Appealing to Large-Language-Model-as-Judge: Annotating and Analyzing the Universe of U.S. Appellate Opinions

Eitan Sapiro-Gheiler
Department of Politics
Princeton University
eitans@princeton.edu

Jonathan P. Kastellec
Department of Politics
Princeton University
jkastell@princeton.edu

1 September 2026

Abstract

We automate annotation of social science texts by manually converting expert codebooks into discrete facts, then automatically extracting those facts using a reproducible, multi-step large language model pipeline. Applying this approach to the U.S. Courts of Appeals, we build a universe-level database of all 440,000 published opinions—twenty times larger than existing hand-coded samples—over 500 times cheaper and 5,000 times faster than hand coding. Our annotations agree with human coders 85–95% of the time on metadata, substantive issues, and litigant types; and 80% on ideological direction (liberal or conservative). Against independent administrative records, they outperform human coders on nearly every available variable, suggesting residual disagreement reflects judgment rather than machine error. Performance depends on our domain-informed task design, not model choice, examples, or fine-tuning. Using our data, we show that partisan polarization is driven largely by dissenting votes and predicts outcomes better than litigant identity.

Keywords: Courts of Appeals, large language models, judicial ideology, panel effects

We thank Dahyun Choi, Nolan McCarty, Arthur Spirling, and participants at the Princeton Politics Institutions Workshop for helpful comments and suggestions. The combined metadata and LLM-annotation database produced in this paper is available at `https://coadata.org`.

1 Introduction

Among high-stakes decisions, judicial behavior is perhaps the most well-documented. Judges publish detailed, technical opinions laying out their ideology and reasoning, which then form the basis of both legal scholarship and empirical study of judicial politics. This corpus is so large that it constrains our ability to conduct comprehensive research on U.S. courts: federal appellate courts alone have produced millions of pages of opinions, adding about 40,000 more pages each year. Empirical scholarship has thus relied on selected samples of opinions annotated at great cost by human coders, centered on two main databases. Beginning in the 1980s, the U.S. Supreme Court Database (Spaeth et al. 2025) offered the first universe-level dataset of cases formally decided by that court.[1] For the U.S. Courts of Appeals—which are the court of last resort for tens of thousands of cases each year—the largest available database, commonly called the "Songer Database" after its creator Donald Songer, covers about 23,000 out of 440,000 published cases and has no coverage at all beyond 2002.[2] Hand-coding the universe of federal appellate cases would be prohibitively time-consuming—even a conservative estimate of half an hour per case requires over a century of full-time employment with a minimum-wage cost over $1.5 million.[3]

In this paper, we propose and implement an approach combining existing expert-written codebooks with modern machine learning tools to overcome the constraints of human annotation. This approach allows us to provide a first look at the universe of published U.S. Courts of Appeals opinions—440,000 cases from the start of the modern system in 1892 to

[1]Formally decided cases include signed opinions (i.e., "merits" cases), per curiam opinions, judgments of the Court, and decrees. They exclude cert decisions and other types of orders—e.g., emergency applications—typically associated with the Court's shadow docket; see Kastellec and Taboni (2026).

[2]The Songer Database, formally the U.S. Courts of Appeals Database, is a stratified-by-circuit-year random sample. The original database (Songer 2008*b*) covered 1925–1996; an update (Kuersten and Haire 2007) extended to 1997–2002; and a "Phase 2" version (Songer 2008*a*) added every case reviewed by the Supreme Court from 1925–2002.

[3]Due to the scale of this task, researchers interested in specific questions about appellate judging have often built their own (usually single-issue) datasets, e.g., Benesh (2002) on confession cases, Farhang and Wawro (2004) on employment discrimination, Cox and Miles (2008) on voting rights, Kastellec (2013) on affirmative action, and Taboni (2026) on cases of first impression.

the present day. At a high level, our strategy is to convert key variables in codebooks written by domain experts into sequences of facts which we extract from opinion text using an open-weights large language model (LLM). Our preferred LLM codes all commonly-used variables from the Supreme Court Database and Songer Database codebooks in less than one week at a cost of about \$3,000; the pipeline we use is fully model-agnostic and can be adapted to accommodate differing budgets or improvements in LLM capabilities. Validated against the Songer Database's 23,000 human-coded cases, our approach matches human coders about 85–90% of the time on summary variables and over 80% of the time on ideological direction (whether an opinion is liberal or conservative), the most judgment-sensitive variable used in empirical work. More strikingly, when our LLM approach and the human coders are compared on an independent administrative-data benchmark—records from the Federal Judicial Center (FJC)—machine coding outperforms human coding on almost all variables, indicating that the residual machine-vs.-Songer disagreement likely reflects reasonable differences of opinion rather than machine error.

Our central methodological finding is that performance is driven by domain-informed task design rather than model choice, access to examples, or fine-tuning. For our key variable of ideological direction, naïvely asking the LLM to directly provide an answer, even in a few-shot setting with balanced examples, produces accuracy around 70% and unstable answers which vary from run to run. Our method instead adapts the process laid out for human coders to identify individual facts which can later be combined to recover the opinion's ideological direction: who the litigants are, what the case is about, and which side won. To allow validation of our results and generalization beyond the specific setting and codebooks used in this paper, we make public the full pipeline code and prompt text.[4]

Two applications show how our universe-level coverage allows analyses that were impossible using existing hand-coded samples. First, we analyze panel effects (Sunstein et al. 2006, Kastellec 2011)—the influence of colleagues' partisanship on a judge's vote—at the

[4]The annotated database itself is available at `https://coadata.org`. Full pipeline code is in a GitHub repository, currently available upon request, which will be made public upon publication of this paper.

individual level across the entire history of the modern appellate system. We find that left-right polarization was essentially absent until the 1960s, while its emergence after that point is almost entirely driven by an increase in ideologically-aligned dissents. Second, using litigant classifications more fine-grained than existing administrative records, we show that the recently documented tendency of Democratic-appointed judges to favor weaker litigants (Cohen 2025) attenuates when that “pro-weak bias” would produce a conservative outcome. These results provide evidence that pro-weak bias is a proxy for political ideology, rather than a substantive preference for underdogs. These findings are robust to formal corrections for the residual error in our machine annotations (Egami et al. 2023, Angelopoulos, Duchi and Zrnic 2023, Kluger et al. 2025), confirming that our approach delivers results accurate enough for empirical work.

The U.S. Courts of Appeals have three properties which make them an ideal setting to introduce and refine our approach to large-scale text annotation. First, they contain a huge volume and variety of legal text, from imperfectly digitized opinions published in the early 1900s to complex modern decisions drawing on centuries of precedent. Second, the presence of existing benchmarks (the Songer Database and the FJC’s records) allows us to benefit from domain knowledge and cleanly compare our approach to the existing gold standard. Third, appellate rulings are highly consequential, meaning that our now-public database has immediate relevance and value. However, we believe the impact of our approach extends far beyond judicial politics. Much of modern empirical social science rests on expert annotation of large corpora of text: floor speeches and legislation, party manifestos, regulatory filings, and more. Recent work has shown that many of these domains are amenable to machine annotation; our results contribute a cross-domain framework that allows accurate, cost-effective annotation without curated examples or extensive fine-tuning. With access to annotated text no longer the limiting constraint, our focus can turn to the rich set of questions now addressable using the full scale and scope of real-world data.

2 Related Literature

Our research connects with substantive predecessors in judicial politics and methodological work using machine learning to collect large-scale information from political texts. On the latter front, Sunstein et al. (2006) and Kastellec (2011) are foundational in the panel effects literature, focusing on the attenuation of partisan outcomes produced by party-mixed panels. We contribute a longer time horizon, causally identified year-by-year estimates, and individual-level analysis that directly compares the same judge on different panels. Cohen (2025) introduces the notion of pro-weak bias by Democratic-appointed judges, defined as an increased probability of supporting the weaker litigant; we clarify that this effect is likely a proxy for broader left-right ideology rather than a direct preference for underdogs. Finally, Cohen and Dehejia (2026) argue that partisan polarization on the Courts of Appeals emerges after 2000, using reversal rates on the near-universe of published and unpublished cases from 1985–2020 (about 400,000 cases). We expand coverage to all published cases since 1892 and code case-level ideology using opinion content rather than judge identity.

Some prior work has used machine learning tools to classify judicial text, generally focusing on narrower tasks compared to our full-codebook annotations. Hausladen, Schubert and Ash (2020) use SVMs and random forests to learn the Songer database's ideological direction codes with 62% accuracy; our LLM-based pipeline outperforms their result (82% accuracy on the same task; see Table 3) and extends it to a much richer set of variables without labeled training data. Choi (2024) finds that GPT-4 performs as well as human coders on a simple Supreme Court classification task, with no gains from fine-tuning; we also attain human-level performance on many more cases and variables using an open-weight model. Ortega, Joshi and Borkowski (2025) use DeepSeek to classify Supreme Court opinions into 15 issue areas with about 80% accuracy; we achieve comparable results (88% accuracy, evaluated via crosswalk to Songer labels; see Table 3) in the harder Courts of Appeals setting and extend the approach to other variables.

More broadly, our work also adds to a growing literature on LLMs as measurement tools in political science (and social science in general). Halterman and Keith (2026) develop a five-stage framework for evaluating LLM-based codebook labeling, finding that off-the-shelf models struggle to follow real-world codebooks in zero-shot settings but improve substantially with instruction-tuning. Our pipeline provides an alternative approach by decomposing codebook variables into facts rather than asking the LLM to internalize codebook logic; in this paradigm, few-shot prompting does not provide additional gains (see Appendix A.4.3). Other recent findings also obtain strong results from well-designed LLM labeling approaches across party manifestos, social media messages, and other political texts (Benoit et al. 2026, Törnberg 2025, Gilardi, Alizadeh and Kubli 2023, Heseltine and Clemm von Hohenberg 2024, Ornstein, Blasingame and Truscott 2025). We take these results, and our strong performance on the text-rich but complex corpus of judicial opinions, as evidence that LLM reliability is driven by expert-informed task design rather than model choice or fine-tuning.

Our project contributes at three levels of increasing generality. First, our applications provide new insights about changes in ideological behavior by judges over time, clarifying when and how modern judicial polarization has affected appellate outcomes. Second, our database enables thorough exploration not only of ideology, but of other critical features of appellate rulings such as litigant win rates, cross-circuit differences, and the prevalence of fine-grained procedural or substantive case types. Finally, by documenting both the output of our AI-based codings and the pipeline that produced them, this paper offers a roadmap for scaling expert-based annotations to test both long-standing and novel hypotheses.

3 Data

In this section we describe the input data we used in our pipeline. (See Table 1 for a summary of the entire pipeline, including LLM labeling of codebook-derived variables.) Cleaned and structured metadata are available as part of our full annotated database at `https://coadata.org`; samples of opinion text are available upon request.

0. Input corpora and ground truth
a. Case text: Caselaw Access Project (1892–2019); CourtListener bulk exports (2019–2025).
b. Ground truth for LLM validation: Songer Courts of Appeals Database (about 23K hand-coded cases, 1925–2002); FJC Integrated Database (about 285K cases, all published opinions 1971–present).
c. Judge database: FJC Biographical Directory; CourtListener bulk exports; ideology scores.
1. Data preparation
a. Ingest and sanitize text: Unicode normalization, punctuation cleanup, OCR corrections.
b. Patch structure: split consolidated hearings, normalize non-standard opinion types, recover dissents and concurrences buried in majority text.
c. Match judges to FJC: staged name matcher, resolve ambiguous matches using source text.
d. Cleanup: stitch fragmented opinions from same author, deduplicate within and across sources using docket numbers and hearing dates.
2. Prompt development
a. Hand-code *Federal Reporter, 3rd Series* vol. 1 (226 cases) on all SCDB and Songer variables.
b. Iterate SCDB and Songer prompts against the 226-case set until headline accuracies plateau.
c. Light-touch refinements after an exploratory run on about 5,000 early-Songer cases.
3. LLM labeling (Qwen3-235B, two tracks × two passes)
a. *SCDB*: first pass codes broad variables; second pass refines area-specific variables.
b. *Songer*: first pass receives SCDB facts, codes its own broad variables; second pass refines.
c. Mechanical ideological direction post-processing.
4. Validation
a. *Songer*: compare LLM codes to Songer's about 23K hand-coded cases.
b. *IDB*: compare LLM performance to Songer performance on variables in FJC IDB.

Table 1: LLM-powered pipeline for the U.S. Courts of Appeals case universe.

3.1 Case Data

We obtained the universe of published U.S. federal appellate opinions from two sources.[5] The Caselaw Access Project (CAP, The President and Fellows of Harvard University 2024) provides machine-readable scans of all cases published in the Federal Reporter from the establishment of the modern appellate court system in 1892 through 2019. We extend coverage to the present using bulk data downloads from CourtListener (CL, Free Law Project 2026), which provides up-to-date published opinions; we use data through the end of 2025.[6] The combined dataset contains approximately 510,000 unique cases. As we explain below, about 440,000 of these are from the Courts of Appeals.

[5]Like most researchers studying the Courts of Appeals—but not all, see e.g. Yung (2010)—we focus solely on published opinions. "Published" is now something of a misnomer: since roughly the turn of the 21st century, essentially all opinions—precedential and not—are disseminated through electronic reporting services and databases. The operative distinction is precedential weight, since "unpublished" opinions are not supposed to carry it. We plan to extend our dataset to non-precedential opinions in future work.

[6]CL also includes the CAP data, but metadata differences make bulk scraping of CAP more practical. Pre-2019 published cases from both sources are identical.

Next, we cleaned raw case data in several automated steps (details in Appendix A.1). In particular, we deduplicated about 7,500 opinions published in more than one volume, split case records from about 700 instances where panels reheard cases into distinct observations, normalized a handful of non-standard opinion types (e.g., plurality opinions) into three standard categories (majority, concurrence, or dissent), and used opinion bylines to detect and properly label about 15,500 separate opinions accidentally embedded in the majority text during digitization.[7] We validate these cleaning steps using both the Songer Database (Songer 2008*b*;*a*, Kuersten and Haire 2007; henceforth simply "Songer") and the Federal Judicial Center Integrated Database (IDB), the administrative dataset covering all federal appellate cases from 1971 onward, or about 2.4 million total (published and unpublished) cases (Federal Judicial Center 2026*b*). Both provide independent records of en banc status and the presence of dissenting or concurring opinions; Songer additionally provides information about panel judges and opinion authors. We match 99.4% of Songer cases[8] to our data, and match 97.8% of cases in the IDB's coverage window (post-1970 cases from the 12 circuit courts[9]). Details of both matching procedures are in Appendix A.2.1.

Both comparisons support the cleaning steps above. Against Songer, we seat exactly the same three judges in about 94% of the 20,223 cases where both sources record a three-judge panel, agree on the majority-opinion author in over 97% of comparable cases, and agree on about 98% of over 40,000 individual judge votes—close to or above Songer's own reported intercoder reliability (about 95%). Against the IDB we recover almost 90% of en banc cases, 94% of dissents, and 85% of concurrences. While we appear to have high false positive rates, detecting more dissents and concurrences than the IDB, almost every one of these appears to be an IDB error: over 99% of the disputed separate opinions carry an author which we resolve to a specific panel judge, and every disputed en banc case has at least one

[7]Deduplication then removes about 3,500 of those as duplicate or fragmentary, for a net gain of roughly 12,000 dissents and concurrences over the raw scrape.

[8]For a handful of cases, we either find no match for the Songer database's Federal Reporter citation (after allowing for small typos) or are unable to disambiguate between multiple candidate matches.

[9]For the rest of this work, unless otherwise stated, we use "circuit courts" to include the 1st–11th circuits and the D.C. Circuit; this excludes the Federal Circuit.

named judge beyond the standard three.[10] Full metadata validation results, including the corrections applied to both benchmarks, are in Appendix A.2.

3.2 Validation Data

We obtain data for LLM validation from two independent sources. First, Songer includes hand-coded values for issue area, ideological direction of the disposition, authority type (e.g., statutory or constitutional), threshold issues (e.g., standing), and litigant typology (e.g., business or individual). Since our primary goal in this work is to produce codings that follow the Songer codebook's rules, in Section 5.1 we validate the LLM's performance by treating the Songer codings as ground truth, setting aside human error or genuine disagreement in the Songer data to examine how well the LLM reproduces human-coded outputs.

While Songer includes inter-coder reliability measures based on 250 cases coded by multiple human coders (which we compare to human-vs.-LLM agreement rates in Section 5.2), our best independent comparison of human and LLM performance comes from the IDB. For cases from 1971–2025, the IDB contains administrative records of litigant characteristics, agency involvement, issue area, disposition (reverse or affirm), and the type of court appealed from. Of course, the IDB may also contain coding errors (as documented in the previous section), but we believe these errors are unlikely to differentially match Songer human-coded outputs more often than LLM-generated outputs or vice versa.[11] Thus, in Section 5.3 we treat the IDB's outputs as ground truth and use them to compare the LLM's performance with that of the hand-coded Songer database.

[10] These cases have clear bylines, e.g., "Before TORRUELLA, Chief Judge, ALDRICH and COFFIN, Senior Circuit Judges, SELYA, CYR, BOUDIN and LYNCH, Circuit Judges" in Veilleux v. Perschau (101 F.3d 1); "GARZA, Circuit Judge, dissenting: I respectfully dissent from the opinion of my esteemed brethren..." in Menchaca v. Chrysler Credit Corp. (613 F.2d 507); and "PREGERSON, Circuit Judge (concurring): As I read the district court's Memorandum and Order..." in Exner v. FBI (612 F.2d 1202).

[11] Two opposing biases in the IDB could favor one side or the other: shared heuristics among human coders could inflate IDB-Songer agreement, but IDB reliance on formal administrative labels rather than holistic judgment could favor our fact-extraction LLM approach. We cannot formally evaluate either, but believe—given the scope and administrative status of the IDB—that neither is fatal.

3.3 Judge Data

For every judge in these cases, we obtain biographical, professional, and ideological variables primarily from the Federal Judicial Center (FJC) Biographical Directory of Article III Federal Judges (Federal Judicial Center 2026*a*), which records each judge's appointment history, demographics, and education. We supplement this data with CL bulk downloads (the same source as our modern case data). In total, our combined judge database covers approximately 6,000 judges, including not only circuit and district judges but also, e.g., magistrate and bankruptcy judges who occasionally sit on appellate panels as visiting judges.

The FJC database contains the party of the appointing president for each judge; we augment that measure of judge ideology with three standard ideology measures. GHP scores (Giles, Hettinger and Peppers 2001) place judges on the DW-NOMINATE scale via the ideology of their appointing president and home-state senators.[12] CF/DIME scores (Bonica and Sen 2017; 2024) estimate ideology from campaign contribution patterns (coverage starts in 1979). Finally, JuDJIS scores (Cope 2026) are dynamic expert-sourced measures of judicial ideology derived from text analysis of third-party evaluations, estimated annually (coverage starts in 1990); we attach them at the judge-case level so each judge receives the score appropriate to the date of a given case.

We link each case to the judge data through a multi-stage fuzzy matching algorithm (see Appendix A.1.3). Of the 512,263 cases in the clean CAP-CL dataset, 442,753 are from one of the circuit courts; the remainder include 10,244 cases from the Federal Circuit and 59,266 cases from other federal courts (e.g., the U.S. Court of Claims or the U.S. Emergency Court of Appeals) published in the Federal Reporter before their decisions moved to specialized reporters or before the courts themselves were abolished. Of the 442,753 appellate cases, our pipeline detects exactly three valid judge names—a standard appellate panel—in 420,659 (95.0%) of them; the remaining 22,094 (5.0%) are split between 8,321 (1.9%) en banc panels

[12] Epstein et al.'s (2007) "Judicial Common Space" scores extend the GHP scores through 2024 as of this writing (Epstein et al. 2024); we independently derive these scores using Voteview data (Lewis et al. 2026), extending coverage through 2025.

Variable	All cases (1,261,977 slots) n	%	Cases with dissent(s) (103,926 slots) n	%	Cases with concurrence(s) (75,174 slots) n	%	1971–2025 (844,524 slots) n	%
Appointment								
Party of appt. president	1,258,885	99.8	103,741	99.8	75,048	99.8	842,340	99.7
ABA rating	707,104	56.0	66,116	63.6	52,715	70.1	686,383	81.3
Demographics								
Gender	1,258,826	99.8	103,734	99.8	75,044	99.8	842,281	99.7
Race/ethnicity	1,258,708	99.7	103,730	99.8	75,040	99.8	842,173	99.7
Birth date (age)	1,233,912	97.8	100,794	97.0	73,147	97.3	830,747	98.4
Birth state	1,257,774	99.7	103,686	99.8	75,019	99.8	842,173	99.7
Law school	1,223,816	97.0	101,571	97.7	74,126	98.6	841,703	99.7
All demographics	1,198,554	95.0	98,609	94.9	72,218	96.1	830,276	98.3
Ideology scores								
GHP score	1,242,955	98.5	102,468	98.6	74,278	98.8	838,795	99.3
CF/DIME score	822,903	65.2	74,266	71.5	59,224	78.8	785,539	93.0
JuDJIS score	425,382	33.7	40,499	39.0	33,999	45.2	425,382	50.4

ABA ratings began in the 1950s; CF/DIME scores begin in 1979 and JuDJIS scores begin in 1990.

Table 2: Judge-slot level coverage. The table depicts data coverage for cases with exactly three valid judge names, treating both our failure to match a judge name to metadata and an absent variable in the metadata for a matched judge as missing data.

with more than three judges and 13,773 (3.1%) panels with fewer than three judges.[13] In 413,870 (98.4%) of standard-panel cases, we are able to match all judge names and all opinion author names to the FJC database.[14]

Table 2 summarizes judge-slot level data coverage of cases with exactly three valid judges. Rows contain judge-level characteristics; columns change the sample from all cases, to cases with dissents, to cases with concurrences, to cases from the IDB coverage period (1971–2025). Every variable except the time-restricted ABA (American Bar Association) rating and ideology scores is present for over 96% of judge slots, and all five demographic variables are jointly present for over 94%. Within the appropriate time range, CF/DIME score coverage (1971–2025) is 93.0%, ABA rating coverage (1971–2025) is 81.3% across 1971–2025, and JuDJIS score (1990–2025) is 80.4%.

[13]Due to the thoroughness of our judge matching pipeline (as described in Appendix A.1.3), we have strong evidence that these are truly non-standard panels. Regardless, their small footprint means our results in Section 6 are robust to their inclusion or omission, and we produce LLM codings for their opinions as well.

[14]This total includes detected per curiam cases, which are not signed by any particular judge.

4 LLM Coding Approach

Access to summary information about every published case—e.g., issue area, litigant types, and judges—is valuable in its own right. But the most prominent gap in existing data is the ideological direction of the disposition—whether an appellate ruling is liberal, conservative, or neither. This variable is often central to testing theories of judicial decision-making. Our goal is to produce direction codings that follow the existing judicial politics literature rather than lay perceptions (or an LLM's version of those perceptions, internalized from broad training data). To this end, we focus on three key desiderata. First, codings should be *rules-based*: driven by human-understandable principles and best practices that can be written down and explained to both LLMs and human users. Second, codings should be *replicable*: to account for developments in LLM capacities, coding standards, or use cases, our pipeline should not depend on closed-source models, hidden hyperparameters, or other runtime features. Third, our codings should be *reliable*: similar cases should be coded consistently, and LLM stochasticity should be minimized. Before discussing these principles in detail, we first provide a technical summary of the LLM tools used.

4.1 Technical Details

All case-level coding is produced by Qwen3-235B-A22B-Instruct-2507 (Yang et al. 2025), an open-weights mixture-of-experts language model released by Alibaba in the summer of 2025, which was then among the leading open-weight models for structured text generation. The model has 235 billion total parameters (22 billion activated per token) and a context window of 128,000 tokens, large enough to fit all but a handful (fewer than 100) of opinions in our sample without truncation. We use the instruction-tuned variant of the model with no fine-tuning, setting temperature to zero and using a fixed seed for reproducibility. As we will describe in Section 4.3, our LLM labeling requires four passes (two each for separate Songer and SCDB tracks); each case is run as four separate API calls with no persistent memory across passes or cases.

We run the pipeline on two backends. The main text uses DeepInfra's OpenAI-compatible API; total cost is approximately \$3,000 (covering about 35 billion input tokens and 1 billion output tokens), and the full run takes approximately one week. To demonstrate that these results are achievable without third-party services, we also use Princeton's HPC infrastructure as a second backend. Details of this alternative run are in Appendix A.4.1.

On both backends, intermediate results are checkpointed so that failures can be resumed without recomputation. On the DeepInfra backend, about 6% of cases fail to parse on at least one pass. More than half of those failures are due to API errors, which occur randomly depending on API load. The remainder are due to unparseable LLM output, typically due to the model emitting excessively long justifications without fully coding the required variables.[15] Both types of errors are resolved by re-querying (with the same prompt) until a valid response is produced;[16] about 0.6% of cases needed more than one re-try. Further details of the LLM coding approach, including more complete descriptions of each prompt and their interactions, are in Appendix A.3. In Appendix A.4, we describe variations on pipeline structure or prompting strategy used as robustness checks. In particular, we show that a few-shot strategy using curated examples does not improve validation accuracy.

4.2 Desiderata Overview

To address *rules-based* coding, we use detailed codebooks from the Supreme Court Database (SCDB) and the Songer Database (Songer). Both codebooks' ideological direction codings operate as a lookup based on issue area and winning side—e.g., in a federal taxation case, a win for the taxpayer should be coded as conservative regardless of taxpayer identity. This approach allows us to task the LLM with extracting factual classifications: issue area (e.g., economics or civil rights), litigant types (e.g., a taxpayer, a business, or a union), and outcome

[15]Longer cases and cases from the 2020s are more likely to fail—around 10% rather than the 6% average.

[16]Although we minimize LLM stochasticity (see next section), LLM responses are still not fully deterministic, so re-querying can produce a valid response when the original failed.

(e.g., reversing or affirming the lower court). We thus avoid well-known LLM issues with anchoring effects, political biases based on pretraining, and reluctance to answer sensitive questions (Santurkar et al. 2023, Röttger et al. 2024). For example, if an environmental group challenges a government regulation, directly asking the LLM whether the outcome is liberal or conservative produces inconsistent responses, since both "an environmental group wins" and "government regulation is upheld" are colloquially liberal.

For *replicability*, we use an open-weight LLM (see the previous section for details), and document all prompting code in the project repository.[17] The model performed comparably on spot-checks to closed-source options (OpenAI's GPT 5.1 and Anthropic's Claude Sonnet 4.6) using samples of about 1,000 cases, at a fraction of the cost.[18] Because it has open weights and fixed hyperparameters, our codings are directly reproducible. Perhaps more importantly, our pipeline is fully model-agnostic: other researchers can re-run it with whichever updated model fits their budget and goals.

Lastly, for *reliability*, we run the LLM separately on SCDB and Songer coding tracks to avoid cross-codebook confusion and produce intra-LLM consistency checks (see Section 5.2). We set temperature to zero and fix a seed to minimize LLM stochasticity (though, as mentioned previously, it cannot be eliminated). Across multiple providers, including a large Princeton HPC run (Appendix A.4.1), individual variables like case disposition, broad issue area, and ideological direction match exactly around 80–85% of the time, but overall accuracy varies by no more than 2–3% with no systematic bias for any particular provider, suggesting no material impact on aggregate performance.

4.3 Direction Coding in Detail

Our approach to ideological direction emphasizes the mechanical nature of the codebook-prescribed task. First, we split coding into two parallel tracks. The SCDB track runs fully

[17] This repository is currently private; please contact the authors for access. We will make it public upon publication of this paper.

[18] Approximate cost using GPT 5.1 would be about $50,000; for Claude Sonnet 4.6, about $120,000.

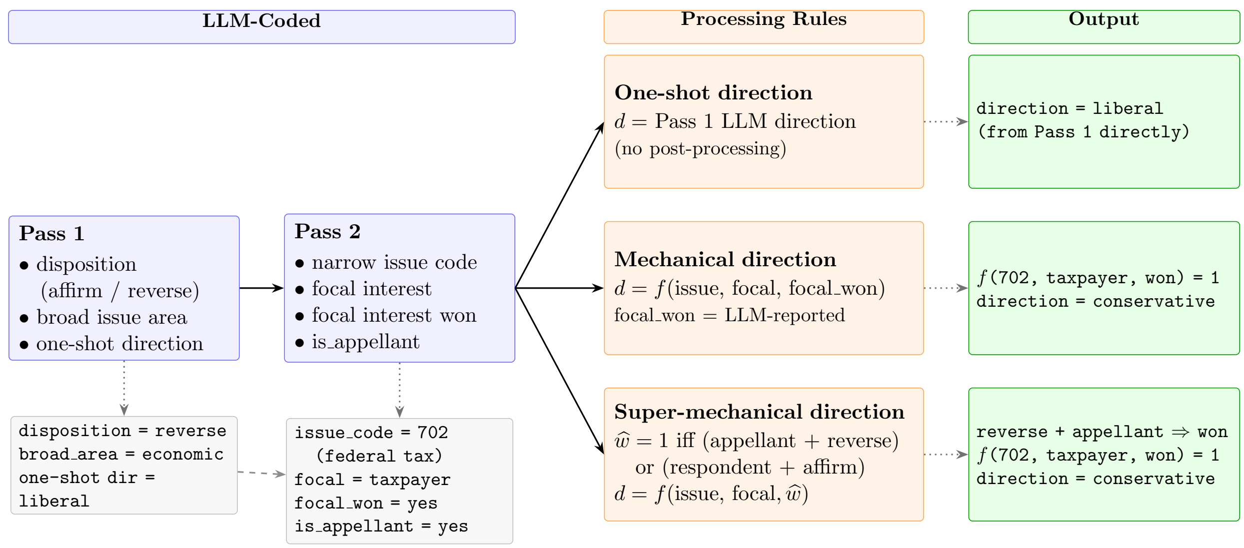

Figure 1: Ideological direction coding pipeline. Each track (SCDB, Songer) runs two LLM passes that produce the LLM-coded signals on the left; three post-processing rules (one-shot / mechanical / super-mechanical) transform those signals into a final direction code. In the main text, we focus on super-mechanical direction, which performs best on Songer-track validation.

independently; the Songer track receives natural language descriptions of the appellant and respondent as extracted by the SCDB first pass, along with the extracted disposition and winning side, to ensure that both tracks are working with the same case facts. However, it does not see the SCDB track's categories for the litigants, or any of the substantive SCDB first-pass variables like ideological direction.

Each track's first pass independently codes basic features such as issue and litigant type. The first pass prompt also uses a detailed lookup table derived from the codebook's instructions to prompt the LLM to code an ideological direction. A short set of worked examples highlights more challenging portions of the table; these examples were generated after preliminary testing on volume 1 of the *Federal Reporter, 3rd Series*, which contains 226 cases from the mid-1990s. We intentionally avoid the vocabulary of liberal versus conservative and ask the LLM to provide a flowchart showing how it used the lookup table to determine the direction; this flowchart acts as a human-readable reasoning trace. In that same testing, the lack of explicit ideological vocabulary and mandatory flowchart improved LLM performance, reducing cases where the LLM's intuitions overrode codebook definitions.

