Trachtenberg, Qian and Warnberg for Law360: How Well Can AI Predict Supreme Court Decisions?

By Mark Trachtenberg, Michael Qian, Chloe Warnberg | November 23, 2026

RELATED PRACTICES Appellate and Supreme Court, AI and Technology, Litigation

Haynes Boone Partners **Mark Trachtenberg** and **Michael Qian** and Associate **Chloe Warnberg** authored an article for *Law360* examining how accurately generative AI can predict U.S. Supreme Court decisions.

Drawing on an analysis of 57 cases from the Court's October 2025 term, they found that the AI tool correctly predicted the winning side in 81 percent of cases, but its misses were often unpredictable. The authors discuss how AI can help inform litigation strategy while emphasizing the importance of human judgment in evaluating case outcomes.

Read the full article below and on the *Law360* website **here**.

Lawyers predict how courts will rule every day. Those predictions inform litigation strategy, settlement decisions and client expectations. Generative artificial intelligence appears to be a powerful new tool to assist with those predictions.

But how good really is AI at analyzing and predicting case outcomes? We ran an experiment: We asked a legal AI tool to predict the outcome of every merits case from the last U.S. Supreme Court term, and then we checked its predictions against the actual outcomes.

The tool was good, but not as good as the best human forecasters, and its mistakes were unpredictable. It called the winning side in 81% of cases. That beats a coin flip and edges out some pre-AI statistical models. But it also whiffed on cases that experienced Supreme Court watchers found easy, could not reliably say which of its own predictions to trust, and over-weighted certain factors in inconsistent ways.

For practitioners tempted to lean on these tools as the current Supreme Court term ramps up, the lesson is to use them as an aid to human judgment, not a substitute.

How We Ran the Study

We analyzed all 57 cases argued and decided on the merits during the Supreme Court's October 2025 term.[1] For the analysis, we used Harvey, a legal AI platform, running the Claude Opus 4.6 model.

We used the same prompt and the same closed universe of materials for each case: all case filings, the oral argument transcript, and all Supreme Court opinions from the 2016 through 2024 terms (so that the model was informed about the justices' jurisprudence). We instructed the tool not to conduct outside research or rely on any other sources, which we verified by predicting selected cases both before and after the court's actual decision and observing no difference.

Our goal was to test the tool's ability to predict each case based on the information available on the

© 2026 Haynes and Boone, LLP

case docket before the court ruled.

The prompt asked for the following predictions, along with a confidence estimate: the prevailing party and disposition, the holding and rationale, and the vote breakdown.

For each case, we ran the prompt five separate times to test whether the tool's predictions were consistent, and in separate threads to ensure that each output was independent of previous runs.

What the Tool Got Right, and How Often

On the headline question — who wins — the tool was correct in 46 of 57 cases, or 81%. We considered a case correctly predicted when at least three of the five runs correctly predicted the prevailing party. We treat the remaining 11 cases, where the majority of runs were wrong, as misses.[2]

We also assessed the tool's consistency. We discovered the same prompt and the same record could yield a different predicted winner from one run to the next. The tool selected the correct prevailing party in all five runs in only 38 of the 57 cases, or 67%. In 33% of the cases, at least one of the five runs picked the wrong winner.

The tool also frequently failed to predict the right vote margin. In 43 of the 57 cases, or 75%, at least one of the five runs predicted the wrong split. The tool frequently predicted division among justices where there ultimately was none. Yet in every case that ended up closely divided in a 5-4 ruling, the tool predicted more lopsided majorities.

How That Compares to Other Forecasts

An 81% success rate is notable, but the real question is whether AI improves upon existing prediction methods. Measured against some pre-AI methods, the AI tool performed well.

A 2004 Supreme Court forecasting project reported that a statistical model correctly predicted 75% of the 2002 term's affirm-or-reverse results, while a panel of human predictors collectively got 59.1% right.[3] A 2017 article published in PLOS ONE reported that a machine learning model applied to every case from 1816 to 2015 predicted 70.2% of case outcomes and 71.9% of individual justice votes.[4] And the court reverses roughly 71% of the merits cases it hears each term, so a predictor that reflexively guessed reverse would already be right most of the time.[5]

The AI tool's 81% outcome accuracy sits modestly above those marks.

