represent the Court saving its most politically or ideologically controversial decisions for the end of the term. Though controversial, these cases tend to track strong regularities in the data, ideological or otherwise.

Though the cell size of course becomes smaller and therefore noisier at the case level, the fundamental pattern appears to be similar. The average case level accuracy in the three penultimate months of the term, March, April, and May, was 0.67—the model was correct in two out of three cases. In the final month of the term, the case level accuracy jumped to 0.9—the model was correct in nine out of ten cases. As a group, at the case level, the June cases tend to be the most predictable cases heard in the term.

V. Normative Implications and Institutional Directions

The results thus far focus on model performance and empirical lessons about legal institutions based on extensions of the model. In this Part, the analysis shifts to highlight normative implications, engaging several risks of accurate prediction for the legal system. Based on the results, it also identifies desirable criteria for low-intervention regulatory approaches that courts might use to manage predictive capacity and proposes several institutional strategies.

A. Normative Implications: Guidance and Over-guidance

That modern models produce both accurate and well-calibrated predictions suggests that they will be used by legal actors to forecast outcomes and adjust legal strategies. Indeed, this is how legal tech firms currently market their services.[145]

The small existing literature is largely celebratory regarding predictive capacity.[146] And it is, indeed, fair to expect a measure of benefit from increased predictive capacity. Primary benefits of prediction include better screening of low-probability claims and stronger ex ante compliance—accurate prediction, indeed, is a premise of classic theoretical perspectives in the strategy of litigation.[147] With respect to the screening of claims, that is, prediction may help align parties' beliefs about the likely outcome of a case, thereby improving the odds that they settle.[148]

---

[145] *See, e.g.,* Lex Machina, LexisNexis, https://www.lexisnexis.com/en-us/products/lex-machina.page (offering litigation analytics to help users "optimize case strategies, manage risks, predict outcomes and gain a competitive edge"); Pre/Dicta, https://www.pre-dicta.com/ (marketing tools to "[a]nticipate litigation outcomes and prepare clients for various scenarios" and to "[r]efine dispositive motion practice, tailoring strategy according to the likelihood of a favorable outcome"); Judge Analytics, Trellis, https://trellis.law/judge-analytics (advertising "actionable judicial analytics for the most strategic state trial court practitioners").

[146] ABDI AIDID & BENJAMIN ALARIE, THE LEGAL SINGULARITY: HOW ARTIFICIAL INTELLIGENCE CAN MAKE LAW RADICALLY BETTER (2023); Benjamin Alarie, *The Path of the Law: Towards Legal Singularity*, 66 U. TORONTO L.J. 443 (2016); Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909 (2013); Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 IND. L.J. 1401 (2017).

[147] Priest & Klein, *supra* note 13.

[148] *Id*.

It may also clarify for one side or the other that their side is a likely loser, again promoting settlement.[149] Increased settlement allows the parties to save time and resources. Institutionally, it also allows courts to focus their limited time and attention on the more challenging cases. The D.C. Circuit, indeed, sought to enhance parties' predictive capacity by releasing panel composition information earlier in the litigation cycle, as they thought this would reduce caseload burden.[150]

Ex ante compliance is another possible benefit of stronger predictive capacity. Part of the reason that people do not comply with the law—and thereby generate legal claims—is that they do not know their legal obligations.[151] Understanding, or at least appreciating, legal obligations is a necessary (but not sufficient) condition for complying with them. For example, a judicial decision holding that parties face tort liability for a given fact pattern works as a policy only because non-parties to the case anticipate liability and take precautions to avoid it.[152] They cannot anticipate and internalize the policy without knowing it. The same logic would apply to regulatory compliance.[153] If predictions can distill complex legal signals into a more reliable and communicable form, its ex-ante effects on primary behavior likely become stronger.

Modern prediction engines may be able to inform people of their legal rights and obligations,[154] thereby allowing parties to internalize the relevant legal rules and deterring them from conduct that would give rise to litigation. This guidance function can be seen as in-hand with an access to justice motivation, whereby legal disputes do not arise, or resolve without lawyers, because the law is plain to citizens via accurate prediction.[155] The promise of these mechanisms is a more law-bound society and a relatively less burdened civil and criminal legal system, with cases disproportionately reflecting instances of genuine legal indeterminacy.

Those benefits must be weighed against several possible downsides of prediction.

---

[149] *See* Anthony Niblett & Albert Yoon, *AI and the Nature of Disagreement*, 382 PHIL. TRANS. A MATH PHYS. ENG. SCI. (2024); Benjamin Minhao Chen & Anthony Niblett, *Bargaining in the Shadow of the Algorithm*, Typescript (Aug. 14, 2025) (showing experimentally that forecasts of legal outcomes narrow belief divergence).

