Judicial Disparities in Seventh Circuit Criminal Appeals

Richard Zhu

Date of Submission: March 13, 2026

Instructor: Dr. Nicolette Bruner

Teaching Assistants: Jacob Palmer and Ryan Fajardo

Abstract

Debates over sentencing disparity often presume that appellate discretion allows judicial ideology to influence whether criminal defendants obtain relief. This thesis tests that premise through a newly constructed, judge-identifying dataset of 1,591 Seventh Circuit sentencing decisions issued between 2020 and 2025. Apparent inter-judge differences narrow substantially once posture and case composition are taken into account. A three-track analysis further shows that the remaining aggregate disparities arise primarily from the pooling of precedential and nonprecedential decisions rather than from stable inter-judge differences within either subset. Published doctrine alone can therefore misstate where disparity resides in appellate sentencing review. The thesis thus offers a track-specific framework for studying appellate sentencing review beyond the surface of precedential doctrines in federal circuit courts.

For nearly four decades, federal sentencing has been shaped by a persistent institutional tension between uniformity and discretion. In the wake of the Supreme Court's "sentencing quartet," *Booker*, *Rita*, *Gall*, and *Kimbrough*, district judges regained substantial authority to depart from or vary around the Guidelines, and a large empirical literature emerged to ask whether that renewed discretion widened inter-judge disparity along ideological, partisan, or demographic lines.[1] The basic concern is that if sentencing outcomes depend materially on which district judge draws the case, then the promise of equal treatment under law is weakened at the point where punishment is imposed. But that concern also implies a second question that has received far less sustained attention: whether appellate review mitigates that disparity or instead reproduces it at a second layer staffed by judges with their own ideological and political commitments.

That opacity is the core problem this thesis addresses. At the circuit-court level, the relevant institutional background includes not only the sentencing quartet but also the law of publication and citation: the longstanding debate over unpublished dispositions, the conflict between *Anastasoff* and *Hart* over their status, and the partial settlement embodied in Federal Rule of Appellate Procedure 32.1.[2] Yet even after that doctrinal debate, legal commentary and empirical

[1] The quartet restructured sentencing review in complementary ways. *United States v. Booker* rendered the Sentencing Guidelines advisory rather than mandatory and directed appellate courts to review sentences for reasonableness. 543 U.S. 220 (2005). *Rita v. United States* held that courts of appeals may apply a nonbinding presumption of reasonableness to within-Guidelines sentences on appellate review. 551 U.S. 338 (2007). *Gall v. United States* rejected rigid proportionality review for variances and required abuse-of-discretion review for all sentencing decisions, whether inside or outside the Guidelines range. 552 U.S. 38 (2007). *Kimbrough v. United States* confirmed that district courts may vary from the Guidelines on policy grounds, including disagreement with the crack-to-powder ratio embedded in the Guidelines. 552 U.S. 85 (2007). Together, these cases enlarged district-court sentencing discretion while assigning appellate courts the task of supervising that discretion through deferential reasonableness review.

[2] The doctrinal significance of unpublished appellate dispositions has long been contested. In *Anastasoff v. United States*, Judge Arnold argued that Article III did not permit a federal court to strip prior decisions of precedential force, reasoning that judges are "not delegated to pronounce a new law, but to maintain and expound the old." 223 F.3d 898 (8th Cir. 2000) (quoting 1 William Blackstone, Commentaries on the Laws of England 69), vacated as moot on reh'g en banc, 235 F.3d 1054 (8th Cir. 2000). One year later, in *Hart v. Massanari*, Judge Kozinski rejected that constitutional argument, writing that Article III does not require courts to "make binding law every time we issue a merits decision" and defending the appellate function as one of "managing precedent to develop a coherent body of circuit law." 266 F.3d 1155 (9th Cir. 2001). The Supreme Court has never resolved that split. See also *Plumley v. Austin*, 574 U.S. 1127 (2015) (Thomas, J., dissenting from denial of certiorari, joined by Scalia, J.). Federal Rule of Appellate Procedure 32.1 resolved only the citation question. It neither requires nor forbids unpublished dispositions, does not regulate when courts may designate dispositions as unpublished or nonprecedential, and "says nothing about what effect a court must give" such decisions; it bars only rules restricting citation of designated unpublished federal dispositions issued on or after January 1, 2007. Fed. R. App. P. 32.1(a); Fed. R. App. P. 32.1 advisory committee's note to 2006 adoption. In the Seventh Circuit, publication itself is institutionally constrained: Circuit Rule 40(e) requires circulation to all active judges before publication of a panel opinion that would overrule circuit precedent,

scholarship still gravitate toward published opinions, even though the ordinary work of appellate review is conducted largely through nonprecedential adjudication. The Seventh Circuit resolves a substantial volume of sentencing appeals, but the routine affirmance in a nonprecedential disposition rarely receives systematic measurement. This thesis addresses that gap. There is no recent circuit-specific account of sentencing review in the Seventh Circuit that separates the full merits stream from its precedential and nonprecedential components while disentangling case mix from institutional design.[3] Without that distinction, published doctrine can exaggerate, mute, or otherwise distort where disparity actually resides.