We use a second pass within each track to refine summary information into finer categories (e.g., 200+ fine-grained issue area codes from eight broad ones) and further mechanize direction coding. The LLM is re-initialized with no system memory of the first pass and seeded with case text, factual information from the first pass (court, date decided, case name, litigants, and outcome), and a custom prompt for each broad issue area featuring a pre-specified list of potential focal interests. For example, a case coded as Songer area 1 (Criminal) automatically provides "criminal defendant" as the focal interest, since according to the codebook all criminal cases have a liberal outcome if and only if the criminal defendant wins. The focal interest list for a case coded as Songer area 7 (Economic) includes "taxpayer," "injured party," and "[economic] underdog," among others, because ideological direction depends on the specific type of economic case. The LLM is asked to choose the most relevant focal interest; code whether that focal interest is the appellant or the respondent; and code whether that focal interest won, lost, or neither. In a post-processing script, we convert these codings to a "mechanical direction" following the track's codebook; e.g., on the Songer track, if the focal interest is "taxpayer" and the outcome is "win," the Songer codebook labels this case as conservative. We also produce a "super-mechanical direction" based on the LLM-coded case disposition; if the focal interest is "taxpayer" and the taxpayer appealed the lower court's decision, a disposition of "reverse" implies that they won and the appropriate code is "conservative." This approach is illustrated in Figure 1.

Of the three direction codings—one-shot from the first pass, mechanical using focal interest and focal win/loss, and super-mechanical using focal interest and disposition—super-mechanical is the best-performing when validating the Songer track, while mechanical is best when validating the SCDB track. Both perform about 10% better than one-shot direction, suggesting that naïve prompting remains a weak approach to case annotation even when that prompt is carefully designed. Since the Songer codebook is our main focus, in the main text we present results using super-mechanical direction coding; the other two methods are discussed in appendices A.5.1 and A.5.2.

5 Validation of LLM Results

As discussed in Section 3.2, we use two external sources for validation, broken down into three validation approaches. First, we treat Songer human codings as ground truth and compare against the LLM outputs on matched cases, treating every LLM discrepancy as an error (Section 5.1).[19] Doing so is straightforward for the LLM's Songer track, where LLM output codes are drawn directly from the Songer codebook; for the SCDB track, we construct a SCDB-to-Songer crosswalk that allows validation of most SCDB variables against Songer equivalents. Second, we compare the level of LLM-to-human disagreement to the human-to-human inter-coder reliability metrics reported in the original Songer data (Section 5.2). We also compute cross-track consistency between the LLM's SCDB and Songer tracks using the same metrics. Finally, we use the IDB to compare LLM performance with that of Songer human coders on an independent administrative dataset (Section 5.3), providing out-of-sample evidence that LLM performance is human-level.

5.1 Direct Validation against Songer

We first treat the approximately 23,000 matched cases from Songer as "ground truth" (GT)—any LLM deviation is considered an error—and implement this comparison using three closely related procedures depending on the precise form of LLM output.

For the Songer track, we compare output directly (excluding "not ascertained" codes). Almost all Songer-track variables are compared via exact match, with two types of exceptions. First, while about 85% of cases have a single GT issue area, the remainder are "two-issue cases" with a second issue area and corresponding direction code that Songer describes as having equal importance. Since our LLM outputs only a single most important issue area and its corresponding direction, we validate by "set membership," a one-to-many

[19] As described in the prior section, the majority of prompt development used only a negligible portion of the Songer dataset (226 cases from one particular volume of the Federal Reporter), and about three-quarters of the Songer dataset was never used for any prompt adjustments.

Variable	Songer	SCDB	N
Issue area			
Broad area	90.2%	87.7%	22,666
Subcode	54.1%	42.3%[a]	22,656
Ideological direction[b]			
3-class (cons/mixed/lib)	81.9%	—	20,748
2-class (drop GT mixed)	82.9%	74.2%	19,097
Parties			
Appellant type	85.7%	84.7%	22,910
Respondent type	84.4%	83.9%	22,903
Other			
Disposition (exact code)	87.0%	87.0%	22,850
Disposition (winning side)[c]	92.9%	92.9%	22,850
Appealed from	55.8%	—	22,912
Threshold issues	76.9%	—	22,888

[a] SCDB issue area subcode N is smaller (15,529) because the crosswalk has no entries for several Songer GT subcodes, notably criminal ones (SCDB subcodes are procedural issues like Miranda rights while Songer subcodes are offense types like murder).

[b] The 3-class SCDB cell is omitted because SCDB has no "mixed" code, making every Songer GT = 2 case an error; SCDB also hardcodes the Interstate Relations and Private Action broad issue areas to "unspecifiable," reducing accuracy by about 6 percentage points compared to dropping those cases.

[c] "Winning side" collapses to the four codes which feed super-mechanical direction: appellant won (reversed, rev. + remanded, vacated + rem., vac.), respondent won (affirmed, dismissed), mixed (aff.-in-part), and neither (no merits decision).

Table 3: Headline validation: LLM vs. Songer hand-coded. The table summarizes the accuracy of our LLM coding for major variables, compared directly to Songer ground truth (first column) or using the SCDB-to-Songer crosswalk (second column). We use super-mechanical ideological direction for both tracks.

mapping where the LLM is correct if it matches *either* Songer code.[20] Separately, uncommon binary variables (e.g., whether or not standing is at issue) are evaluated by false-positive and false-negative rates since exact-match accuracy would be dominated by true negatives.

For the SCDB track, where no hand-coded appellate data exists, we construct SCDB-to-Songer crosswalks: thorough mappings from SCDB variables to their Songer equivalents. Issue area, direction, appellant/respondent categories, disposition, and other variables are each mapped through crosswalks. Some SCDB subcodes are validated by the many-to-one or one-to-many set membership approach. For example, SCDB issue code 80030, bankruptcy, may

[20]In Appendix A.5.1, we discuss approaches that require the LLM's direction to match the "more appropriate" issue area's direction.

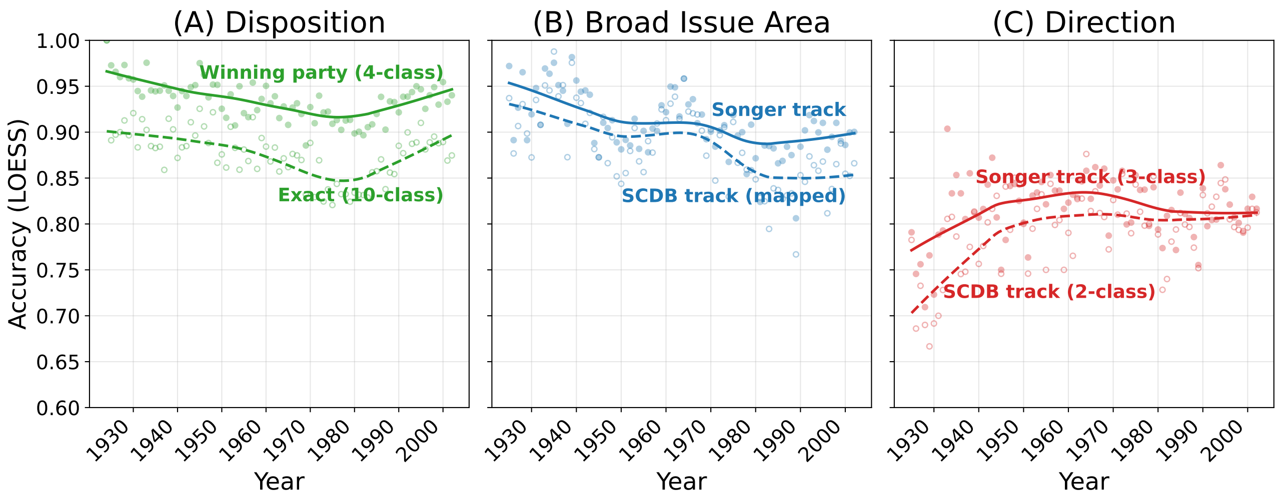

Figure 2: Per-year accuracy of three headline LLM-coded variables (case topic, disposition, direction) against Songer ground truth, by decision year. Lines are LOESS smooths (span 0.5); points are raw yearly accuracies.

correspond to Songer issue codes 741 (Chapter 7 bankruptcy), 742 (Chapter 11 bankruptcy), or 743 (other bankruptcy), and is valid if any of those three are the Songer GT code.

Finally, recall that we created three novel variables—focal interest, focal-won, and focal-was-appellant—to generate mechanical and super-mechanical ideological direction (see Section 4.3). We validate all three by deriving their implied values from the GT issue area, subcode, and direction.[21] For the SCDB track, we define valid SCDB focal interests for each Songer issue area subcode—e.g., Songer subcode 261 (Juveniles) allows SCDB focal interest "child"—then apply the same implied-value logic. For two-issue cases, these mappings are applied to the GT issue area matching the LLM's, if any.

Table 3 summarizes headline accuracy for major variables.[22] For most of the categories, including broad area, ideological direction, disposition, and litigant type, we achieve upwards of 80% accuracy. Issue area subcodes reach only about 50% accuracy, but this is unsurprising; there are more than 200 such categories in Songer and SCDB, some of which are near-identical to others or have only a handful of cases to validate against.

Figure 2 shows the three headline accuracies (case topic, disposition, and direction) over time. Points depict accuracy by year, while lines are LOESS smooths. Exact disposition

[21]Descriptions of these mappings are in Appendix A.3.6.
[22]Validation tables for variables not included here are in Appendix A.5.

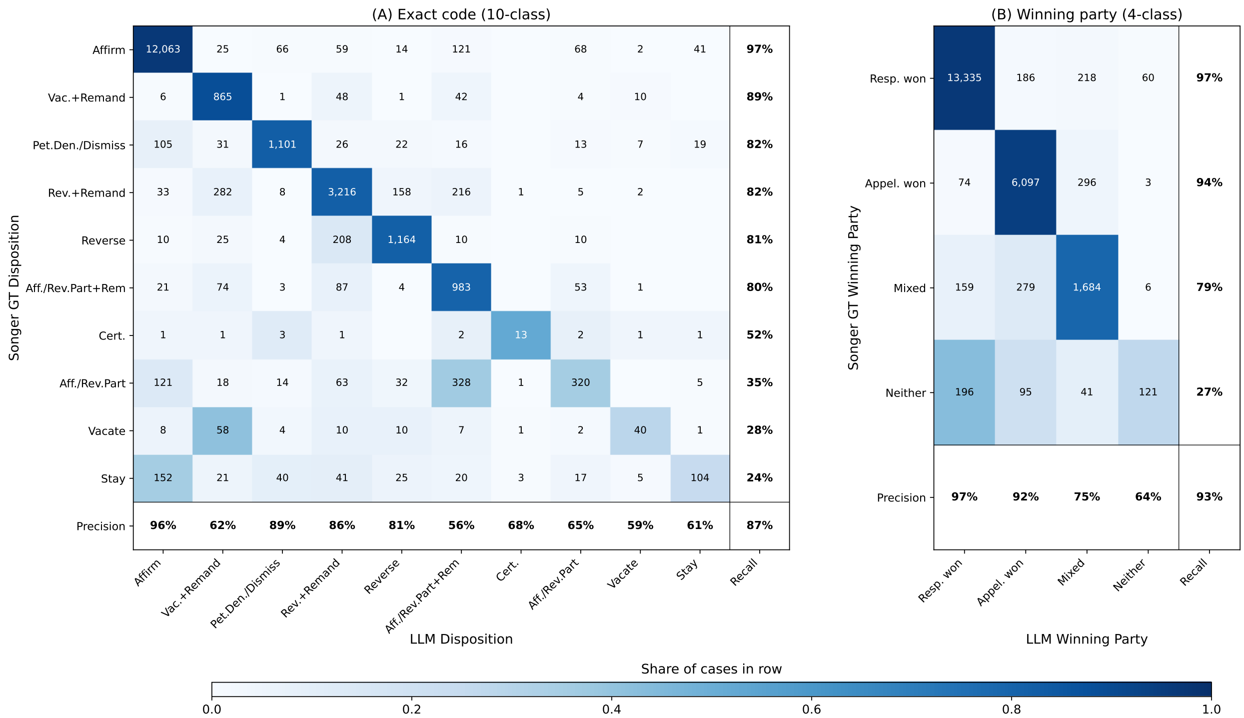

Figure 3: Disposition confusion matrices. Rows are Songer GT, columns are LLM output; cell shading is row-normalized. Per-row recall (right edge) and per-column precision (bottom edge) are bolded. Panel (A) shows exact codes (overall accuracy 87.0%). Panel (B) collapses to the winning side categorization that feeds super-mechanical direction (overall accuracy 92.9%).

accuracy begins around 90%, dips by about 5%, and returns to 90% by 2000; winning side is consistently about 5 percentage points better. Issue area accuracy declines by about 7 percentage points from the start of the series, then stabilizes around 1980, reflecting mid-century increases in case complexity. Ideological direction improves sharply before the 1940s, then stabilizes around 80%. Despite its mechanical dependence on issue area, this stability is possible because issue area errors do not automatically imply direction errors. For example, the LLM may incorrectly code an economics case about unions as a labor case, but if the underlying codebook instructions are similar, they may produce the same direction code.

Since ideological direction, our primary variable of interest, depends on disposition and issue area, we examine each in detail before turning to direction itself. Figure 3 presents confusion matrices for case disposition. Each row of a confusion matrix shows the count of different LLM labels for a given GT coding, with density shading within each row. Dark

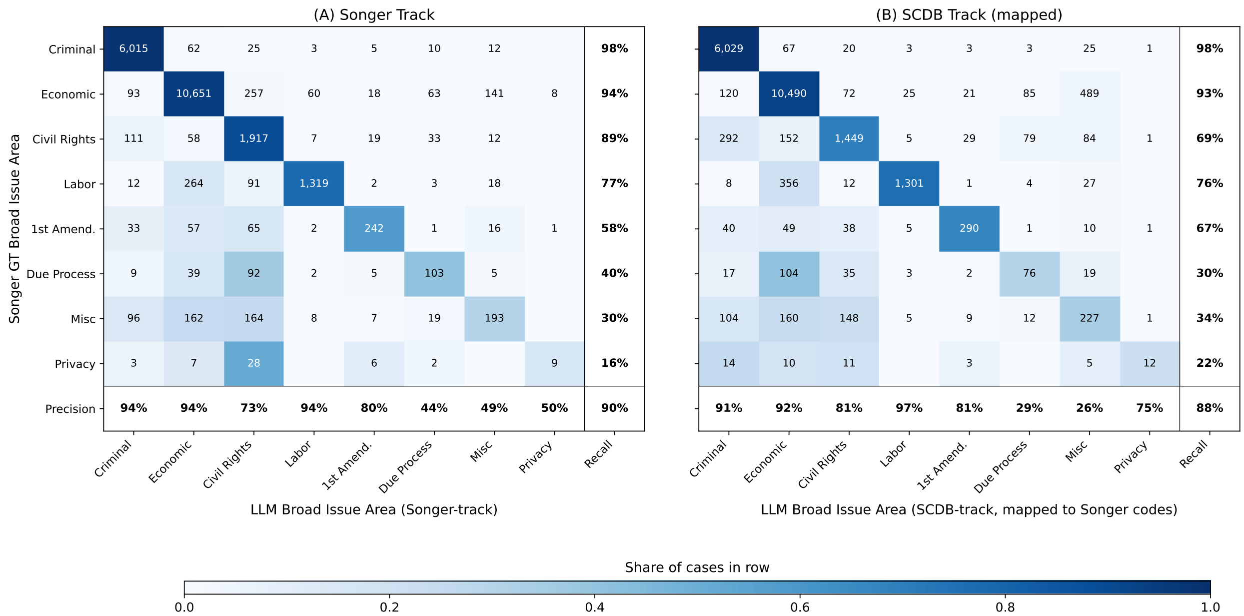

Figure 4: Issue area confusion matrices. Rows are Songer ground truth, columns are LLM output; cell shading is row-normalized. Panel (A) is Songer track, panel (B) is SCDB track mapped to Songer codes. Criminal and economic cases dominate and are classified accurately (>90% recall); due process, miscellaneous, and privacy cases are rarer and more challenging (<50% recall).

diagonal cells thus capture a large share of correct codings (GT matching LLM output). Rows are sorted by recall (the right edge), the share of correctly-classified GT cases in each row. Precision (the bottom edge) is the column-wise share of LLM classifications matching GT. Panel (A) compares the ten exact Songer disposition codes, with high accuracy on the most frequent categories. Weak recall (below 80%) concentrates on procedural outcomes (e.g., granting of cert) and "affirmed in part and reversed in part," which the LLM often codes as indicating a remand as well—suggesting ambiguity in complex dispositions. Panel (B) collapses to the four-class winning side categorization that feeds super-mechanical direction. Overall agreement rises above 90% since most exact-code errors are within the same winning side (e.g., reversed vs. reversed-and-remanded). When mistakes occur, the LLM tends to favor the appellant on mixed dispositions and the respondent on procedural ones.

Figure 4 shows broad issue area confusion matrices for both tracks, with rows in both panels ordered by Songer-track recall for easy comparison. Two-issue cases are placed in their sole LLM-matching row if possible and their first code's row otherwise. Songer-track

Issue area	N	Right, total	Right, all four	First wrong input: Wrong area	Wrong focal interest	Wrong winner	Wrong appellant
Criminal	6,082	92.9	89.4	1.8 (1.1)	0.2 (0.1)	4.5 (2.3)	4.1
Civil Rights	2,074	84.6	70.8	11.4 (9.5)	1.7 (0.9)	6.6 (3.5)	9.5
Due Process	267	79.8	35.2	55.8 (42.7)	1.5 (1.1)	1.5 (0.7)	6.0
Economic	9,610	78.9	63.3	6.0 (3.2)	16.3 (10.3)	4.2 (2.1)	10.1
1st Amend.	411	71.5	45.5	41.1 (23.8)	0.5 (0.0)	4.1 (2.2)	8.8
Privacy	59	71.2	10.2	78.0 (57.6)	3.4 (3.4)	3.4 (0.0)	5.1
Labor	1,680	67.4	38.3	21.8 (14.9)	2.8 (1.6)	20.6 (12.6)	16.4
Misc.	499	54.9	20.6	61.5 (26.3)	11.8 (7.2)	1.0 (0.8)	5.0
Overall	20,684	81.9	67.8	9.5 (5.8)	8.4 (5.2)	5.7 (3.1)	8.6

Table 4: Super-mechanical direction error decomposition, Songer track. Rows are ordered by total direction accuracy. "First wrong input" columns are mutually exclusive, with each case sorted by first wrong input in the order of broad issue area → focal interest → winning side → focal-was-appellant. Cells show total % (within-cell correct-direction %), but "Wrong appellant" has no parenthetical since right inputs with wrong appellant identity *necessarily* produces wrong direction. By construction, the four non-parenthetical wrong input columns plus "Right, all four" sum to 100%; the three correct-direction parentheticals plus "Right, all four" equal "Right, total." The denominator excludes cases with one or more GT or LLM inputs missing, leading to slight differences with Table 3 in the Overall row.

accuracy is highest in the most numerous issue areas (criminal, economic, civil rights, and labor). SCDB-track is similar overall, but with notably better recall on First Amendment cases (by about 10 percentage points) and worse recall on civil rights and due process cases (69% vs. 89% Songer-track and 30% vs. 40% Songer-track, respectively).

Next, we decompose ideological direction errors into errors in issue area, focal interest, winning side, or focal-interest-was-appellant. When all four inputs are correct, direction is necessarily correct; even when any one is wrong, direction may still be correct (as in the example of an economic case mis-classified as a labor case but subject to similar codebook rules). Tables 4 and 5 report this decomposition.

The tables reveal three distinct error patterns. First, broad issue areas with fewer than 500 GT cases—Due Process, First Amendment, Privacy, and Miscellaneous—show significant rates of wrong-area classifications (all but First Amendment above 50%). Many of these still lead to a correct direction, likely because a related issue area was chosen. SCDB-track Civil Rights also has a high wrong area rate, probably because the SCDB codebook

Issue area	N	Right, total	Right, all four	First wrong input			
				Wrong area	Wrong focal interest	Wrong winner	Wrong appellant
Criminal	5,674	94.4	91.7	1.8 (1.0)	0.4 (0.3)	2.8 (1.4)	3.3
Civil Rights	1,828	83.9	55.7	30.9 (24.9)	2.2 (1.5)	3.8 (1.8)	7.3
1st Amend.	342	77.5	54.4	28.9 (20.2)	0.9 (0.9)	4.1 (2.0)	11.7
Economic	7,590	73.7	62.0	8.2 (4.0)	12.6 (6.0)	3.9 (1.8)	13.3
Due Process	246	68.7	26.0	66.3 (39.8)	3.3 (2.4)	0.8 (0.4)	3.7
Labor	1,472	65.6	35.8	22.1 (17.2)	1.4 (1.3)	20.2 (11.3)	20.5
Privacy	48	60.4	12.5	75.0 (43.8)	4.2 (4.2)	2.1 (0.0)	6.2
Misc.	470	49.8	16.2	61.9 (24.7)	15.7 (8.3)	1.9 (0.6)	4.3
Overall	17,672	80.1	66.7	12.5 (7.8)	6.4 (3.2)	4.8 (2.4)	9.6

Table 5: Super-mechanical direction error decomposition, SCDB track (crosswalked). We use the same construction as Table 4, but rows are independently ordered by their own "Total right." Focal interest correctness uses the codebook map from Songer issue area to valid SCDB focal interests. As in previous analyses of SCDB direction, we drop GT direction = 2 since SCDB has no "mixed" code. Since SCDB issue areas 11 and 14 are hardcoded as "unspecifiable," they have no focal interest and are also dropped, raising accuracy compared to Table 3 where they were included as automatic mistakes. Additionally, as in Table 4, the handful of cases with one or more GT or LLM inputs missing are also excluded, leading to further differences with Table 3 in the Overall row.

treats Criminal as a procedural issue area rather than a substantive one (motivated by its focus on the Supreme Court). Second, the two most diverse issue areas—Economic and Miscellaneous—have about 10-15% of cases with focal interest errors. Seeing this step as the dominant error is unsurprising since the former covers anything from taxes (focal interest: "taxpayer") to torts (focal interest: "injured") and the latter is a catch-all including diverse topics like federalism, Native American law, and international law. Finally, Labor is unique in having a large rate of wrong winner choice; many labor cases can have the same focal interest on opposite sides of the codebook's rules (e.g., a union being sued by an individual member vs. defending workers against management). This difficulty in choosing a side is also reflected in the higher rate of wrong appellant choice for labor cases. Overall, though, the vast majority of errors concentrate on the broad issue area, which supports our two-pass approach to prompt design: the second-pass prompt, focusing on within-issue-area fact extraction, is highly effective.

Variable	Goodman–Kruskal's γ			Kendall's τ-b		
	Songer ICR	LLM vs. GT	Inter-LLM	Songer ICR	LLM vs. GT	Inter-LLM
Broad issue area	0.98	0.90 (0.91)	0.93	0.97	0.84 (0.86)	0.88
Issue subcode[a]	0.95	0.78 (0.79)	0.60	0.95	0.77 (0.78)	0.58
Direction[b]	0.94	0.86 (0.88)	0.59	0.89	0.67 (0.70)	0.47
Disposition[c]	0.93	0.88	0.96	0.90	0.84	0.81
Appellant type	0.97	0.87	0.95	0.94	0.77	0.91
Respondent type	0.98	0.80	0.91	0.98	0.74	0.88
Appealed from	0.92	0.66	—	0.87	0.54	—
Threshold issues	0.96	0.87	—	0.93	0.75	—
Constitutionality	0.93	0.96	—	0.53	0.42	—

[a] Inter-LLM validation is computed only for cases where each SCDB subcode maps to exactly one Songer subcode (6,761 cases).
[b] LLM vs. GT and Inter-LLM use super-mechanical direction. Inter-LLM drops cases where GT = 2, since SCDB has no mixed code, and scores cases in SCDB issue areas 11 and 14 (hardcoded as "unspecifiable") as disagreements.
[c] SCDB and Songer tracks share the same disposition, but the SCDB first pass outputs both disposition and winning side. We use our standard crosswalk from disposition to winning side for the Inter-LLM columns, which in this case measure within-track, within-pass consistency.

Table 6: Inter-coder reliability (ICR). Songer ICR comes from the codebook's 250-case reliability sample. LLM-vs-GT compares LLM coding to Songer ground truth. Inter-LLM compares the LLM's Songer-track and SCDB-track outputs on the same cases, and is only available for variables coded in both tracks.

5.2 Inter-Coder Reliability

The previous section set aside the possibility of errors in the published Songer codings. However, the Songer codebook provides a 250-case reliability sample documenting how often multiple human coders agreed on case labels.[23] We compare three types of inter-coder reliability (ICR): (1) Songer's reported human-to-human value; (2) LLM vs. Songer GT; and (3) Songer-track vs. SCDB-track LLM outputs.[24] For two-issue cases, the Songer ICR sample reports agreement on each code separately, so the "LLM vs. GT" matches that approach by comparing to the first code only, with the either-code value in parentheses.

Table 6 reports Goodman–Kruskal's γ (symmetric) and Kendall's τ-b (penalizes ties, more conservative) across all three comparisons. Songer human coders show high agreement with each other: γ is at least 0.92 for all nine variables in Column 1. In Column 2, LLM-vs.-

[23] That sample's selection process, complete coder outputs, and choice of published value are not documented. We treat comparisons to the IDB in the next section as a stronger independent benchmark.

[24] For ideological direction, the one-shot and mechanical variants provide additional internal-consistency checks; other variables are coded once per case.

human agreement is slightly below that ceiling for issue area (0.08 lower), ideological direction (0.08 lower), disposition (0.05 lower), and appellant type (0.10 lower), but meaningfully lower for respondent type (0.18 lower). τ-b gaps are generally larger but follow the same pattern; in particular, the human-to-human vs. human-to-LLM gap for ideological direction increases to 0.22. More disagreement on ideological direction is expected given its compositional structure. The either-match credit for two-issue cases raises LLM vs. GT measures slightly, but does not close the gap to Songer ICR.

Column 3 addresses our goal of reliable coding by checking whether our Songer and SCDB tracks, run independently on the same cases, produce the same (crosswalked) codings. Agreement is high where LLM outputs are closest to case facts (disposition and litigant types, all over 0.90 γ and 0.80 τ-b) or the codebooks have similar definitions (broad issue area, 0.93 γ / 0.88 τ-b) but drops sharply for direction (0.59 γ / 0.47 τ-b). These values reflect both codebook differences (like the hardcoded SCDB issue areas) and direction's compositional structure: the parallel tracks can disagree at any stage of the chain, and those disagreements compound. Other than issue subcode and ideological direction, the inter-LLM values close most of the gap to the Songer ICR scores.

5.3 Validation using the IDB

Our most conservative approach to Songer errors uses independent administrative data from the IDB as ground truth.[25] Since the IDB uses its own coding schemes, we validate via set-membership crosswalks—each IDB code maps to a set of valid Songer codes (details in Appendix A.6.1). This crosswalk is applied identically to our LLM outputs and the published Songer codings. For two-issue cases, *either* Songer code may match, giving Songer a small but principled advantage over our LLM's single output.

Table 7 reports the results of this comparison.[26] The "LLM" and "Songer" columns

[25] As noted in sections 3.1 and 3.2, the IDB is not error-free but is unlikely to be substantially biased in favor of either source.

[26] Additional validation results using the IDB are in Appendix A.6.

Check	LLM	Songer	N (LLM)
Issue area (many-to-one)			
Nature of suit to Songer issue area	87.0%	85.6%	175,244
Nature of suit to SCDB issue area	78.0%	85.8%	175,244
Criminal appeal to Songer criminal issue	99.2%	96.5%	72,616
Criminal appeal to SCDB criminal issue	99.1%	—	72,616
Agency appeal to Songer issue area	57.2%	83.2%	25,207
Agency appeal to SCDB issue area	59.8%	83.3%	25,207
Disposition (one-to-many)			
Overall disposition agreement	88.9%	88.3%	278,960
Litigant identification			
Federal government as appellant	86.8%	83.9%	13,278
Federal government as respondent	95.3%	94.1%	79,533
Pro se appellant coded as individual	97.1%	93.6%	8,480
Pro se respondent coded as individual	49.4%	33.3%	1,133
Other variables			
Appeal type to court appealed from	89.0%	96.6%	284,550
Jurisdictional basis to authority type[a]	79.5%	61.3%	175,617

[a] Songer ground truth restricted to $N = 2,846$ cases where the constitutional-basis or federal-law flags were coded; about 88% of Songer cases have these flags uncoded.

Table 7: LLM vs. Songer ground-truth accuracy against the IDB. For issue area, we follow the IDB in separating civil cases (validated via nature of suit crosswalks), criminal cases (a single broad issue area in both LLM tracks), and administrative cases (validated via agency appeal crosswalks). The remaining variables (i.e., dispositions, litigants, appeal types) span all three IDB case types.

compare pipeline and human accuracy for each crosswalked variable; "N (LLM)" is about twenty times larger than the corresponding number of IDB-matched Songer cases because of Songer's lower coverage. On most checks, the LLM matches or exceeds Songer performance. It is slightly stronger on the key ideological direction inputs: disposition (88.9% LLM vs. 88.3% Songer) and issue area (87.0% LLM vs. 85.6% Songer on civil cases; 99.2% Songer-track LLM/99.1% SCDB-track LLM vs. 96.5% Songer on criminal case identification). The LLM is also stronger at all four litigant identification tasks, with weakness for both the LLM and Songer on coding pro se respondents as individuals driven by a definitional mismatch on multi-litigant cases (both code the lead respondent only while the IDB codes *all* respondents). However, it trails Songer on nature of suit (NOS) $\rightarrow$ SCDB issue area (less precise than mapping to Songer issue areas), agency appeals (the LLM emphasizes the underlying claim—e.g. civil rights or labor—rather than the IDB's regulatory framing), and appeal type.

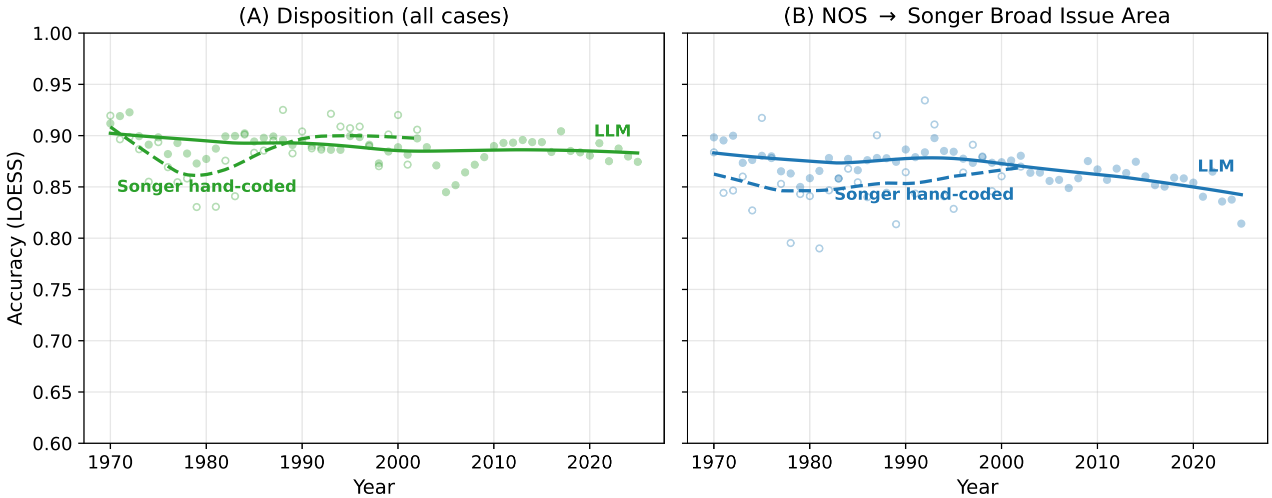

Figure 5: Per-year accuracy of LLM and Songer hand-coded variables against IDB ground truth, by year. Lines are LOESS smooths (span 0.5); points are yearly accuracies. Solid lines and points are the LLM results; dashed lines and open points are Songer hand-coded data. Panel (A), disposition, covers all cases in the IDB. Panel (B), issue area, covers IDB civil cases only.