[150] Samuel P. Jordan, *Early Panel Announcement, Settlement, and Adjudication*, BYU L. REV. 55, 58 (2007) (quoting Judge Edwards as saying that releasing panel composition "might lead some parties to settle their claims to avoid certain panels. We were happy to accommodate those who might thus settle and thereby reduce our caseload," even as he contested the view that panel composition, in fact, bore a relationship to case outcomes).

[151] *See, e.g.*, Richard Craswell & John E. Calfee, *Deterrence and Uncertain Legal Standards*, 2 J.L. ECON. & ORG. 279 (1986) (showing that uncertainty about the content of legal standards affects compliance); Louis Kaplow, *A Model of the Optimal Complexity of Legal Rules*, 11 J.L. ECON. & ORG. 150 (1995) (show that legal complexity can depress compliance); Isaac Ehrlich & Richard A. Posner, *An Economic Analysis of Legal Rulemaking*, 3 J. LEGAL. STUD. 257 (1974) (analyzing the information costs that legal rules impose on regulated parties).

[152] *E.g.,* GUIDO CALABRESI, THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS (1970).

[153] *E.g.,* Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715 (2015).

[154] This is a common theme in the more affirming parts of the literature on predictive technology. E.g., Casey & Niblett, *supra* note 146 (arguing that prediction enables context-specific, personalized "microdirectives" regarding legal obligations that allows escape from both rules and standards); ABDI AIDID & BENJAMIN ALARIE, *supra* note 146, at 200 (pointing to "a deepened understanding of real-time legal rights and obligations" emerging from predictive technology); Omri Ben-Shahar & Ariel Porat, *Personalizing Negligence Law*, 91 N.Y.U. L. REV. 627 (2016).

[155] Rebecca L. Sandefur, *Access to What?*, 148 DAEDALUS 49 (2019) (framing access to justice around substantive resolution, rather than access to legal services).

A first theoretical downside is that, if relied on too strongly, parties may comply with the law as it is rather than as it ought to be.[156] Predictive models, that is, must be trained on historical data. There is a natural risk therefore that they will predict what history indicates the law to be. Especially in times of rapid social, moral, and technological change, however, it is possible that the ground shifts under the feet of the historical data. If parties rely on historically based predictions when making litigation choices, they threaten the ability of the legal system to innovate relative to the new social, moral, or technological environment.[157] For the system to innovate, it is important that parties litigate cases they may believe will be unlikely to succeed under the current law so that courts can reconsider historical positions. The theoretical literature on legal evolution adopts different mechanisms to explain why law innovates, but common to every theory is a need for a flow of cases that provide opportunity to update old law.[158]

To take one historical example, Thurgood Marshall doubted that a majority existed to overrule *Plessy v. Ferguson*.[159] And though the Court had more assiduously required equality in separate-but-equal schemes in the years leading up to *Brown*, they also refused to revisit the fundamental question of separate-but-equal. It is plausible—even likely—that a predictive model would maintain that they would lose *Brown*. Yet that model, of course, would be wrong. It would miss the rapid changes in social understanding and the shifting ground under the doctrine, and the new openness of the Court to reconsider positions it historically maintained. Over reliance threatens to reduce the opportunities for law to innovate—for law to evolve, parties must litigate cases that models predict likely to be losers.[160]

Of course, *Brown* proceeded against concerns from Marshall and others that the Court was not ready to overturn *Plessy*. The litigants thought that, despite questions about viability, the case was worth pursuing. This clarifies that the live concern is that over-reliance on prediction will put pressure on the flow of (potentially) innovative cases, not eliminate it, but thereby harming the ability of the legal system to adapt to changes in the environment. The live concern, that is, is not the cessation of learning, but instead the slowing of learning. This is, in fact, almost a

---

[156] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242 (2019) (fearing a "paradigm of adjudication that favors standardization above discretion").

[157] For related points, *see* David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. 1001 (2021) (worrying that predictions "will progressively drain the system of its flexibility, its adaptive capacity, and its dialogic core"); Frank Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68 U. TORONTO L.J. 63 (2018).

[158] Posner, *supra* note 24 (arguing judges drive efficiency by applying cost benefit analysis); Rubin, *supra* note 24 (contending that inefficient rules disproportionally produce litigation); Priest, *supra* note 24 (similar to Rubin); Gennaioli & Shleifer, *supra* note 24 (modeling legal innovation by allowing judges to distinguish fact patterns through new cases).

