LLMs as Supreme Court Super Forecasters?

Amit Haim[*] Tel Aviv University

Aniket Kesari Fordham University

Johannes Kruse Max Planck Institute Bonn

February 23, 2026

Abstract

Can large language models meaningfully anticipate judicial decision-making in high-stakes legal institutions? This article examines whether contemporary large language models (LLMs) can forecast outcomes of U.S. Supreme Court cases and what such performance reveals about the structure and predictability of judicial behavior. Using Supreme Court cases from the 2023–2024 and 2024–2025 Terms, we evaluate predictions generated by three leading proprietary LLMs—GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro—based solely on information available prior to decision. The models correctly predict between 76 and 82 percent of case outcomes, a level of accuracy comparable to established statistical models and crowdsourced forecasts. Prediction accuracy varies systematically across justices, with higher performance for ideologically consistent justices and lower accuracy for those occupying the Court's ideological center. When applied to cases decided before the models' knowledge cutoff dates, accuracy exceeds 90 percent, underscoring the importance of temporal separation in evaluating predictive claims. Taken together, these findings suggest that while LLMs can synthesize legal and institutional cues in ways that rival existing forecasting approaches, their performance reflects underlying regularities in Supreme Court decision-making rather than a fundamental advance in predictive power. The results contribute to broader debates about judicial behavior, institutional predictability, and the appropriate role of artificial intelligence in legal analysis.

[*] Corresponding author: amithaim@tauex.tau.ac.il

1 Introduction

In June 2025, *The Economist* published SCOTUSbot, an “AI tool to predict Supreme Court decisions”. This public debut of a legal prediction system built on large language models (LLMs) reflects a long-standing fascination with whether the decisions of the U.S. Supreme Court can be forecast with any reliability. From the pioneering *Supreme Court Forecasting Project* in 2004 Ruger et al. [2004] to more recent machine learning ensembles trained on centuries of case data [Katz et al., 2017], the Court has served as a proving ground for forecasting models of legal decision-making.

Interest in prediction has also spilled beyond academic literature. Crowdsourced platforms like *FantasySCOTUS* have harnessed the wisdom of the crowds to forecast case outcomes, and real-money markets have briefly traded on blockbuster cases such as *NFIB v. Sebelius*. Journalists regularly speculate on likely outcomes after oral arguments, parsing the tone and frequency of Justices’ questions.

The rapid rise of LLMs adds a new and largely untested dimension to this literature. Unlike prior statistical models, which relied on carefully coded features from the Supreme Court Database or simple metadata about cases, LLMs can ingest and synthesize the full text of briefs, amicus filings, and even oral arguments. This raises a natural question: can LLMs serve as “superforecasters” [Tetlock and Gardner, 2016] of Supreme Court decisions, combining the pattern-recognition power of machine learning with the interpretive capacities of human legal experts? Early experiments suggest that modern LLMs achieve accuracies in the 70–80% range when asked to predict judicial outcomes, comparable to the best models and crowdsourced forecasts, but their performance on genuinely out-of-sample cases remains underexplored. To address this question, we leverage the fact that models have a “knowledge cutoff” date. For example, GPT 4o’s original model cutoff date was October 2023, Claude Sonnet 3.5 in April 2024, and Gemini 1.5’s was November 2023. By using these older versions of the models that could not have seen the real-world decisions, we gain insight into how useful LLMs might be for forecasting future events.

This paper benchmarks the forecasting ability of three leading closed-source LLMs – OpenAI’s GPT-4o, Anthropic’s Claude 3.5 Sonnet, and Google’s Gemini 1.5 Pro – on all argued cases during the 2023–2024 and 2024-2025 Terms. Using a retrieval-augmented generation pipeline, we provide each model with the filings available to the Justices prior to decision and elicit simulated votes and draft syllabi. We then compare model outputs to authoritative ground-truth outcomes from the Supreme Court Database, as well as to baseline benchmarks from prior academic models and crowd-sourced forecasts. Our findings contribute to ongoing debates about the promise and limitations of AI in law: while LLMs can now perform on par with some of the best statistical and crowd-based methods, they do not necessarily exceed these methods.

2 Background

2.1 Quantitative Models of Judicial Prediction

A rich empirical literature has developed quantitative models to forecast U.S. Supreme Court case outcomes and individual justice votes. Early work in political science often modeled the Court's decisions as functions of case facts or ideological factors. A landmark effort was the *Supreme Court Forecasting Project*, which prospectively compared an algorithmic model to human experts for all cases in the 2002 Term [Ruger et al., 2004]. The Project's statistical model - a logistic regression using six general case features (e.g., issue area, circuit of origin, lower court ruling) - correctly predicted 75% of case outcomes (affirm/reverse) in that term, significantly outperforming the legal expert panel's 59.1% accuracy. Notably, the model achieved this accuracy despite disregarding case-specific doctrinal details, suggesting that broad contextual factors capture much of the Court's behavior. The model excelled in predicting the pivotal swing votes of moderate justices (e.g. Justices O'Connor and Kennedy), whereas experts did better on the most ideologically predictable justices. This result was a simple statistical model out-predicting seasoned Court watchers, which was a striking demonstration of the potential of quantitative methods in legal forecasting.