Figure 5 compares performance over time on the two IDB-validatable direction inputs: disposition and issue area. LLM accuracy (solid LOESS lines) smoothly declines by a few percentage points, consistent with modern case complexity; Songer performance (dashed LOESS lines) dips and rebounds. The LLM stays above average Songer disposition accuracy even at its lowest point and above average Songer issue-area accuracy until the 2010s. This issue-area result is especially strong given that two-issue Songer cases get two matching chances to the LLM's one; using only the first Songer code reduces Songer accuracy near-uniformly by about two percentage points, leaving it below even the LLM's weakest performance. Given strong LLM performance when benchmarked against within-human variation or an external source of ground truth, our results suggest that a substantial share of LLM "errors" in Section 5.1 are better understood as reasonable differences of opinion.

6 Applications

Our LLM-coded universe of cases provides two broad benefits for downstream applications. First, an order of magnitude more data than existing hand-coded samples permits previously infeasible approaches, such as computing year-by-year effects. Second, our output variables are more fine-grained than existing proxies like administrative records and contribute new measurements of key features like ideological direction. In this section, we present two analyses of ideological decision-making on appellate courts to demonstrate both benefits.

We begin by extending work on panel effects, the well-documented phenomenon of individual judges on three-judge panels acting not just based on their own preferences, but also based on the characteristics and preferences of the other two judges on the panel. Our universe-level coverage allows year-by-year analyses and decompositions across majority and dissenting opinions. Next, we use our fine-grained litigant and ideological direction codings to clarify recent work on "pro-weak bias"—the tendency by Democratic-appointed judges to favor the weaker litigant in a broad subset of cases (Cohen 2025). This application incorporates all of our main LLM-coded variables—issue area, litigant type, disposition, and ideological direction—to varying degrees. Across both applications, the qualitative conclusions presented are robust to formal corrections for LLM measurement error; a thorough discussion of these corrections is in Appendix A.8.

6.1 Panel Effects

Standard appellate panels include three judges, and a strong body of evidence indicates that this group's composition can substantively alter case outcomes (Sunstein et al. 2006, Kastellec 2011, Beim and Kastellec 2014). This phenomenon of "panel effects" dates back to Revesz (1997), which found that adding a Democratic appointee to a panel shifts its overall voting behavior in a liberal direction (and analogously for Republican appointees). Related work documents analogous compositional effects for gender (Farhang and Wawro 2004, Boyd, Epstein and Martin 2010) and race (Cox and Miles 2008, Kastellec 2013), generally using hand-selected subsets of cases.

Studying panel effects across the full modern Courts of Appeals era was previously quite difficult: Songer lacks data outside of 1925–2002 and is too sparse to allow meaningful estimates within circuit $\times$ year cells (the level at which judges are effectively randomly assigned to panels). By contrast, our dataset contains a median of around 250 cases per cell compared to Songer's maximum of 60 in *any* cell, supporting comparisons of individual judges as their colleagues' characteristics vary. We focus on ideological voting, defining a judge's vote as *aligned* when their appointing party matches the vote's direction and misaligned otherwise. That is, alignment occurs when a Democratic appointee joins a liberal majority or dissents from a conservative one; vice versa for a Republican appointee.[27] Since the Songer codebook is the standard for appellate court data, we use the best-performing Songer-track ideological direction variable (super-mechanical direction) to code a vote's direction; in Appendix A.7, we replicate the analysis with the best-performing SCDB-track ideological direction variable (mechanical direction) and find highly similar results.

Let $Y_{ijct} = 1$ whenever, on case i (on circuit c in year t), judge j casts an aligned vote (0 otherwise). Define a "(party-)unanimous panel" as one with all Democratic or all Republican appointees (DDD or RRR); "mixed panels" have at least one of each (DDR or DRR). On mixed panels, we use separate indicators for judges of the majority party (D on DDR, R on DRR) and the minority party (R on DDR, D on DRR). Our baseline specification is

$$Y_{ijct} = \beta_1 \, \text{MajorityParty}_{ijct} + \beta_2 \, \text{MinorityParty}_{ijct} + \alpha_j + \gamma_{ct} + \varepsilon_{ijct},$$

where α_j is a judge fixed effect[28] and γ_{ct} is a circuit$\times$year fixed effect. The omitted category is unanimous panels, and standard errors are two-way clustered by circuit$\times$year and judge. The reported estimates are β_1 and β_2.

Table 8 reports the results of this analysis in Column (1) (we return to the model in Column (2) shortly). Consistent with the existing panel effects literature, judges on mixed panels are less likely to vote their ideology as compared to judges on unanimous panels.

[27] Our liberal/conservative definition comes from the Songer codebook, which is in line with modern left-vs.-right judicial ideology. Historically, Democratic and Republican appointees often disagreed along other dimensions; we capture only variation in ideological voting *as defined by Songer*.

[28] Judge fixed effects are a stronger control for idiosyncratic ideological voting than the usual battery of demographic controls, and are only feasible because of the scale of our data. The coefficients β_1 and β_2 can thus be interpreted as the effect of switching *each judge* from a unanimous to a majority-party or minority-party panel, averaged (with weights) over judges.

	Aligned vote rate	
	(1) Role only	(2) Role × alignment
Party-minority on mixed panel	−0.0498***	0.0268***
	(0.0033)	(0.0020)
Party-majority of mixed panel	−0.0242***	0.0065***
	(0.0023)	(0.0008)
Case direction aligned w/ party		0.9616***
		(0.0015)
Party-minority × Aligned case		−0.0243***
		(0.0026)
Party-majority × Aligned case		−0.0023
		(0.0012)
Baseline (omitted cell)	0.5663	0.0192
Observations (judge–cases)	1,220,661	1,220,661
of which on unanimous panels	393,492	393,492

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Standard errors in parentheses, two-way clustered by circuit×year and judge. Column 1's omitted cell is a judge on a party-unanimous panel; column 2's omitted cell is a judge on a party-unanimous panel hearing a misaligned case.

Table 8: Panel-role effects on aligned voting. The dependent variable equals one when a judge's vote is *aligned* with their appointing party (a Democratic appointee joining a liberal-direction majority or dissenting from a conservative one; Republican appointees with the converse), using Songer super-mechanical direction. Column (1) reports the basic role specification. Column (2) augments Column (1)'s specification by interacting with an indicator for case alignment.

The pooled effect across years shows that a party-majority judge on a mixed panel (e.g., a Republican appointee on a DRR panel) casts about 2.4 percentage points fewer ideologically aligned votes, or about a 4% decrease from the baseline level of 56.6% ideologically aligned votes on a unanimous panel. The effect on party-minority judges is about twice as large.

Panel A of Figure 6 estimates the same specification year-by-year across 1892–2025 and plots the predicted vote rates for the baseline (unanimous panels) as well as the two coefficients of interest (minority-mixed and majority-mixed) smoothed with a LOESS line. Prior to 1960, there was essentially no difference between judge roles, and the rate of ideologically aligned voting was around 50% (i.e., not correlated at all with Songer-defined ideology). If anything, the share of ideologically aligned voting slightly decreased from 1892 through 1960. From that point onward, we see the previously-established increase in ideologically aligned voting for all three judge roles, as well as the separation between them: judges vote their ideology most often on unanimous panels, then as the party majority on mixed panels, and

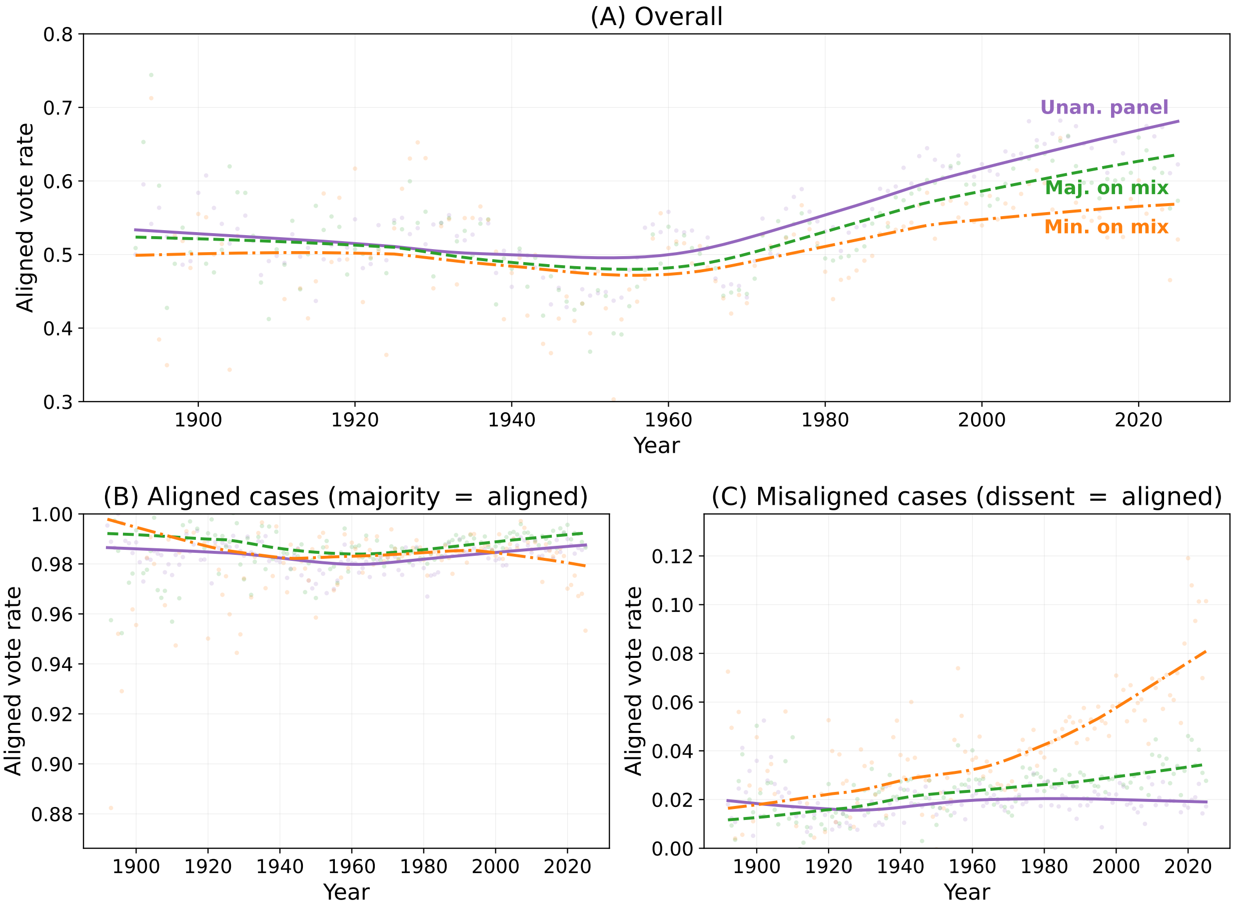

Figure 6: Aligned-vote rates over time. Each panel plots the estimated aligned-vote rate by year for judges on a party-unanimous panel (DDD or RRR; omitted baseline, purple), party-majority judges on mixed panels (R on DRR or D on DDR; green), and party-minority judges on mixed panels (orange). Lines are LOESS smooths (span 0.5). Panel (A) pools all cases. Panels (B) and (C) split by case alignment.

least often (though still more than before 1960) as the party minority on mixed panels.

Mechanically, an increase in aligned voting occurs either because judges dissent[29] less often from aligned majorities—e.g., Democratic-appointed judges used to dissent from liberal opinions for non-partisan reasons, but they no longer do so—or because judges dissent more often from misaligned majorities—e.g., Democratic-appointed judges are now more willing to vote their ideology against a conservative majority ruling.[30] To decompose these channels, we

[29]Dissents are far more common in published cases than in unpublished appellate cases, and publication itself may be endogenous to the presence of a dissent. This analysis should be taken as a decomposition of the causal effect for published cases established in the pooled sample (different panel types have different rates of ideological voting, and ideological voting on all panel types has increased over time).

[30]These shifts may come from changing *ideology* (Democratic-appointed judges more often support

interact the party-majority and party-minority indicators with case-level alignment. Similar to the vote-level specification, a case i is aligned relative to, for example, a Democratic-appointed judge j when it has a liberal outcome and misaligned when it has a conservative outcome. Under this decomposition, $Y_{ijct} = 1$ when judge j joins the majority disposition on aligned case i, or if judge j dissents on misaligned case i. The full specification is

$$\begin{aligned}Y_{ijct} = {} & \beta_1 \, \text{MajorityParty}_{ijct} + \beta_2 \, \text{MinorityParty}_{ijct} + \beta_3 \, \text{AlignedCase}_{ijct} \\ & + \beta_4 \, (\text{MajorityParty} \times \text{AlignedCase})_{ijct} + \beta_5 \, (\text{MinorityParty} \times \text{AlignedCase})_{ijct} \\ & + \alpha_j + \gamma_{ct} + \varepsilon_{ijct},\end{aligned}$$

with the same fixed effects and standard error clustering as before. The omitted category is now unanimous panels on *misaligned* cases; β_3 measures the increased share of aligned votes on unanimous panels with *aligned* cases. Column 2 of Table 8 reports these coefficients and the pooled estimates for each (role, alignment) cell, while Panels B and C of Figure 6 plot the year-by-year baseline-plus-coefficient lines for the aligned-case and misaligned-case subsamples respectively. The reported estimates are β_1, β_2, β_3, β_4, and β_5.

Column (2) shows that the panel-composition effect in Column (1) is driven by misaligned majorities. Focusing on that subset, judges rarely dissent on unanimous panels (only 2% of the time). On mixed-panel misaligned cases, majority-party judges dissent about 1/3 more than the baseline (just under 0.7 percentage points more) and minority-party judges more than double their dissent rate (about 2.6 percentage points more). For aligned cases, the difference between judge roles (given by the sum of corresponding coefficients) is less than 0.5 percentage points over the baseline of about 98%—essentially no panel effects at all.

Panels (B) and (C) of Figure 6 estimate the specification from Column (2) of Table 8 year-by-year with a LOESS line for smoothing, showing that misaligned majorities also drive the increase in ideological votes over time. On aligned cases, party-majority judges cast only

codebook-liberal policies) or changing *behavior* (Democratic-appointed judges always supported codebook-liberal policies, but are now more likely to vote for them).

incrementally more aligned votes than on unanimous panels. Party-minority judges have become less likely to support aligned majorities over time, but the effect is small—aligned voting falls from nearly 100% to just under 98%. The effects on misaligned cases are much larger. Dissents by party-majority judges rise from half the rate of unanimous-panel judges to nearly twice as common. By the 1980s party-minority judges dissent more than three times as often as unanimous-panel judges, more than offsetting the slight decline in aligned-majority votes.

6.2 Pro-Weak Bias

Our next application highlights our LLM-built dataset's finer litigant labels. In recent work, Cohen (2025) uses IDB administrative data to show that panels with more Democratic-appointed judges are systematically more likely to favor structurally weaker litigants—e.g. an individual vs. a corporation. That approach uses litigant type and the case disposition as inputs, with the outcome being a set of coefficients on panel composition (see, e.g., Table 4 of Cohen 2025). Our database extends both inputs.

Extending disposition is straightforward, simply allowing analysis outside the IDB coverage window (1971–2025). Our litigant-power extension is more substantive: besides a finer ranking (e.g. state government $<$ federal government), the LLM sometimes identifies the weaker side even without a structural difference. For some civil cases (primarily economic), both codebooks introduce an explicit "underdog" label, e.g., a small business vs. a national chain or a tenant vs. a landlord. The underdog signal recovers about $20,000$ Songer-track and $92,000$ SCDB-track cases[31] that would drop out of a structural ranking (e.g. both sides were businesses). To address concerns that panel type biases underdog labeling (e.g., panels with more Democratic appointees are more likely to discuss underdogs) we regress underdog-label-presence on panel type indicators with circuit $\times$ year fixed effects. No coefficient is

[31] We recover more SCDB-track cases because the underdog label is valid for essentially all SCDB economic and labor cases (41% underdog-valid overall), and has no mixed direction code. Songer (about 17% underdog-valid) subdivides economic cases more finely and allows a mixed direction with an explicit no-underdog label.

statistically significant at the 5% level and point-estimated coefficients indicate less than 0.7 percentage points of difference over a baseline of 15% (Songer) or 26% (SCDB).

Table 9 compares specifications with varying degrees of LLM involvement. The first column reproduces headline estimates from Cohen (2025) for reference. The second uses the same inputs as that work, including only published cases but using the full IDB window of 1971–2025[32] in keeping with this paper's scope. The next two columns replace the litigant typology from Cohen (2025) with the LLM's ranking (from either the Songer or SCDB track) while keeping the IDB's case disposition. Finally, the last two specifications use LLM codings for both case disposition and litigant power (Songer-track LLM codings in the first, SCDB-track LLM codings in the second). Each specification is as similar as possible to the main specification in Cohen (2025). Specifically, we estimate

$$\text{ProWeak}_i = \alpha_{ct} + \alpha_a + \alpha_d + \alpha_s + \beta_1 \text{DRR}_i + \beta_2 \text{DDR}_i + \beta_3 \text{DDD}_i + \mathbf{x}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where i indexes a case decided in circuit c and year t; α_{ct} is a circuit×year fixed effect; α_a is an appeal-type fixed effect; α_d is a district-court-of-origin fixed effect; and α_s is a subject-matter fixed effect[33] $\mathbf{x}_i$ collects panel-level controls (at-least-one woman, at-least-one minority, panel mean tenure); RRR is the omitted baseline; and standard errors are clustered by circuit×year. The reported estimates are β_1, β_2, and β_3.

The IDB replication reproduces Cohen (2025)'s pattern at somewhat larger magnitudes: each Democratic appointee increases pro-weak outcomes by 2, 4, and 5 percentage points respectively. The third and fourth columns, pairing IDB outcomes (affirm, reverse, etc.) with LLM litigant types (e.g. underdogs, state vs. federal government), track the replication closely, with smaller magnitudes for the SCDB track. Finally, the full-LLM columns extend

[32] Cohen (2025) restricts to 1985–2020, but includes unpublished cases.

[33] α_a, α_d, and α_s are IDB-derived and apply to the four IDB-bearing columns. For the full-LLM columns we replace α_a and α_d with a single court-appealed-from fixed effect (Songer track only; not coded by the SCDB track) and replace α_s with an LLM-coded broad issue area fixed effect (Songer-track or SCDB-track as appropriate). The intercept implied by these fixed effects is absorbed into them, hence no separate β_0 in the specification.

	Cohen (2025)[a]	IDB Replication	IDB outcome + LLM types		Full LLM	
			Songer	SCDB	Songer	SCDB
DRR	0.012***	0.0239***	0.0267***	0.0135**	0.0208***	0.0079
	(0.002)	(0.0040)	(0.0032)	(0.0052)	(0.0027)	(0.0041)
DDR	0.035***	0.0633***	0.0717***	0.0388***	0.0564***	0.0282***
	(0.003)	(0.0053)	(0.0042)	(0.0064)	(0.0037)	(0.0051)
DDD	0.071***	0.1161***	0.1196***	0.0783***	0.0913***	0.0509***
	(0.005)	(0.0078)	(0.0065)	(0.0092)	(0.0056)	(0.0070)
N	554,060	125,022	217,796	80,226	383,436	159,392
Sample window	1985–2020	1971–2025	1971–2025	1971–2025	1892–2025	1892–2025

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Baseline = RRR panel. Controls (suppressed): at-least-one woman, at-least-one minority, panel mean tenure. Fixed effects: circuit×year, district-court-of-origin, appeal-type, and subject-matter (NOS/offense subdivision) for the IDB-bearing columns; circuit×year and LLM-coded broad issue area (Songer also adds court-appealed-from) for the full-LLM columns. Standard errors clustered by circuit×year.

[a] Reproduced from Cohen (2025) Table 4, column 7 ("All cases"), covering 1985–2020 with N = 554,060. The remaining columns are our own estimates.

Table 9: Pro-weak outcomes by panel type. "IDB Replication" uses the structural typing of weak vs. strong litigant from Cohen (2025) and IDB's case-disposition field. "IDB outcome + LLM types" keeps IDB's disposition but replaces the structural litigant typology with the LLM's litigant codes. "Full LLM" uses the LLM's disposition and litigant codes.

analysis back to 1892 and find point estimates slightly smaller than their corresponding IDB-LLM columns. Cross-model differences suggest that variable definitions account for some of the pro-weak effect's magnitude, but confirm that it is large across all sources.

Pro-weak bias and our ideological direction codings sometimes disagree on what counts as liberal. Leading examples include tax cases, where the government is stronger than an individual taxpayer but a government victory is coded as liberal; environmental regulation, where the government is stronger than a polluting business; and union antitrust, where a union is stronger than an individual worker. In our data, this clash between the two definitions occurs on about 20% of cases, allowing a test of whether pro-weak bias operates as a preference for weaker litigants *per se*. Reduced sample size is only a mild issue for our comprehensive data, but would make Songer-only analysis impossible.[34]

Our specification interacts panel composition with an indicator, Clash_i, equal to 1 when the codebook flags the pro-weak outcome in case i as conservative and 0 when the codebook

[34] The sparsity of Songer data does affect standard LLM-bias corrections; see Appendix A.8.2.

	IDB outcome + LLM types		Full LLM	
	Songer	SCDB	Songer	SCDB
DRR	0.0332***	0.0259***	0.0278***	0.0176***
	(0.0032)	(0.0062)	(0.0028)	(0.0048)
DDR	0.0893***	0.0741***	0.0704***	0.0525***
	(0.0046)	(0.0076)	(0.0040)	(0.0060)
DDD	0.1495***	0.1295***	0.1128***	0.0831***
	(0.0072)	(0.0107)	(0.0063)	(0.0083)
Clash	0.0324***	−0.0368**	−0.0301***	−0.0879***
	(0.0086)	(0.0117)	(0.0058)	(0.0083)
DRR × Clash	−0.0465***	−0.0543***	−0.0367***	−0.0343***
	(0.0088)	(0.0127)	(0.0063)	(0.0091)
DDR × Clash	−0.1230***	−0.1445***	−0.0845***	−0.0872***
	(0.0100)	(0.0137)	(0.0072)	(0.0100)
DDD × Clash	−0.1895***	−0.2136***	−0.1135***	−0.1098***
	(0.0129)	(0.0182)	(0.0089)	(0.0118)
N	191,851	68,804	345,299	140,905
Sample window	1971–2025	1971–2025	1892–2025	1892–2025

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Baseline = RRR panel. Controls (suppressed): at-least-one woman, at-least-one minority, panel mean tenure. Fixed effects: circuit×year, district-court-of-origin, appeal-type, and subject-matter (NOS/offense subdivision) for the IDB-bearing columns; circuit×year and LLM-coded broad issue area (Songer also adds court-appealed-from) for the full-LLM columns. Standard errors clustered by circuit×year.

Table 10: Pro-weak outcomes interacted with codebook direction. The Clash indicator equals 1 when the pro-weak outcome is conservative under the codebook (e.g. taxation cases, regulatory enforcement) and 0 when pro-weak is liberal. Reported main-effect coefficients are evaluated at Clash = 0; the panel-composition effect on clashing cases is the sum of the main effect and the corresponding interaction.

flags it as liberal.[35] We follow the approach of Table 9 but add interaction terms:

$$\text{ProWeak}_i = \alpha_{ct} + \alpha_a + \alpha_d + \alpha_s + \sum_k \beta_k \, \mathbf{1}_k(i) + \delta \, \text{Clash}_i + \sum_k \theta_k \, \mathbf{1}_k(i) \cdot \text{Clash}_i + \mathbf{x}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where $\mathbf{1}_k(i)$ is the panel-composition indicator $k \in \{\text{DRR, DDR, DDD}\}$ and θ_k is the corresponding interaction with Clash_i.

Attenuation of pro-weak bias on clash cases implies that our codebook-driven direction codes better capture judicial behavior than the blunter litigant-type measure. Full attenuation or reversal implies that pro-weak behavior is merely a proxy for left-right ideology.

[35] Cases from areas without a uniform ideological direction for pro-weak outcomes are dropped, covering almost 90% of cases when IDB determines issue area (most IDB-to-codebook maps are one-to-many) but only about 10% of LLM-determined cases.

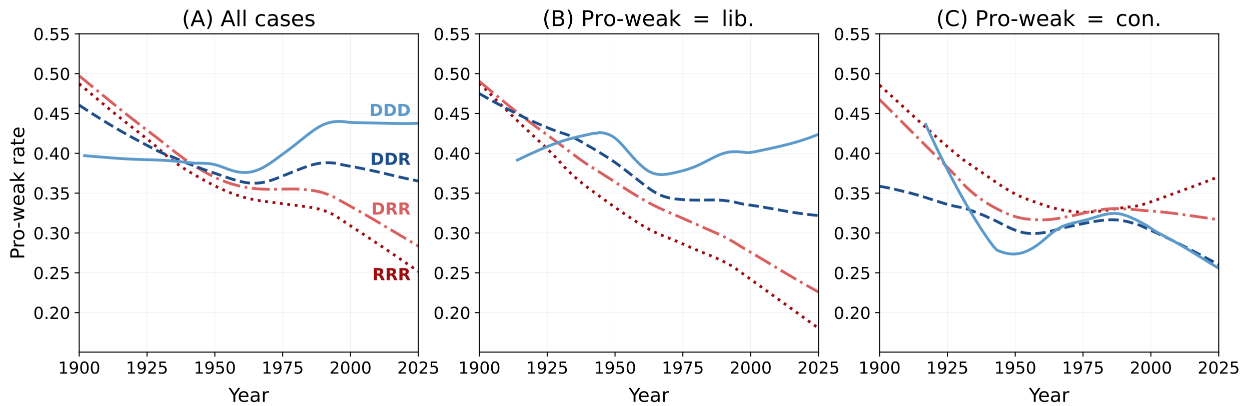

Figure 7: Pro-weak outcomes over time. Each panel plots the estimated pro-weak rate by year for the four panel compositions (RRR / DRR / DDR / DDD) using the full-LLM Songer specification (second-to-last column in tables 9 and 10). Lines are LOESS smooths (span 0.5). Panel (A) pools all cases. Panel (B) restricts to cases where the codebook calls the pro-weak outcome liberal (Clash = 0, about 80% of cases). Panel (C) restricts to the cases where the codebook calls the pro-weak outcome conservative (Clash = 1, about 20% of cases—e.g., taxation, environmental regulation). Data are sparse prior to about 1950, leading to noisier estimates.

Our results show strong support for the former (weaker) result, but only mixed support for the latter. The interaction terms in Table 10 are generally large, negative, and statistically significant: clear evidence that pro-weak bias attenuates on clash cases. While in that table the monotone increase in pro-weak bias with more Democratic appointees essentially vanishes, this stronger result does not persist under LLM-bias corrections (albeit with limited statistical power; see Appendix A.8.2). Given that pro-weak bias retains some independent explanatory power, ensemble approaches integrating it with judge ideology and our LLM-driven codings are a natural direction for future work.

Paralleling Figure 6 from Section 6.1, Figure 7 estimates the full-LLM Songer clash specification year-by-year, omitting the individual point estimates and showing only LOESS-smoothed lines for visual clarity. Panel (A) pools all cases as in Table 9; panels (B) and (C) follow Table 10 and separate non-clashing cases (center) and clashing cases (right). In the pooled sample, pro-weak outcomes decline with minimal differences between panel types until the 1950s, suggesting a secular change driven by non-ideological features (e.g., the types of cases appealed or the magnitude of litigant-power gaps). From the 1950s onward, and especially after the 1970s, we see the pattern in Cohen (2025): more liberal panels make more pro-weak decisions. That pattern is even clearer when restricting to pro-weak = liberal cases (center panel), where differences between panel types are visible in the early 1900s, though at half the magnitude of the same effects in the 2000s. When pro-weak = conservative (right panel), that pattern reverses: after about 1985, pro-weak outcomes increase on all-Republican panels and decrease on (both types of) majority-Democratic panels.[36] As with ideological voting, the magnitude and scope of pro-weak bias are largest in the modern era, though these results show that some pro-weak bias on cases with clear ideological valence existed even in the mid-1900s.

7 Conclusion

In this paper, we introduced a pipeline for extending expert-driven codings from small, hand-curated samples to corpus-level scale. We demonstrated it by creating the first universe-level database of published opinions issued by the U.S. Courts of Appeals: approximately 440,000 cases from 1892 to the present, with clean opinion text linked to judge metadata. Building on the two standard codebooks for judicial politics research—the Supreme Court Database and Songer Courts of Appeals Database—we used an open-weight LLM to extract case and litigant metadata, the substantive and procedural issues in each opinion, and the ideological direction of the appellate outcome. The resulting database offers universe-level coverage of the Courts of Appeals comparable to the gold-standard Supreme Court Database but for orders of magnitude more cases.

Our central methodological contribution is a documented, adaptable, and validated approach under which codebook-driven LLM coding approaches human coder performance. Validated directly against Songer hand codings, the starting point for much prior research on the Courts of Appeals, our LLM achieves 85–90% accuracy on broad issue area, disposi-

[36]Because this sample is much smaller, the overall time trend is less clear. The reversed ranking emerges in the 1930s, but then compacts again in the 1970s before separating in the modern period.

tion, and litigant type, and roughly 80% accuracy on ideological direction. Using the IDB as independent ground truth, the LLM matches or exceeds Songer human coder accuracy on disposition, broad issue area, and litigant identification, providing strong evidence that much of the LLM's disagreement with the Songer database reflects reasonable differences of opinion rather than model failure. Accuracy is largely stable across validation windows of over 50 years, covering both now-archaic cases from the early 1900s and technical modern opinions. While human coding on the scale necessary for these kinds of datasets remains infeasible, these results indicate that LLM coding is now a viable substitute when the relevant judgments can be decomposed into fact extraction guided by an explicit codebook.

The two applications in Section 6 illustrate the possibilities created by access to case-level ideological direction and fine-grained litigant types at scale. Our year-by-year panel-effects estimates, infeasible with previous data, show that ideologically aligned voting (along the modern left-vs.-right judicial ideology spectrum) was essentially constant with no difference between panel types until the late twentieth century. The post-1960 emergence of both panel effects and ideological voting is driven by dissents from ideologically misaligned majorities, masking a smaller *decrease* in ideologically aligned majority votes by party-minority judges. We also use our finer litigant type and ideological direction codings to extend Cohen (2025), showing that Democratic appointees' support for structurally weaker litigants is also strongest after the 1960s and is muted where such a pro-weak outcome is coded conservative by the Songer codebook. We expect that future work using our database will further enhance our understanding of political ideology on appellate courts.

These successes come with important limitations, which suggest natural directions for future work. Our coverage is restricted to the selected sample of published appellate opinions; extending it to unpublished opinions and other courts (e.g., district courts, state supreme courts) is a natural next step, though it will likely require codebook adaptations. Our direction codings inherit the lookup-table approach of the Songer and SCDB codebooks (Harvey and Woodruff 2013); researchers who disagree with that paradigm can implement their own coding choices on our raw summary data. Validation against external ground truth is possible only in bounded windows (1925–2002 for Songer, 1971–present for the IDB). While we find limited variation in performance over time, rarer variables like threshold issues (e.g., standing) are more sensitive and accordingly less trustworthy outside those windows. Finally, our compositional approach to ideological direction means that flaws in any single input may propagate; future work on ensemble estimation may combine our LLM codings with party-based proxies, IDB outcomes, and other signals.

We believe further extensions of our results, alongside new discoveries made by other researchers and practitioners using this database (available at `https://coadata.org`), will contribute greatly to our understanding of high-stakes outcomes throughout the judicial branch. As AI-powered methods for natural language understanding continue to develop, we also expect that our framework will guide further innovation in expert-informed text analysis at the scale necessary to test impactful hypotheses across social science more broadly.

References

Angelopoulos, Anastasios N., John C. Duchi and Tijana Zrnic. 2023. "PPI++: Efficient Prediction-Powered Inference." *arXiv preprint, https://arxiv.org/abs/2311.01453* .