[159] Cass R. Sunstein, *Did Brown Matter?* THE NEW YORKER (Apr. 26, 2004) (noting skepticism of many the then-sitting justices about overturning Plessy, often based on stare decisis); Michael Klarman, *Brown v. Board of Education: Law or Politics?* Typescript (Dec. 2002) (observing that at the time of Brown, "Justices were unenthusiastic about confronting so quickly the question they had deliberatively evaded in the 1950 segregation cases … Even Thurgood Marshall had doubts whether the justices were yet prepared to invalidate school segregation").

[160] See GUIDO CALABRESI, A COMMON LAW FOR THE AGE OF STATUTES 125 (1982) (arguing that statutory precedents become stale "misfits" with conditions). This point connects with a broader concern about the distributional consequences of algorithmic predictive capacity. E.g., Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633 (2017). *Cf.* Cass R. Sunstein, *On Analogical Reasoning Commentary*, 106 HARV. L. REV. 741 (1992).

necessary feature of increased settlement, generally understood to be a benefit of accurate prediction. The point just is that, whatever the benefits, increased settlement implies a tradeoff of reduced learning in the legal system, particularly in the limit.

This reduced learning, moreover, need not focus on landmark or would-be landmark cases. Reducing the flow of more incremental questions similarly harms system learning. Most learning in fact likely arises through incremental cases in which courts distinguish, rather than overrule, earlier decisions. Such distinguishing decisions would not necessarily garner attention from many outside that litigation space—no front-page *New York Times* stories—yet all the same, over time, in aggregate, they likely represent the primary channel through which the legal system learns and adapts to new circumstances.

A second, related downside of strong predictive models again rests on over-reliance: that turning to settlement instead of litigation reduces the opportunities for courts to articulate the contours of law through judicial opinions. Law is in substantial measure a justificatory enterprise, wherein law is both constrained and developed through the norm that courts explain their reasoning when they issue decisions.[161] Observers have long worried that arbitration or settlement will deplete the legal system of these public justifications, over time degrading the quality of the precedential system and legal system as a whole.[162] Over reliance on prediction fits in the same family of concerns.

As one of the more recent entries in the literature on the "disappearing trial" put it,[163] the shift from trials to arbitration and settlement threatened "values of democratic participation, legitimacy," among other values.[164] The acute concern with prediction, though, is that—either through personalized law, settlement, or litigation engineering—it will diminish our precedential system. As one scholar put it bluntly, "a trial is a prerequisite to precedent, and precedent is a cornerstone of our common law system."[165] And there is indeed empirical evidence of this degradation in some domains. Issacharoff and Marotta-Wurgler show that adjudications involving electronic consumer contracts declined in state courts, hobbling the "elaboration" of

---

[161] EDWARD H. STIGLITZ, THE REASONING STATE (2022) (arguing that the distinctive feature of legal and administrative systems is their ability to generate credible public reasons, and showing experimentally that public reasons both constrain and legitimize conduct); Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633 (1995) (arguing that reasons constrain conduct by committing the reason-giver to adhere to those reasons in similar future circumstances); FULLER, *supra* note 21, at 35 (illustrating a failed legal system that renders dispositions without reasons); Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797 (2021); Micah Schwartzman, *Judicial Sincerity*, 94 VA. L. REV. 987, 1004 (2008) (arguing that "the principle of legal justification is based on the idea that legal and political authorities act legitimately only if they have reasons that those subject to them can, in principle, understand and accept"); Hildebrandt, supra note 28, at 29 (observing that "[t]hough it feeds on legal argumentation and decision making, [legal prediction] does not add its own arguments, merely providing numerical re-articulation and mathematical functions to connect the dots).

[162] Owen M. Fiss, *Against Settlement*, 93 YALE L.J. 1073 (1983); J. Maria Glover, *Mass Arbitration*, 74 STAN. L. REV. 1283 (2022); Leandra Lederman, *Precedent Lost: Why Encourage Settlement, and Why Permit Non-Party Involvement in Settlements*, 75 NOTRE DAME L. REV. 221 (1999).

[163] Marc Galanter, *The Vanishing Trial: An Examination of Trials and Related Matters in Federal and State Courts*, 1 J. EMPIRICAL LEGAL STUD. 459 (2004); Robert Burns, THE DEATH OF THE AMERICAN TRIAL (2009); Shari Diamond & Jessica Salerno, *Reasons for the Disappearing Jury Trial: Perspectives from Attorneys and Judges*, 81 LA. L. REV. 119 (2020).

[164] Glover, *supra* note 162, at 3055; *see also* ALEXANDRA D. LAHAV, IN PRAISE OF LITIGATION (2017).