Against the best human prognosticators, however, the tool lagged. Crowdsourced predictions from FantasySCOTUS achieved 94.74% accuracy across the 2025 term's cases — well ahead of our tool.[6] This suggests that human judgment — at least when aggregated — outperforms more mechanical prediction methods.

The 11 Misses

One interesting aspect of the exercise was studying where the tool went wrong, particularly in the 11 cases in which it predicted the wrong bottom-line outcome. The misses do not fall into a clear pattern, which is itself the first lesson. It is hard to know ahead of time how much confidence to place in any given AI case prediction.

© 2026 Haynes and Boone, LLP

Indeed, the tool was generally poor at gauging its own reliability. Its stated confidence levels — typically between 60% and 85% — bore little relationship to whether its predictions were correct. Divergence among the five runs was also a weak warning sign: Split runs did correlate with a higher chance of error, but the cases in which the tool generated the same incorrect prediction in all five runs still accounted for roughly half of the misses.

In contrast, all of FantasySCOTUS' misses were in cases where human predictors were closely divided. The crowd revealed which calls were close; the AI tool did not.

Adding to the unpredictability of the tool's misses, many came in cases that the court itself evidently did not consider a close call. In six cases — Case v. Montana, Villarreal v. Texas, Urias-Orellana v. Bondi, Hain Celestial Group Inc. v. Palmquist, Enbridge Energy LP v. Nessel and Sripetch v. U.S. Securities and Exchange Commission — the tool predicted a divided court in nearly every run.

Yet, in reality, the court was unanimous — in the opposite direction from what the tool predicted. Of course, close legal questions can lurk behind unanimous rulings. But it is still striking that the AI tool was so often unable to predict outcomes on which all nine justices could agree.

One possible explanation for at least some of the misses, which has also been observed in at least one similar study,[7] is that AI leans too heavily on ideology — predicting outcomes from the ideological salience of the dispute and the court's perceived center of gravity rather than from the legal question actually presented. We saw this in our study, but only a few times: Just three of the misses were cases in which the tool incorrectly predicted 6-3 or 7-2 results with only Supreme Court Justices Sonia Sotomayor, Elena Kagan and/or Ketanji Brown Jackson in dissent.

Other times, the AI tool missed cases that did break down along those conventional lines. Consider Fernandez v. U.S., a decision holding that prisoners cannot obtain compassionate release based on the invalidity of their convictions. That case did result in a divided court, with a six-justice majority siding with the government, and Justices Sotomayor, Kagan and Jackson disagreeing with its holding. This was an easy call for humans; FantasySCOTUS reports that 100% of its participants called it correctly. Perhaps this was because the result fell along familiar lines for a case about collateral attacks on criminal convictions.

Yet the AI tool appeared to disregard ideology here, and predicted the reverse outcome across all five runs. When asked to identify the source of its error, the tool said it overrelied on the petitioner's framing, misread oral argument signals, and misjudged how the justices would weigh competing legal principles. Ideology, in other words, was an unreliable guide to when the AI tool would err.

A few other patterns emerged, though again none explains every miss. The tool sometimes appeared to overweigh what it described as sympathetic facts, including in Landor v. Louisiana Department of Corrections, where prison officials shaved a Rastafarian prisoner's hair despite binding precedent protecting his religious objection, and Urias-Orellana v. Bondi, where petitioners fled cartel violence in El Salvador.

The tool also showed a pro-challenger, anti-government tilt. In 10 of the 11 misses, a governmental or institutional party actually won, while the tool predicted the private challenger would prevail, and in 10 of the 11 the court affirmed, while the tool predicted reversal.

© 2026 Haynes and Boone, LLP

Lastly, the tool often misinterpreted oral argument signals — in both directions. Sometimes it treated pointed questions from individual justices as signals of their votes, when those questions actually reflected devil's advocacy or boundary-testing. Other times it missed what human observers considered to be clear signs from oral argument of a likely result.

In Case v. Montana, for example, SCOTUSblog correctly predicted after argument that the court "appeared ready to side with police officers," with the justices' comments signaling "fairly broad agreement" against the legal rule petitioner proposed.[8] Likewise in Fernandez v. U.S., favorable questions to petitioner were few and far between, making it easy for humans at SCOTUSblog to diagnose that a "majority of the justices appeared skeptical" of his position.[9] Yet the AI tool discounted those signals.