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic. 2023. "Prediction-powered inference." *Science* 382(6671):669–674.

Beim, Deborah and Jonathan P. Kastellec. 2014. "The Interplay of Ideological Diversity, Dissents, and Discretionary Review in the Judicial Hierarchy: Evidence from Death Penalty Cases." *The Journal of Politics* 76(4):1074–1088.

Benesh, Sara C. 2002. *The U.S. Court of Appeals and the Law of Confessions: Perspectives on the Hierarchy of Justice*. New York, NY: LFB Scholarly Publishing.

Benoit, Kenneth, Scott De Marchi, Conor Laver, Michael Laver and Jinshuai Ma. 2026. "Using Large Language Models to Analyze Political Texts through Natural Language Understanding." *American Journal of Political Science* .

Bonica, Adam and Maya Sen. 2017. "A Common-Space Scaling of the American Judiciary and Legal Profession." *Political Analysis* 25(1):114–121.

Bonica, Adam and Maya Sen. 2024. "Common-space Measures of Judicial Ideology for Federal Judges (2024 update)." `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CK3EEL`. Data last updated 30 April 2025.

Boyd, Christina L., Lee Epstein and Andrew D. Martin. 2010. "Untangling the Causal Effects of Sex on Judging." *American Journal of Political Science* 54(2):389–411.

Choi, Jonathan H. 2024. "How to Use Large Language Models for Empirical Legal Research." *Journal of Institutional and Theoretical Economics* 180(2):214–233.

Cohen, Alma. 2025. "The Pervasive Influence of Political Composition on Circuit Court Decisions." *Journal of Legal Analysis* 17(1):14–41.

Cohen, Alma and Rajeev Dehejia. 2026. "Judges Judging Judges: Polarization in the U.S. Courts of Appeals." *NBER Working Paper No. 32920* .

Cope, Kevin L. 2026. "An Expert-Sourced Measure of Judicial Ideology." *Political Analysis* 34(2):258–277.

Cox, Adam B. and Thomas J. Miles. 2008. "Judging the Voting Rights Act." *Columbia Law Review* 108(1):1–54.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: design-based supervised learning for social science applications of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc. pp. 68589–68601.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal and Chad Westerland. 2007. "The Judicial Common Space." *Journal of Law, Economics, & Organization* 23(2):303–325.

Epstein, Lee, Andrew D. Martin, Jeffrey A. Segal and Chad Westerland. 2024. "The Judicial Common Space." `https://epstein.wustl.edu/jcs`. Data last updated 19 April 2024.

Farhang, Sean and Gregory Wawro. 2004. "Institutional Dynamics on the U.S. Court of Appeals: Minority Representation Under Panel Decision Making." *Journal of Law Economics & Organization* 20(2):299–330.

Federal Judicial Center. 2026*a*. "Biographical Directory of Article III Federal Judges, 1789–Present." `https://www.fjc.gov/history/judges`. Data updated nightly; exported data current through 31 December 2025.

Federal Judicial Center. 2026*b*. "Federal Court Cases: Integrated Database." `https://www.fjc.gov/research/idb`. Data updated regularly; exported data current through 31 December 2025.

Free Law Project. 2026. "CourtListener." `https://www.courtlistener.com`. Data updated regularly; exported data (bulk download) current through 31 December 2025.

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120(30):e2305016120.

Giles, Micheal W., Virginia A. Hettinger and Todd Peppers. 2001. "Picking Federal Judges: A Note on Policy and Partisan Selection Agendas." *Political Research Quarterly* 54(3):623–641.

Halterman, Andrew and Katherine A. Keith. 2026. "Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts." *Political Analysis* 34(2):188–204.

Harvey, Anna and Michael J Woodruff. 2013. "Confirmation bias in the United States Supreme Court judicial database." *The Journal of Law, Economics, & Organization* 29(2):414–460.

Hausladen, Carina I., Marcel H. Schubert and Elliott Ash. 2020. "Text Classification of Ideological Direction in Judicial Opinions." *International Review of Law and Economics* 62:105903.

Heseltine, Michael and Bernhard Clemm von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11(1).

Kastellec, Jonathan P. 2011. "Panel Composition and Voting on the U.S. Courts of Appeals Over Time." *Political Research Quarterly* 64(2):377–391.

Kastellec, Jonathan P. 2013. "Racial Diversity and Judicial Influence on Appellate Courts." *American Journal of Political Science* 57(1):167–183.

Kastellec, Jonathan P. and Anthony R. Taboni. 2026. "A Database of the United States Supreme Court's Shadow Docket, 1993–2025." *Journal of Law and Courts* 14(1):220–237.

Kluger, Dan M. Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang and Stephen Bates. 2025. "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling." *arXiv preprint, https://arxiv.org/abs/2501.18577* .

Kuersten, Ashlyn K. and Susan B. Haire. 2007. "Update to the United States Courts of Appeals Database, 1997–2002." http://www.songerproject.org/us-courts-of-appeals-databases.html. Supplement to the original U.S. Courts of Appeals Database, covering 1997–2002. Data last updated 13 July 2011.

Lewis, Jeffrey B., Keith T. Poole, Howard Rosenthal Howard, Barney Chen, Adam Boche, Aaron Rudkin, Luke Sonnet, Erik Hanson, William Lewis, Felipe Nunes, Fabio Souto and Jonah Wood. 2026. "Voteview: Congressional Roll-Call Votes." https://voteview.com. Data updated regularly; exported data current through 31 December 2025.

Ornstein, Joseph T., Elise N. Blasingame and Jake S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13(2):264–281.

Ortega, John E., Dhruv D. Joshi and Matt P. Borkowski. 2025. "Large-Language Memorization During the Classification of United States Supreme Court Cases." *arXiv preprint, https://arxiv.org/abs/2512.13654* .

Revesz, Richard L. 1997. "Environmental Regulation, Ideology, and the D.C. Circuit." *Virginia Law Review* 83(8):1717–1772.

Röttger, Paul, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze and Dirk Hovy. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. Lun-Wei Ku, Andre Martins and Vivek Srikumar. Association for Computational Linguistics pp. 15295–15311.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org pp. 29971–30004.

Songer, Donald R. 2008*a*. "Phase II Courts of Appeals Database." http://www.songerproject.org/us-courts-of-appeals-databases.html. Appeals cases subsequently reviewed by the U.S. Supreme Court. Data last updated 21 October 2008.

Songer, Donald R. 2008*b*. "The United States Courts of Appeals Database." http://www.songerproject.org/us-courts-of-appeals-databases.html. Original U.S. Courts of Appeals database, covering 1925–1996. Data last updated 21 October 2008.

Spaeth, Harold J., Lee Epstein, Michael J. Nelson, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger and Sara C. Benesh. 2025. "Supreme Court Database, Version 2025 Release 01." http://scdb.psu.edu. No data used; codebook current through 31 December 2025.

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

Taboni, Anthony R. 2026. "The Path of Law: Legal Uncertainty and Issues of First Impression in the U.S. Courts of Appeals." *American Political Science Review* 120(2):624–640.

The President and Fellows of Harvard University. 2024. "Caselaw Access Project." https://case.law. Data last updated 5 September 2024.

Törnberg, Petter. 2025. "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages." *Social Science Computer Review* 43(6):1181–1195.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong

Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou and Zihan Qiu. 2025. "Qwen3 Technical Report." *arXiv preprint, https://arxiv.org/abs/2505.09388*. Model access: https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507.

Yung, Corey Rayburn. 2010. "Judged by the Company You Keep: An Empirical Study of the Ideologies of Judges on the United States Courts of Appeals." *Boston College Law Review* 51:1133–1208.

Appendix

A.1 Data Pipeline

This appendix describes the full data processing pipeline from raw case text to LLM-ready data. The pipeline has five main stages, enumerated below; complete documentation is available in the project repository.[37]

1. Scraping with judge name extraction
2. Buried opinion detection and reclassification
3. Basic judge matching
4. Panel judge and author match fixes
5. Fragmented opinion cleanup and deduplication

A.1.1 Scraping with Judge Name Extraction

Cases come from two sources:

- **Caselaw Access Project** (1892–2019): bulk-downloaded published opinions in JSON format from the *Federal Reporter*, *Federal Reporter, 2nd Series*, and *Federal Reporter, 3rd Series*. Each volume contains full opinion text and metadata (case name, date, court, docket, cited judges).
- **CourtListener** (2019–2025): bulk-downloaded data exports, filtered to federal appellate courts. We stream compressed files, split combined opinion blocks into majority/dissent/concurrence via byline detection, and emit pipeline-compatible output.

A shared extraction pipeline reads structured metadata and cleans text via NFKD Unicode normalization. Likely judge names are extracted from bylines by looking for the common pattern of ALL-CAPS judge names, e.g., "Before ALICE, BOB, and CAROL, Circuit Judges," using stopword filtering, name prefix/suffix handling, and OCR correction. Per curiam opinions are caught by exact match against known variants plus regex and fuzzy-matching fallbacks for garbled text. Judicial designations (Circuit, District, Senior, Chief) are extracted alongside each name and used for disambiguation.

[37] This repository is currently private; please contact the authors for access. We will make it public upon publication of this paper.

A.1.2 Buried Opinion Detection and Reclassification

Dissents and concurrences are frequently embedded inside majority text, especially in older CAP volumes. Buried opinion detection scans for byline patterns and validates each candidate to ensure it is a true byline rather than a narrative reference, manifest of included opinions, or parenthetical citation. If buried bylines are found, the text is split to properly separate the buried opinion. A reclassification step then corrects existing labels that disagree with the text: e.g. if CAP codes a dissent but the byline designates it as a concurrence. Additional steps normalize non-standard opinion types into the majority, concurrence, or dissent taxonomy and split consolidated cases with multiple listed panels into separate records.

A.1.3 Basic Judge Matching

Each judge name extracted during scraping by the simple ALL-CAPS filter is matched to the FJC Biographical Directory using a multi-stage algorithm that applies progressively relaxed strategies, including RapidFuzz fuzzy matching from 100% down to 70%, with circuit and date validation throughout. Circuit validation handles the Fifth/Eleventh Circuit split (1981) and the Eighth/Tenth Circuit split (1929). Since CAP-scanned historical cases carry pervasive OCR artifacts in judge names, two curated correction tables, generated via data review using Claude Code, run before each matching attempt.

When multiple FJC candidates match a surname, a unified disambiguator narrows the field, considering only judges whose tenure overlaps with the case date. It first checks for a full-name match (First Middle Last) in the extracted name string or opinion text; a unique hit resolves immediately, correctly identifying visiting judges from other circuits. Otherwise, candidates are filtered through a four-tier hierarchy: same-circuit appellate judges, same-circuit district judges, other-circuit judges, then the remainder. Within the best non-empty tier, the disambiguator applies suffix and first-name checks, case text name search, structured and text-derived designations (circuit/district/senior/chief), and FJC status checks. Each step narrows the candidate set or resolves the match; if none succeeds, the match fails rather than selecting an uncertain candidate.

A.1.4 Panel Judge and Author Match Fixes

After initial matching, panel judge patching applies the matcher across five sequential phases to repair partial or failed matches. Each phase targets a specific failure mode and preserves matches resolved in earlier phases.

1. *Per curiam clearing.* Removes "PER CURIAM" text incorrectly placed in judge name slots, freeing those slots for re-matching with actual judge names.
2. *Validation and backfilling.* Cleans invalid names from judge slots using stopword and pattern filtering, backfills empty slots from extra judges extracted during scraping, applies OCR corrections, and attempts to split concatenated names (e.g., "L HAND

AUGUSTUS N HAND" into two entries). Names are validated against FJC surnames active in the case year.

3. *Direct matching and source disambiguation.* Runs the multi-stage fuzzy matcher on remaining unresolved slots. For slots that still fail, a source-text disambiguator searches the panel line, author bylines, head matter, and opinion text for clues—initials (e.g., "K. MOORE"), suffixes (e.g., "WOOD Jr."), designations (e.g., "Chief Judge,"), and senior-status indicators—to narrow multiple FJC candidates to one.
4. *Re-scraping from case text.* When name-level matching fails entirely, the pipeline re-extracts judge names from the case body ("Before:" panel lines and opinion bylines), then retries matching and disambiguation on the freshly extracted names. Successfully re-scraped names replace the original corrupted entries.
5. *Failure classification.* Remaining unresolved slots are classified by failure reason for downstream filtering. If a case has no judges with valid FJC candidates, the entire panel is replaced with a sentinel value.

After panel judges are patched, opinion authors are linked to the panel through a four-stage matcher. As a final fallback, when the author byline is malformed, the pipeline re-extracts a byline from opinion text and retries the full cascade from stage 1.

1. *Pre-processing.* The raw author byline is cleaned via OCR correction, stopword removal, possessive stripping, and junk-pattern filtering. If the byline is missing or invalid, the pipeline attempts re-extraction from the case text.
2. *Structured panel match.* The cleaned author name is matched against panel judges by exact and fuzzy surname comparison (threshold $\geq$70), with full-name verification and designation-based tie-breaking for shared surnames.
3. *FJC database lookup.* When neither panel-match stage succeeds—typically because the author is a visiting judge not on the panel—the cascade falls back to direct FJC lookup. A unique FJC match identifies the author as a visiting judge and adds the matched judge to the case's extra-judges metadata.
4. *Per curiam resolution.* Authors identified as per curiam—via exact match against the list of known variants or fuzzy detection—are recorded with a per curiam flag rather than a judge identifier.

When the pipeline detects per curiam in the authorship slot, the case is counted as fully matched: the authorship text was found and parsed correctly, and the absence of a named FJC author is the right answer (the court itself is the author). The 98.4% all-judges-and-authors-matched figure cited in Section 3.3 therefore counts per-curiam-flagged cases on the success side, not as missing data.

A.1.5 Fragmented Opinion Cleanup and Deduplication

A final step addresses fragmented and duplicate records produced by OCR artifacts, source-data errors, or the buried-opinion extractor double-counting opinions the source's own structured fields already capture. The cleanup drops empty placeholder records and redundant "combined" blocks and deduplicates within each judge—same opinion type by the same author is concatenated. Deduplication tracking resets at rehearing opinions so a judge's separate opinions before and after rehearing stay distinct. After all merges and removals, the dissent, concurrence, majority, and total opinion counts—and the en banc panel-size indicator—are recomputed from the cleaned record set.

Because the same case can appear multiple times—either within a source (e.g., a case reprinted in successive Federal Reporter volumes) or across sources (CAP and CourtListener overlap for cases decided around 2019)—we run a merge-based deduplication pipeline that preserves opinion content rather than simply dropping the weaker record. The process first runs within each source, then runs across sources.

Cases sharing circuit, decision date, and a common docket token are grouped as candidate duplicates. The case with the longest majority opinion is the *base*. Each other case's opinions are compared against the base: subsumed opinions are discarded, novel opinions are appended, and a strict-subset panel on the base is upgraded to the superset. A second pass groups by circuit, date, panel, and case name (rather than docket) to catch formatting variants. Cases that share a docket but were decided on different dates—typically a rehearing—are flagged as such rather than merged. We treat rehearings as separate voting events: the court issues a separate opinion on a separate date, sometimes with a different panel, and our sources (CAP, CL, IDB) themselves split or merge such cases inconsistently. Rehearings are uncommon (about 3% of observations), but to alleviate concerns that our splitting behavior affects estimates of panel effects and dissenting behavior in Section 6.1, we re-run that analysis with all flagged rehearings dropped and find no meaningful changes—all less than half the size of the coefficient standard errors, or about 3% of the coefficient magnitudes at most.

A.1.6 Judge Data Assembly

The FJC master file (`fjc_master.csv`) is constructed by merging six external data sources:

1. **FJC service records**: Appointment history, court assignments, commission and termination dates, appointing president, party, ABA rating, and seat identifier. Historical "Circuit Court" designations are normalized to "Court of Appeals." Assignment and reassignment rows with missing president/party fields are forward-filled from the judge's prior appointment.
2. **FJC demographics**: Birth and death dates (constructed from year/month/day components, with missing months defaulting to July and missing days to the 15th), gender, race or ethnicity, and birth state.
3. **FJC education**: Law school and degree, selected as the highest-sequence entry that is not "Read Law."
4. **CF ideal points** (Bonica and Sen 2017): Campaign-finance-based ideology scores, merged via the `fjc_nid` field.
5. **GHP scores**: Independently derived from Voteview NOMINATE data (Lewis et al.

2026) using the Giles, Hettinger and Peppers (2001) method. For each judge, the score is the average first-dimension DW-NOMINATE score of the home-state senator(s) from the appointing president's party during the relevant Congress; when no same-party senator exists, the appointing president's own NOMINATE score is used. State assignments for circuit judges come from a Wikipedia-derived dataset of duty stations (i.e. the de facto state for any given seat in a circuit) matched by name, circuit, and commission year; for district judges, the state is determined by the first two characters of the FJC seat identifier.

When new or updated judge attributes become available, a dedicated refresh script updates the judge metadata attached to each case without re-running the full matching pipeline. The refresh preserves all case-specific fields (match stage, age, seniority, JUDJIS score).

Manual corrections. Two hand-curated tables supplement the automated merge: a single FJC manual person-id override (consolidating duplicate FJC records for the same judge) and 13 CourtListener-to-FJC matching overrides for judges that automated name matching cannot resolve due to date mismatches, name changes, or typos.

A.2 Metadata Validation

Unlike the LLM annotations validated in Appendix A.5, the case-level metadata our pipeline extracts from opinion text—who sat on a panel, who dissented or concurred, and who wrote each opinion—is factual and can be checked against external sources that derived it separately. Both Songer and the IDB contain limited metadata information and cover a portion of our universe-level metadata. Songer covers a 5% sample of 1925–2002 and is a per-judge record, allowing us to directly test whether our metadata seats the right panelists and assigns authorship and separate opinions correctly. The IDB covers 1971–2025 completely; it is a per-case record without judge identities and can only confirm, e.g., the existence of a dissent rather than its author. While neither is error-free—and indeed Songer's own reported inter-coder reliability is around 95% on judge identification—both provide a sanity check that our pipeline correctly reconstructs crucial factual information about cases and panels.

A.2.1 Matching Cases to External Sources

Songer records identify a case by its Federal Reporter citation, so we match on reporter, volume, and beginning page. We infer the reporter series from the Songer year (F. through 1924, F.2d 1925–1993, F.3d from 1993, trying both at the ambiguous 1993 boundary). Exact-page matches are taken first; failures widen to ± 5 pages. Where several of our cases remain candidates, we disambiguate on circuit, then year (± 1), then fuzzy docket number.

The IDB has no citation field, so matching runs through dockets and dates in a four-stage cascade with progressively relaxed criteria. Stage 1: circuit, exact docket, exact decision date. Stage 2: ± 7 days within the same calendar year, optionally with fuzzy case-name verification (Levenshtein token-set ratio $\geq$75%). Stage 3: circuit and docket within the same year, no strict date requirement. Stage 4 (case-name-required): fuzzy caption match, where both IDB litigant names must appear in the pipeline caption with token-set-ratio $\geq$75%. When several IDB candidates survive a stage we prefer, in order, published status (PUBSTAT$\in \{2, 4, 6\}$), the best litigant-name match against the caption, the closest JUDGDATE (margin $\geq$30 days, distance $\leq$365 days), and finally DKTDATE under the same margin rule. Where no tiebreaker resolves, all candidates are kept as a *set match* with the best as primary.

Because dockets are reused across rehearings and consolidations, the IDB match is not one-to-one in either direction. Of about 285,000 matched cases, roughly 800 are one-to-many (the IDB split a case we hold as one record, typically a consolidation) and about 5,000 IDB rows are claimed by more than one of our cases, covering some 11,000 cases; about 92% of the latter show a clean rehearing signature, with one claimant on the exact decision date and others sharing the docket but not the date. For many-to-one collisions we drop stage-4 (name-only) claimants whenever a docket-matched claimant exists on the same row; legitimate rehearing matches at stages 1–3 are preserved. For one-to-many set matches, validation OR-aggregates across the candidate set, so a case counts as positive if any candidate evidences the signal.

The cascade yields 284,691 matched cases, 97.8% of post-1970 published circuit-court cases (the IDB does not cover the Federal Circuit). A further 12,444 Songer cases match to the IDB through the same chain, which gives an independent Songer-versus-IDB benchmark alongside our own. For each IDB variable we therefore report three figures: *Songer* (the hand-coded database scored on the same IDB predicate), *raw* (treating the IDB as ground truth), and *corrected* (adjusting for IDB data-quality problems documented below).

A.2.2 Songer Results: Panels, Authorship, and Votes

Songer identifies judges by its own numeric codes rather than by name, so every comparison in this subsection runs through a crosswalk mapping each Songer judge code to an FJC identifier. We build it from the Songer codebook's judge appendix, resolving each code to a judge and recording per-code year windows where a code's referent changes; codes that resolve to no judge, and codes whose identity we could only establish by reading our own data, are flagged and reported separately rather than scored.

Some of the disagreements below are demonstrably Songer's. Where the two sources share only two of three judges, 24.0% of the judges only Songer names were not on the bench in the case year, against 0.0% of the judges only we name—the latter by construction, since our matcher gates every seat on the judge's FJC service window.[38] Such errors are usually a choice between two judges sharing a surname, with 25 pairs in 130 cases with this pattern. For example, in *Power Mfg. Co. v. Tindall*, 100 F.2d 463 (8th Cir. 1938), Songer records Walter Henry Sanborn, who died in 1928, while we seat John Benjamin Sanborn, commissioned to that court in 1932. Separately, 140 cases share *no* judge with Songer even though citation, year, and Songer's own circuit field agree. Spot-checking the opinion text shows Songer's panel is wrong, which suggests row misalignment in the published database rather than a judge-identification failure on either side.

Panel composition. Table A.1 separates the comparison by panel size before asking which judges each source names; most disagreement turns out to be about the former rather than the latter. Within the 20,223 cases where both sources record a three-judge panel—91.2% of the sample, and the population the paper's analyses are about—we seat exactly the same three judges 93.9% of the time, share two of three in a further 1,012 cases, and share none in 115.

Size disagreements run mostly in one direction. On 1,190 cases (5.4%) we seat three judges where Songer records fewer. Spot-checking the opinion text suggests we are almost always right to seat a full panel, so we read these as Songer under-recording rather than as our seating extra judges.[39] The mirror-image block is smaller: 151 cases where Songer records

[38] A non-zero figure on our side would indicate that the tenure gate had failed, which is why we compute and report it rather than assume it.

[39] For example, *McKenzie v. Lee*, 246 F.3d 494 (2001), where the text names Jones, DeMoss and Barzilay and Songer records two judges; and *Vera-Valera v. INS*, 147 F.3d 1036 (1998), naming Schroeder, Kleinfeld and Wallach (the latter two sitting by designation from the Court of International Trade).

Panel agreement with Songer	Cases	Share	Author agrees
Both sources seat three judges (20,223; 91.2%)			
3 of 3 judges shared	18,998	85.6%	99.0% (14,203)
2 of 3 judges shared	1,012	4.6%	76.6% (607)
1 of 3 judges shared	98	0.4%	38.3% (47)
0 of 3 judges shared	115	0.5%	0.0% (63)
Both seat the same other size (168; 0.8%)			
all judges shared	155	0.7%	96.4% (83)
not all shared	13	0.1%	100.0% (3)
We seat three, Songer disagrees on size (1,209; 5.5%)			
Songer seats fewer	1,190	5.4%	93.1% (361)
Songer seats more (Songer says en banc)	19	0.1%	58.3% (12)
We seat fewer, Songer three	151	0.7%	98.0% (101)
Other size mismatch	3	0.0%	—
Songer code unresolvable	429	1.9%	97.1% (243)
All three-judge cases	22,183	100.0%	97.3% (16,111)

Table A.1: Agreement with the Songer database on panel composition, opinion authorship, and per-judge votes. Panel rows are grouped by the two sources' recorded panel sizes. Agreement is bounded above by Songer's own reliability: its reported intercoder agreement is 95.2% on the per-judge vote field and 96.4% on whether an opinion is per curiam or not.

three judges and we seat fewer, and in 145 of them (96.0%) every judge we did seat is one Songer also lists, so the gap is one of recall—a missing third judge—rather than a wrongly seated one.

Authorship. We agree with Songer on the majority-opinion author in 97.3% of 16,111 comparable cases, and the right-hand column of Table A.1 shows that almost all of the disagreement is inherited from panel disagreement rather than from misreading a byline. Given a panel both sources agree on we name the same author 99.0% of the time. The 0.0% on the same-size-no-overlap band (63 comparable cases) is informative in its own right: a genuinely wrong author would still agree with Songer occasionally by chance, so a clean zero across that many cases indicates row-alignment errors rather than authorship errors. Per curiam detection, scored against Songer's `signed` field on 20,764 cases, agrees on 97.7% of them, with recall 90.4% and precision 97.0% treating per curiam as the positive class (TP 3,517, FP 110, FN 373).

Per-judge votes. Comparing each seated judge's recorded vote, agreement is 98.3% across 41,479 judge-slots. Scored as a binary variable, we recover 81.8% of the judge-votes Songer codes as dissents at a precision of 91.4% (TP 2,250, FP 211, FN 502).[40] Concurrence votes can only be compared in the KH update, where concurrences are coded separately from majorities; there we recover 88.5% at a precision of 61.4% (TP 108, FP 68, FN 14).

[40] A judge who concurs in part and dissents in part is counted as doing both, as with the IDB comparison in Appendix A.2.3.2.

A.2.3 IDB Results: En Banc, Dissent, and Concurrence

A.2.3.1 En Banc Detection

We flag en banc cases as cases with more than three judges listed in the opinion text. IDB records en banc status via `ENBANC` (2008+) and `JDGCODE2`=0000 (1974–2007); blanks in the older file are "not en banc" per the IDB codebook. The two together cover the 284,639 matched cases with an ENBANC-comparable IDB row.

	IDB en banc	IDB not en banc
Our data, en banc	2,816 (TP)	2,702 (FP)
Our data, not en banc	339 (FN)	278,782 (TN)

Raw: Recall 89.3%, Precision 51.0%.

Songer benchmark: Recall 92.2%, Precision 70.2% (TP=177, FP=75, TN=12,177, FN=15).

Corrected (FN only when exactly 3 judges extracted; FP only when no extra judges matched): Recall 95.0%, Precision 100.0%.

False negatives (339). These divide into three groups. 149 list exactly three judges despite formally being en banc, and so enter analysis as apparent panel cases; these are the only false negatives the correction below retains, because they are the only ones where our panel is both usable and wrong. A further 124 are extraction failures in which fewer than three judges could be recovered from the opinion text at all, and 66 carry no judges at all—these 190 are dropped by downstream filters.

False positives (2,702). Every one has at least one extra FJC-matched judge beyond the standard three—genuine en banc panels the IDB did not record. Songer is subject to the same principled correction: of the 75 cases where Songer says en banc and the IDB does not, 71 name a real judge in service at the time, raising Songer's precision from 70.2% to 97.8%.[41]

A.2.3.2 Dissent Detection

We count dissenting opinions from the opinion text. Mixed opinions (which concur in part and dissent in part) are counted as both dissents and concurrences.

[41] Songer records en banc status by listing a fourth judge rather than in a dedicated field, and its judge slots are not always contiguous—20 cases have a gap, and on 5 of the 505 cases naming four or more judges the fourth slot itself is empty while later slots are filled. We therefore read any slot past the third rather than the fourth alone.

	IDB has dissent	IDB no dissent
Our dis>0	13,313 (TP)	15,229 (FP)
Our dis=0	891 (FN)	255,231 (TN)

Raw: Recall 93.7%, Precision 46.6%.

Songer benchmark: Recall 94.6%, Precision 56.4% (TP=914, FP=707, TN=10,771, FN=52).

Corrected (FP only when dissent-author match failed): Recall 93.7%, Precision 99.4%.

False negatives (891). 155 have a concurrence detected instead (our label reads as a concurrence despite the IDB suggesting it contains a dissent); the remaining 736 have no separate opinion extracted at all—likely brief dissent notations without a written opinion or text-extraction failures.

False positives (15,229). 15,146 (99.5%) have a dissenting opinion whose author is FJC-matched; only 83 do not. These look like genuine dissents the IDB missed rather than spurious detections on our side.

The same comparison, for Songer. Scored against the IDB, Songer recovers 94.6% of dissents at a raw precision of 56.4%. Correcting its false positives on the same principle as our own raises that to 88.4%, short of our 99.4%. The shortfall is Songer's own internal inconsistency: its case-level dissent count and its per-judge vote fields agree on only 82.0% of the cases where either indicates a dissent; on 108 of the disputed cases it names no dissenting judge at all.

A.2.3.3 Concurrence Detection

	IDB has concurrence	IDB no concurrence
Our conc>0	7,198 (TP)	14,874 (FP)
Our conc=0	1,301 (FN)	261,291 (TN)

Raw: Recall 84.7%, Precision 32.6%.

Songer benchmark: Recall 84.5%, Precision 72.1% (TP=93, FP=36, TN=1,939, FN=17).

Corrected (FP only when concurrence-author match failed): Recall 84.7%, Precision 99.3%.

False negatives (1,301). 335 have a dissent detected instead (the mirror of the dissent false-negative pattern); the remaining 966 have no separate opinion extracted.

False positives (14,874). 14,824 (99.7%) have an FJC-matched concurring author; only 50 do not. As with dissents, these read as genuine concurrences the IDB missed.

The same comparison, for Songer, restricted to 1997–2002 (the KH window): Songer recovers 84.5% of concurrences at a raw precision of 72.1%, corrected to 87.7% on the same principle. Songer's concurrence count and its per-judge votes agree on only 77.4% of 2,130 cases, so the corrected figure should be read with that ceiling in mind.

A.2.3.4 Publication Status and Disposition

Of the 282,473 matched cases whose source the IDB does not itself record, it marks 90.9% as published (PUBSTAT$\in \{2, 4, 6\}$) and 9.1% as unpublished; that gap likely reflects circuit-level publication rules or IDB coding conventions rather than cases wrongly in our universe. 99.0% of matched cases are merits terminations (DISP$\in \{1, 2, 3\}$) against 0.9% procedural, confirming the published-opinion sample does not inadvertently include procedurally terminated cases.

A.2.3.5 Set vs. Resolved Validation

The above results use set-aggregation: for cases matched to multiple IDB candidates, a signal counts as positive if *any* candidate evidences it. As a robustness check we re-ran using only the tiebreaker-selected primary. The results are essentially identical, with less than 1 percentage point of difference in any (corrected or uncorrected) precision or recall metric. We report the set-aggregated results throughout.

A.3 LLM Prompt Architecture

This appendix describes the four prompts used in the LLM coding pipeline, the inter-pass data flow, and the direction lookup tables. The complete prompt code (approximately 6,300 lines across four prompt-builder files plus an 838-line direction post-processor) is available in the project repository.[42]

Each case runs through four sequential LLM passes in two parallel tracks (SCDB and Songer). Every pass uses a system prompt that assigns the model a role, specifies a JSON output schema, and provides variable-by-variable coding instructions adapted from the relevant codebook. Coding is kept as mechanical as possible: direction is defined by lookup tables with worked and negative examples, not by ideological labels.

After the two passes, mechanical and super-mechanical direction are computed via a post-processing script (see Figure 1 in Section 4.3 of the main text). For each case, the script:

1. Reads (1) the focal interest, (2) whether the focal interest was the appellant, (3) whether the focal interest won, and (4) case disposition (from the SCDB first pass).
2. Combines (1) and (3) for mechanical direction; combines (1), (2), and (4) for super-mechanical direction (replacing "focal interest won" with the combination of "focal interest was appellant" and "appellant won" via case disposition).
3. Looks up the appropriate direction code in the issue-area-specific mapping table (Section A.3.5).
4. Compares the mechanical code against the first-pass LLM-coded direction to flag disagreements for quality assurance.