Building on such efforts, researchers have increasingly turned to modern machine learning (ML) techniques to improve prediction of Supreme Court decisions. Katz et al. [2017] developed a general ML model that predicts both case outcomes and each justice's vote. Using nearly two centuries of Supreme Court data, their ensemble model (based on a random forest classifier) achieved about 70.2% accuracy at the case-outcome level and 71.9% at the individual justice-vote level. Notably, performance was even higher for more recent periods (e.g. post-1900), reflecting improved data and perhaps more stable voting patterns. The model was built on the Supreme Court Database (SCDB) and integrated numerous features: case origin, issue area, lower-court outcomes, timing of decision, and historical voting tendencies of each justice. The authors report that a random forest (an ensemble of decision trees) outperformed other algorithms, including support vector machines and feed-forward neural networks. In other words, an ensemble approach leveraging the wisdom of crowds among many decision trees proved most effective, beating even neural network models in this context. This suggests that complex nonlinear interactions in judicial behavior are well-captured by ensemble methods. Indeed, the random forest model significantly outperformed various baselines and earlier models, providing the first generalized, out-of-sample predictor for Supreme Court decisions. It consistently beat models and even modestly improved on in-sample optimized benchmarks, indicating real predictive signal rather than overfitting. The model's overall accuracy (around 70–72%) approaches the upper bound of predictability observed in this domain and has been touted as evidence that computers can rival or surpass legal experts in forecasting Supreme Court outcomes.

Beyond traditional classifiers, scholars have explored creative quantitative approaches. Guimerà and Sales-Pardo [2011], for example, applied network analysis to voting patterns: treating justices' votes as a complex social network, they inferred "justice blocs." Their model could predict a particular justice's vote given the votes of other justices in the same case, uncovering hidden alignments. Remarkably, this method outperformed both legal experts and content-based algorithms, highlighting that inter-justice dynamics can be highly predictive. Other researchers have incorporated textual analysis, for instance using the text of oral arguments or briefs, to anticipate decisions. Black et al. [2012] showed that the party who faces more questions at oral argument is more likely to lose, and Jacobi and Rozema [2017] found that subtle vocal cues, such as the justices' tone or emotional arousal, correlate with eventual voting patterns. These findings suggest that machine learning on rich data (not just case metadata but also transcripts and audio) could further improve predictions by capturing doctrinal and contextual cues that static databases might miss. Overall, the quantitative literature demonstrates that Supreme Court decisions are far from random: from simple regressions to complex ML ensembles, numerous models have achieved approximately 60–75% accuracy in predicting case outcomes, substantially better than chance. This is compelling evidence that despite the legal complexities and occasional surprises, the Court's votes contain discernible patterns-whether rooted in ideology, issue context, or interpersonal dynamics-that can be learned by algorithms.

2.1.1 The Supreme Court Forecasting Project and Its Legacy

The *Supreme Court Forecasting Project* is a foundational study in this field [Ruger et al., 2004]. In this project, political scientists and legal academics collaborated to pit a data-driven model, developed by Andrew Martin and Kevin Quinn, against expert intuition. The forecasting model used was relatively sparse - relying on just a half-dozen case attributes (circuit of origin, issue area, type of petitioner, etc.); yet it captured broad predictors identified by social science theories of Court behavior. On the other hand, the Project gathered predictions from 83 legal experts (professors, practitioners, and Court watchers), each focusing on cases in their area of doctrinal expertise. The stark result was that the algorithm beat the experts overall, correctly forecasting 75% of the 2002 Term's case outcomes versus 59% for the human predictors. This difference was statistically significant and challenged the conventional wisdom that seasoned lawyers, steeped in case facts and oral argument nuances, would outperform a formulaic approach. The authors noted that the model's edge came largely from correctly calling outcomes in ideologically mixed or centrist-influenced cases, whereas many experts struggled to predict swings by Justices O'Connor and Kennedy. By contrast, experts did well on easy cases (e.g. unanimous outcomes or ideologically extreme justices). These findings suggested that experts sometimes "read too much" into nuanced case details or misleading cues, whereas a blunt algorithm focusing on historical patterns could more reliably predict the decisive center of the Court. The Project's success was widely noted, and it has been cited as early evidence that statistical models can capture the reckonability of judicial behavior better than many experts expected.

The Supreme Court Forecasting Project also spurred debate and further research. Its authors acknowledged that purely formal features (like issue area or petitioner type) are proxies that omit rich legal reasoning, yet the strong performance hinted that much of Supreme Court decision-making might follow broader ideological or institutional regularities. Subsequent work sought to enrich these models. Early attempts with multilayer perceptrons and support vector machines did not dramatically improve accuracy beyond the logistic baseline, and in Katz et al. [2017] comprehensive study the ensemble (random forest) approach proved superior. However, the incorporation of additional data has yielded gains. Modern models factor in justice-specific voting histories, measures of each justice's propensity to vote in a liberal or conservative direction, and even conditional probabilities (such as how likely each justice is to vote to reverse a lower court, given the Court's overall reversal rate). By updating predictions as new information arrives (e.g. updating features after oral argument timing or as the term progresses), these models function in a semi-online learning fashion, continually refining forecasts for pending cases. The result is a more nuanced prediction that can adjust to context. For instance, recognizing that a long delay in decision might signal internal disagreement and a higher chance of a narrow, unanimous outcome.

2.1.2 Martin-Quinn Scores and Ideological Drift

Accurate prediction of Supreme Court behavior often hinges on quantifying judicial ideology. A key tool in this regard is the Martin-Quinn scoring system [Martin and Quinn, 2002]. Martin-Quinn (MQ) scores position each justice on a latent ideological spectrum (typically liberal-conservative) for each Supreme Court term. These scores are estimated from justices' voting records using a Bayesian item-response theory model, treating each case vote as revealing something about a justice's ideal point. Crucially, MQ scores are dynamic-a justice's ideal point can evolve over time, allowing researchers to track shifts in ideology across years or decades. This dynamism contrasts with one-time labels (like calling a nominee "liberal" or "conservative" at confirmation); instead, MQ scores are based on actual judicial behavior and updated as new votes are cast. They have become the standard metric in empirical studies for comparing justices' ideological leanings and for identifying the pivotal median justice in any given term. For example, MQ scores scientifically substantiated the oft-made claim that Justice Anthony Kennedy was the de facto swing vote on the late-Rehnquist and early-Roberts Courts.