The AI tool's misses were therefore unpredictable. The reasons for those misses varied. And nothing in the output exposed which mistake the tool might be making in any given case. A user cannot easily see a miss coming — except by verifying with human judgment.

A Closer Look: Urias-Orellana v. Bondi

Urias-Orellana v. Bondi offers a helpful example of the gap between AI and human predictors. The court held that federal appellate courts must apply deferential, substantial evidence review to the Board of Immigration Appeals' determination as to whether undisputed facts amount to persecution under the asylum statute.

The outcome of the case was predictable to human observers — 87% of FantasySCOTUS participants got it right. The dispute had narrowed before argument, ultimately in the government's favor. The government's brief conceded de novo review of purely legal questions about the persecution standard, and petitioners did not raise that kind of challenge in this case.

At oral argument, justices across the ideological spectrum appeared skeptical of petitioners' position. Post-argument human predictions strongly favored the government.[10]

The AI tool got it wrong anyway. All five runs predicted a 6-3 or 7-2 win for the petitioners, each at roughly 70% confidence. AI disregarded oral argument signals: While it noted the skepticism of multiple justices at oral argument, it counted several of those justices in petitioners' favor anyway, and overlooked that most of the court was converging on the same point.

The tool also gave more weight to broad doctrinal trends and sympathetic facts, rather than to the arguments that justices across the ideological spectrum focused on during oral argument, and that ultimately were the basis for the unanimous ruling.

What This Means for Practice

A few implications deserve emphasis for lawyers and the courts.

First, run-to-run inconsistency is currently a feature of AI tools. If an advocate uses AI to advise a client on likely outcomes, or if a judge or clerk uses it to prepare an opinion, the outcome may depend on which run happens to surface. One should not assume a single output represents the model's consistent view, and even convergence across runs is not a reliable confidence signal.

Second, the cases the tool missed were disproportionately cases that sophisticated court watchers

© 2026 Haynes and Boone, LLP

predicted with confidence. Used as a supplement — to surface arguments, stress-test a theory of the case, or generate a first-draft analysis for further scrutiny — these tools can be genuinely additive. Used as an oracle, they can and will fail, sometimes in cases where the answer seemed clear to practitioners.

The bar should treat AI prediction as a supplemental tool whose output always needs careful scrutiny, not as a substitute for human judgment.

Disclosure: Qian and Warnberg served on the legal team representing the petitioner in Sripetch v. SEC. Qian also served on the legal team representing the petitioner in Villarreal v. Texas, and assisted in the filing of an amicus brief on behalf of Agudath Israel of America supporting the petitioner in Landor v. Louisiana Department of Corrections and Public Safety.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] We excluded Hamm v. Smith , which the Court dismissed as improvidently granted.

[2] The eleven misses were Case v. Montana , Urias-Orellana v. Bondi , Fernandez v. United States, T.M. v. Univ. of Maryland Med. Sys. Corp. , Landor v. Louisiana Dep't of Corrections and Public Safety , United States Postal Service v. Konan , Villarreal v. Texas, Rutherford v. United States , Sripetch v. SEC , Hain Celestial Group Inc. v. Palmquist, Enbridge Energy LP v. Nessel .

[3] https://www.jstor.org/stable/4099370.

[4] https://doi.org/10.1371/journal.pone.0174698.

[5] https://ballotpedia.org/SCOTUS_case_reversal_rates_%282007_-_Present%29.

[6] https://fantasyscotus.net/.

[7] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7096518.

[8] https://www.scotusblog.com/2025/10/court-hears-arguments-on-when-police-may-enter-a-home-without-a-warrant/.

[9] https://www.scotusblog.com/2025/11/justices-evaluate-limits-of-the-compassionate-release-statute/.

[10] See, e.g., Kelsey Dallas, Court debates asylum determinations, SCOTUSblog (Dec. 2, 2025), https://www.scotusblog.com/2025/12/court-debates-asylum-determinations/; Oral Arguments in Urias-Orellana v. Bondi Raise Key Questions about how Federal Courts Interpret Asylum Protections, Quarles (Dec. 11, 2025), https://www.quarles.com/newsroom/publications/oral-arguments-in-urias-orellana-v-bondi-raise-key-questions-about-how-federal-courts-interpret-asylum-protections.

© 2026 Haynes and Boone, LLP