[165] Lederman, *supra* note 162, at 221.

common law.[166] Another scholar went further, contending that the rise of arbitration would mean that for "entire categories of cases … common law doctrinal development will cease."[167] These empirical and normative takes are consistent with the classic theoretical literature, which sees precedent as a depreciating asset that become obsolescent and requires continual renewal.[168]

This, again, is the almost-necessary converse of the benefits that might flow from accurate prediction. Predictive technology promises individualized, context specific determinations of legal obligations that enhance ex ante compliance and encourage settlement. If so, the legal system will have fewer opportunities to innovate, as above, but also fewer opportunities to accrete towards a fully textured system of public obligations.[169]

Over time, moreover, this coarsened legal articulation may degrade the quality of the data used to train the predictive models and thereby model performance. Strategic behavior induced by accurate predictions, that is, may erode the foundations of the model. It is possible to imagine an equilibrium in which model performance degrades to such an extent that parties make poor judgments about judicial outcomes, and the legal system—and therefore model—see more cases useful to legal articulation. The theoretical direction of these processes, therefore, may not be to a floor of meaningless; but neither is it assured to be healthy.

A third form of over-reliance focuses on judges rather than the parties.[170] As Justice Roberts recently worried, modern prediction engines may come to influence judicial decisions themselves—going against a confident algorithmic prediction of who wins, he thought, might take "a great deal of courage."[171] Thus, under Justice Roberts' concern even if parties overcome the incentives to settle, and they bring a case a model predicts they will lose, *judges* may not be able to overcome the model's prediction.[172] Particularly as models must be trained on historical data, this aspect of over-reliance further risks stasis in the legal system, with doctrine failing to meet the

---

[166] Samuel Issacharoff & Florencia Marotta-Wurgler, *The Hollowed out Common Law*, 67 UCLA L. REV. 600 (2020).
[167] Myriam Gilles, *The Day Doctrine Died: Private Arbitration and the End of Law*, 2016 U. ILL. L. REV. 371, 377 (2016).
[168] William M. Landes & Richard A. Posner, *Legal Precedent: A Theoretical and Empirical Analysis*, 19 J. L. ECON. 249 (1976).
[169] Ezra Friedman & Abraham Wickelgren, *Chilling, Settlement, and the Accuracy of the Legal Process*, 26 J.L. ECON. & ORG. 144 (2010).
[170] A normative margin that this Article does not engage is the prospect of judges using AI, of AI-based judges, or the legal or normative implications of authorities using predictive technology. For consideration of that topic, see, e.g., Allison Koenecke et al., *Tasks and Roles in Legal AI: Data Curation, Annotation, and Verification*, arxiv: 2504:01349 (2025); Anthony J. Casey & Anthony Niblett, *Problems with Probability*, 73 U. TORONTO L.J. 1 (2023); Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237 (2018); Anthony J. Casey & Anthony Niblett, *Will Robot Judges Change Litigation and Settlement Outcomes?*, MIT COMPUTATIONAL LAW REPORT (2020); Benjamin Alarie, Anthony Niblett & Albert Yoon, Regulation by Machine, Typescript (Dec. 2, 2016); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014).
[171] John G. Roberts, Jr., Chief Justice, A Conversation with John G. Roberts, Jr., Chief Justice of the United States, Address at the Baker Institute for Public Policy, Rice University (Mar. 17, 2026), https://www.msn.com/en-us/money/news/chief-justice-john-roberts-warns-ai-will-make-it-really-tough-for-young-lawyers/ar-AA1YYNFH.
[172] For classic related social science work, see Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124 (1974); *see also* NUNO GAROUPA & TOM GINSBURG, JUDICIAL REPUTATION: A COMPARATIVE THEORY (2015) (theorizing how judges respond to reputational incentives shaped by the audiences that observe their work).

requirements of the day's social, moral, and economic realities—though now due to over-dependence and social incentives around judicial behavior.

In sum, predictive technology is likely to produce both upsides and downsides. Legal tech companies and the small literature on this point observes many of the upsides—more settlement and stronger ex ante compliance. However, predictive technology promises three downsides largely unaccounted for in the literature—reduced learning in the legal system, poorer articulation of legal rules, and judicial over-reliance.

B. Institutional Directions: Managing Predictive Capacity

Given the tradeoffs discussed earlier, it is unlikely that humans will be best off with total predictive capacity or the absence of predictive capacity: the former leads to reduced innovation and a degraded legal system; the latter leads to reduced ex ante compliance, over-use of litigation, and generally a legal system that under-guides citizen conduct. The solution to the optimal level of predictive capacity is in the interior. The question then is how to best regulate predictive technology.

The models used to generate predictions depend on data from human institutions, and that critical resource can be calibrated to human needs.[173] The information-as-regulation approach is attractive because it represents a light-touch strategy that is within the control of the judiciary. Some parts of this strategy, indeed, can be implemented administratively, with little change required on the part of individual judges. Others require changes in judicial norms and culture. A critical consideration throughout the discussion is to remain within the core norms, values and practices of existing judicial institutions. For instance, unsigned judicial decisions with no reasoning—though now common in the emergency docket[174]—provide little information and would limit predictive capacity. Yet this practice sits outside of judicial norms.[175] Safe-guarding judicial norms protects against the possibility that steps taken to regulate predictive capacity themselves exert a negative force on rule of law values.