One important application of Martin-Quinn scores has been to investigate ideological drift-the tendency of justices' views to change during their tenure on the bench. Epstein et al. [2007] found that almost every justice exhibited significant ideological drift over their career. Several justices appointed as moderates or conservatives drifted leftward (e.g. Blackmun, Stevens, Souter), while a few drifted to the right (e.g. Black, Frankfurter). The magnitude of these shifts was sometimes striking and not fully anticipated at the time of appointment. These findings confirm that ideological stability is the exception, not the norm. This has implications for prediction, as models must account for the fact that a justice's past behavior may not perfectly predict future votes if their preferences are in flux. In addition, MQ scores illuminate periods of Court stability versus volatility and quantify the impact of membership changes (e.g., how the replacement of one justice shifts the Court's ideological center).

2.1.3 Expert Judgment and Court Watchers

Quantitative models notwithstanding, a longstanding tradition in Supreme Court prognostication relies on the qualitative judgment of legal experts, journalists, and court watchers. Before the recent boom in data-driven models, most Supreme Court predictions were made by seasoned observers drawing on doctrinal knowledge, intuition, and close readings of each case's context. Even today, major news outlets and blogs regularly publish forecasts of high-profile decisions based on oral arguments and insider cues. Experienced correspondents often venture predictions about who will win a case or how specific justices will vote, parsing the justices' questions at oral argument or the nuances of precedent.

Outside formal studies, expert court watchers have had well-publicized prediction successes and failures. A commonly cited indicator is the pattern of questioning at oral arguments. Black et al. [2012] found that the attorney who faces more questions from the justices is more likely to lose. Many practitioners and journalists have internalized this heuristic-during arguments, they literally keep score of how many and what kinds of questions each side receives. Similarly, Jacobi and Rozema [2017] measured vocal pitch and emotional arousal in justices' voices and reported that these auditory cues could predict many of the justices' eventual votes. Qualitative observers often intuit these signals: a justice's skeptical or hostile tone toward one side is a red flag for that side's prospects. Indeed, media coverage after high-profile arguments is replete with analysts reading the tea leaves. Such intuition is correct more often than not; practiced court watchers systematically outperform random guessing, though typically not as high as advanced algorithms. However, qualitative prediction has clear limitations. It can be anecdotal and prone to bias: observers might see what they expect or desire to see. Famous surprises (e.g. *NFIB v. Sebelius*, 2012) underscore that no matter how experienced, human prognosticators face uncertainty and can be blindsided by strategic or personal factors.

2.1.4 Crowdsourcing and Prediction Markets

In addition to individual experts, the collective wisdom of crowds has emerged as a notable approach. *FantasySCOTUS*, launched in 2009, harnesses thousands of enthusiasts to forecast case results [Blackman et al., 2012]. During its inaugural term, over 5,000 members submitted more than 11,000 predictions. Overall, the crowd's consensus was correct in a bit over half of the cases, modestly exceeding chance. Crucially, the top-performing individuals were remarkably accurate, correctly forecasting about 75% of cases-on par with statistical models. Katz et al. [2017] analyzed more than 600,000 crowd predictions and showed that an optimal combination of crowd forecasts could achieve about 80.8% accuracy - the highest performance recorded in any Supreme Court prediction study to date. This indicates that with enough data and proper weighting, crowdsourcing can robustly outperform not only chance but many expert-based or algorithmic methods.

Real-money prediction markets have also traded on Supreme Court outcomes. *Intrade*, an Irish trading platform, famously offered contracts on the 2012 Affordable Care Act case [NFI, 2012]. Before oral arguments, the market implied only a 35% chance of invalidation of the mandate, but after tough questioning, odds shot above 60%. The market settled around 75% odds that the mandate would be struck down [Perry]. The actual outcome - the mandate upheld - proved the market wrong. This episode illustrates both the responsiveness and fallibility of prediction markets: they quickly incorporate new information but can misinterpret signals. In general, markets tend to outperform chance and often rival expert opinion, but they require sufficient liquidity and knowledgeable participants, which are scarce for lower profile cases.

2.1.5 Using LLMs in Prediction

With the advent of modern LLMs, researchers and practitioners have begun exploring whether these models' language understanding can translate into improved predictions. Early results indicate that LLMs, in their current form, perform in the same ballpark - i.e. middling accuracy that plateaus around 70–80% on binary outcome prediction. For instance, Nay (2023) tested GPT-family models on a set of U.S. court cases involving fiduciary duty questions. He found that OpenAI's latest model (GPT-4) could correctly predict the case outcome about 78% of the time, whereas its 2020-era predecessor performed no better than random guessing.

Anecdotal evidence suggested Anthropic's Claude to act as a "SCOTUSbot" and decide 37 real pending Supreme Court cases. Claude's predictions matched the actual outcomes in 27 of 37 cases – roughly 73% accuracy, mirroring the performance of specialized ML models and far from a perfect record [Unikowsky, 2025].

In a recent project reported by *The Economist* (2025), a method based on OpenAI's o3 reasoning model could often guess the correct result and even produce plausible draft opinions, but it was prone to error on the trickier splits. Intriguingly, the evaluators had to run each case through the model multiple times and observed variation in its answers. This underscores the earlier point about prompt sensitivity – the stochastic nature of LLM outputs means an AI legal prediction is not a single fixed response but a distribution of possible answers, some of which will be wrong.

2.2 Case Complexity Measurement

A growing strand of research emphasizes the importance of case complexity in understanding and predicting Supreme Court behavior. Complexity captures how difficult or multifaceted a dispute is, both doctrinally and strategically. Scholars have developed three main families of measures: issue- and brief-based, opinion- and docket-based, and network- and argument-based.