Cases where the focal interest is "neither" or cannot be determined receive direction code 0 (not ascertained).

As noted in Section 4.1, about 6% of cases needed at least one re-try, with fewer than 1% requiring more than one. A provenance table attached to each row of LLM output documents the exact source of the final output, tracking the number of re-tries. The full distribution of total attempts is in Table A.2.

In the remainder of this section, we provide a detailed description of the LLM pipeline to supplement the ideological-direction-focused overview in Section 4 of the main text.

A.3.1 SCDB First Pass

The SCDB first pass receives the full opinion text and codes the broad SCDB variables. The system prompt opens:

[42] This repository is currently private; please contact the authors for access. We will make it public upon publication of this paper.

Attempts	Retries	Cases	Share	Cumulative
1	0	481,021	93.92%	93.92%
2	1	28,043	5.48%	99.39%
3	2	2,013	0.39%	99.78%
4	3	475	0.09%	99.88%
5	4	539	0.11%	99.98%
6	5	83	0.02%	100.00%
7	6	4	0.00%	100.00%
8	7	3	0.00%	100.00%

Table A.2: LLM parse-retry distribution. Total $N = 512,181$ LLM-coded cases.

> You are an expert legal coding assistant trained to classify U.S. federal Courts of Appeals cases using an adaptation of the Supreme Court Database (SCDB) codebook. [...] You are working with U.S. Courts of Appeals, not the Supreme Court. Apply the SCDB coding rules as closely as possible, treating the Court of Appeals as "the Court" and assuming that the lower court is, e.g., a federal district court whose decision is being reviewed.

The user message follows a fixed template:

```
[CASE_METADATA]
Case name: Smith v. Jones
Court: U.S. Court of Appeals, Ninth Circuit
Date decided: 2003-04-15
Opinion type: majority

[CASE_TEXT]
<opinion text>
[END_CASE]
```

The JSON output schema has twelve top-level keys:

- appellant and respondent (with name, description, and codebook-driven litigant type)
- case disposition;
- winning side, which narrows the disposition to a simple win/lose/neither for the appellant
- issue area (from the fourteen SCDB issue area codes); tiebreaking rules resolve ambiguous cases, including a checklist for distinguishing Economic Activity from Private Action and priority rules for cases spanning Criminal Procedure and Civil Rights.
- procedural holding: an indicator if the ruling was procedural rather than substantive
- legal provisions: a list of (area code, provision text) pairs across nine law-area categories
- an indicator if the ruling was about constitutionality
- method
- who appealed
- one-shot ideological direction using the lookup table in Section A.3.5.1

- Justification: a free-text explanation including a mandatory ideological direction derivation flowchart

The centerpiece of the first-pass prompt is the direction lookup table (reproduced in Section A.3.5). The table maps (issue area, benefiting side) pairs to direction codes 1 or 2, with code 3 reserved for indeterminate cases. Each row specifies the benefiting side in plain language (e.g., "accused / convicted person / criminal defendant" → code 2 for Criminal Procedure). Exception sub-rules handle takings clause cases, union antitrust, worker-vs-union disputes, and arbitration. Worked and negative examples illustrate mappings that the LLM struggled with in preliminary testing (e.g., taxpayer winning = code 1, not code 2). The prompt also mandates a structured derivation in the justification field: `DIRECTION: Area [X]; [who BENEFITED]; code [N]`.

A.3.2 SCDB Second Pass

The second SCDB pass refines the first-pass issue area into a specific five-digit SCDB issue area subcode and identifies the focal interest and outcome. The prompt is assembled dynamically by combining a common header with the area-specific template for the issue area assigned in the first pass. The second pass receives structured context from the first pass—the petitioner and respondent (name, description, category), the case disposition, and the winning side—but does *not* see the first-pass direction code:

```
You are refining the coding for a case previously
classified in Issue Area 1 (Criminal Procedure).

PARTIES:
Petitioner: Smith (federal criminal defendant
appealing conviction)
Respondent: United States (appellee)

CASE OUTCOME:
Disposition: affirmed
winning side: Respondent won | the petitioner's
appeal was unsuccessful
```

We made several key design choices in this section to push the LLM towards mechanical fact identification, in keeping with our overall strategy for ideological direction coding. As in the previous setting, the prompt explicitly instructs the model not to use ideological labels like "liberal" or "conservative" for any output, and to focus instead on identifying focal interests. For each issue area, we derived a narrow set of focal interest types from the SCDB codebook's ideological direction definitions. In areas where the codebook identifies a single unambiguous focal interest, e.g., Criminal Procedure (only "accused"), we provided only that single option. In others, such as SCDB issue areas 2–5, we follow the codebook in presenting multiple claimant types (e.g., "civil rights," "child," "indigent,") but no opposing side. When asking the model to code whether the focal interest benefited, was harmed, or received a mixed outcome, we require three mandatory cross-checks: (1) if the focal interest is the appellant, and the winning side code indicates the appellant won, this coding should match that; (2) if the case involves taxation, the focal interest must be "taxpayer" and this coding must be based on whether the taxpayer won or lost; and (3) if this coding is "mixed," the model must verify that the focal interest received meaningful benefit on some issues *and* meaningful defeat on others, not merely that the case is complex or the coder is uncertain.

The JSON output schema has six top-level keys:

- issue code and label (the five-digit refinement of the first-pass's broad issue area)
- focal interest, and additional (secondary) focal interests for interpretability
- whether the focal interest was the appellant
- whether the focal interest won or lost

When the focal interest was coded as "underdog", the LLM was also asked to provide a brief description of the specific underdog (e.g., "individual consumer," "injured seaman," "debtor").

A.3.3 Songer First Pass

The Songer first pass codes the broad Songer variables *independently* of the SCDB codings. It receives only litigant names, descriptions, disposition, and winning side from the SCDB first pass; it deliberately does not see SCDB issue area or direction, preserving Songer-track independence for cross-validation. The system prompt opens:

```
You are an expert legal coding assistant trained to classify U.S. federal Courts
of Appeals cases using the Songer U.S. Courts of Appeals Database codebook.
```

The user message has the same shape as the SCDB second pass, with SCDB first-pass litigant and outcome information as factual context above the case text. We share litigant and outcome but not issue area or direction because the former are objective facts observable in the opinion which ensure comparability of outputs across the two tracks, while the latter are classificatory judgments that could bias the Songer coding if leaked from the SCDB track. The JSON output schema has ten top-level keys:

- issue area (one of eight Songer broad areas)
- procedural holding
- one-shot ideological direction (defined by a Songer-specific lookup table in the same style as the SCDB one; see Section A.3.5.2).
- authority type (constitutional, statutory, administrative review, court rules, common law, or other)
- appellant and respondent types and categories

- type of issues considered (criminal, civil-private, civil-government plaintiff, civil-government defendant, or diversity)
- the type of lower court judgment appealed from
- threshold issues (jurisdiction, standing, mootness, ripeness or exhaustion, timeliness, immunity, failure to state a claim, frivolous case, political question, late appeal, frivolous appeal, and other trial/appellate thresholds)
- justification: as in the SCDB first pass, a combination of free text and the mandatory ideological direction flowchart.

A.3.4 Songer Second Pass

The Songer second pass refines the broad issue area and litigant types into more fine-grained subcodes and identifies the focal interest and outcome for mechanical direction conversion. It draws context from both the SCDB first pass (litigant descriptions, disposition, winning side) and the Songer first pass (broad issue area, type of issues, litigant categories):

```
You are refining the coding for a case previously
classified in Songer Area 1 (Criminal).

PARTIES:
Appellant: Smith | category 7 (natural person);
federal criminal defendant appealing conviction
Respondent: United States | category 3 (federal
government); appellee

CASE OUTCOME:
Disposition: affirmed
winning side: Respondent won | the appellant's
appeal was unsuccessful

FIRST-PASS SONGER CODING:
Issue area: Code 1 | Criminal
Type of issues: 1 (criminal)
```

Like the SCDB second pass, the prompt is assembled dynamically from a common header and per-area templates, and features many of the same design choices: no ideological labels, narrow sets of focal interests (we can in fact force only a single choice for five of the eight categories), and mandatory cross-checks. We also refine the type of issues with a tailored set of binary variables (e.g., search and seizure or death penalty for criminal; due process for civil; judicial review for government; and many more). The JSON output schema is thus more complex and issue-area-dependent than for SCDB, but includes the same six general keys as well as the Songer-specific additions of litigant type refinements and type of issue indicators.

Area	focal interest	Outcome	Dir.
1 (Crim. Pro.)	accused / criminal defendant	any relief	2
		no relief	1
2 (Civ. Rts.)	civil rights claimant	prevailed	2
		lost	1
3 (1st Amend.)	person asserting 1st Amend. rights	prevailed	2
		lost	1
4 (Due Proc.)	person challenging deprivation	prevailed	2
		lost	1
4 (Due Proc.)	*Takings exception*: govt/anti-owner	prevailed	2
5 (Privacy)	person asserting privacy/autonomy	prevailed	2
		lost	1
6 (Attorneys)	attorney	prevailed	2
		lost	1
7 (Unions)	*See sub-rules below*		
8 (Econ. Act.)	*See sub-rules below*		
9 (Jud. Power)	broader court access/review	prevailed	2
		lost	1
10 (Federalism)	federal authority	prevailed	2
		lost	1
11 (Interstate)	*all cases*		3
12 (Fed. Tax.)	taxpayer	won	1
		lost	2
13 (Misc.)	*See sub-rules below*		
14 (Priv. Act.)	*all cases*		3

Table A.3: SCDB direction lookup table. Reproduced from the SCDB-track first-pass prompt.

A.3.5 Direction Lookup Tables

The tables below reproduce the full direction lookup tables from the LLM prompts. These tables define the mechanical mapping from (issue area, focal interest, outcome) to direction codes. In the prompt, we emphasize that these codes have no inherent ideological meaning; they are defined entirely by the tables. These tables may have slight differences with the focal interests and rules in the second-pass prompts, which are intentional. While the second-pass prompts were broad-issue-area specific and could thus tailor their instructions more carefully, the first-pass prompts had to balance detail with length, and therefore exhibit slight simplifications compared to the detailed rules and focal interest sets described in Appendix A.3.6.

A.3.5.1 SCDB Direction Table

The SCDB track uses three direction codes: 1 (conservative), 2 (liberal), and 3 (unspecifiable). For each issue area, the prompt specifies a focal interest and maps the outcome to a code; Table A.3 reproduces the mapping for all areas except 7, 8, and 13, which we describe in detail below.

Area 7 (Unions) sub-rules.

- **7A—Standard union/labor (default):** Focal = union/workers. Won → 2. Lost → 1.
- **7B—Union antitrust exception:** Focal = competition/antitrust enforcement. Won → 2. Lost → 1.
- **7C—Worker vs. union** (individual member suing union): Focal = union. Union prevailed → 2. Worker prevailed → 1.
- **7D—Arbitration:** Focal = pro-trial position. Won → 2.

Area 8 (Economic Activity) sub-rules.

- **8A—Tax:** Focal = taxpayer. Won → 1. Lost → 2. (Same pattern as Area 12.)
- **8B—Government regulation:** Focal = govt regulation. Upheld → 2. Struck down → 1.
- **8C—Environmental/consumer protection:** Focal = protection interest. Won → 2. Lost → 1.
- **8D—Torts/personal injury:** Focal = injured party. Won → 2. Lost → 1.
- **8E—Patents/copyrights/trademarks:** Focal = IP rights holder. Won → 2. Lost → 1.
- **8F—Bankruptcy:** Focal = debtor. Won → 2. Lost → 1.
- **8G—Antitrust/securities enforcement:** Focal = enforcement. Won → 2. Lost → 1.
- **8H—Government benefits:** Focal = claimant. Won → 2. Lost → 1.
- **8I—Other commercial:** If clear underdog exists, focal = underdog (won → 2, lost → 1). If no clear underdog (business vs. business of comparable stature) → 3.

Area 13 (Miscellaneous) sub-rules.

- **13A—Incorporation of territories:** → 2.
- **13B—Executive authority vis-à-vis congress/states:** → 2.
- **13C—Judicial authority vis-à-vis legislative authority:** → 2.
- **13D—Legislative veto:** → 1.
- **13E—None of the above:** → 3.

A.3.5.2 Songer Direction Table

The Songer track uses four direction codes: 0 (not ascertained), 1 (conservative), 2 (mixed), and 3 (liberal). Table A.4 reproduces the mapping for all areas except 6, 7, and 9, which we describe in detail below.

Area 6 (Labor) sub-rules.

- **6A—Standard union/labor:** Focal = union/workers. Won → 3. Lost → 1.
- **6B—Government enforcing labor laws:** Focal = govt enforcement. Won → 3. Lost → 1.
- **6C—Executive branch vs. union:** Focal = executive branch. Won → 3. Lost → 1. (Reversed from 6A.)
- **6D—Worker vs. union:** Focal = union (not individual worker). Union won → 3. Worker won → 1. (Reversed from intuition.)
- **6E—Rival unions:** Focal = union opposed by management. Won → 3. Lost → 1.
- **6F—Injured workers/consumers:** Focal = injured party. Won → 3. Lost → 1.
- **6G—Other labor:** Clear underdog? Focal = underdog (won → 3, lost → 1). No underdog → 0.

Area 7 (Economic) sub-rules.

Area	focal interest	Outcome	Dir.
1 (Criminal)	defendant/accused	any relief	3
		no relief	1
2 (Civ. Rts.)	rights claimant	prevailed	3
		lost	1
3 (1st Amend.)	person asserting 1st Amend. protections	prevailed	3
		lost	1
4 (Due Proc.)	person asserting due process rights	prevailed	3
		lost	1
5 (Privacy)	person asserting privacy/autonomy	prevailed	3
		lost	1
6 (Labor)	*See sub-rules below*		
7 (Economic)	*See sub-rules below*		
9 (Misc.)	*See sub-rules below*		

Table A.4: Songer direction lookup table. Reproduced from the Songer-track first-pass prompt.

- **Step 1—Tax:** Focal = taxpayer. Won → 1. Lost → 3. (Counterintuitive: taxpayer winning = 1.)
- **Step 2—Government regulation/benefits:** Regulation upheld → 3. Struck down → 1. Benefits claimant won → 3. Lost → 1.
- **Step 3—Torts/personal injury:** Focal = injured plaintiff. Won → 3. Lost → 1.
- **Step 4—Patents/copyrights/trademarks:** Focal = IP rights holder. Won → 3. Lost → 1.
- **Step 5—Bankruptcy:** Focal = debtor. Won → 3. Lost → 1.
- **Step 6—Antitrust:** Focal = enforcement. Won → 3. Lost → 1.
- **Step 7—Commercial catch-all:** Clear underdog? Focal = underdog (won → 3, lost → 1). No underdog → 0.

Area 9 (Miscellaneous) sub-rules.

- **9A—Interstate conflict:** → 0.
- **9B—Federalism:** Focal = federal power. Won → 3. Lost → 1.
- **9C—Attorneys:** Focal = attorney. Won → 3. Lost → 1.
- **9D—Selective service:** Focal = inductee. Won → 1. Lost → 3.
- **9E—Magistrates/referees:** Focal = challenged official. Won → 3. Lost → 1.
- **9F–9I—Indian law**
  - **9F—Criminal:** Focal = defendant. Won → 3. Lost → 1.
  - **9G—Civil:** Focal = Indian or tribal party. Won → 3. Lost → 1.
  - **9H—vs. Gov't:** Focal = Indian or tribal party. Won → 1. Lost → 3. (reverse of 9G)
  - **9I—Tribal Regulation:** Focal = regulator. Won → 3. Lost → 1.
- **9J—Immigration:** Focal = immigrant. Won → 1. Lost → 3.
- **9K—International law:** Focal = US/US firms. Won → 3. Lost → 1. Neither → 0.
- **9L—National security:** Focal = individual challenging gov't. Won → 1. Lost → 3.
- **9M—Executive privilege:** Focal = executive. Won → 3. Lost → 1.
- **9N—Other/not ascertained:** → 0.

A.3.6 Second-Pass Outcome Validation Rules

This section documents how we derive ground-truth focal interest, appellant status, and outcome from Songer hand-coded variables, and how we map SCDB focal interests to Songer focal interests for cross-scheme validation.

A.3.6.1 Deriving Songer focal interest and Outcome from Ground Truth

Our LLM prompt asks the model to identify a focal interest, whether the focal interest was the appellant, and whether that interest prevailed. The Songer codebook does not record these variables directly, but they are implied by the combination of issue area, case type subcode, disposition, and direction. We first describe the focal interest options, then explain how focal interest is derived from the case outcome, as well as how the valid focal interest options and GT direction determine the focal-interest-won and focal-interest-is-appellant variables.

For areas 1–5, each area specifies a single forced focal interest. Hand-coded direction determines the outcome, since a win for the focal interest is always coded as liberal (except for reverse-discrimination subcodes `223`, `224`, `234`, and `235` under civil rights, where it is coded as conservative).

For area 6 (Labor), the permitted focal interests are: `union`, `government`, `management`, and `worker`. These are not dependent on subcodes, and mechanically determine direction. A win for `union` or `government` is always coded as liberal; a win for `management` is always coded as conservative; a win for `worker` is coded liberal unless the issue area subcode is `609`, in which case it is coded conservative.

For area 7 (Economic), we separate the broad issue area into fifteen groups of issue area subcodes, described in Table A.5. The groups are mutually exclusive except for subcode `732` (disputes over government contracts), which is both in the "commercial disputes" group and the "government contracts/seizure" group. Most groups have a single substantive focal interest, which determines direction within that group (the civil-RICO group, subcode `760`, allows two—`injured_party` or `underdog`—both liberal on a win); some groups also have `neither` as an option, which forces direction code 0.

Finally, for area 9 (Miscellaneous), we again separate the broad issue area by subcode, this time into 16 fully mutually exclusive groups of issue area subcodes, described in Table A.6. Most groups have a single substantive focal interest, which determines direction within that group; subcode `920` adds `neither`, which forces direction code 0; and subcodes `901`, `999`, and `000` only include `neither` and force direction 0.

Each Songer GT issue area subcode therefore determines a small set of valid focal interests. Most subcodes have a single valid focal interest, with an implied correct value of focal-interest-won given by the GT direction code (and, for the reverse-discrimination civil rights cases, GT issue area subcode). Subcodes that include one substantive focal interest and `neither` permit only substantive label when GT direction $\in \{1, 2, 3\}$, only `neither` when GT direction $= 0$, and both when GT direction is absent. Subcode `732` allows the combined focal interests of its two groups (so a 732 case with GT direction $= 3$ admits

Subcode group	Subcode(s)	focal interest options	Direction (focal wins)
Tax	701–706	`taxpayer`	1 (cons.)
IP	710–713	`ip_rights_holder` or `neither`	3 (lib.)
Torts	720–730, 780	`injured_party` or `neither`	3
Commercial	731–740	`underdog` or `neither`	3
Bankruptcy	741–743	`debtor`	3
Antitrust/mergers	744–746	`antitrust_enforcer`	3
Private securities	747	`underdog` or `neither`	3
Government regulation	748, 752, 755–759	`regulated_party`	1
Individual benefits	750–751	`claimant`	3
Government contracts/seizure	732, 773–774	`private_party`	1
Environmental/consumer	753–754, 765–766	`environmental_or_consumer`	3
Civil RICO	760	`injured_party`, `underdog`, or `neither`	3
Admiralty (personal injury)	761	`injured_party` or `neither`	3
Admiralty other, property, other econ.	762–764, 770, 772, 799	`underdog` or `neither`	3
Eminent domain	771	`property_owner`	1

Table A.5: Songer area 7 (Economic) focal interest groups. "Direction (focal wins)" is the code implied by the substantive focal interest winning (`neither` always carries direction 0). Groups are mutually exclusive except for subcode `732` (disputes over government contracts), which intentionally appears in both the "Commercial" group and the "Government contracts/seizure" group. Subcode `760` (civil RICO) admits either `injured_party` or `underdog` as the substantive focal interest, both winning on direction 3.

both `(underdog, won)` and `(private_party, lost)` as valid LLM outputs). Subcodes in broad issue area 6 allow all four focal interests, but require consistency between focal-interest-won and direction—for example, if direction is liberal, then `government` that wins and `management` that loses are both valid pairs. If the LLM emits `government` that loses, we mark the focal interest as correct and focal-interest-won as incorrect. To determine whether the focal interest was the appellant, we use the GT disposition to determine whether the appellant won or lost; combined with the implied focal interest outcome above, we then know that the focal interest was the appellant when appellant-won and focal-won agree, and the respondent when they disagree.

A.3.6.2 Mapping SCDB focal interests to Songer focal interests

SCDB focal interest lists differ slightly from Songer; in particular, SCDB broad issue areas 1-5 (Criminal, Civil Rights, First Amendment, Due Process, Privacy) contain a list of the possible claimants, rather than the single "rights claimant" type for each category.[43] To accommodate these differences, we create a separate list of valid SCDB focal interests for each Songer issue area subcode, guided by the valid Songer focal interests as described in Appendix A.3.6.1. The full mapping includes at least one valid SCDB focal interest for each Songer subcode and ensures that each SCDB focal interest the LLM can emit (except for the SCDB-specific "judicial review" focal interest) is targeted by at least one Songer subcode; it is available in the project repository.[44] The focal-interest-won and focal-

[43]Although, as with the Songer focal interests, we try to maintain all of them on the same side of the case, e.g., various types of civil rights claimants but never the government rejecting the claim.

[44]This repository is currently private; please contact the authors for access. We will make it public upon publication of this paper.

Subcode group	Subcode(s)	focal interest options	Direction (focal wins)
Interstate conflict	901	neither (forced)	0
Federalism	902	federal_power	3 (lib.)
Attorneys	903	attorney	3
Selective service	904	inductee	1 (cons.)
Magistrate / referee authority	905, 906	official	3
Indian law — criminal	910	defendant	3
Indian law — commercial/property	911, 912	indian_or_tribal	3
Indian law — vs. fed./state authority	913, 914	indian_or_tribal	1
Indian law — tribal regulation	915	tribal_regulation	3
Indian law — other	916	indian_or_tribal	3
International law	920	us_interest or neither	3
Immigration	921	immigrant	1
National security / Patriot Act	922, 923	individual	1
14th Amend. congressional power	924	congressional_power	3
Executive privilege	925	executive	3
Other / not ascertained	999, 000	neither (forced)	0

Table A.6: Songer area 9 (Miscellaneous) focal interest groups. All 16 groups are mutually exclusive at the subcode level. "Direction (focal wins)" is the code implied by the substantive focal interest winning. neither is the only focal interest for the three "forced-neither" subcodes (901, 999, 000), and is also a valid alternative on subcode 920 when no clear US interest exists. The focal interest indian_or_tribal intentionally has two different implied directions, following the codebook: a focal win is liberal in rows 911/912 and 916 but conservative in row 913/914.

interest-is-appellant variables are handled in the same way as their Songer counterparts, albeit respecting differences in SCDB direction codings (e.g., union antitrust, worker vs. union) when determining the direction-implied winner.

A.4 Alternative Pipelines

As noted in the main text, we used several supplementary approaches besides the headline DeepInfra-hosted run (see Section 4.1) to examine cross-infrastructure and cross-provider performance, shed light on irreducible LLM stochasticity, and explore whether in-prompt examples could lift accuracy. These results show some small variation but generally match headline accuracy claims to within a few percentage points.

A.4.1 Princeton HPC

To verify that the pipeline's implementability does not depend on third-party infrastructure, we used Princeton high-performance computing (HPC) infrastructure to reproduce outcomes on almost all pre-2010 cases.[45] Specifically, we served an FP8-quantized checkpoint of the same model via vLLM on a single node (4× NVIDIA A100 80GB GPUs, tensor-parallel size 4) with a 32,000-token context window. This window fits about 80% of opinions without truncation; the remainder are truncated with a marker. A full pass takes about six days of wall-clock time given a fixed cap on concurrent slots. The HPC run produced 408,938 cases, amounting to essentially complete coverage of *Federal Reporter* and *Federal Reporter, 2nd Series*, plus the first 732 volumes of the *Federal Reporter, 3rd Series*.

Tables A.7 and A.8 compare main and HPC accuracy on the common subset of cases coded by both runs.[46] In both tables, each group of columns imposes further restrictions: "Paired" includes all cases with common output, "Accuracy (vs. GT)" restricts to cases with common output matched to a ground truth source (Songer hand codings in Table A.7, the IDB in Table A.8), and "Disagreements only" restricts to cases with common output matched to a ground truth source *and* different outputs between Main and HPC. For that last group, rows whose validation uses a many-to-one crosswalk can sum to more than 100% because multiple LLM codes can simultaneously be valid (e.g., on a two-issue case, main can match one code and HPC can match the other, leading both to win despite inter-LLM disagreement). The two runs produce the same code about 85–95% of the time on almost all variables, with only case type subcodes (around 75% agreement; fine-grained, with hundreds of possible choices) substantially below that range. Accuracy compared to appropriate ground truth is within 1–2 percentage points on the bulk of variables, including our main targets—issue area, disposition, and direction. The main exception is an advantage ranging from about 5 to about 15 percentage points for the main run on some litigant-typology rows in Songer and IDB validation; this gain comes from a minor prompt refinement which improved DeepInfra performance but was not propagated to the HPC run.

[45] SLURM queue constraints prevented us from completing about 300 volumes of the *Federal Reporter, 3rd Series* or CL data, as well as from re-running cases that failed to parse on the first attempt.

[46] Because of this restriction, values for the main run differ slightly from the corresponding tables in the main text.

Variable	Paired (no GT) N	Paired (no GT) Agree	Accuracy (vs GT) N	Accuracy (vs GT) Main	Accuracy (vs GT) HPC	Disagreements only N_{dis}	Disagreements only Main wins	Disagreements only HPC wins
Broad issue area (Songer)	400,288	95.5%	20,303	90.4%	91.1%	953	37.7%	51.8%
Broad issue area (SCDB)	406,136	92.3%	20,533	88.0%	88.4%	1,349	63.8%	69.7%
Issue area subcode (Songer)	403,296	75.6%	20,410	54.0%	55.6%	5,068	22.1%	28.4%
Issue area subcode (SCDB)[a]	405,597	76.2%	14,053	42.4%	43.0%	2,886	22.8%	25.5%
Direction, 3-class (Songer)[b]	397,565	85.2%	18,476	80.6%	79.2%	2,261	51.7%	40.9%
Direction, 2-class (Songer)[b]	397,565	85.2%	17,068	85.3%	84.6%	1,900	52.5%	45.9%
Direction, 2-class (SCDB)[b]	405,300	86.2%	17,314	76.7%	77.1%	2,083	46.9%	50.7%
Appellant type (Songer)	400,282	91.7%	20,529	85.6%	80.4%	1,956	70.6%	16.1%
Appellant type (SCDB)	406,120	91.3%	20,773	84.4%	78.8%	2,053	70.9%	13.7%
Respondent type (Songer)	400,282	87.2%	20,523	84.4%	80.1%	2,680	58.5%	25.8%
Respondent type (SCDB)	406,117	89.3%	20,767	83.7%	79.6%	2,367	58.2%	22.2%
Disposition (exact code)	406,134	93.6%	20,741	86.8%	87.4%	1,099	28.8%	40.7%
Disposition (winning side)[c]	406,134	96.4%	20,741	92.7%	93.0%	554	38.4%	48.2%
Appealed from	400,282	84.0%	20,531	55.4%	57.3%	2,779	21.0%	34.9%
Threshold issues (TYPEISS)	400,282	86.8%	20,509	76.7%	76.6%	2,006	42.7%	41.8%

[a] Accuracy and disagreement columns restricted to cases where both pipelines emitted a crosswalk-included SCDB issue area subcode; the paired column has no such restriction.
[b] Direction rows use *super-mechanical* direction (the same variant used in Table 3), since both Main and the few-shot variant compute it. Both Songer direction rows have the same sample for the paired column; only the accuracy and disagreement blocks have matched GT, and can therefore drop GT = 2 cases.
[c] Same four-way collapse as Table 3, note c.

Table A.7: Headline accuracy: main run vs. Princeton HPC run, following Table 3. "Paired" columns compare Main to HPC whenever both produced output. "Accuracy" columns are vs. Songer GT, restricted to cases where both runs produced output. "Disagreements only" columns restrict to cases where Songer GT exists and Main disagrees with HPC. There, SCDB issue area (both broad and subcode) sums to more than 100% because multiple LLM codes can be valid. Similar logic applies to two-issue cases' Songer broad issue area and direction, as well as winning side disposition. For all other rows, the remainder reflects both LLM runs being wrong.

Variable	Paired (no GT) N	Paired (no GT) Agree	Accuracy (vs GT) N	Accuracy (vs GT) Main	Accuracy (vs GT) HPC	Disagreements only N_{dis}	Disagreements only Main wins	Disagreements only HPC wins
Nature of suit → Songer issue area[a]	400,288	95.5%	118,087	86.5%	85.5%	8,597	58.2%	45.4%
Nature of suit → SCDB issue area[b]	406,136	92.3%	117,410	79.2%	79.7%	11,949	60.3%	65.0%
Criminal appeal → Songer criminal issue	400,288	99.1%	46,997	99.1%	99.0%	217	67.7%	32.3%
Criminal appeal → SCDB criminal issue	406,136	98.9%	47,295	99.0%	99.2%	223	21.5%	78.5%
Agency appeal → Songer issue area[a]	400,288	95.5%	16,641	63.1%	65.6%	809	18.5%	70.6%
Agency appeal → SCDB issue area[b]	406,136	92.3%	17,171	66.4%	66.6%	235	34.5%	48.1%
Disposition → IDB outcome	406,134	93.6%	186,929	88.6%	88.4%	9,236	55.6%	51.6%
U.S. as appellant	400,282	97.2%	10,107	86.7%	70.5%	1,904	92.9%	7.1%
U.S. as respondent	400,282	95.3%	62,091	95.4%	93.2%	2,616	75.7%	24.3%
Pro se appellant → individual	400,282	93.9%	3,871	96.1%	95.9%	57	57.9%	42.1%
Pro se respondent → individual	400,282	93.0%	575	49.6%	41.6%	104	72.1%	27.9%
Appeal type → court appealed from	400,282	84.0%	187,585	89.4%	91.8%	21,984	65.7%	86.1%
Jurisdictional basis → authority type	394,030	83.3%	116,767	78.9%	79.9%	20,436	62.0%	67.8%

[a] Paired columns match with those of the other "a" row, since the LLM-side variable for both rows is Songer broad issue area. Their Accuracy and Disagreement columns differ because each row's IDB-side variable differs.
[b] Paired columns match with those of the other "b" row, since the LLM-side variable for both rows is SCDB broad issue area. Their Accuracy and Disagreement columns differ because each row's IDB-side variable differs.

Table A.8: IDB accuracy: main vs. HPC, following Table 7 with the same structure as Table A.7. Disagreement win rates for rows using many-to-one validation—nature of suit → Songer/SCDB issue area, agency appeal → Songer/SCDB issue area, disposition → IDB outcome, appeal type → court appealed from, and jurisdictional basis → authority type—may sum to more than 100%. The remaining rows (criminal appeal, U.S. as appellant/respondent, pro se appellant/respondent) are exact binary checks where the remainder reflects both LLM runs being wrong.

A.4.2 Other Inference Providers

During prompt development, we also ran the pipeline using additional inference providers (Together.ai and OpenRouter) and evaluated two closed-source models (OpenAI's GPT 5.1 and Anthropic's Claude Sonnet 4.6) on the first few volumes of the *Federal Reporter, 3rd Series*. These exploratory runs informed prompt design and provider selection but did not generate full-corpus validation tables and are not reported here. Closed-source model cost estimates referenced in Section 4.1 (about \$50K for GPT 5.1, about \$120K for Claude Sonnet 4.6) are extrapolations from the number of tokens used for the headline DeepInfra run, scaled at provider list prices. Prompt caching could reduce these costs by 25–50%, but even that substantial reduction would leave an order-of-magnitude gap between our open-weights approach and the commonly-used closed-source alternatives.