1. Panel Disclosure Timing

For models to generate reliable predictions, the same features used to train a model must be available at the time of inference. The same is true for the Holmesian human predictor. That basic insight motivated the D.C. Circuit to experiment with changing the timing of when they released the identities of judges empaneled for a case.[176] The D.C. Circuit sought to increase predictive capacity, hoping to induce parties to settle, so they allowed access to panel composition earlier than had been true, but this lever of calibration could be adjusted to any time between filing and the day of oral argument.

---

[173] Zachary Clopton & Aziz Huq, *The Necessary and Proper Stewardship of Judicial Data*, 76 STAN. L. REV. 893 (2024).
[174] *See* VLADECK, *infra* note 184.
[175] *Id.*
[176] *See* Jordan, *supra* note 150.

The identity of the judge or panel deciding a case is, of course, a salient predictor of case outcomes.[177] It is also the main piece of information that the courts control post-filing and that can be withheld from or released to litigants, consistent with judicial norms. By virtue of their own knowledge or the natural course of litigation, they will know most features relevant to prediction—but they will not know the identities of the deciding judges until the date of argument, unless told them earlier by the court. The circuits, in fact, adopt a wide range of disclosure policies. The D.C. Circuit, as mentioned, informs parties of panels very early during litigation, generally about two months before argument.[178] The Eighth Circuit releases panel composition "approximately one month" before argument.[179] Many other circuits disclose panel composition one to two weeks in advance of argument.[180] Other circuits disclose judges' identities only the morning of argument.[181] Manipulation of this timing, therefore, is well within normal judicial parameters and likely substantially shapes predictive capacity.

A virtue of this form of information regulation, moreover, is that it can be done administratively, without modification of prevailing judicial culture, norms, or work practices. Related administrative tools of prediction calibration would include making docket information freely machine-accessible so that parties could train models more easily.

2. Authorship

Earlier or later panel disclosure modulates model performance at inference time. Separately, it is also possible to modulate model performance by changing the information available to train a model. A key lever, again, relates to information regarding the identities of the decision-makers—through authorship rather than empanelment.

Courts adopt a wide range of practices with respect to authorship, from seriatim opinions to a norm of unsigned or per curiam opinions. The more the model can learn about the individual judges' preferences and personalities through their writing and rationales, the more accurate the model will become. Seriatim opinions, therefore, in principle reveal the most information about future judicial behavior and enhance predictive capacity; unsigned opinions reveal the least and diminish predictive capacity. Seriatim opinions were the practice in the early Supreme Court.[182] The same was true of the House of Lords and is true of the Supreme Court of the United Kingdom,[183] as well as other commonwealth jurisdictions. The current United States Supreme

[177] SEGAL & SPAETH, *supra* note 35; Matthew Spitzer & Eric Talley, *Left, Right, and Center: Strategic Information Acquisition and Diversity in Judicial Panels*, 29 J.L. ECON. & ORG. 638 (2013).
[178] LAUREL HOOPER, DEAN MILETICH & ANGELIA LEVY, CASE MANAGEMENT PROCEDURES IN THE FEDERAL COURTS OF APPEALS, FEDERAL JUDICIAL CENTER 56 (2011).
[179] UNITED STATES COURT OF APPEALS FOR THE EIGHTH CIRCUIT, INTERNAL OPERATING PROCEDURES 6 (Feb. 26, 2024).
[180] HOOPER, *supra* note 178, at 68 (First Circuit); *id.* at 80 (Second Circuit); *id.* at 91 (Third Circuit, "no later than 10 days" before argument); *id.* at 115 (Fifth Circuit); *id.* at 127 (Sixth Circuit); *id.* at 175 (Ninth Circuit); *id.* at 194 (Tenth Circuit); *id.* at 209 (Eleventh Circuit).
[181] *Id.* at 104 (Fourth Circuit); *id.* at 142 (Seventh Circuit); *id.* at 217 (Federal Circuit).
[182] WHITE, *supra* note 30.
[183] M. Todd Henderson, *From Seriatim to Consensus and Back Again: A Theory of Dissent*, 1 SUP. CT. REV. 283 (2007) (noting an "almost thousand year history" of English courts writing seriatim opinions).

Court, on the other hand, commonly issues unsigned or per curiam decisions.[184] Though the practice of unsigned opinions is criticized, it is important not to confuse concerns about attribution with concerns about the quality of explanation. In recent practice, many emergency docket decisions tend to be both unsigned and poorly explained, but those two features need not necessarily be joined.