First, issue- and brief-based indicators capture complexity ex ante. Lauderdale and Clark [2022] develop a latent complexity measure extracted from merits briefs, demonstrating that pre-treatment measures avoid the endogeneity of post-decision proxies. Counts of *amici curiae* and their diversity likewise serve as ex ante signals of complexity [Paul M. Collins, 2004].

Second, opinion- and docket-based proxies infer complexity from how the Court writes and manages cases. Longer majority opinions, more separate writings, splintered vote splits, and longer decision times are used as indicators. These correlate with difficulty and bargaining costs [**?**], but are post-treatment and thus problematic for explaining votes. Amici activity is also informative: a larger and more diverse set of amici can increase the information load.

Third, network- and argument-based indicators assess complexity through inter-justice dynamics and doctrinal entanglement. Citation-network studies compute centrality of precedents to map how legally "entangled" a dispute is, with higher centrality indicating greater doctrinal interdependence [Fowler and Jeon, 2008]. Oral argument analytics - such as interruptions, hostile questioning, or emotional arousal-also capture facets of decision difficulty and coalition strain [Jacobi and Rozema, 2017].

Best practices favor ex ante measures when the goal is forecasting votes, while post-treatment proxies are more appropriate for characterizing doctrinal outcomes. In predictive modeling, higher complexity generally reduces confidence and increases the likelihood of divided outcomes.

2.3 Out-of-Sample Challenges with LLMs

Another obstacle is domain shift. LLMs trained on broad internet text often falter when confronted with data from a substantially different domain or time period. When moving to specialized domains like biomedical or legal text, performance often degrades sharply [Sun et al., 2023]. One reason is that models may lack the domain-specific knowledge and vocabulary, but even continual learning on new domain data risks catastrophic forgetting of prior knowledge. The net effect is poor generalization across domains: an LLM might perform well on generic tasks but struggle with the unique language and conventions of legal pleadings. Relatedly, models trained on static historical data are inherently stale with respect to future developments. If an evaluation involves a time shift – predicting examples drawn from after the training period – performance can drop due to emergent phrases, events, or rules that the model has never seen. In practice, this means true out-of-sample tests must account for temporal and contextual shifts, which many benchmarks do not capture.

Compounding these issues are evaluation design flaws that can give an overly rosy picture of out-of-sample ability. A prominent concern is training–test contamination: LLMs with billions of training tokens have often ingested portions of common benchmark datasets or facts about "held-out" examples. Studies have found that for popular NLP benchmarks, a surprising number of test questions or their close paraphrases appear in models' training data. This leakage allows models to answer by memory rather than generalization, inflating scores. Carlini et al. [2022] demonstrate that large LLMs can memorize and regurgitate training passages verbatim, especially as model size grows. Similarly, Yao et al. [2024] show that even without exact memorization, duplication of test content in training leads to overly optimistic performance estimates. Therefore, careful benchmark design is needed to ensure evaluations truly reflect novel inputs − for example, by holding out data chronologically (to prevent any information about future examples in training) and by periodically refreshing test sets [Kiela et al., 2021]. In summary, the ML literature indicates that achieving robust out-of-sample prediction with LLMs is challenging due to overfitting, prompt fragility, domain/context shifts, and pitfalls in evaluation design. Addressing these issues is crucial, especially in high-stakes domains like law where the test examples may differ in form and context from anything seen before.

It is also important to emphasize the problem of contamination and leakage in legal evaluations. Some of the most striking claims about LLM performance on law-related tasks have later been questioned because the test items were likely present in the training data. This has been particularly prominent in the case of professional exams. GPT-4 was reported by OpenAI to have passed the Uniform Bar Exam at a level placing it in the top decile of human test-takers. Yet subsequent scrutiny pointed out that many UBE questions are publicly released and widely circulated on the internet. Given the opacity of GPT-4's training data, it is impossible to rule out that these questions, or close paraphrases, were seen during training. If so, the model's high score would reflect memorization rather than genuine reasoning ability. Martínez [2024] and others have stressed that without strict controls on temporal cutoffs and training exposure, such evaluations cannot be interpreted as true out-of-sample tests. More broadly, Hidayat et al. [2025] show through controlled experiments that even small amounts of benchmark leakage can inflate LLM scores dramatically, while Balloccu et al. [2024] document widespread malpractice in evaluating closed-source LLMs, where contamination risks are ignored. For legal prediction tasks, where case law and exam questions are publicly available, the risk of leakage is particularly acute. If models "know" the outcome of a case because summaries, media coverage, or briefs were in their training set, then their apparent predictive success is illusory.

Finally, many successful legal outcome prediction models (including commercial tools such as LexisNexis's Lex Machina) achieve their accuracy by leveraging metadata and statistics more than deep "legal reasoning." These systems might factor in the identity of the judge, the circuit, or the type of litigant to find correlations with outcomes. While effective, this approach does not truly understand the legal issue. LLMs, in theory, hold the promise of more text-centered reasoning, reading briefs and precedents to determine the outcome. However, to date even LLM-based tools often still fall back on surface cues. The Economist's SCOTUSbot, for example, changed its prediction on a case after "hearing" oral argument transcripts that indicated a particular lean from the justices. This suggests the model was picking up on signals (like the tone of questions) that human court watchers also use – a useful strategy, but one which might not generalize if the signals mislead or if a future case lacks clear cues. In sum, legal decision prediction remains a challenging testbed for true out-of-sample performance. Current LLMs and ML models can capture broad patterns and have moderate success forecasting decisions under controlled conditions, but they struggle with the hard cases – those involving evolving doctrines, unique fact constellations, or shifts in the Court's composition or priorities. Achieving reliability in these scenarios likely requires hybrid strategies – combining LLMs with up-to-date legal knowledge bases, continual training on new data, or expert-in-the-loop systems – to ensure the model's predictions keep pace with the law's development. The research community is increasingly aware that rigorous temporal hold-out testing (e.g. training on past cases, testing on a future term) is essential to measure progress. Such evaluations force models to confront the unpredictable "unknown unknowns" of the legal world – precisely where genuine intelligence, or lack thereof, is revealed. As legal AI moves forward, the lessons from general ML –

avoid overfitting to yesterday's data, expect performance to drop on out-of-distribution problems, and continuously validate on fresh, real-world examples – will be vital in guiding the development of models that can prospectively predict legal outcomes with greater fidelity.