A separate diagnostic isolates same-provider, same-prompt LLM variability. Due to a batching error, we re-coded a sample of 2,145 *Federal Reporter* cases (1892–1924) under identical pipeline code, identical prompts, and identical model and provider (Qwen3-235B-A22B-Instruct-2507 via DeepInfra, 128K context, all four passes), and compared the two cold runs at the field level. Broad classifications are highly stable: Songer issue area agrees 96.7% across runs, SCDB issue area 85.8%, and primary party plus win/loss 85–88%. Direction is moderately stable in the low 80s: SCDB and Songer mechanical and super-mechanical direction all land in the 80–85% range, with SCDB second-pass direction at 84.4%. Granular issue subcodes are noisier (76–77%), and the most stochasticity-sensitive variables are the fine-grained Songer party subcategories (53.9% and 48.5% across runs, where adjacent categories are confusable and the model flips between them). The implication is that aggregate analyses on broad areas, directions, and focal interests should be robust to single-run noise, while analyses depending on litigant subcategories or specific issue subcodes are best validated against multi-run ensembles or majority voting.

A.4.3 Few-Shot Examples

Our main pipeline uses zero-shot prompting, working from the premise that LLMs naturally perform well at fact extraction and worked-example anchoring can introduce subtle area-, direction-, or litigant-side biases. To check whether this design choice harmed accuracy, we carefully developed a few-shot prompting variant of the pipeline, which we ran on a sample of about 5,000 cases spanning all topics and time periods. Specifically, we searched the Songer dataset for appropriate example cases, which we integrated into the prompt as 3,500-character case text excerpts[47] followed by example output. Since these examples were drawn from Songer, they had hand-coded target output for the majority of our variables (determined via crosswalk for SCDB examples) which we supplemented with adjustments made by Anthropic's Claude Opus 4.7 subject to review by the authors. Examples are introduced with explicit framing language clarifying that they illustrate output format only,

[47] We included the first 2,000 characters and last 1,500 characters of the opinion; for reference, this would be about the 10th percentile of opinion length (median length is about 15,000 characters).

not a coherent sequence of related coding decisions.

The first-pass SCDB and Songer prompts were augmented with six shared cases: two criminal, two economic, one civil rights, and one labor; with two liberal, two conservative, and two mixed (Songer) or unspecifiable (SCDB) dispositions. Because the main variation in second-pass SCDB and Songer prompts is driven by broad issue area, we selected three within-issue-area examples for each second-pass SCDB area (except area 11, Interstate Relations, where we only found one suitable case) and for all but two Songer issue areas: Songer issue areas 7 (Economic) and 9 (Miscellaneous) received 9 and 10 examples, respectively, since these are the most diverse issue areas. Second-pass examples included equal numbers of liberal, conservative, and mixed/unspecifiable ideological directions (except of course SCDB area 11, which had only an unspecifiable direction example, and Songer area 9, which had one extra mixed direction example).

Because appended examples increase per-case input-token cost substantially, we evaluated few-shot performance on 22 reporter-volume partitions: *Federal Reporter* (vols. 100, 200, 300), *Federal Reporter, 2nd Series* (vols. 100 through 900 by 100), and *Federal Reporter, 3rd Series* (vol. 1 plus vols. 100 through 900 by 100), totaling 4,578 cases spanning the full chronological range of the corpus.

Tables A.9 and A.10 compare the main zero-shot run against the few-shot variant on this 4,578-case subset, following the same approach as the HPC comparison in Appendix A.4.1. The two pipelines agree on the same code about 85-95% of the time across most variables. The exceptions are case type subcode (about 65-70%) and super-mechanical ideological direction (about 80-85%), unsurprising since the former has hundreds of possible codes and the latter is a composite of other variables. Accuracy against Songer ground truth is within 3 percentage points on every headline row but exact disposition, where Main is better by about 4 percentage points, with no consistent direction of improvement: few-shot does better on 2-class SCDB ideological direction but worse on 2-class and 3-class Songer ideological direction, and better on winning side disposition but worse on broad issue area. Against IDB ground truth the two pipelines are generally indistinguishable on all twelve rows (differences below 1 percentage point on accuracy, except for a few rows where Main is better by 2–4 percentage points; mixed wins on the head-to-head disagreement subsets). We read these results as evidence that the zero-shot baseline already captures the bulk of the signal worked examples might add, and that any marginal gains in some variables from few-shot prompting are not worth the additional complexity of example selection and input-token cost.

Variable	Paired (no GT) N	Paired (no GT) Agree	Accuracy (vs GT) N	Accuracy (vs GT) Main	Accuracy (vs GT) Few-shot	Disagreements only N_{dis}	Disagreements only Main wins	Disagreements only Few-shot wins
Broad issue area (Songer)	4,382	94.8%	229	95.2%	95.2%	9	55.6%	55.6%
Broad issue area (SCDB)	4,466	86.8%	233	91.8%	91.0%	30	76.7%	70.0%
Issue area subcode (Songer)	4,442	66.9%	233	53.2%	54.9%	60	23.3%	30.0%
Issue area subcode (SCDB)[a]	4,452	68.0%	161	43.5%	47.8%	49	24.5%	38.8%
Direction, 3-class (Songer)[b]	4,365	84.3%	209	84.2%	81.3%	19	63.2%	31.6%
Direction, 2-class (Songer)[b]	4,365	84.3%	192	85.9%	82.8%	19	63.2%	31.6%
Direction, 2-class (SCDB)[b]	4,436	83.2%	195	78.5%	79.5%	25	44.0%	52.0%
Appellant type (Songer)	4,382	94.7%	231	84.4%	83.5%	8	62.5%	37.5%
Appellant type (SCDB)	4,466	94.1%	235	83.0%	81.7%	8	62.5%	25.0%
Respondent type (Songer)	4,382	90.3%	231	84.4%	84.8%	20	35.0%	40.0%
Respondent type (SCDB)	4,466	92.0%	235	84.3%	85.1%	17	29.4%	41.2%
Disposition (exact code)	4,466	86.8%	235	88.5%	84.3%	27	59.3%	22.2%
Disposition (winning side)[c]	4,466	94.7%	235	92.3%	93.6%	7	14.3%	57.1%
Appealed from	4,382	83.9%	231	62.8%	60.6%	33	39.4%	24.2%
Threshold issues (TYPEISS)	4,382	87.0%	230	77.4%	74.3%	25	56.0%	28.0%

[a] Accuracy and disagreement columns restricted to cases where both pipelines emitted a crosswalk-included SCDB issue area subcode; the paired column has no such restriction.
[b] Direction rows use *super-mechanical* direction (the same variant used in Table 3), since both Main and the few-shot variant compute it. Both Songer direction rows have the same sample for the paired column; only the accuracy and disagreement blocks have matched GT, and can therefore drop GT = 2 cases.
[c] Same four-way collapse as Table 3, note c.

Table A.9: Headline accuracy: main run vs. few-shot variant on a 4,578-case sample of CAP (mod-100 volumes), following Table 3 with the same structure as A.7.

Variable	Paired (no GT) N	Paired (no GT) Agree	Accuracy (vs GT) N	Accuracy (vs GT) Main	Accuracy (vs GT) Few-shot	Disagreements only N_{dis}	Disagreements only Main wins	Disagreements only Few-shot wins
Nature of suit → Songer issue area[a]	4,382	94.8%	1,484	86.5%	86.3%	113	56.6%	54.0%
Nature of suit → SCDB issue area[b]	4,466	86.8%	1,499	78.1%	78.5%	253	60.1%	62.5%
Criminal appeal → Songer criminal issue	4,382	99.0%	639	99.5%	99.4%	1	100.0%	0.0%
Criminal appeal → SCDB criminal issue	4,466	97.8%	643	99.2%	96.7%	16	100.0%	0.0%
Agency appeal → Songer issue area[a]	4,382	94.8%	228	56.1%	56.6%	14	35.7%	42.9%
Agency appeal → SCDB issue area[b]	4,466	86.8%	231	59.3%	57.1%	15	53.3%	20.0%
Disposition → IDB outcome	4,466	86.8%	2,432	88.7%	87.1%	287	73.2%	59.9%
U.S. as appellant	4,382	98.9%	125	91.2%	89.6%	6	66.7%	33.3%
U.S. as respondent	4,382	97.4%	753	96.3%	96.8%	8	25.0%	75.0%
Pro se appellant → individual	4,382	96.1%	86	94.2%	95.3%	1	0.0%	100.0%
Pro se respondent → individual	4,382	95.5%	17	47.1%	47.1%	0	—	—
Appeal type → court appealed from	4,382	83.9%	2,424	89.2%	88.4%	295	76.9%	70.5%
Jurisdictional basis → authority type	4,277	84.8%	1,451	78.4%	78.9%	240	64.6%	67.5%

[a] Paired columns match with those of the other "a" row, since the LLM-side variable for both rows is Songer broad issue area. Their Accuracy and Disagreement columns differ because each row's IDB-side variable differs.
[b] Paired columns match with those of the other "b" row, since the LLM-side variable for both rows is SCDB broad issue area. Their Accuracy and Disagreement columns differ because each row's IDB-side variable differs.

Table A.10: IDB accuracy: main vs. few-shot on the 4,578-case mod-100 sample of CAP, following Table 7 with the same structure as A.7.

A.5 Supplemental Songer Validation Results

This appendix presents validation results for every LLM-coded variable in our pipeline output. We organize them into four tables: direction outputs (the headline ideological direction codes themselves), direction inputs (the case-level variables that feed the mechanical and super-mechanical ideological direction computations), other categorical variables (case-type subcode, litigant-type categories, applfrom, method, litigant 5-digit decoded fields), and other binary variables (threshold indicators, type-of-issue subtype indicators, authority basis indicators, and binary 5-digit-decoded fields). All comparisons use 22,921 Songer ground-truth cases linked via Federal Reporter citation crosswalk. Per-row N falls below this match count whenever the relevant Songer GT field is uncoded or marked "not ascertained."

A.5.1 Direction Outputs

As described in Section 4.3 and shown in Figure 1, we produce three separate ideological direction codes for each track: one-shot, mechanical, and super-mechanical. There are a few design choices in our validation approach, many of which interact with two-issue cases (as noted in Section 5.1, about 15% of Songer cases contain two issue area codes and a corresponding ideological direction code for each). In the main text, we take the straightforward approach of counting the LLM as accurate when it matches any of the Songer GT direction codes for its case, regardless of their origin. We provide a more detailed breakdown of performance, as well as a comparison across direction validation methods, in Table A.11. To do so, it will be useful to first define *shared-rule groups*: these are codebook-defined sets of issue area codes which have the same rule for coding ideological direction. For example, all cases in Songer broad issue area 1 (Criminal) are in the same shared-rule group, since the Songer codebook states that all criminal cases are coded as liberal if the criminal defendant is favored, conservative if they are disfavored, and mixed otherwise. However, Songer broad issue area 7 (Economic) contains multiple shared-rule groups which depend on issue area subcodes, e.g., tax cases are coded as conservative if the taxpayer wins, tort cases are coded as liberal if the injured plaintiff wins. These shared-rule groups are the backbone of mechanical direction coding, allowing us to automatically map issue area codes and outcomes to ideological direction; they are described as part of second-pass validation in Appendix A.3.6, and fully documented in the LLM data postprocessing script.

Turning to the table, we first note the three blocks of rows. The first covers the three Songer-track direction approaches. The second covers the three SCDB-track direction approaches, dropping cases where Songer GT codes direction = 2 (mixed) because SCDB has no equivalent code. However, we keep cases the LLM codes as SCDB broad issue areas 11 and 14 (Interstate Relations and Private Action, respectively) even though these cases are hardcoded to direction = "unspecifiable" and are therefore automatic errors. The third block drops these cases entirely to allow a clearer picture of SCDB accuracy without this codebook mismatch. A more sophisticated approach is to assign cases coded in those two areas a two-step precision credit. First, we compute the within-area precision: what share of SCDB 11 cases, crosswalked to Songer broad issue area 9, actually have Songer GT code 9? Then, of cases in the same shared-rule group as the selected case, what share have a Songer GT ideological direction code that matches the LLM's output? This approach, and the analogous one for SCDB 14 cases, essentially treats the LLM's direction code as wrong for sure with a probability equal to its broad-category error rate, and a random guess within the appropriate shared-rule group otherwise. The resulting accuracies are close to the values shown in the third block of rows.

We now turn to the columns. The first block covers cases where the LLM's issue area code is in the same shared-rule group as at least one of the Songer GT issue area codes. We then break this block down into cases with a single GT issue area code, two codes and a match to the first only, two codes and a match to the second only, and a match to both (because both GT issue area codes are in the same shared-rule group). In the first three sets, there is a single appropriate GT ideological direction code, and we report the share of LLM codes that match that GT code. In the final set, we allow the LLM to match any of the GT direction codes (i.e., either code if there are two different ones) since with both GT issue area codes in the same shared-rule group, we have no principled way of knowing if one of them is closer to the LLM's issue area code.

Next, the second block covers cases where the LLM's issue area code is not in the same shared-rule group as any of the Songer GT issue area codes. There are three potential approaches here, with slight modifications depending on whether there are one or two Songer GT issue area codes. First, we can treat all of these as failures; this approach is the most conservative, but ignores the fact that different shared-rule groups may have similar coding rules—e.g., a criminal sentence appeal and a civil rights petition by a prisoner fall in separate shared-rule groups (Songer broad issue areas 1 and 2 respectively) but have a qualitatively similar rule of treating a win by the prisoner as liberal. Second, reported in the *Exact* columns, we can treat these as successes whenever the LLM matches any of the GT direction codes—this approach is the loosest, and may credit the LLM with a "lucky guess" when its issue area code was wholly unrelated to the GT issue area code(s) but direction happened to match. Third, reported in the *Base* columns, we can take an in-between approach similar to our credit for hardcoded SCDB areas, treating the direction for a wrong shared-rule group as a random guess within that shared-rule group and assigning accuracy according to the base rate of direction codes in that group.

		At least one correct shared-rule group								No correct shared-rule group					
		Single GT		Multi: CT1 only		Multi: CT2 only		Multi: CT1=CT2		Single GT			Multi-issue		
Variable	Total	N	Acc	N	Acc	N	Acc	N	Acc	N	Exact	Base	N	Exact	Base
Songer — super-mech.	81.5%	14,991	86.3%	943	80.9%	436	75.2%	969	83.7%	2,960	58.6%	43.3%	438	72.4%	41.9%
Songer — mech.	79.9%	14,991	84.5%	943	81.4%	436	76.6%	969	82.8%	2,960	56.7%	46.2%	438	73.3%	45.4%
Songer — one-shot LLM	72.3%	14,991	74.9%	943	68.6%	436	67.7%	969	78.3%	2,960	58.5%	45.6%	438	75.6%	46.1%
SCDB (all) — super-mech.	74.0%	13,057	84.4%	609	79.1%	286	78.7%	1,163	83.6%	3,519	34.6%	32.0%	453	43.3%	30.8%
SCDB (all) — mech.	76.6%	13,057	86.1%	609	83.1%	286	82.5%	1,163	91.7%	3,519	38.1%	32.6%	453	51.9%	32.4%
SCDB (all) — one-shot LLM	74.7%	13,057	83.3%	609	82.4%	286	73.1%	1,163	91.3%	3,519	39.0%	31.5%	453	50.8%	32.0%
SCDB (excl. hardcoded) — super-mech.	79.7%	13,057	84.4%	609	79.1%	286	78.7%	1,163	83.6%	2,298	53.0%	49.0%	297	66.0%	46.9%
SCDB (excl. hardcoded) — mech.	82.6%	13,057	86.1%	609	83.1%	286	82.5%	1,163	91.7%	2,298	58.4%	49.9%	297	79.1%	49.4%
SCDB (excl. hardcoded) — one-shot LLM	80.5%	13,057	83.3%	609	82.4%	286	73.1%	1,163	91.3%	2,298	59.7%	48.2%	297	77.4%	48.8%

Table A.11: Validation of all ideological direction methods decomposed by shared-rule group matches. The first block reports the number of cases and direction accuracy when the LLM's issue area code matches at least one Songer GT issue area code's shared-rule group, broken down by the type of match. The second block reports the number of cases, exact match rate, and base-rate credit score when the LLM's issue area code matches none of the Songer GT issue area codes' shared-rule groups. The *Total* column checks within-shared-rule-group matches for cases in the first block and exact matches for cases in the second block. Rows are grouped into three blocks: Songer-track, SCDB-track with hardcoded broad issue areas (Interstate Relations and Private Action) included, and SCDB-track with hardcoded broad issue areas dropped.

With these thorough preliminaries covered, we can summarize the table's results. Super-mechanical is the best-performing method for Songer-track ideological direction, but the worst for SCDB-track direction, with one-shot about one percentage point ahead and plain mechanical generally about 3 percentage points ahead. Table A.12 will show that this cross-track difference is mostly driven by a gap in appellant identification. For about 65-70% of cases, there is a single GT shared-rule group, which the LLM correctly identifies; for about 10% it correctly identifies (at least) one GT shared-rule group. Direction accuracy is generally higher in the former case than the latter, especially when the matching shared-rule group is given by the second GT issue area code; accuracy on that subset is about 5-10 percentage points lower.

For the remaining 20-25% of cases, the LLM does not match any of the GT shared-rule groups. As expected, it is more likely to match the exact GT direction code when there are two codes to match, but even with a single GT direction code it is always more likely to match exactly than a random guess would imply (by a margin of around 5-15 percentage points, with the smallest gaps for SCDB super-mechanical). This result strongly suggests that the LLM is choosing shared-rule groups with similar ideological direction rules as discussed earlier rather than simply guessing, and argues in favor of the simple "does the LLM match any GT direction code" validation approach used in the main text. Finally, note that to compute the *Total* column, if the LLM matches at least one shared-rule group, we force it to match that shared-rule group's direction code. In the main text, for two-issue cases we allow it to match the direction code from an unmatched shared-rule group, increasing overall accuracy slightly.

A.5.2 Direction Inputs

In Table A.12, we assess accuracy for the inputs to ideological direction: shared-rule group (the partial refinement of broad issue area that feeds mechanical and super-mechanical direction), disposition (including the winning side collapse that feeds super-mechanical direction), and our novel second-pass variables (focal interest, focal interest won, and focal interest is appellant). These results complement the similar breakdown of direction accuracy by error type in Tables 4 and 5 of the main text, but with a focus on the actual variables coded by the LLM rather than the (interpretable) error type. For the three novel variables, GT values are derived following the procedure in Appendix A.3.6, which produces two separate chains of values for two-issue cases. Since the focal interest options are directly affected by the chosen shared-rule group, we report accuracy only within chains where the shared-rule group was correct, treating wrong shared-rule group as an automatic error in downstream variables. In the parenthetical term, we follow Appendix A.5.1: to partially credit the LLM for shared-rule group errors that are still conceptually close we add accuracy as given by the base rate of the LLM's emitted value in its chosen shared-rule group.

When the LLM chooses the correct shared-rule group, it never outputs a focal interest which is not appropriate for that shared group. The parenthetical credit here is small, because focal interests are almost always unique to particular shared-rule groups. However, that

Variable	Accuracy	N
Shared-rule group		
Songer (exact match)	82.0%	22,656
SCDB (mapped)	75.9%	22,656
Disposition		
Exact (10-code)	87.0%	22,850
winning side (4-code collapse)	92.9%	22,850
focal interest		
Songer (exact match)	82.0% (83.3%)	22,656
SCDB (many-to-one, mapped)[a]	75.9% (77.0%)	22,656
focal interest won?		
Songer	69.2% (77.2%)	20,878
SCDB	63.7% (70.6%)	20,878
focal interest was appellant?[b]		
Songer	73.5% (83.2%)	18,415
SCDB	68.9% (79.2%)	18,415
SCDB first-pass winning side		
Mapped to Songer disposition (3-code collapse)	86.4%	22,153

[a] SCDB focal interest uses a per-area focal interest vocabulary (e.g. "accused," "underdog," "govt") mapped many-to-one to Songer's focal interest; see Appendix A.3.6.2.
[b] GT is-appellant derived from focal interest win and disposition: focal won with disposition $\in \{1, 8\}$ (affirmed/dismissed) implies focal is respondent; focal won with disposition $\in \{2, 3, 4, 7\}$ (reversed/vacated/remand) implies focal is appellant; mirror logic for focal lost. Cases with partial dispositions $\in \{5, 6\}$ drop because the winner is ambiguous.

Table A.12: Validation: direction inputs.

correction becomes much more significant for the next two variables, where it adds about 7-10 percentage points to each. Across all three of the novel variables, the Songer-track LLM outperforms the SCDB-track LLM, with gaps of about 5 percentage points throughout.

A.5.3 Other Categorical Variables

Table A.13 reproduces many of the results from Table 3, but adds some detail about the precise validation structure as well as a handful of new variables. Worth noting is that Songer GT codes litigant category twice over—once in a broad category, and once in a five-digit code, of which the first digit is a broad category; there is occasional mismatch between these two GT codings. The remainder of the five-digit code can be decomposed into per-category sub-fields that our LLM emits as separate variables. For instance, if our LLM's first pass coded the appellant as a business, the second pass codes what type of business; in the Songer database, this coding simply appears as part of the 5-digit appellant code. In Table A.14, we validate each of our LLM sub-categorizations by decoding the GT 5-digit code at the corresponding digit position, restricted to the subset where the LLM and GT agree on the broad category (digit 1).[48]

[48]Table 3 in Section 5.1 of the main text already validates the broad appellant and respondent categories. An incorrect broad litigant category directly implies incorrect sub-fields, so we focus on correct broad categories to capture the most detail possible about where the LLM makes additional mistakes.

Variable	Accuracy	N
Issue area subcode		
Songer (exact match)	54.1%	22,656
SCDB (mapped, set-membership)	42.3%	15,529
Litigant category (broad, 9 codes)		
Appellant — Songer	85.7%	22,910
Respondent — Songer	84.4%	22,903
Appellant — SCDB (mapped)	84.7%	22,909
Respondent — SCDB (mapped)	83.9%	22,901
Litigant category (from Songer 5-digit code, first digit only)		
Appellant — Songer	84.6%	22,907
Respondent — Songer	84.2%	22,892
Songer-only variables (no SCDB equivalent)		
Threshold issue type	76.9%	22,888
Court appealed from	55.8%	22,912
Method of decision and appeal initiator		
Method — Songer (collapsed to 3 categories)	77.7%	18,973
Method — SCDB (mapped, set-membership)	97.9%	18,973
Appeal initiator — SCDB (mapped)	70.8%	22,877

Table A.13: Validation: other categorical LLM variables not in the direction tables. Variables not validated here: Songer authority code (the LLM emits a single code 1–5 but Songer GT records authority as three separate binary indicators—see Table A.17 authority block); SCDB legal provisions (free-text response; per-indicator derivations also in Table A.17); and free-response "other focal interest" fields from the SCDB and Songer second passes.

Full confusion matrices for categorical variables are easily generated from LLM output, but are omitted here. Key qualitative patterns include litigant-type confusions dominated by business ↔ individual (sole proprietors named instead of business entities); method confusion concentrates on en banc and rehearing (rare proceedings); and who-appealed misclassifies government-agency-initiated appeals as plaintiff-initiated.

A.5.4 Other Binary Variables

Threshold and type-of-issue subtype variables are rare events (1–5% base rate), so we report false positives, false negatives, precision, and recall (rather than simply accuracy). Across both groups the LLM codes conservatively, with low false-positive rates—it rarely flags a feature that wasn't there—but it misses many genuine instances, and because these features are rare even a low false-positive rate translates into moderate or low precision. The 12 threshold flags average 31.6% precision / 31.4% recall; the 52 type-of-issue subtype indicators average 31.6% precision / 26.6% recall, with the per-indicator distribution shown in Figure A.1.

The type-of-issue indicators partition into four groups by the case's primary type-of-issue. Recall is lowest in two regimes: routine procedural notes that the LLM does not surface as headline issues (e.g., trial or post-trial procedure, judge discretion), and administrative-

Appellant side			Respondent side		
Sub-field	Acc.	N	Sub-field	Acc.	N
Business: scope	31.8%	5,072	Business: scope	35.4%	5,047
Business: sector	65.8%	4,516	Business: sector	64.2%	4,411
Business: subtype	44.7%	4,368	Business: subtype	44.4%	4,238
Organization: type	94.5%	725	Organization: type	94.0%	367
Organization: subtype	55.3%	725	Organization: subtype	39.3%	364
Federal govt: branch	50.8%	1,438	Federal govt: branch	75.9%	8,826
Federal govt: entity	48.4%	1,174	Federal govt: entity	64.8%	4,169
Sub-state govt: type	16.8%	309	Sub-state govt: type	26.0%	569
Sub-state govt: entity	50.0%	304	Sub-state govt: entity	50.5%	564
State govt: branch	27.5%	436	State govt: branch	42.5%	1,376
State govt: entity	47.2%	426	State govt: entity	51.9%	1,351
Misc: litigant type	89.1%	311	Misc: litigant type	81.3%	369
Misc: litigant subtype	54.2%	301	Misc: litigant subtype	38.4%	357

Table A.14: Validation: litigant categorical fields decoded from Songer 5-digit appellant and respondent codes. Each row is restricted to the subset where the LLM's broad-category coding matches the parenthetical category. Categories 6 (gov't level not ascertained) and 9 (not ascertained) have no sub-fields. Category 7 (natural person) sub-fields are binary and appear in Table A.17.

record indicators that require explicit codebook vocabulary the opinion text does not always use (de novo review, erroneous review, agency discretion, administrative law judge). Full results for these are in Table A.16.

Table A.17 reports authority-basis indicators (constitutional, federal statute, and procedural) derived two ways—from the LLM's single Songer authority code (1–5) collapsed to the three Songer binaries, and from the SCDB legal-provisions field's area code (1/2 → constitutional, 3 → federal statute, 4 → procedural)—plus the SCDB procedural-holding indicator and the natural-person binaries.

Indicator	N	GT+	FP	FN	Prec.	Recall
Jurisdiction	22,896	1,680	11.4%	32.1%	32.0%	67.9%
Standing	22,897	426	1.7%	40.8%	39.1%	59.2%
Mootness	22,896	247	0.8%	43.7%	43.7%	56.3%
Ripeness / exhaustion	22,897	378	1.1%	61.9%	37.5%	38.1%
Timeliness	22,897	747	2.1%	48.5%	45.3%	51.5%
Immunity	22,897	477	0.7%	69.6%	47.4%	30.4%
Frivolous case	22,897	73	0.2%	83.6%	20.7%	16.4%
Frivolous appeal	22,897	84	0.2%	82.1%	24.6%	17.9%
Political question	22,896	41	0.2%	80.5%	18.6%	19.5%
Other trial threshold	22,897	704	0.9%	92.6%	20.0%	7.4%
Late appeal	22,897	189	0.2%	89.4%	32.3%	10.6%
Other appellate threshold	22,895	968	0.2%	98.8%	18.2%	1.2%

Table A.15: Threshold validation: Songer threshold indicators (12 total) are coded as rare-event binaries, showing precision (for LLM-flagged cases, how often they are also flagged by Songer) and recall (how often the LLM flags cases that Songer flagged).

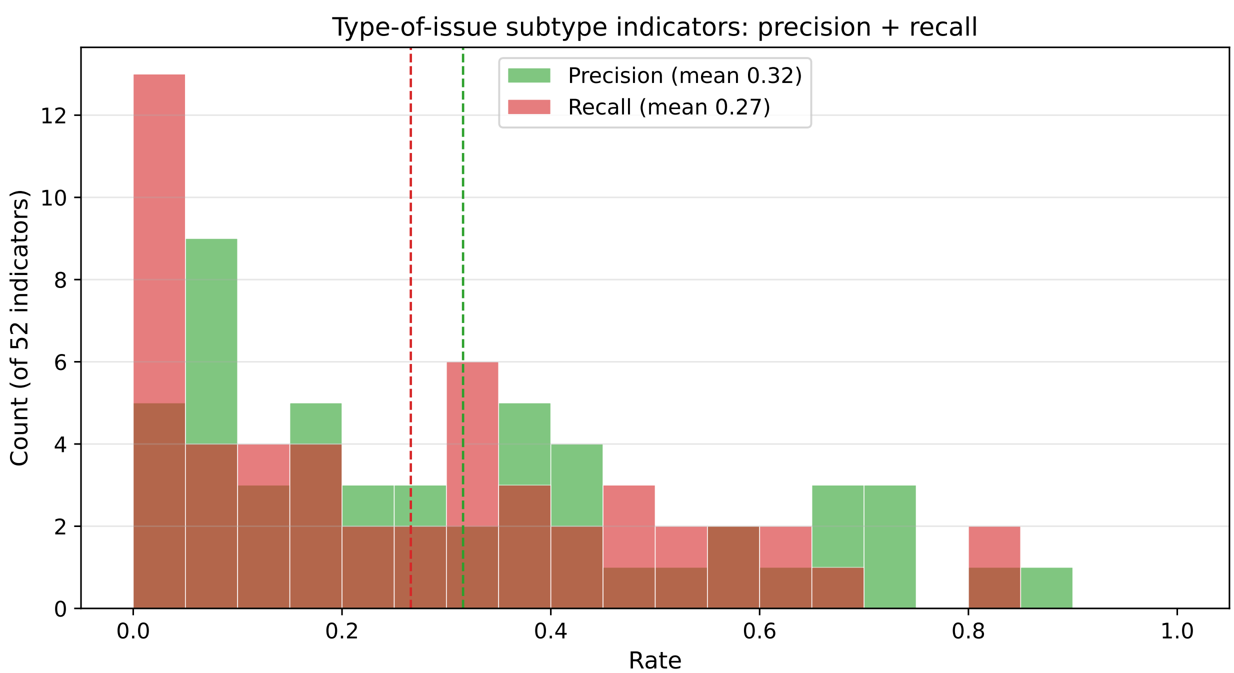

Figure A.1: Distribution of per-indicator precision and recall across the 52 Songer type-of-issue subtype indicators (Table A.16); each count is over the 52 indicators. The LLM's conservative coding yields mostly low recall and only moderate precision.