A practice of per curiam decisions masks useful information that could be used to train predictive models, thereby diminishing predictive capacity. This form of model throttling requires deeper changes in judicial norms and practices than changing the timing of panel disclosure. It also impinges more deeply into the decision-making process itself, thereby plausibly changing not only predictive capacity but also the outcome of judicial decisions. Part of the reason that some courts shifted to unsigned opinions, indeed, is precisely to change the nature and perception of decisions—to make the court speak more "institutionally" rather than individually.[185] It is important to recognize the possibility of such substantive tradeoffs. From the perspective of model throttling, the main advantage of the authorship lever is that it can be applied even when the identity of the judges is fixed. The United States Supreme Court, for instance, does not decide in panels—for most matters, it decides cases as a full court—so litigants know who will decide the case well in advance.

3. Decision Structure

An even deeper—and therefore more freighted—intervention that affects model capacity is the structure of the decision itself. Rather than simply changing the attribution of authorship, many courts regulate the ability of members to dissent. In other jurisdictions, courts suppress dissents, on the theory that doing so forces members to decide as an institution and for their decisions to be regarded as such by the public.[186] This may enhance legitimacy of judicial decisions.[187]

Unlike the United States and most common law areas, in Europe many courts from the civil law tradition do not publish dissents. The Court of Justice of the European Union suppresses dissents, publishing only an unsigned majority opinion.[188] The French Conseil d'Etat, France's chief administrative court, likewise does not allow for dissents.[189] The German Federal Constitutional Court, likewise, did not publish dissents until 1971.[190] Other jurisdictions in this group include Austria, Belgium, and Italian high courts.[191] Though many of these jurisdictions

---

[184] STEPHEN VLADECK, THE SHADOW DOCKET: HOW THE SUPREME COURT USES STEALTH RULINGS TO AMASS POWER AND UNDERMINE THE REPUBLIC (2023) (noting that emergency docket decisions tend to be unsigned).
[185] Karl M. ZoBell, *Division of Opinion in the Supreme Court: A History of Judicial Disintegration*, 44 CORNELL L. Q. 186, 193 (1959) (noting Justice Marshall's innovation of the "opinion of the Court," and observing that, "[f]or the first time, the Court as a judicial unit had been committed to an opinion—a ratio decidendi—in support of its judgments").
[186] MITCHEL LASSER, JUDICIAL DELIBERATIONS: A COMPARATIVE ANALYSIS OF TRANSPARENCY AND LEGITIMACY (2004); Koen Lenaerts, *Rule of Law and the Coherence of the Judicial System of the European Union*, 44 COMMON MARKET L. REV. 1625 (2007).
[187] Lasser, *supra* note 186.
[188] MARSHA C. ERB, DISSENT AVERSION AT THE COURT OF JUSTICE OF THE EUROPEAN UNION (2014).
[189] JOHN BELL & FRANCOIS LICHERE, CONTEMPORARY FRENCH ADMINISTRATIVE LAW 72 (2022).
[190] Katalin Kelemen, *Dissenting Opinions in Constitutional Courts*, 14 GERMAN L. J. 1345 (2013).
[191] *Id.*

wrestle with whether to allow dissents, it is well within the scope of judicial norms to make public only a majority, unsigned opinion. Such a practice removes perhaps the single most important predictors from model training—individualized vote and opinion data that allows construction of abstract representations of judges' preferences and personalities. For jurisdictions that currently allow dissents, such as the United States, this control over the decision structure would of course constitute a dramatic step.

VI. Conclusion

This Article has introduced the Experience Model, a fine-tuned transformer trained on fifteen years of unstructured Supreme Court data, and tested it against the most recent completed term. The model achieves 74 percent vote-level accuracy and 82 percent case-level accuracy in October Term 2024. For the test term, this performance exceeds algorithmic benchmarks and, at the case level, human crowds. Beyond classification, the model produces well-calibrated probability estimates, allowing it to serve not merely as a forecasting tool but as a research instrument for studying the Court.

Four empirical lessons emerge from the model's predictions. Algorithmic predictability, first, varies substantially across areas of law, offering an empirical methodology for assessing whether specific domains satisfy the guidance function that legal theorists from Hart to Fuller identify as central to the rule of law.[192] Criminal procedure proves the least predictable area—excluding it would boost case-level accuracy to 87 percent—suggesting a domain where outcomes may reflect judicial discretion more than the regularity of legal rules.[193] Regardless of whether that unpredictability reflects genuine doctrinal indeterminacy or the distinctive litigation incentives of criminal defendants, the normative implication is similar: this is a high-stakes area where citizens face the loss of liberty, and unpredictability is most costly.