3 Methods

3.1 Data and Materials

Our analysis draws on two primary sources of data: docket filings from the Supreme Court of the United States and outcome metadata from the Supreme Court Database (SCDB). To assemble the filings dataset, we scraped the official SCOTUS website for the 2022–2023, 2023–2024, and 2024–2025 Terms. The scraper collected all docket entries associated with argued cases, including briefs, amicus curiae submissions, and other filed documents. We then filtered the collected files to retain only substantive filings. Procedural documents, appendices, scheduling orders, and miscellaneous correspondence were excluded. The retained corpus therefore consisted of party briefs and amicus briefs filed in cases that were ultimately resolved with a signed or per curiam opinion on the merits.

We collected all available docket entries for argued cases across the 2022-2023, 2023-2024, and 2024-2025 Terms. After filtering, we retained substantive legal documents for each case, typically comprising party briefs, reply briefs, and amicus curiae submissions. Documents were classified as substantive using a dual approach: pattern matching against keywords (petition, brief, appendix, merits, amicus, reply, supplemental, joint appendix, opinion, judgment, transcript) and LLM-based classification for ambiguous cases. We explicitly excluded documents matching procedural patterns (motion, order, extend, extension, certificate, proof of service, rehearing, waiver, distributed, record, compliance). This filtering process ensured that our corpus consisted exclusively of documents containing substantive legal arguments relevant to case outcomes.

Each PDF document was processed using PyMuPDF (fitz) for text extraction with page-level granularity. Text was segmented into chunks of 1,500 characters with 300-character overlap using LangChain's RecursiveCharacterTextSplitter, with separators prioritizing paragraph and sentence boundaries. This chunking strategy balanced context preservation with retrieval precision, resulting in approximately 50-200 chunks per case depending on filing volume.

Each case in this filtered corpus was linked by docket number to its record in the SCDB. The SCDB provided the authoritative outcome label for each case and each individual justice's vote. Because the SCDB is updated term by term and coded with consistent outcome categories (e.g., affirm/reverse, petitioner/respondent win), it provided a reliable ground truth against which to evaluate predictions. Linking filings to SCDB records also allowed us to verify case disposition type and exclude cases lacking a merits opinion.

3.2 Experimental Pipeline

To evaluate model performance on this corpus, we designed a retrieval-augmented generation (RAG) pipeline with hybrid search capabilities. Documents were embedded using OpenAI's textembedding-3-large model with 1,536 dimensions and indexed in a FAISS vector store. We implemented a dual retrieval strategy: BM25 for keyword-based search and FAISS for semantic similarity, retrieving the top 20 chunks per document type (petitioner brief, respondent brief, amicus brief, etc.). Retrieved chunks were then re-ranked using the cross-encoder model ms-marco-MiniLM-L-6-v2, with the top 50 passages selected for final context. The search queries were specifically crafted to capture comprehensive case information, including "Supreme Court case [docket] detailed factual background procedural history legal issues constitutional questions" and similar formulations.

Each justice was modeled with a detailed judicial philosophy profile incorporating their interpretative methodology and key decision factors. For instance, Justice Thomas was characterized as a "strict originalist and textualist who interprets the Constitution according to original public meaning," with key factors including "original meaning, textualism, limited government, states rights" and a conservative lean score of 0.90. Justice Sotomayor was modeled as a "pragmatic liberal emphasizing real-world impacts on vulnerable populations," with factors including "practical impacts, fairness, vulnerable populations, broad remedial powers" and a conservative lean score of 0.15. These profiles were derived from analysis of historical voting patterns and published opinions, providing the model with justice-specific reasoning frameworks.

The prompt design required each model not only to cast a binary vote (affirm/reverse or petitioner/respondent win) but also to provide a textual justification for the vote. In this way, the pipeline elicited both quantitative predictions and qualitative reasoning. After collecting the nine simulated votes for each case, we tallied the votes to obtain a predicted Court judgment. As an additional step, we also asked each model to draft a syllabus of the predicted majority opinion. This element of the design allowed us to evaluate whether the models could reproduce not only outcomes but also the genre of reasoning and summary typical of official SCOTUS opinions.

3.3 Models

Weimplemented the pipeline with three leading closed source LLM systems: OpenAI's GPT-4o (model version gpt-4o with 128k context window), Google's Gemini 1.5 Pro (with 1M context window), and Anthropic's Claude 3.5 Sonnet (model version claude-3-5-sonnet-20240620 with 200k context window). Each model was accessed via its respective API with specific configurations: temperature set to 0.6 to balance consistency with diversity across replicates, maximum output tokens of 4,096 for vote predictions and up to 16,384 for syllabus generation, and a timeout of 60 seconds with 2 maximum retries for robustness. The RAG retrieval layer utilized batch processing with 30-document batches for FAISS index creation to manage API rate limits. All models received identical retrieved context and prompts, with the only variation being the model-specific token limits

3.4 Evaluation

Predictions were evaluated against ground truth outcomes from the SCDB. We assessed performance at both the justice-vote level and the case-outcome level. Justice-level accuracy was defined as the proportion of individual votes correctly predicted. Case-level accuracy was defined as the proportion of cases in which the model correctly predicted the final disposition of the Court (i.e., the majority outcome). Because each justice's vote was simulated three times, we also calculated a confidence measure based on the proportion of consistent predictions across runs. Finally, we compared the generated syllabi with the official Court syllabi for qualitative assessment, though our primary evaluation focused on accuracy metrics derived from SCDB data.