Indicator	N	GT+	FP	FN	Prec.	Recall
Prosecutorial misconduct	6,013	418	0.9%	85.2%	56.4%	14.8%
Insanity defense / competence	6,066	85	0.7%	60.0%	44.2%	40.0%
Improper jury influence	6,055	114	0.9%	89.5%	18.5%	10.5%
Jury instructions	6,008	909	3.1%	63.7%	67.6%	36.3%
Jury composition / bias	6,066	216	1.1%	76.4%	44.0%	23.6%
Death penalty	6,069	75	0.3%	82.7%	41.9%	17.3%
Sentencing	6,058	1,151	2.0%	44.0%	87.0%	56.0%
Defective indictment	6,063	530	0.8%	80.2%	69.1%	19.8%
Confession admissibility	6,054	384	5.5%	46.1%	39.9%	53.9%
Search and seizure	6,055	708	4.6%	16.9%	70.6%	83.1%
Other evidence admissibility	6,000	1,214	6.1%	65.8%	58.9%	34.2%
Plea bargaining	6,065	184	2.9%	50.5%	34.7%	49.5%
Inadequate counsel	6,062	406	3.8%	39.4%	53.6%	60.6%
Right to counsel	6,066	240	0.4%	92.9%	39.5%	7.1%
Insufficient evidence	6,067	1,447	8.9%	33.9%	70.1%	66.1%
Indigent-defendant rights	6,068	42	0.2%	85.7%	30.0%	14.3%
Entrapment	6,068	107	0.6%	45.8%	61.1%	54.2%
Procedural dismissal	6,067	122	8.5%	68.0%	7.2%	32.0%
Other criminal issue	6,055	1,436	9.6%	77.5%	42.1%	22.5%
Due process	16,822	668	16.3%	38.5%	13.5%	61.5%
Executive order / regulation	16,822	472	1.1%	95.8%	9.9%	4.2%
State / local law	16,822	2,901	16.6%	71.4%	26.4%	28.6%
Weight / sufficiency of evidence	16,821	4,590	19.6%	54.4%	46.6%	45.6%
Pre-trial procedure	16,822	645	0.7%	97.1%	15.2%	2.9%
Trial procedure	16,818	1,585	0.8%	82.8%	70.1%	17.2%
Post-trial procedure	16,822	702	0.8%	94.3%	23.3%	5.7%
Attorneys' fees	16,822	681	0.3%	69.8%	81.1%	30.2%
Trial-judge discretion	16,841	1,317	2.8%	94.2%	15.2%	5.8%
Alternative dispute resolution	16,841	293	0.1%	99.7%	4.3%	0.3%
Injunction	16,822	771	4.3%	51.9%	35.2%	48.1%
Summary judgment	16,819	1,417	26.3%	16.2%	22.7%	83.8%
Federal vs. state law	16,822	413	3.8%	69.5%	16.9%	30.5%
Domestic vs. foreign law	16,821	46	0.1%	100.0%	0.0%	0.0%
Treaty / international law	16,822	56	0.3%	92.9%	8.2%	7.1%
Conflict of laws (other)	16,822	65	5.0%	73.8%	2.0%	26.2%
Discovery	16,822	265	0.3%	86.4%	39.6%	13.6%
Other civil issue	16,821	1,205	39.2%	60.4%	7.2%	39.6%
Substantial-evidence review	11,599	878	1.5%	61.4%	67.5%	38.6%
De novo (agency) review	11,599	85	0.6%	98.8%	1.4%	1.2%
Clearly-erroneous review	11,599	301	0.1%	99.7%	5.9%	0.3%
Arbitrary-and-capricious review	11,587	440	2.1%	65.5%	39.8%	34.5%
Deference to agency discretion	11,598	421	0.2%	99.3%	12.0%	0.7%
Reviewability	11,599	226	3.4%	82.7%	9.1%	17.3%
Agency statutory interpretation	11,599	1,366	1.4%	96.0%	28.4%	4.0%
Agency notice	11,599	176	0.2%	97.2%	20.8%	2.8%
Administrative law judge	11,598	380	0.2%	98.9%	18.2%	1.1%
Agency information acquisition	11,599	62	0.1%	98.4%	6.2%	1.6%
Agency disclosure / FOIA	11,598	78	0.4%	65.4%	34.6%	34.6%
Opportunity to comment	11,598	56	0.2%	100.0%	0.0%	0.0%
Adequacy of agency record	11,598	117	0.2%	98.3%	7.7%	1.7%
Diversity jurisdiction	3,111	90	14.6%	43.3%	10.4%	56.7%
Choice of law	3,111	96	16.2%	59.4%	7.4%	40.6%

Table A.16: Type-of-issue validation: Songer type-of-issue subtype indicators (52 total) are coded as rare-event binaries, showing precision (for LLM-flagged cases, how often are they correct, i.e., flagged by Songer) and recall (how often does the LLM correctly flag cases that Songer flagged). The histogram of these per-indicator values appears in Figure A.1.

Indicator	N	GT+	FP	FN	Prec.	Recall
Authority basis + related indicators[a]						
Constitutional (Songer authority code)	22,480	2,206	17.1%	31.5%	30.4%	68.5%
Federal statute (Songer authority code)	22,473	6,267	37.5%	37.8%	39.0%	62.2%
Procedural (Songer authority code)	22,472	16,753	15.6%	86.0%	72.4%	14.0%
Constitutional (SCDB legal provisions)	22,873	2,233	41.0%	8.6%	19.4%	91.4%
Federal statute (SCDB legal provisions)	22,866	6,336	81.3%	0.9%	31.8%	99.1%
Procedural (SCDB legal provisions)	22,865	17,041	16.1%	77.2%	80.6%	22.8%
Procedural holding (SCDB)[b]	22,866	17,042	7.3%	93.2%	72.9%	6.8%
Unconstitutionality (SCDB)[c]	22,874	2,233	1.2%	85.9%	55.5%	14.1%
Litigant binary (decoded from the Songer 5-digit litigant code, cat 7 = natural person)[d]						
Appellant gender (female)	10,246	1,410	1.0%	8.3%	93.4%	91.7%
Appellant citizenship (non-citizen)	489	339	8.0%	8.0%	96.3%	92.0%
Respondent gender (female)	1,983	489	5.0%	7.8%	85.9%	92.2%
Respondent citizenship (non-citizen)	74	41	6.1%	24.4%	93.9%	75.6%

[a] Only about 10% of Songer cases have any of the three legal authority types we match to (constitutional, federal statute, and procedural) explicitly coded—the remainder have all three left as "not specified" (0).
[b] The very high FN rate reflects a semantic gap: the LLM indicator means "primary holding is procedural" (narrow) while Songer's authority-code-derived indicator means "has any procedural authority basis" (broad).
[c] This row compares SCDB unconstitutionality $\in \{1, 2\}$ (facial or as-applied strike-down) to Songer's constitutionality > 0 (case holding rests on constitutional analysis), but the two concepts differ in scope—SCDB narrowly identifies strike-downs while Songer broadly identifies constitutional reasoning. Low recall is expected, while moderate precision largely reflects Songer's sparse legal-authority-is-constitutional coding (see note a), which leaves many constitutional-analysis cases unmarked and thus counted as false positives.
[d] Binary (collapsed) accuracy for the natural-person rows, where the digit-exact-match accuracy understates performance: appellant gender 98.0%, respondent gender 94.4%, appellant citizenship 92.0%, respondent citizenship 83.8%. Gender collapses Songer codes 1,2 (male) vs. 3,4 (female).

Table A.17: Validation: other binary LLM variables (FP/FN per indicator). Per-indicator rows for the 12 threshold and 52 type-of-issue subtype variables are in tables A.15 and A.16 respectively; this table holds authority-basis indicators and the natural-person binaries. Songer rarely codes gender and citizenship, leading to a dramatically reduced validation sample.

Keyword set in scan window	Songer `treat`
{affirm, revers, remand}	6
{affirm, revers}, no `remand`	5
{affirm, vacat, remand}	6
{affirm, vacat}, no `remand`	5
{revers, remand}, no `affirm`	3
{vacat, remand}, no `affirm`, no `revers`	4
{revers} alone	2
{vacat} alone	7
{affirm} alone (incl. `enforc` folded in)	1
{dismiss} or {den}, no other dispositive stem	8
{grant, cert} together (e.g. "petition for certiorari granted")	0
{cert} alone	9
{grant} alone (no `cert`)[a]	no map
{affirm, dismiss}, no other dispositive stem	flagged

Table A.18: Regex disposition rules. The same rules apply in both plain and strict variants and to both scan strategies (conclusion, walkback); only the per-stem matching differs.

A.5.5 Regex Disposition Detection

Of all our key variables, disposition is the most amenable to automatic extraction without using an LLM. It is generally expressed in a sentence close to the end of the opinion which uses a limited vocabulary, often in all-caps, e.g., AFFIRMED, REVERSED, VACATED. To further validate the LLM's coding, we attempt to detect the disposition using a simple regular-expression string-matching approach (we describe this output as "regex disposition"). We restrict the analysis to CAP cases from the twelve valid appellate circuits (1–11 and DC; see Section 3), excluding the Federal Circuit and other specialized federal courts. For each opinion we strip footnotes, split into sentences, and search for disposition stems: `affirm`, `revers`, `remand`, `vacat`, `dismiss`, `grant`, `den(y|ied|ies|ial)`, `enforc`, `cert(if|iorari)`. Hits are mapped to disposition buckets via the rules in Table A.18. We use two complementary scan strategies, as well as a hybrid approach which prefers the first and falls back to the second if necessary.

- Conclusion: if a conclusion header is found in the last 50% of the opinion (a labeled "Conclusion" / "Disposition" / "Decision" section, or a bare terminal roman-numeral section like "IV." on its own line), scan the entire conclusion block for hits.
- Walkback: search from the last sentence through up to five sentences; anchor at the first sentence whose hits map to a non-empty disposition bucket; then take the union of hits from the anchor through the end of the opinion (e.g., to catch the trailing-remand pattern "Affirm in part, reverse in part. We remand part as well.").

We also test two matching variants: plain (case-insensitive stem matching) and strict (matching the all-caps form only).

Table A.19 reports coverage and accuracy for each of the six regex configurations: the three

Configuration	Coverage	Exact match		winning side	
		vs LLM	vs Songer	vs LLM	vs Songer
Plain conclusion	22.6%	83.1%	81.7%	87.3%	85.9%
Plain walkback	85.3%	88.1%	86.6%	92.2%	90.0%
Plain hybrid	87.4%	87.3%	86.0%	91.5%	89.8%
Strict conclusion	12.6%	91.0%	92.4%	95.0%	95.1%
Strict walkback	25.4%	90.6%	91.2%	94.2%	93.8%
Strict hybrid	26.3%	91.4%	92.4%	95.2%	95.1%

Table A.19: Performance of regex disposition. As in Table 3, "exact match" compares to all ten Songer codes; "winning side" collapses to the four categories that feed direction coding. The LLM-coded disposition agrees with Songer GT at 87.0% exact and 92.9% winning-side.

scan strategies (conclusion-only, walkback-only, and hybrid) crossed with the two matching variants (plain, strict). Coverage is the fraction of the CAP corpus which the scan classifies into a Songer disposition bucket.[49] The four agreement columns score classified cases on two metrics. As with validation in the main text, "exact" is the accuracy when mapped to the appropriate Songer code, while "winning side" collapses those ten codes to appellant won, respondent won, mixed, or neither—the categorization which feeds super-mechanical direction coding. For each of those metrics, we report accuracy against the LLM-coded disposition and the Songer human-coded ground truth.

A conclusion section was detected in about one-quarter of the CAP dataset. When a conclusion was detected, it contained at least one disposition keyword 84.2% via the plain scanner and 47.1% via the strict scanner. The walkback method finds disposition keywords almost 90% of the time; it rarely fails to detect a keyword in a case where the conclusion section succeeds (which would occur only if the keyword was in the conclusion but more than five sentences from the end). As expected, the strict methods have far lower coverage, detecting all-caps keywords in only about one-eighth (conclusion) or one-quarter (walkback or hybrid) of cases. Performance for the plain methods is slightly below the LLM, while performance for the strict methods is slightly above, but both gaps are small—around 5% in either direction, with all methods agreeing with the LLM around 85–90% of the time.

Table A.20 reports per-disposition coverage, as well as recall conditional on assigning a disposition. The upper section uses the ten exact Songer codes; Plain hybrid is comparable to the LLM overall with slightly worse coverage, with the only large gaps in recall on pure vacate dispositions and stays, both of which are rare. Strict hybrid generally outperforms the LLM, but has far worse coverage. The lower section collapses to the four-class winning-side categorization that feeds super-mechanical direction; here, the LLM distances itself from plain hybrid regex and closes the gap to within a couple percentage points of strict hybrid regex. Only the "Neither" bucket (non-merits dispositions) remains weak; the LLM actually significantly outperforms both regex methods here, since non-merits dispositions often lack clear regex-findable keywords.

[49] For the remaining cases, none of the target patterns in Table A.18 were present in the text.

Bucket	GT N	Plain hybrid hit %	Plain hybrid recall	Strict hybrid hit %	Strict hybrid recall	LLM benchmark hit %	LLM benchmark recall
Exact code (10-class, Songer abbrev.)							
Affirm	12,443	89.0%	93.9%	17.6%	98.0%	99.9%	96.8%
Vac.+Remand	977	88.2%	83.3%	21.3%	88.0%	100.0%	88.5%
Pet.Den./Dismiss	1,341	81.1%	86.5%	12.3%	89.1%	99.6%	82.3%
Rev.+Remand	3,900	81.7%	84.0%	17.4%	92.3%	100.0%	82.1%
Reverse	1,430	88.5%	86.5%	16.5%	94.1%	99.9%	81.3%
Aff./Rev.Part+Rem	1,224	87.5%	71.1%	34.8%	81.7%	99.9%	80.1%
Cert.	26	69.2%	55.6%	15.4%	50.0%	96.2%	52.0%
Aff./Rev.Part	901	80.1%	36.4%	15.5%	62.9%	100.0%	35.4%
Vacate	142	84.5%	65.0%	12.0%	76.5%	99.3%	28.4%
Stay	433	66.1%	0.7%	5.3%	0.0%	98.6%	24.4%
Total	22,817	86.3%	86.0%	17.9%	92.4%	99.9%	87.0%
winning side (4-class)							
Respondent won	13,784	88.2%	95.8%	17.1%	98.6%	99.9%	96.6%
Appellant won	6,449	84.2%	90.5%	17.6%	97.4%	100.0%	94.2%
Mixed	2,125	84.4%	61.4%	26.6%	80.4%	100.0%	79.1%
Neither	459	66.2%	4.3%	5.9%	7.4%	98.5%	26.8%
Total	22,817	86.3%	89.8%	17.9%	95.1%	99.9%	92.9%

Table A.20: Per-disposition performance for regex disposition. The upper block uses the ten Songer disposition codes, sorted by LLM conditional recall to match Figure 3, Panel (A); the lower block collapses both axes to the four-class winning-side categorization, matching that figure's Panel (B).

Overall, we see that a strict regex, which only matches all-caps dispositions (similar to the all-caps judge name extraction used in text scraping; see Appendix A.1.1) is highly accurate but lacks sufficient coverage to be viable, while the looser plain regex is a reasonable and low-cost substitute for LLM coding. As a measure of LLM accuracy, these regex approaches suggest some small room for LLM improvement, especially on simpler dispositions like pure vacate rulings or dismissals, but broadly agree with the LLM and further support its strong performance. As with ideological direction, where the LLM approach can be combined with litigant-based or judge-based measures, LLM-coded disposition and regex-extracted disposition are complementary signals: where they agree the disposition is essentially certain, and where they disagree the LLM can help close coverage gaps and shed light on more complex disposition text.

A.6 Supplemental IDB Validation Results

Appendix A.2 used the IDB to validate metadata our pipeline reads directly from the opinion text. This appendix provides details for the validation of substantive LLM-coded variables—issue area, disposition, litigant identity, and ideological direction.

A.6.1 IDB-to-Songer Crosswalk Construction

IDB uses coding schemes which do not correspond one-to-one to SCDB or Songer, so validation requires crosswalk maps from IDB codes to sets of plausible values in our schemes. We use two check types:

1. **Set-membership** (issue area, disposition, appealed-from, authority type): an IDB code maps to a *set* of plausible LLM codes—e.g., IDB NOS 470 (RICO) $\rightarrow$ Songer broad issue area $\in \{6, 7\}$ (Labor for union racketeering, Economic otherwise). Accuracy is the share of cases whose LLM code falls in the set.
2. **Precision/recall** (criminal appeal, federal government litigant, pro se litigant): one source flags a binary condition (e.g., IDB USAPT=1 means the federal govt is the appellant); we report the share of IDB-flagged cases the pipeline also flags.

The most complex IDB variable is nature of suit (NOS). Our NOS $\rightarrow$ Songer crosswalk uses a two-tier rule. First, most NOS codes have an unambiguous Songer area based on the label itself (NOS 442 "Civil Rights: Jobs" $\rightarrow$ Songer broad issue area 2 "Civil Rights"; NOS 870 "Federal Tax Suits" $\rightarrow$ Songer 7 "Economic"); these we hand-assign. For codes without a principled single-area match—typically residual "other" buckets (NOS 890, NOS 190)—we fall back on Songer GT: whenever $\geq$10 cases and $\geq$10% of hand-coded NOS-cases fall into a given Songer broad issue area, that area enters the set. NOS codes with fewer than 10 Songer GT cases default to the hand-assigned mapping. The NOS $\rightarrow$ SCDB crosswalk is derived from the Songer crosswalk by applying the same Songer broad category $\rightarrow$ SCDB broad category mapping we use elsewhere: 1$\rightarrow$\{1\}, 2$\rightarrow$\{2\}, 3$\rightarrow$\{3\}, 4$\rightarrow$\{4\}, 5$\rightarrow$\{5\}, 6$\rightarrow$\{7\}, 7$\rightarrow$\{8, 12, 14\}, 9$\rightarrow$\{6, 9, 10, 11, 13\}.

For disposition, our IDB OUTCOME $\rightarrow$ disposition crosswalk uses the same 10% threshold we applied to residual "other" NOS codes. Notably, IDB OUTCOME 2 (reversed) maps to Songer 2 (reversed), 3 (reversed-and-remanded) and 4 (vacated-and-remanded), reflecting IDB's ambiguity in coding "reversed and remanded." The same two Songer targets also apply for IDB OUTCOME 6 (remanded).

The remaining variables are more straightforward. The name of a federal agency in IDB maps to SCDB regulatory areas, e.g. Social Security Administration to issue 8 (Economic), or IRS to issue 12 (Tax). IDB APPTYPE describes the source of an appeal and maps to expected Songer `applfrom` codes (e.g., administrative review $\rightarrow$ 12; civil $\rightarrow$ any district-court code). IDB JURIS maps to expected Songer authority types (federal question $\rightarrow$

constitutional or statutory). IDB also includes flags for federal government involvement and pro se appellants/respondents; these have Songer codebook equivalents whose definitions are slightly different from, but often comparable to, the IDB's.

A.6.2 Binary Confusion Matrices

Table 7 in the main text reports recall for each binary check—the share of IDB-positive cases the LLM also flags. The 2×2 matrices below add false positives and the derived precision and recall. Denominators differ across checks because each matrix requires valid LLM and IDB codes only for its specific variables.

Criminal appeal (Songer)

	IDB +	IDB −
LLM +	72,059	24,998
LLM −	557	186,995

Precision 74.2% *Recall* 99.2%

Criminal appeal (SCDB)

	IDB +	IDB −
LLM +	71,940	30,956
LLM −	676	181,037

Precision 69.9% *Recall* 99.1%

Federal government as appellant

	IDB +	IDB −
LLM +	11,523	6,902
LLM −	1,755	264,427

Precision 62.5% *Recall* 86.8%

Federal government as respondent

	IDB +	IDB −
LLM +	75,823	50,381
LLM −	3,710	154,692

Precision 60.1% *Recall* 95.3%

Pro se → individual appellant

	IDB +	IDB −
LLM +	8,230	179,106
LLM −	250	96,601

Precision 4.4% *Recall* 97.1%

Pro se → individual respondent

	IDB +	IDB −
LLM +	560	40,961
LLM −	573	242,092

Precision 1.3% *Recall* 49.4%

Compared to near-perfect recall for both tracks, precision on the criminal checks is moderate (about 75% for Songer and 70% for SCDB), likely because the LLM understandably struggles with cases on the criminal/civil rights boundary, such as petitions by prisoners. While a minority of these cases—e.g., habeas corpus petitions or challenges to a sentence—are designated as criminal appeals cases by the Songer codebook, most others are explicitly designated as civil rights cases, not criminal appeals. For example, Songer issue area subcode 204 in the civil rights broad issue area covers non-habeas prisoner petitions. Since the SCDB codebook categorizes criminal appeals by procedural issue, it may have more defensible mislabelings due to genuine codebook differences, but is also more likely to struggle with cases on the boundary between criminal appeals and civil rights or due process claims dealing with criminal procedure.

Federal government identification generally has strong recall, with the appellant role more challenging for both the LLM and Songer GT (see Table 7 in the main text) because many appellants are listed using their names, e.g., "John Johnson, Secretary of Agriculture" rather than "Secretary of Agriculture of the United States." This pattern is less common when a federal official is the respondent. Precision is significantly (20-35 percentage points) lower than recall because of definitional mismatch: the Songer codebook category we validate against includes all federal agencies regardless of name, while the IDB only includes litigants whose name literally contains "United States."

Definitional differences are again an issue for the pro se → individual litigant tables. For these, precision must be disregarded entirely: the IDB codes only litigants without counsel, while the LLM follows the Songer codebook and codes any individuals, e.g., criminal defendants with a lawyer. Thus precision mechanically punishes the LLM for detecting many individuals who are not representing themselves. As mentioned briefly in Section 5.3, a similar issue drives low recall for pro se respondents. While the IDB flags cases where *any* litigant is pro se, the LLM follows the Songer codebook and codes only the properties of the first litigant even if the case has multiple litigants on one or both sides. The reasoning behind this approach is that the first litigant listed is generally the lead for multi-litigant cases, and thus the most important. However, this lead role means they are less likely to represent themselves than the other litigants, resulting in low recall for both the LLM and Songer GT. By contrast, multi-appellant cases are much more likely to have an attorney representing all of the appellants together, which is why pro se appellant recall is much higher.

A.6.3 IDB-Mechanical Direction: A Post-Songer Audit

As an audit of super-mechanical direction outside the Songer GT window (1925–2002), we construct an *IDB-mechanical* direction variable that substitutes IDB's administrative disposition for the LLM's. The recipe: (i) derive an appellant-won indicator from IDB (affirmed → lost, reversed/remanded → won, aff-in-part → mixed); (ii) combine with the LLM's coding of whether the focal interest is the appellant to determine whether the focal interest won; (iii) map the IDB's NOS code to the appropriate rule in the Songer codebook, following a refined version of the NOS → Songer crosswalk (e.g., NOS 870, Federal Tax Suits, maps specifically to the Songer codebook rule for tax cases rather than the general Economic area). Table A.21 shows agreement between IDB-mechanical and super-mechanical direction.

Two findings stand out. First, on the 8,203 Songer-GT cases where both methods have the necessary inputs to produce a direction, IDB-mechanical is slightly less accurate than super-mechanical (77.6% compared to 79.4%); when the two disagree on a Songer-coded case, super-mechanical is right about 10% more often. IDB's 7-bucket disposition is coarser than Songer's, and the LLM reads partial-disposition nuances the administrative record compresses. Second, the two methods agree 84.5% on post-2002 cases, actually slightly higher than agreement during the Songer-IDB overlap window of 1971–2002. This agreement rate provides evidence that super-mechanical does not further diverge from administrative ground truth on modern opinions. The approximately 15% disagreement in modern cases is dominated by disposition-granularity mismatches; as discussed earlier, IDB lumps reversed-in-

	Comparison		Songer GT validation		
Era	N	Agreement	N w/ GT	IDB-mech acc.	Super-mech acc.
Songer window (year $\leq$ 2002)	130,566	82.4%	8,203	77.6%	79.4%
Post-Songer (year $\geq$ 2003)	64,942	84.5%	—	—	—

Table A.21: IDB-mechanical and super-mechanical direction by era. The *Comparison* block counts cases where both methods produce a non-null direction and reports the share where they agree. The *Songer GT validation* block restricts further to cases with a non-zero Songer hand-coded direction and reports each method's accuracy against that ground truth. On the Songer window row, IDB-mechanical and super-mechanical disagree on $1,471$ of $8,203$ cases (17.9%); super-mechanical matches the GT 52.1% of the time vs. IDB-mechanical's 42.0% (allowing matches to either GT code for two-issue cases).

whole with reversed-in-part-and-remanded, which Songer, and therefore our LLM, separates.

	(1) Role only		(2) Role × alignment	
	Songer	SCDB	Songer	SCDB
Party-minority on mixed panel	−0.0498***	−0.0515***	0.0268***	0.0288***
	(0.0033)	(0.0035)	(0.0020)	(0.0021)
Party-majority of mixed panel	−0.0242***	−0.0255***	0.0065***	0.0071***
	(0.0023)	(0.0023)	(0.0008)	(0.0008)
Case direction aligned w/ party			0.9616***	0.9604***
			(0.0015)	(0.0015)
Party-minority × Aligned case			−0.0243***	−0.0254***
			(0.0026)	(0.0027)
Party-majority × Aligned case			−0.0023	−0.0025*
			(0.0012)	(0.0012)
Baseline (omitted cell)	0.5663	0.5764	0.0192	0.0197
Observations (judge–cases)	1,220,661	1,103,721	1,220,661	1,103,721
of which on unanimous panels	393,492	351,570	393,492	351,570

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Standard errors in parentheses, two-way clustered by circuit×year and judge. Column block (1)'s omitted cell is a judge on a party-unanimous panel; column block (2)'s omitted cell is a judge on a party-unanimous panel hearing a misaligned case.

Table A.22: Comparison of panel effects estimates using Songer super-mechanical direction (reproduced from Table 8) and SCDB mechanical direction. Each column block mirrors one column of the main-text table.

A.7 Panel Effects with SCDB Mechanical Direction

The main-text panel-effects analysis in Section 6.1 uses Songer super-mechanical direction, the best-performing direction variant on Songer-track validation (see Section 5.1 and Appendix A.5.1). For symmetry, we repeat that analysis using SCDB mechanical direction—the best-performing SCDB-track variant—to produce parallel versions of Table 8 and Figure 6 from Section 6.1 (Table A.22 and Figure A.2, respectively) in this appendix. Recall that the SCDB track does not have a mixed direction code (only liberal, conservative, and unspecifiable) and hard-codes two broad issue areas (Interstate Relations, Private Action) as all-unspecifiable, meaning that they are automatically dropped from this analysis. The SCDB-track sample contains $1,103,721$ judge–case observations, about 9.6% fewer than the Songer-track sample.

The Songer and SCDB columns of Table A.22 agree closely: every role coefficient is within about 10% of its Songer counterpart, all signs and significance levels match, and the role × alignment interactions reproduce the same direction and magnitude. Figure A.2 likewise replicates most of the dynamics from Figure 6. While both panel differences and the overall share of aligned votes rise starting around the 1960s, as in the Songer-track figure, the decrease in ideological voting pre-1960 is sharper, especially for the two mixed-panel roles. Those have almost-identical aligned vote rates about 10 percentage points above the aligned vote rate of unanimous panels in the 1890s, before declining and crossing under the unanimous panel rate in the 1930s to produce the modern order of unanimous, then majority on

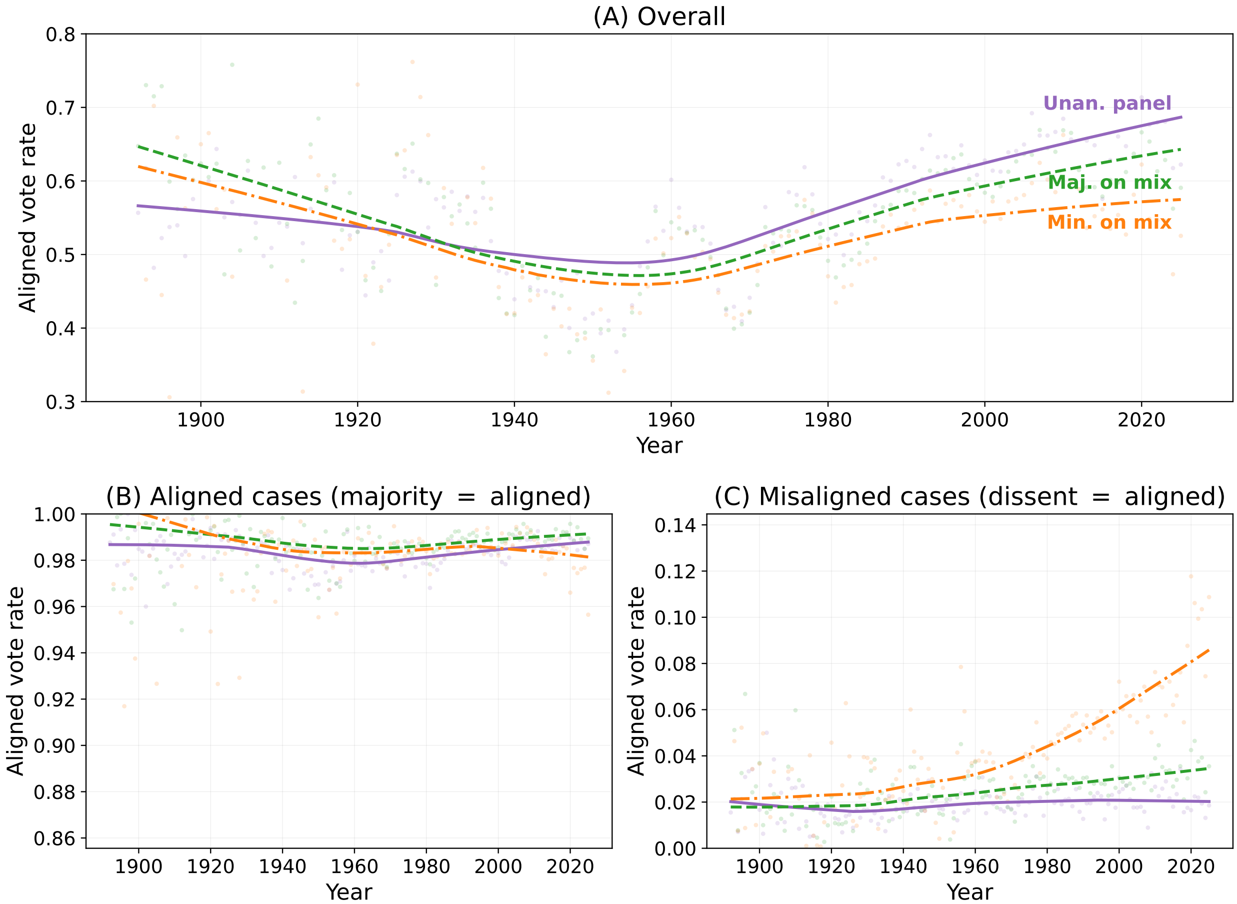

Figure A.2: SCDB-mechanical mirror of Figure 6. Same axes, panels, and series, but using SCDB mechanical direction rather than Songer super-mechanical direction.

mixed, then minority on mixed. Behavior on the subsets of aligned and misaligned cases is essentially identical to the Songer track, even including the slight decrease in aligned-majority votes by panel-minority judges. Broadly, we read this cross-track agreement as evidence that the panel-effects findings are not artifacts of the Songer codebook in particular but reflect a substantive pattern in appellate voting.

A.8 Adjusting for LLM Measurement Error

The applications in Section 6 use LLM-coded variables as regression inputs (ideological direction in both sections, as well as disposition and litigant typing in §6.2). As we have shown in Section 5, LLM coding is highly accurate but not perfect, and the residual classification error can in principle affect both point estimates and confidence intervals. In this appendix we apply three standard corrections for LLM-augmented labels—design-based supervised learning (DSL) (Egami et al. 2023), tuned PPI++ (Angelopoulos, Duchi and Zrnic 2023), and Predict-Then-Debias bootstrap (PTDB) (Kluger et al. 2025)—using the Songer hand-labeled cases to calibrate the effect of LLM error on our estimates.

These corrections are well-suited to our setting; they assume a (possibly stratified) random sample of labeled data, which approximately matches the Songer dataset's fixed-N-per-circuit-year sampling procedure. All three corrections accommodate this stratification explicitly: DSL and PPI++ inverse-probability-weight the labeled fits by the per-circuit-year coverage rate, and PTDB resamples within circuit × year strata under the same weights.[50] However, there are some caveats. Because these corrections rely on a labeled subsample, the LLM-versus-GT bias estimates they compute are based only on the Songer sample period of 1925–2002, even though our pooled estimates cover 1892–2025 (the full timespan of the modern Courts of Appeals system). The corrected coefficients we report still apply to the full pooled sample, but we are implicitly extrapolating the bias correction beyond the years covered by the Songer sample. Additionally, these corrections are not feasible for any time-series (i.e., year-by-year) estimates, since there are too few Songer-labeled cases per circuit × year to estimate corrected coefficients for our time-series results.[51] Despite these caveats, we believe that since the results of this section broadly confirm the qualitative conclusions in the main text, they also provide solid ground to believe that even the portions of our analysis that cannot be explicitly corrected for LLM bias are unlikely to be significantly distorted.