The model's calibrated ex ante win probabilities, second, provide a novel test of the Priest-Klein hypothesis using the theoretically relevant but previously unobservable quantity: the ex ante probability of victory, i.e., the probability plausibly experienced by litigants at the time of their settlement decisions.[194] The mass of predicted reversal probabilities concentrates near fifty percent, with few sure-losers on either side. The distribution is shifted rightward, consistent with the certiorari process adding a reverse-bias on the classic selection mechanism.[195] With more data, this methodology could be extended to compare litigated and settled cases across the lower courts.

An ablation study, third, reveals that oral argument substantially improves predictive accuracy, but does so heterogeneously across justices. Argument data boosts accuracy for Justices Jackson and Sotomayor by more than ten percentage points while producing negligible effects for Justices Alito, Thomas, and Barrett. This is, to my knowledge, the first study to

[192] Part IV.A.
[193] Part III.B; Part IV.A.
[194] Part IV.B.
[195] *Id*.

measure the marginal predictive contribution of oral argument at the individual justice level. Prior work examines either the marginal contribution at the case level or justice-level predictions from argument data alone, but not both.

Finally, cases decided in the so-called June Crunch—often characterized as the term's most important and complex—turn out to be the easiest to predict. The model correctly predicts 90 percent of June cases. End-of-term cases, though controversial, tend to divide along well-established ideological and other voting lines rather than presenting genuinely difficult legal questions that generate unusual coalitions.

These results are from a single test term, of course, and performance is not assured in the future. Changes in Court composition, shifts in the political valence of legal issues, or evolving collegial dynamics could erode the patterns the model has learned. Indeed, the model is currently deployed for live prediction of the Court's current term, as an on-going test of its capability.[196] Extension to the federal circuits, where larger dockets provide greater statistical power and courts are more constrained by precedent and institutional hierarchy, represents a natural and promising next step in this predictive agenda.

That promise of growing capability, however, only sharpens the normative tradeoffs implied by predictive technology. Some capacity enhances law's guidance function, promotes ex ante compliance, and supports efficient resolution of disputes.[197] Most observers would see those as normative benefits. But over-reliance risks suppressing the flow of innovative cases that allow doctrine to adapt, depleting the system of the public justifications through which law accretes, and anchoring judicial reasoning to historical patterns.[198] The optimal level of predictive capacity is in the interior—a complete lack of predictive capacity is likely degenerate; so is exceptionally predictive capacity.

Legal institutions can regulate predictive capacity by managing the flow of data needed to train and conduct inference on models. Panel disclosure timing, authorship norms, and the structure of decisions, including whether to publish dissents, modulate the information available for model training and inference. These tools operate within existing judicial norms and can be calibrated over time to the needs of a healthy legal system. The challenge and opportunity is to manage the conditions under which it operates, titrating information to preserve both the guidance that accurate forecasting enables and the productive uncertainty on which the legal system relies.

[196] *See* reasonablemachines.io.
[197] Part V.A.
[198] *Id*.

VII. Appendix

Model Architecture

The Experience Model is a transformer-based classifier fine-tuned to predict individual justices' votes from the materials available prior to decision. The base of the prediction system reported in this Article is a Llama 3.1 70b. The model is fine-tuned using low-rank adaptation (LoRA) applied to all linear layers, enabling efficient training on a single GPU. During fine-tuning, the model develops high-dimensional internal representations that summarize doctrinal, institutional, and personality characteristics that correlate with justices' eventual votes. The model processes each justice-case pair independently.

Data and Feature Construction

The dataset consists of merits docket cases from October Terms 2010 through 2024, totaling approximately one thousand decisions and nine thousand individual votes. For each case, signals are extracted from the lower court decision, party and amicus briefs, and oral argument transcripts. These raw, unstructured documents are processed into a structured representation that includes: (1) features capturing the legal and doctrinal stakes, such as the connection between a doctrinal choice and institutional values noted in the legal process literature; (2) oral argument features, including the frequency and direction of each justice's questioning toward each party and the sequence of interruptions; and (3) case-level metadata, such as lower court identity, issue area, and procedural posture, which earlier generations of prediction engines likewise include. The features are assembled into a structured JSON file for model input.

Training Procedure

Training proceeds by minimizing cross-entropy loss over justice votes.[199] The training set consists solely of cases from October Term (OT) 2010 through OT 2023; OT2024 is withheld, with no model exposure to that term. This temporal cordoning departs from standard k-fold cross-validation because cases within a term share common shocks—such as the posture of the executive toward the Court, public mood, or collegial dynamics among the justices—that would leak term-specific information into training.