This evaluation design provides a rigorous test of the ability of LLMs to generalize from substantive filings to actual case outcomes. By restricting the corpus to merits cases, linking filings to SCDB outcomes, and implementing a temporally out-of-sample regime, we ensured that the task reflects a realistic and challenging setting for predictive legal modeling.

4 Results

Our experiment yielded two key findings regarding the performance of LLMs in forecasting Supreme Court decisions for the 2023-2024 term.

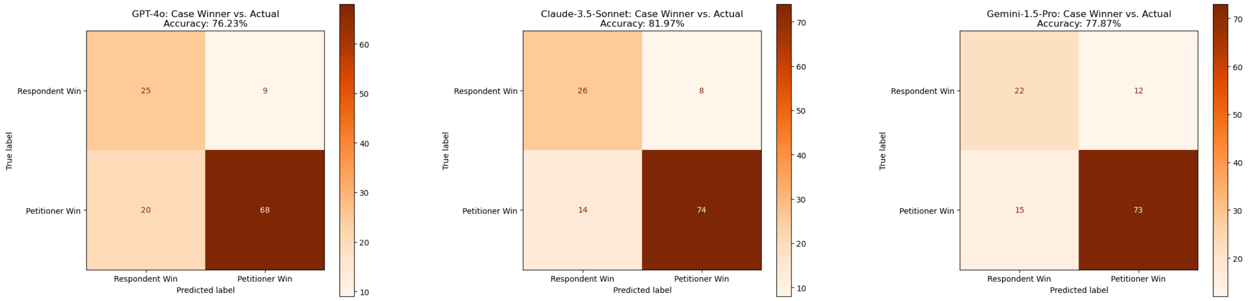
Table 1: Comparison of SCOTUS Forecasting Accuracy

Model / Method	**Term**	**Accuracy (%)**
FantasySCOTUS Crowd	2023–24	83.1
Claude 3.5 Sonnet	**2023–24**	**81.97**
GPT-4o	**2023–24**	**76.23**
Gemini 1.5 Pro	**2023–24**	**77.87**
Martin–Quinn Statistical Model	2002–03	75.0
Ruger–Kim Legal Experts Panel	2002–03	59.1

First, at the case-outcome level, all three models performed similarly to established historical benchmarks. As shown in Figure 1, Claude 3.5 Sonnet achieved the highest overall accuracy at 81.97%, followed by Gemini 1.5 Pro at 77.87%, and GPT-4o achieved the highest accuracy at 76.23%. As seen in Table 1, compared to Ruger and Kim's expert forecasts for the 2002-2003 term (59% accurate), and Quinn and Martin's decision-tree based forecasting model for the same term, (75% accurate), LLMs fare comparably to the Martin-Quinn approach. That being said, all three models do slightly *worse* than the consensus forecast on FantasySCOTUS for the 2023-2024 term.

Second, the models demonstrated an ability to predict individual justice votes, with performance varying across the bench. The heatmaps in Figure 2 reveal that Claude 3.5 Sonnet was the most consistent model on this task, with its accuracy ranging from a high of 87% for Justice Thomas to a low of 79% for Chief Justice Roberts. All models showed a general pattern of higher accuracy for justices with more predictable ideological leanings (e.g., Thomas, Sotomayor) and found it more challenging to predict the votes of justices near the Court's ideological center, such as Chief Justice Roberts.

In addition to forecasting judge votes, we also leverage a unique capability of LLMs relative

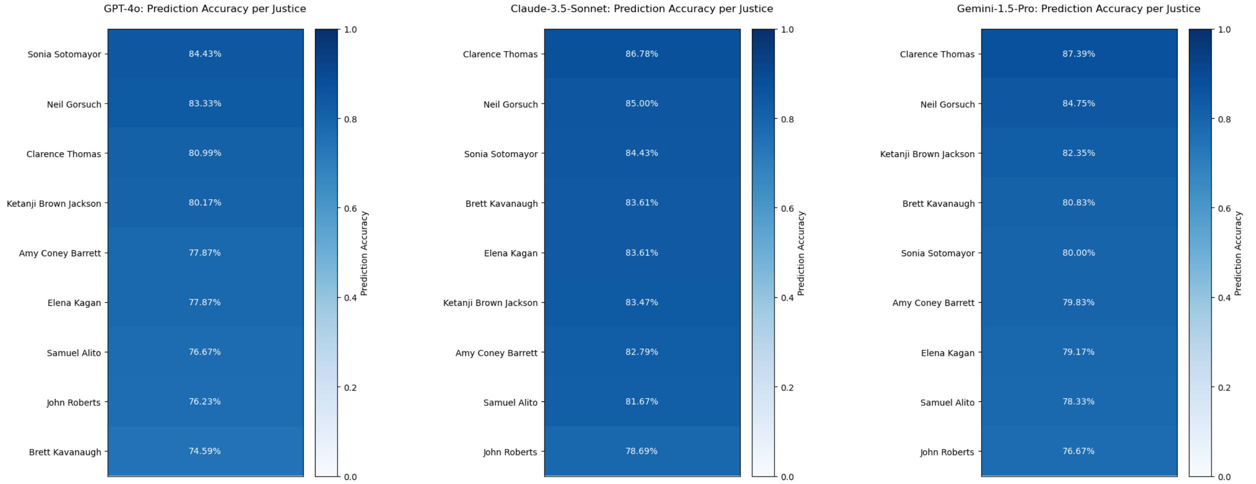


(a) GPT-4o (76% Accuracy)

(b) Claude 3.5 Sonnet (82% Accuracy)

(c) Gemini 1.5 Pro (78% Accuracy)

Figure 1: Case Outcome Confusion Matrices by Model (2023–2024 and 2024–2025 Terms). *Notes*: Each matrix shows the predicted winner (Petitioner or Respondent) against the actual winner.