Before presenting the bias-corrected results, we first provide a brief description of each correction method. Readers interested in more complete details or implementation code should refer to the cited works and software packages.

- **Design-based supervised learning (DSL)**, from Egami et al. (2023). DSL uses the labeled subset to predict the LLM-vs-GT discrepancy as a flexible function of row covariates, then applies that prediction to every row to construct a doubly-robust moment equation:

$$m_i^{DR} = m_i^{\text{pred}} + (m_i^{\text{orig}} - m_i^{\text{pred}}) \cdot \frac{\mathbf{1}_i^L}{p_i},$$

  where m_i^{pred} uses the SL-predicted GT label, m_i^{orig} uses the actual Songer GT (zero on

[50] Each correction also admits simpler or alternative variants—circuit-year-only and judge-only DSL clusterings, unweighted PPI++ or untuned classical PPI (Angelopoulos et al. 2023), and a circuit × year cluster-bootstrap PTDB variant—which give very similar results throughout and are available upon request.

[51] The shortage of hand-labeled data at this level of granularity was one of the motivations for this paper.

unlabeled rows), and p_i is the per-row sampling probability. To accommodate Songer's stratified design (approximately fixed N per circuit-year), we set $p_i = n^L_{c,t}/n_{c,t}$ within each circuit-year cell (c,t) rather than imposing equal sampling probabilities. The variance estimator is the empirical second moment of the doubly-robust scores. When the predictor collapses to a constant shift across covariates, DSL reduces to a constant rectifier, which is the approach used by PPI++ (described below). We implement DSL via the `dsl` R package (available at `https://naokiegami.com/dsl/index.html`) with a minor hand-coded adjustment to allow two-way standard error clustering for the specifications corresponding to Section 6.1 of the main text. Circuit-year-only and judge-only clusterings implemented using the vanilla package give very similar results and are available upon request.

- **Tuned PPI++**, from Angelopoulos, Duchi and Zrnic (2023). PPI++ refines the original prediction-powered inference (PPI) approach of Angelopoulos et al. (2023), which re-estimates each specification on the Songer-labeled subsample L—once with LLM labels ($\hat\beta^L_{\text{LLM}}$) and once with Songer labels ($\hat\beta^L_{\text{GT}}$)—and corrects the main text's full-sample $\hat\beta$ with a (constant) *rectifier*, $\hat\beta_{\text{PPI}} = \hat\beta + (\hat\beta^L_{\text{GT}} - \hat\beta^L_{\text{LLM}})$. PPI++ adds per-coefficient power tuning:

$$\hat\beta_{\text{PP++}}(\lambda) = \hat\beta^L_{\text{GT}} + \lambda \cdot (\hat\beta - \hat\beta^L_{\text{LLM}}), \qquad \lambda^*_k = \frac{\widehat{\text{Cov}}(\psi^L_{y,k},\, \psi^L_{f,k})}{\widehat{\text{Var}}(\psi^L_{f,k})},$$

where ψ^L_y, ψ^L_f are the (potentially weighted) least squares influence functions of the GT and LLM regressions on L and $\widehat{\text{Cov}}$, $\widehat{\text{Var}}$ apply the relevant specification's cluster structure (Cameron–Gelbach–Miller two-way clustering for §6.1, one-way for §6.2). $\hat\beta_{\text{PP++}}(\lambda)$ is consistent for the GT coefficient at every $\lambda \in [0,1]$ and interpolates between classical PPI ($\lambda = 1$) and the labeled-only estimate $\hat\beta^L_{\text{GT}}$ ($\lambda = 0$); λ^* is chosen to minimize its asymptotic variance. The corresponding analytic clustered standard error comes from the PPI++ sandwich formula

$$\widehat{\text{Var}}\big(\hat\beta_{\text{PP++}}(\lambda^*)\big)_k = \widehat{\text{Var}}\big(\psi^L_{y,k} - \lambda^*_k \psi^L_{f,k}\big) + (\lambda^*_k)^2 \cdot \widehat{\text{SE}}(\hat\beta)^2_k,$$

combining the labeled-rectifier variance with the main text's full-sample clustered SE under the standard PPI independence assumption (small n_L/N). As with DSL, we fit the labeled regressions $\hat\beta^L_{\text{LLM}}, \hat\beta^L_{\text{GT}}$ by inverse-probability-weighted least squares (weights $1/p_i$, with p_i the per-circuit-year label-coverage rate), so the rectifier targets the population quantity under Songer's stratified design. Because no existing PPI++ implementation supports fixed effects with cluster-robust standard errors, we manually implement the estimator as an extension of our existing `pyfixest` implementation following the approach in the `ppi-python` package (documented at `github.com/aangelopoulos/ppi_py`) where possible. We report the power-tuned λ^* estimator; the untuned classical-PPI estimate ($\lambda = 1$) is very similar and available upon request.
- **Predict-Then-Debias bootstrap (PTDB)**, from Kluger et al. (2025). PTDB combines the rectifier from classical PPI with a bootstrap-estimated debiasing matrix to

optimize standard error tightness. The estimator is

$$\hat{\beta}_{\text{PTDB}} = \hat{\beta}^L_{\text{GT}} + \widehat{W}(\hat{\beta} - \hat{\beta}^L_{\text{LLM}}),$$

where the diagonal entries of the tuning matrix $\widehat{W}$ play the same role as λ^* in PPI++, but are estimated jointly with the inferential uncertainty by resampling the labeled and unlabeled samples. We resample 1,000 times within circuit $\times$ year using inverse-probability weights (matching both other corrections and the design under which Songer was originally drawn). Confidence intervals are bootstrap percentile intervals on $\hat{\beta}_{\text{PTDB}}$ rather than analytic sandwich SEs, which are robust to mis-specification of the cluster structure. When $\widehat{W}$ collapses to zero (e.g. when there are too few labeled clusters to estimate the predictor-error covariance), PTDB returns the labeled-only estimate $\hat{\beta}^L_{\text{GT}}$ with bootstrap CIs around it; we flag the rare cells where this happens in the relevant table notes. We implement PTDB via the `PTDBoot` R package. Bootstrapping clusters defined by circuit $\times$ year without inverse-probability weights gives very similar results, which are available upon request.

A.8.1 Panel Effects (Section 6.1)

Table A.23 reports DSL, PPI++, and PTDB corrected coefficients alongside the main text's reported numbers for the role-only and role-by-alignment specifications of our panel effects analysis in Table 8; the first two rows correspond to coefficients from Column (1) of that table, while the remaining five correspond to coefficients from Column (2). Overall, the qualitative result in the main text holds—panel effects decrease ideological voting, especially by party-minority judges, but party-minority judges tend to vote their ideology by dissenting from ideologically misaligned majorities. Across all three corrections, panel effects decrease ideological voting on mixed panels—by about 2–5 percentage points for party-majority judges and about 4–7 percentage points for party-minority judges. The magnitudes, and the gap between roles, vary slightly across corrections, but ordering and statistical significance are always preserved (with the exception of the PPI- and PTDB-corrected party-majority $\times$ aligned interactions, which gain significance). Under all three correction methods, the increase in ideologically aligned dissents from misaligned majorities is consistently about 3 times as large for party-minority judges as for party-majority judges, compared to 4 times as large in the main text. Overall, the corrected results show mostly minor differences from the main text, with DSL showing larger changes but still pointing in the same direction as the uncorrected results.

A.8.2 Pro-Weak Bias (Section 6.2)

In Section 6.2, we use LLM-generated case disposition, litigant typing, and ideological direction across various specifications in tables 9 and 10. Each additional LLM channel introduces additional measurement error, leading to more substantial effects on statistical significance even when the point estimate remains similar. However, these results should be interpreted

	Main text $\hat{\beta}$ (SE)	DSL $\hat{\beta}$ (SE)	PPI++ $\hat{\beta}$ (SE)	PTDB $\hat{\beta}$ (SE)
Role-only specification (col. 1 of Table 8)				
Party-minority on mixed	-0.0498^{***}	-0.0721^{***}	-0.0446^{***}	-0.0376^{***}
	(0.0033)	(0.0131)	(0.0075)	(0.0053)
Party-majority of mixed	-0.0242^{***}	-0.0520^{***}	-0.0178^{**}	-0.0152^{***}
	(0.0023)	(0.0108)	(0.0067)	(0.0046)
Role × alignment specification (col. 2 of Table 8)				
Party-minority on mixed	0.0268^{***}	0.0264^{***}	0.0312^{***}	0.0267^{***}
	(0.0020)	(0.0045)	(0.0026)	(0.0022)
Party-majority of mixed	0.0065^{***}	0.0093^{**}	0.0113^{***}	0.0102^{***}
	(0.0008)	(0.0034)	(0.0020)	(0.0018)
Aligned-case main effect	0.9616^{***}	0.9405^{***}	0.9614^{***}	0.9603^{***}
	(0.0015)	(0.0037)	(0.0017)	(0.0012)
Party-min × Aligned	-0.0243^{***}	-0.0162^{**}	-0.0238^{***}	-0.0178^{***}
	(0.0026)	(0.0062)	(0.0028)	(0.0016)
Party-maj × Aligned	-0.0023	-0.0046	-0.0040^{*}	-0.0029^{*}
	(0.0012)	(0.0042)	(0.0016)	(0.0015)

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
1,220,661 judge-rows in the full sample and 54,090 in the labeled subsample. DSL and PPI++ standard errors are two-way clustered on circuit × year and judge. The PTDB column resamples 1,000 times within circuit × year strata with inverse-probability weights. Alternative DSL clustering, classical PPI, and the cluster-bootstrap PTDB variant are available upon request.

Table A.23: Robustness of panel effects (Table 8, Section 6.1) to LLM measurement error. Main text $\hat{\beta}$ is the full-sample LLM regression; DSL, PPI++, and PTDB replace it with a corrected estimate. Most DSL-corrected estimates are about 1.4–2.1 times larger than those in the main text (4 of 7 coefficients); the exceptions are confined to the role × alignment specification, where the party-minority and alignment main effects are essentially unchanged and the party-minority interaction is about a third smaller. The PPI++ and PTDB corrections are generally slightly smaller for the role-only specification and slightly larger for the role × alignment specification. This pattern suggests that LLM measurement error varies with covariates, which only the DSL approach incorporates in its correction.

with caution, especially for SCDB-driven coefficients, since the matched sample of labeled cases for these estimates is substantially smaller than in Section 6.1. As a result, the corrected estimators are mechanically underpowered.

A.8.2.1 Original specifications

Tables A.24 and A.25 present DSL, PPI++, and PTDB corrections for the four columns of Table 9 and Table 10 respectively. On the pooled sample of all cases, results from both Songer-track specifications (mixed with IDB and full-LLM) are almost entirely preserved. Across all three correction methods, each additional Democrat consistently increases pro-weak bias, and all coefficients remain statistically significant at the 5% level. Magnitudes are only modestly attenuated: PPI++ coefficients retain roughly 80% or more of their main text magnitudes, while PTDB is somewhat more conservative. Results from both SCDB-track specifications generally maintain the increasing pattern of point estimates, but only four of those corrected point estimates (DSL and PTDB) are statistically significant at the 5% level. Part of this change is mechanically driven by increased standard errors from the smaller labeled sample (about one-third of the corresponding Songer-track specifications). However, it also suggests that the SCDB codebook may be less well-suited for this application.

The qualitative pattern of corrections changes when we examine the corrected Table 10, which separates cases by whether pro-weak outcomes are treated as liberal or conservative in the appropriate codebook (dropping cases where the broad area contains both). Here, the LLM bias corrections usually increase the magnitude of the point estimates (29 of 40 corrected cells). This increase is near-universal for DSL (13 of 14) but only occurs for about half of the PPI++ and PTDB (16 of 26) coefficients. Intuitively, LLM error in the clash indicator may be biasing coefficients towards zero. As with the pooled corrections, Songer-track coefficients see smaller changes in magnitude and statistical significance. Magnitudes increase (with positive values for main effect coefficients, negative for interacted coefficients) as more Democrats are added to the panel regardless of correction method, and main effect coefficients remain statistically significant at the 1% level across all correction methods. Corrected main effect coefficients are generally comparable to their main text counterparts or somewhat larger, by up to about 80%. While interacted coefficients also tend to be larger in magnitude (i.e., more negative), they rarely retain statistical significance: only 4 of 18 corrected Songer-track interaction estimates are significant at the 5% level, and all of those are DSL-corrected. Significant coefficients are scattered across panel types; their increase in magnitudes is evidence that this change is likely driven by reduced sample size, unlike in the pooled case where LLM bias attenuated effect sizes as well as statistical significance.

This power issue is even more problematic for the two SCDB tracks, where the labeled sample includes at most 783 cases (due to SCDB having fewer broad issue areas that separate between liberal and conservative pro-weak outcomes). In the IDB-SCDB specification almost no coefficients are statistically significant; in the full-LLM SCDB specification, several PPI++ and PTDB coefficients—the larger-panel effects and their clash interactions—do reach significance, though the very small labeled sample warrants caution. For the clash-interacted full-LLM SCDB specification, the small sample leads to a degenerate DSL fit. Estimates are implausibly large (up to 9 times larger than any other correction method) and standard errors are even more inflated. Given the extremely small sample size, SCDB-track PPI++ and PTDB results should thus be interpreted with caution, while DSL results should largely be disregarded. The Songer-track corrections, which have enough data to fit all three correction methods, generally support the main text's conclusion that pro-weak bias is attenuated on clash cases, especially on liberal-majority panels.

A.8.2.2 Liberal-majority robustness specifications

The difficulty of achieving appropriate statistical power for SCDB-driven estimates, especially on the clash-interaction specifications, makes it difficult to assess the qualitative consequences of LLM bias for the conclusions supported by the uncorrected estimates: pro-weak bias is attenuated or even reversed on clash cases. To recover statistical power, we collapse the panel-composition specifications into a single liberal-majority indicator with conservative-majority panels (RRR and DRR) as the omitted baseline. This roughly doubles the treatment-cell sample size at the cost of some granularity in our estimates. We present those results, which include both new estimates in the style of the main text and corrected estimates as in the rest of this appendix, in tables A.26 and A.27.

	Main text $\hat{\beta}$ (SE)	DSL $\hat{\beta}$ (SE)	PPI++ $\hat{\beta}$ (SE)	PTDB $\hat{\beta}$ (SE)
IDB-out + LLM-types Songer — N_{full} = 217,868, N_{lab} = 8,850				
DRR	0.0267***	0.0330***	0.0244***	0.0270***
	(0.0032)	(0.0063)	(0.0072)	(0.0073)
DDR	0.0717***	0.0851***	0.0682***	0.0585***
	(0.0042)	(0.0077)	(0.0083)	(0.0079)
DDD	0.1196***	0.1362***	0.1207***	0.0993***
	(0.0065)	(0.0122)	(0.0121)	(0.0114)
IDB-out + LLM-types SCDB — N_{full} = 80,281, N_{lab} = 3,204				
DRR	0.0135**	0.0048	−0.0036	−0.0003
	(0.0052)	(0.0163)	(0.0176)	(0.0165)
DDR	0.0388***	0.0346*	0.0157	0.0122
	(0.0064)	(0.0164)	(0.0188)	(0.0184)
DDD	0.0783***	0.0619*	0.0441	0.0312
	(0.0092)	(0.0243)	(0.0264)	(0.0237)
Full-LLM Songer — N_{full} = 383,451, N_{lab} = 17,042				
DRR	0.0208***	0.0198**	0.0173*	0.0163*
	(0.0027)	(0.0072)	(0.0077)	(0.0080)
DDR	0.0564***	0.0481***	0.0453***	0.0361***
	(0.0037)	(0.0083)	(0.0084)	(0.0082)
DDD	0.0913***	0.0841***	0.0801***	0.0633***
	(0.0056)	(0.0110)	(0.0112)	(0.0105)
Full-LLM SCDB — N_{full} = 159,408, N_{lab} = 6,349				
DRR	0.0079	−0.0026	0.0052	0.0045
	(0.0041)	(0.0155)	(0.0175)	(0.0160)
DDR	0.0282***	0.0145	0.0077	0.0018
	(0.0051)	(0.0165)	(0.0179)	(0.0170)
DDD	0.0509***	0.0480*	0.0375	0.0262
	(0.0070)	(0.0209)	(0.0237)	(0.0217)

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Standard errors for DSL and PPI++ are clustered at the circuit × year level. DSL, PPI++, and PTDB all use inverse-probability weights within circuit × year (PTDB samples 1,000 times). Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.24: Robustness of pro-weak bias over all cases (Table 9, Section 6.2) to LLM measurement error. Songer-track results maintain the ranking of point estimates and statistical significance, with some attenuation of magnitudes depending on the correction method. Almost all SCDB-track coefficients are not statistically significant, though the point estimates are usually ordered appropriately.

Despite the change in specification, the patterns observed in tables A.24 and A.25, and the statistical power issues for the clash specification, largely persist: shifts in magnitudes and statistical significance are smaller on Songer-track specifications, SCDB-track specifications mostly lose statistical significance, and corrected magnitudes fall for the pooled specifications and rise for the clash-interacted specifications. There are a few differences worth noting, though. In Table A.26, PTDB is more conservative relative to the other correction methods than it was in Table A.24, reducing point-estimated magnitudes for Songer-track estimates by substantially more than DSL or PPI++ and shrinking SCDB-track estimates to near zero. In Table A.27, the PPI++ and PTDB-corrected interacted coefficients are more mixed than in Table A.25: most still grow in magnitude, but both corrections shrink the full-LLM Songer interaction to roughly a third of its main text value. Clash interactions are statistically significant at the 5% level for the mixed IDB-Songer specification across all three correction methods, but not statistically significant at that level for most other specifications and correction methods. Overall, DSL is the only correction to preserve significance on both Songer-track specifications, while PPI++ retains significance only for IDB-Songer and PTDB does so for IDB-Songer and full-LLM SCDB.

The key claim advanced by Table 10 and Figure 7 is that for cases where a pro-weak outcome is coded as conservative according to the Songer codebook, pro-weak bias on liberal-majority panels is attenuated. The table suggests that pro-weak bias may be eliminated entirely on these cases, while the figure suggests that at least for modern cases, it may even be reversed (with conservative-majority panels exhibiting pro-weak bias rather than liberal-majority ones). We can formally test these hypotheses via a one-sided test that the interaction coefficient is negative—meaning that pro-weak bias is attenuated—or that the sum of interacted and main effect coefficients is negative—meaning that pro-weak bias is eliminated or reversed. Results from the former test are in Table A.28; results from the latter are in Table A.29. Both tables reproduce the estimated coefficient of interest, labeled $\hat{\beta}$, from the appropriate specification, and provide the one-sided p-value for the null hypothesis that the coefficient is zero against the hypothesis that it is negative.

The results in the main text actually support the stronger hypothesis—reversal of pro-weak bias on clash cases—with p-values all rejecting the null at the 1% level. The results in A.28 provide some support for attenuation: the corrected IDB-Songer track coefficients show attenuation for all three methods, as does the DSL-corrected full-LLM Songer coefficient and the PPI- and PTDB-corrected full-LLM SCDB coefficients. Additionally, IDB-SCDB falls just short of conventional significance under both PPI++ ($p = 0.104$) and PTDB ($p = 0.138$). However, the corrected coefficients in Table A.29 all fail to reject the null (of no pro-weak bias at all on clash cases) at the 10% level, except for PPI++ and PTDB correcting full-LLM SCDB ($p = 0.059$ and $p = 0.095$ respectively). Of course, these conclusions only apply to the pooled estimates; as Figure 7 in the main text showed, pro-weak bias and its reversal are mostly present from the 1970s onward. However, the failure of the IDB-mixed specifications to reject the null suggests that those time-restricted results are likely also affected by LLM bias, leaving ample room for ensemble measures of ideology which use ground truth data such as the IDB to correct for LLM errors.

	Main text $\hat{\beta}$ (SE)	DSL $\hat{\beta}$ (SE)	PPI++ $\hat{\beta}$ (SE)	PTDB $\hat{\beta}$ (SE)
IDB-out + LLM-types Songer — $N_{\text{full}} = 191{,}923$, $N_{\text{lab}} = 4{,}282$				
DRR	0.0332***	0.0424**	0.0349**	0.0337**
	(0.0032)	(0.0150)	(0.0119)	(0.0115)
DDR	0.0893***	0.1179***	0.1109***	0.0992***
	(0.0046)	(0.0154)	(0.0131)	(0.0131)
DDD	0.1495***	0.1765***	0.1677***	0.1444***
	(0.0072)	(0.0290)	(0.0215)	(0.0193)
DRR × Clash	−0.0465***	−0.0359	−0.0192	−0.0169
	(0.0088)	(0.0663)	(0.0518)	(0.0514)
DDR × Clash	−0.1230***	−0.1539*	−0.1032	−0.0969
	(0.0100)	(0.0630)	(0.0542)	(0.0531)
DDD × Clash	−0.1895***	−0.2159	−0.1807	−0.1649*
	(0.0129)	(0.1162)	(0.0941)	(0.0801)
IDB-out + LLM-types SCDB — $N_{\text{full}} = 68{,}860$, $N_{\text{lab}} = 742$				
DRR	0.0259***	−0.0183	0.0322	0.0294
	(0.0062)	(0.1460)	(0.0720)	(0.0843)
DDR	0.0741***	0.0034	0.0722	0.0632
	(0.0076)	(0.1301)	(0.0853)	(0.0867)
DDD	0.1295***	0.1578	0.3299*	0.3066
	(0.0107)	(0.1762)	(0.1456)	(0.1770)
DRR × Clash	−0.0543***	0.0506	−0.1265	−0.1198
	(0.0127)	(0.4181)	(0.1417)	(0.1669)
DDR × Clash	−0.1445***	0.0517	−0.1604	−0.1513
	(0.0137)	(0.3639)	(0.1534)	(0.1687)
DDD × Clash	−0.2136***	−0.2181	−0.4667	−0.4545
	(0.0182)	(0.4834)	(0.2386)	(0.2942)
Full-LLM Songer — $N_{\text{full}} = 345{,}313$, $N_{\text{lab}} = 4{,}586$				
DRR	0.0278***	0.0504***	0.0347**	0.0377***
	(0.0028)	(0.0143)	(0.0113)	(0.0112)
DDR	0.0704***	0.0997***	0.0794***	0.0791***
	(0.0040)	(0.0157)	(0.0130)	(0.0129)
DDD	0.1128***	0.1526***	0.1276***	0.1291***
	(0.0063)	(0.0223)	(0.0191)	(0.0175)
DRR × Clash	−0.0367***	−0.1024*	−0.0390	−0.0419
	(0.0063)	(0.0480)	(0.0494)	(0.0492)
DDR × Clash	−0.0845***	−0.1369**	−0.0281	−0.0334
	(0.0072)	(0.0507)	(0.0534)	(0.0541)
DDD × Clash	−0.1135***	−0.2047**	−0.1206	−0.1334
	(0.0089)	(0.0752)	(0.0883)	(0.0818)
Full-LLM SCDB — $N_{\text{full}} = 140{,}921$, $N_{\text{lab}} = 783$				
DRR	0.0176***	−0.6114	0.0793	0.0839
	(0.0048)	(1.1065)	(0.0647)	(0.0525)
DDR	0.0525***	−0.6477	0.1455*	0.1474*
	(0.0060)	(1.3100)	(0.0674)	(0.0665)
DDD	0.0831***	−0.9918	0.2061	0.1913
	(0.0083)	(2.1705)	(0.1268)	(0.1017)
DRR × Clash	−0.0343***	1.8485	−0.1602	−0.1573
	(0.0091)	(3.3379)	(0.1142)	(0.1020)
DDR × Clash	−0.0872***	1.9417	−0.2565*	−0.2619*
	(0.0100)	(3.9094)	(0.1265)	(0.1097)
DDD × Clash	−0.1098***	3.1579	−0.3686	−0.3594*
	(0.0118)	(6.5070)	(0.1908)	(0.1676)

$^{*}p < 0.05$ $\quad ^{**}p < 0.01$ $\quad ^{***}p < 0.001$

Standard errors for DSL and PPI++ are clustered at the circuit × year level. DSL, PPI++, and PTDB all use inverse-probability weights within circuit × year (PTDB samples 1,000 times). Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.25: Robustness of pro-weak bias in clash and no-clash cases (Table 10, Section 6.2) to LLM measurement error. The small labeled sample for the *Full-LLM SCDB* column makes DSL unable to accurately learn LLM bias as a function of covariates, while the constant rectifier of the other two methods remains viable.

	Main text $\hat{\beta}$ (SE)	DSL $\hat{\beta}$ (SE)	PPI++ $\hat{\beta}$ (SE)	PTDB $\hat{\beta}$ (SE)
IDB-out + LLM-types Songer — $N_{\text{full}} = 217{,}868$, $N_{\text{lab}} = 8{,}850$				
Liberal-majority panel	0.0597***	0.0694***	0.0581***	0.0449***
	(0.0032)	(0.0062)	(0.0059)	(0.0057)
IDB-out + LLM-types SCDB — $N_{\text{full}} = 80{,}281$, $N_{\text{lab}} = 3{,}204$				
Liberal-majority panel	0.0354***	0.0343**	0.0243	0.0162
	(0.0046)	(0.0119)	(0.0130)	(0.0125)
Full-LLM Songer — $N_{\text{full}} = 383{,}451$, $N_{\text{lab}} = 17{,}042$				
Liberal-majority panel	0.0463***	0.0387***	0.0383***	0.0280***
	(0.0029)	(0.0058)	(0.0056)	(0.0053)
Full-LLM SCDB — $N_{\text{full}} = 159{,}408$, $N_{\text{lab}} = 6{,}349$				
Liberal-majority panel	0.0258***	0.0223*	0.0104	0.0027
	(0.0036)	(0.0111)	(0.0119)	(0.0103)

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Controls and fixed effects match Table 9 in the main text. Standard errors for DSL and PPI++ are clustered at the circuit × year level. DSL, PPI++, and PTDB all use inverse-probability weights within circuit × year (PTDB samples 1,000 times). Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.26: Pro-weak bias by liberal majorities (DDD or DDR) compared to a baseline of conservative majorities (DRR or RRR), in the style of tables 9 and A.24.

	Main text $\hat{\beta}$ (SE)	DSL $\hat{\beta}$ (SE)	PPI++ $\hat{\beta}$ (SE)	PTDB $\hat{\beta}$ (SE)
IDB-out + LLM-types Songer — $N_{\text{full}} = 191{,}923$, $N_{\text{lab}} = 4{,}282$				
Liberal-majority panel	0.0756***	0.1020***	0.0915***	0.0819***
	(0.0037)	(0.0144)	(0.0097)	(0.0100)
Lib-majority × Clash	−0.1066***	−0.1363*	−0.1091**	−0.1008*
	(0.0069)	(0.0576)	(0.0387)	(0.0403)
IDB-out + LLM-types SCDB — $N_{\text{full}} = 68{,}860$, $N_{\text{lab}} = 742$				
Liberal-majority panel	0.0661***	0.0356	0.0901	0.0861
	(0.0055)	(0.0761)	(0.0591)	(0.0647)
Lib-majority × Clash	−0.1232***	−0.0160	−0.1417	−0.1367
	(0.0092)	(0.2154)	(0.1124)	(0.1253)
Full-LLM Songer — $N_{\text{full}} = 345{,}313$, $N_{\text{lab}} = 4{,}586$				
Liberal-majority panel	0.0577***	0.0750***	0.0607***	0.0600***
	(0.0032)	(0.0123)	(0.0097)	(0.0095)
Lib-majority × Clash	−0.0691***	−0.0869*	−0.0238	−0.0264
	(0.0051)	(0.0383)	(0.0389)	(0.0371)
Full-LLM SCDB — $N_{\text{full}} = 140{,}921$, $N_{\text{lab}} = 783$				
Liberal-majority panel	0.0458***	0.3553	0.0730	0.0763
	(0.0044)	(0.3660)	(0.0477)	(0.0450)
Lib-majority × Clash	−0.0708***	−1.0706	−0.1648	−0.1795*
	(0.0068)	(1.1569)	(0.0897)	(0.0788)

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Standard errors for DSL and PPI++ are clustered at the circuit × year level. DSL, PPI++, and PTDB all use inverse-probability weights within circuit × year (PTDB samples 1,000 times). Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.27: Pro-weak bias by liberal majorities (DDD or DDR vs. DRR or RRR) on clash and no-clash cases in the style of tables 10 and A.25. Reported coefficients are the liberal-majority main effect (Clash = 0) and the liberal-majority × clash interaction; the net effect on clash cases is their sum.

	Main text specification		DSL correction		PPI++ correction		PTDB	
	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p
IDB-out + LLM-types Songer	−0.1066	< 0.001	−0.1363	0.009	−0.1091	0.002	−0.1008	0.006
IDB-out + LLM-types SCDB	−0.1232	< 0.001	−0.0160	0.470	−0.1417	0.104	−0.1367	0.138
Full-LLM Songer	−0.0691	< 0.001	−0.0869	0.012	−0.0238	0.270	−0.0264	0.238
Full-LLM SCDB	−0.0708	< 0.001	−1.0706	0.177	−0.1648	0.033	−0.1795	0.011

$^*p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Reports the one-sided p-value against $H_0 : \beta_{\text{libmaj}\times\text{clash}} = 0$ in favor of $H_1 : \beta_{\text{libmaj}\times\text{clash}} < 0$. Computed from the coefficient and standard error of the libmaj × clash term in Table A.27; the PPI++ column uses analytic clustered SEs at $\lambda = \lambda^*$, while the PTDB column uses bootstrap-implied SEs from the stratified-IPW variant. Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.28: Formal one-sided test that pro-weak bias by liberal majorities is attenuated on clash cases ($\beta_{\text{noclash}} > \beta_{\text{clash}}$) under each of the four column specifications and under DSL, PPI++, and PTDB corrections.

	Main text specification		DSL correction		PPI++ correction		PTDB	
	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p	$\hat{\beta}$	p
IDB-out + LLM-types Songer	−0.0310	< 0.001	−0.0344	0.226	−0.0176	0.293	0.0070	0.579
IDB-out + LLM-types SCDB	−0.0570	< 0.001	0.0196	0.555	−0.0516	0.240	−0.0242	0.388
Full-LLM Songer	−0.0114	0.007	−0.0119	0.343	0.0369	0.869	0.0523	0.934
Full-LLM SCDB	−0.0251	< 0.001	−0.7154	0.184	−0.0918	0.059	−0.0692	0.095

$^*p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$
Reports the one-sided p-value against $H_0 : \beta_{\text{libmaj}} + \beta_{\text{libmaj}\times\text{clash}} = 0$ in favor of $H_1 : \beta_{\text{libmaj}} + \beta_{\text{libmaj}\times\text{clash}} < 0$ — whether the *total* pro-weak effect of a liberal-majority panel on clash cases (main effect plus interaction) is negative. Computed from the joint variance of the two coefficients: analytic vcov for the main-text specification, the PPI++ sandwich extended to the linear combination $\hat{\beta}_{\text{libmaj}} + \hat{\beta}_{\text{libmaj}\times\text{clash}}$, and the DSL joint vcov from `dsl`'s fit; the PTDB column uses the bootstrap-implied SE of the sum from the stratified-IPW variant. Classical PPI and the cluster-bootstrap PTDB variant are available upon request.

Table A.29: Formal one-sided test that pro-weak bias by liberal majorities *vanishes or reverses* on clash cases ($\beta_{\text{libmaj}} + \beta_{\text{libmaj}\times\text{clash}} < 0$) — stronger than the attenuation test in Table A.28. Note that because PTDB re-optimizes tuning weights for the sum, the $\hat{\beta}$ values in the PTDB column block do not precisely equal the sum of the individual coefficients from Table A.27. In contrast, $\hat{\beta}$ for the other column blocks *is* mechanically equal to the sum of the individual coefficients.