The training process follows standard fine-tuning best practices.[200] The challenge is to make the model learn aggressively enough to extract generalizable patterns from the training data, without overfitting—that is, without the model "memorizing" the training data rather than

[199] Fine-tuning required three to four hours on a single H100 GPU.

[200] Fine-tuning follows the LoRA framework of Hu et al., adapted for the Llama 3.1 series following Meta's recommendations. Edward J. Hu et al., LoRA: *Low-Rank Adaptation of Large Language Models*, 10 INT'L CONF. ON LEARNING REPRESENTATIONS (2022); Aaron Grattafiori et al., *The Llama 3 Herd of Models* (July 31, 2024) (unpublished manuscript), https://arxiv.org/abs/2407.21783.

learning its generalizable features.[201] In practice, model performance appears more sensitive to the quality of the training data than to modest changes in training parameters.[202]

The Experience Model employs Monte Carlo (MC) dropout to approximate an ensemble at inference. Dropout remains active during inference, and multiple stochastic forward passes are performed for each justice–case pair. The predicted probability for a given justice-case pair is the mean of the forward-pass outputs, which can be interpreted as an approximate posterior predictive mean under the variational framework of Gal and Ghahramani.[203]

Vote Aggregation via Monte Carlo Simulation

Individual vote-level probabilities are aggregated to case-level outcome probabilities using Monte Carlo simulation.[204] Let p_i denote the calibrated probability that justice i votes to reverse. In each trial t = 1, …, T, each justice's vote is drawn as a Bernoulli random variable with parameter p_i. This creates nine votes for that trial, and a case level outcome can be assessed on that basis. The case-level probability of reversal is the fraction of the T trials that result in reversal; I set T=10,000. The main parameter choice in the aggregation step concerns the degree of dependence in voting behavior. If, say, two justices each vote to reverse with probability 0.50, do they tend to vote together or independently? Dependence is modeled using a one-factor Gaussian copula, calibrated on validation data, which allows a common shock to propagate through the votes of the justices.[205]

Evaluation Metrics

Consider first classification metrics. Accuracy is the fraction of correctly predicted votes (or cases). Because the reverse class is more prevalent (approximately 65 percent of votes), accuracy alone can be misleading (a naive always-reverse classifier achieves 65 percent accuracy). We therefore also report the F1 score, defined as the (harmonic) mean of precision and recall. Specifically, Precision = TP / (TP + FP), Recall = TP / (TP + FN), F1 = 2 × (Precision × Recall) / (Precision + Recall), where TP, FP, and FN denote true positives, false positives, and false negatives, respectively, with respect to each class. We report per-class F1 as well as the macro F1 (unweighted average across classes) and weighted macro F1 (weighted by class frequency).

Now consider calibration metrics. Calibration refers to the degree to which the model's confidence is aligned with reality. If a model says that a justice will vote "reverse" with 0.8 probability, does the justice in fact vote to reverse 80 percent of the time?

---

[201] Fine-tuning required about three hours on a single H100 GPU.

[202] E.g., in one early experience, I found that some of the training data was being malformed before training. Fixing that bug resulted in a two to four percentage point increase in accuracy, which is very substantial in this difficult task.

[203] Yarin Gal & Zoubin Ghahramani, *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, *in* INTERNATIONAL CONFERENCE ON MACHINE LEARNING 1050 (2016).

[204] One case is dropped from the results, Lab. Corp. of Am. Holdings v. Davis, 604 U.S. ___ (2025) (per curiam), because it was a one-line per curiam decision saying only that cert was improvidently granted.

[205] For an example of similar modeling, see David X. Li, On Default Correlation: A Copula Approach, Risk Metrics Group (1999), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=187289.

The Brier score measures the mean squared error between predicted probabilities and observed binary outcomes: $Brier = \frac{1}{N}\sum(p_i - y_i)^2$, where p_i is the predicted probability of reversal and $y_i \in \{0,1\}$ is the observed outcome. A lower score indicates better calibration, or alignment between predicted probabilities and outcomes. A naïve always-reverse model ($p_i = 1$ for all votes) would yield a Brier score of 0.35; the Experience Model achieves 0.17 at the vote level and 0.16 at the case level.

The expected calibration error (ECE) partitions predictions into M equal-width bins and computes the weighted average absolute deviation between each bin's mean predicted probability and observed frequency: $ECE = \sum \frac{n_m}{N}|\text{acc(m)} - \text{conf(m)}|$, where n_m is the number of observations in bin m, $acc(m)$ is the average accuracy in that bin, and $conf(m)$ is the average confidence in that bin. We use M = 10 bins at the vote level and M = 3 at the case level, where fewer observations make finer binning noisy. The ECE can be interpreted as approximately the average error in the predicted probability. The Experience Model achieves an ECE of 0.058 at the vote level and 0.09 at the case level.