(a) GPT-4o

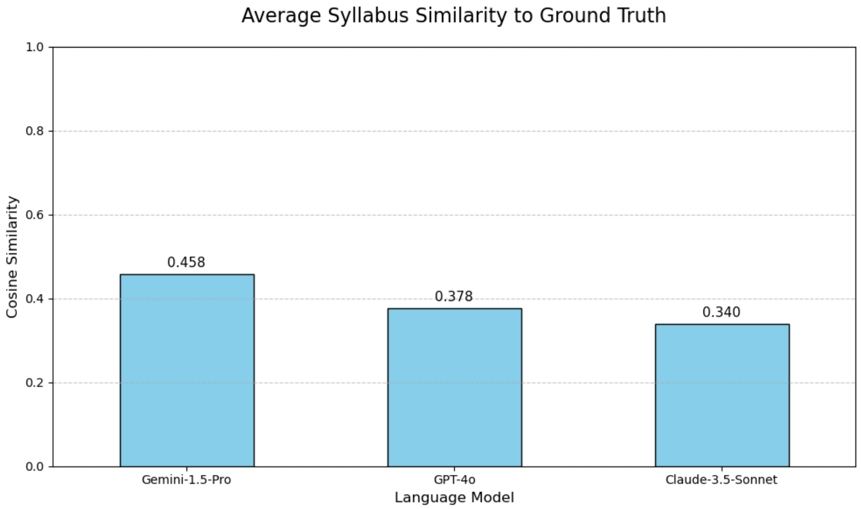
(b) Claude 3.5 Sonnet

(c) Gemini 1.5 Pro

Figure 2: Heatmaps of Justice-Level Vote Prediction Accuracy by Model. *Notes*: Darker shades indicate higher accuracy for a given justice.

to other methods: the ability to generate text. As each LLM makes predictions about how each Justice will vote, we also prompt it to ask it to write a syllabus representing the majority opinion. Figure 3 shows the average cosine similarity between each model's generated syllabi and the real-world syllabi written by the Supreme Court. We see that the models range from cosine similarities of .34 to about .46, indicating that generated syllabi have some similarity to real world outputs. Interestingly, the most accurate model, Claude 3.5, has the lowest average cosine similarity with the eventual real syllabi, whereas Gemini 1.5 is the most similar on average.

Figure 3: Cosine Similarity between Generated and Actual Syllabi)



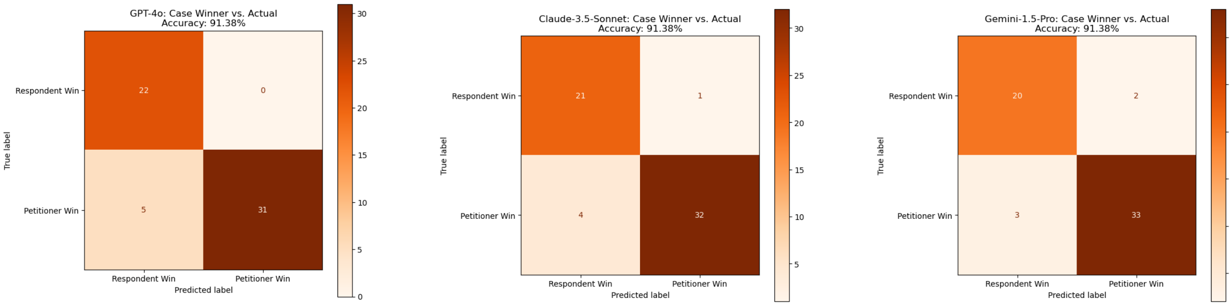
Notes: Barplot showing the average cosine similarity between syllabi generated by each LLM and the actual syllabi written by the Supreme Court.

One major concern with using LLMs for legal prediction is the problem of memorization. Models might look more capable than they actually are if used to make predictions on historical data because they have memorized the outcomes of those legal decisions. For example, a LLM asked to predict the vote outcomes in a famous case like *Brown v. Board of Education* has likely not only been trained on the text of the opinion itself, but text from various sources describing the history of that decision and the significance of the 9-0 vote. By memorizing this information, the predictions are more akin to recalling information rather than making predictions about the future. To assess whether this type of memorization is a problem for Supreme Court forecasting, Figure 4 shows the results of having each LLM make predictions about cases in the 2022–2023 Supreme Court term. These cases were all plausibly in the training data as they were decided before each model's knowledge cutoff date.

Here, we see that making predictions on cases that were decided before the knowledge cutoff date dramatically "improves" accuracy. Each model achieved a 91.38% accuracy (though with different proportions of getting the right petitioners and respondents) when used to predict case outcomes in the 22–23 term. This represents a 10-15% increase over predictions on the newer cases, suggesting that memorization does indeed give a misleading indicator of model performance when the training data likely includes the real-world outcomes.

5 Discussion

The results of this study offer evidence that large language models are a viable tool for court forecasting. Notably, all three models perform similarly to consensus votes and other statistical models. Performing roughly as well as both a "wisdom of the crowds" and data-driven approaches



(a) GPT-4o (91.38% Accuracy)

(b) Claude 3.5 Sonnet (91.38% Accuracy)

(c) Gemini 1.5 Pro (91.38% Accuracy)

Figure 4: Case Outcome Confusion Matrices by Model (2022–2023 Term). *Notes*: Each matrix shows the predicted winner (Petitioner or Respondent) against the actual winner.

suggests that LLMs pick up on relevant features of both the legal complexities involved with each case the Supreme Court hears, and the individual ideological preferences of the Justices.

There are a few limitations that are worth noting. The first is regards to generalizability. While Supreme Court forecasting is popular because of the high stakes nature of many of the cases the top court hears, there is lots of information available both about these cases and the Justices themselves. While we guard against the problem of leakage about case outcomes by purposely using older models with cutoff dates prior to the final decisions, it is possible that the models pick up on other information like how lower court judges ruled. That being said, court forecasters also have access to this same information.

Another limitation concerns the tradeoffs between accuracy and transparency. Our study suggests that LLMs perform comparably with other approaches such as crowdsourcing and simple statistical techniques. Yet, LLMs can be considerably less transparent than these other techniques, and without achieving additional accuracy above these approaches, it is unclear if LLM-based legal forecasting is worth the costs. In this case, there are both literal costs in terms of monetary API costs, but also with regards to reproducibility. While a decision tree classifier can be tested even decades in the future, the models used in this study are likely to be deprecated over time, making it difficult to perform exercises like forecasting Supreme Court decisions even further into the future.

6 Conclusion

Overall, this study shows that LLMs are increasingly capable of rivaling the best Supreme Court forecasters, though still not exceeding them. These results underscore both the promise and the limits of LLMs in legal forecasting. On the one hand, models can synthesize unstructured legal text in ways that rival hand-coded features or the heuristics of human experts. On the other hand, their performance remains uneven, particularly for ideologically centrist justices and closely divided cases.

Going forward, future work should look at whether LLMs are capable forecasters in other types of legal prediction scenarios: lower court decisions, audits, and inspections for example. SCOTUS predictions provide a nice starting point because of their long history in both the academic literature and hobbyist community. More broadly, the question of whether LLMs can *ever* be "superforecasters" is of broad interest in both legal and other domains. At least in this initial study, we do not yet see LLMs as being capable of exceeding the wisdom of the crowds, though their comparable performance suggests that they may be possible substitutes in scenarios where such data is expensive or difficult to generate.

References

References

National federation of independent business v. sebelius, 2012. URL https://www.supremecourt.gov/opinions/11pdf/11-393c3a2.pdf. 567 U.S. 519 (2012).

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Long Papers)*, pages 67–93, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.5. URL https://aclanthology.org/2024.eacl-long.5/.

Ryan C. Black, Timothy R. Johnson, and Justin Wedeking. *Oral Arguments and Coalition Formation on the U.S. Supreme Court: A Deliberate Dialogue*. University of Michigan Press, Ann Arbor, 2012.

Josh Blackman, Adam Aft, and Corey Carpenter. FantasySCOTUS: Crowdsourcing a prediction market for the supreme court. *Northwestern Journal of Technology and Intellectual Property*, 10 (3):327–362, 2012.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. *CoRR*, abs/2202.07646, 2022. URL https://arxiv.org/abs/2202.07646. Carlini et al. (2022).

Lee Epstein, Andrew D. Martin, Kevin M. Quinn, and Jeffrey A. Segal. Ideological drift among supreme court justices: Who, when, and how important? *Northwestern University Law Review*, 101(4):1483–1542, 2007.

James H. Fowler and Sangick Jeon. The authority of supreme court precedent. *Social Networks*, 30 (1):16–30, 2008.

Roger Guimerà and Marta Sales-Pardo. Justice blocks and predictability of U.S. supreme court votes. *PLOS ONE*, 6(11):e27188, 2011. doi: 10.1371/journal.pone.0027188.

Naila Shafirni Hidayat, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. Simulating training data leakage in multiple-choice benchmarks for LLM evaluation. *arXiv preprint arXiv:2505.24263*, 2025. URL https://arxiv.org/abs/2505.24263.

Tonie Jacobi and Kyle Rozema. Judicial conflicts and voting agreement: Evidence from interruptions at oral argument. *Journal of Empirical Legal Studies*, 14(2):192–229, 2017.

Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12(4):e0174698, 2017. doi: 10.1371/journal.pone.0174698.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/. Kiela et al. 2021.

Benjamin E. Lauderdale and Tom S. Clark. Measuring supreme court case complexity. *Journal of Law, Economics, & Organization*, 38(1):92–118, 2022. doi: 10.1093/jleo/ewab012.

Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via markov chain monte carlo for the U.S. supreme court, 1953–1999. *Political Analysis*, 10(2):134–153, 2002. doi: 10.1093/pan/10.2.134.

Eric Martínez. Re-evaluating GPT–4's bar exam performance. *Artificial Intelligence and Law*, pages 1–24, 2024. doi: 10.1007/s10506-024-09396-9. URL https://link.springer.com/article/10.1007/s10506-024-09396-9.

Jr. Paul M. Collins. Friends of the court: Examining the influence of amicus curiae participation in U.S. supreme court litigation. *Law & Society Review*, 38(4):807–832, 2004.

Mark J. Perry. Intrade odds at 77.7 for the individual mandate to be ruled unconstitutional. URL https://www.aei.org/carpe-diem/intrade-odds-at-77-7-for-the-individual-mandate-to-be-ruled-unconstitutional/.

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 decisionmaking. *Columbia Law Review*, 104(4):1150–1210, 2004.

X. Sun et al. Feature distribution matching for federated domain generalization. *Proceedings of Machine Learning Research (PMLR)*, 189:—, 2023.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction.* Crown Publishers, 2016. ISBN 9780804136718. New York, 352 pages.

Adam Unikowsky. In ai we trust — part ii. Substack, 2025. URL https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii. Accessed: YYYY-MM-DD.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17864–17875, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.990. URL https://aclanthology.org/2024.emnlp-main.990/.