This thesis makes two contributions. First, it argues that appellate sentencing review is best understood as an institutional process, not simply as the aggregate expression of individualized judicial preferences. Second, it assembles and analyzes the full text of 1,591 Seventh Circuit decisions through a reproducible coding pipeline that permits comparison across adjudicatory tracks and across judges on a common footing. A validated large-language-model (LLM) extraction workflow codes case-level variables from each decision, and a Bayesian hierarchical model estimates inter-judge tendencies while accounting for case difficulty and uneven exposure. The result is a more reliable basis for comparison than the published-opinion sample or raw rate comparisons alone.

Substantively, the thesis finds that Seventh Circuit sentencing review is more cohesive and more institutionally constrained than published doctrine alone would suggest. Once posture and case difficulty are taken into account, inter-judge differences in measured leniency narrow substantially, and most residual variation tracks tenure and case mix more than persistent ideological division. The point is not that judges lack preferences. It is that the institutional conditions

create an inter-circuit conflict, or establish a new rule or procedure. See 7th Cir. R. 40(e). That compromise matters here, as this thesis argues that any account of appellate sentencing review that ignores the nonprecedential docket mischaracterizes the institutional environment in which most sentencing appeals are actually decided.

[3] For purposes of this study, the relevant appellate universe consists of Seventh Circuit criminal sentencing appeals resolved between January 1, 2020, and December 31, 2025. The full merits stream refers to the court's merits dispositions within that universe, as distinct from collateral, procedural, or other non-merits matters. Within that stream, the analysis distinguishes among the aggregate merits docket, the nonprecedential merits subset ("Routine"), and the published precedential subset ("Doctrine"), so that publication regime can be separated from case mix and judge identity in evaluating apparent disparity.

of appellate review, especially deferential standards of review, panel decisionmaking, and the routinized use of nonprecedential dispositions, limit how often those preferences become visible in ordinary criminal appeals.

Literature Review

Federal sentencing scholarship and appellate-behavior scholarship have developed on partly separate tracks. Sentencing studies usually follow defendants and sentence imposition in district courts, while appellate studies usually follow votes, panels, publication practices, and opinion production in the courts of appeals. Because these fields ask fundamentally different questions, a blind spot has emerged at their intersection: the appellate review of criminal sentences has rarely been analyzed with the same judge-level granularity as district-court outcomes. To fill this gap, this chapter proceeds in three parts. First, it reviews the district-court literature to establish the baseline mechanics of sentencing disparity. Second, it turns to the appellate-behavior literature, exploring how panel dynamics and institutional design shape judicial outcomes beyond a simple attitudinal model.[4] Finally, it examines the machinery of appellate triage and selective publication, demonstrating why a complete account of sentencing disparity requires looking beyond published precedent to the high-volume dockets where most appeals are actually resolved.

District-court sentencing scholarship establishes the baseline for this inquiry. A robust literature demonstrates that defendant characteristics—such as race, sex, and offense type—remain strongly associated with federal sentencing outcomes even after controlling for legal variables.[5]

[4] J. Woodford Howard Jr., *Courts of Appeals in the Federal Judicial System: A Study of the Second, Fifth, and District of Columbia Circuits* (Princeton University Press, 1981); Frank B. Cross, *Decision Making in the U.S. Courts of Appeals* (Stanford University Press, 2007); Richard A. Posner, *The Federal Courts: Challenge and Reform* (Harvard University Press, 1996).

[5] These case-level disparities are well-documented. Spohn reports that race and sex remain significant predictors of sentence length regardless of criminal history, and Sentencing Commission reports show longer average sentences for Black and Hispanic males. See Cassia Spohn, "The Effects of the Offender's Race, Ethnicity, and Sex on Federal Sentencing Outcomes in the Guidelines Era," *Law and Contemporary Problems* 76, no. 1 (2013): 75–104; United States Sentencing Commission, *The Influence of the Guidelines on Federal Sentencing: Federal Sentencing Outcomes, 2005–2017* (U.S. Sentencing Commission, 2020); United States Sentencing Commission, *2023 Demographic Differences in Federal Sentencing* (U.S. Sentencing Commission, 2023), https://www.ussc.gov/research/research-reports/2023-demographic-differences-federal-sentencing. See also Kristin Finklea and Lisa N. Sacco, *Cocaine: Crack and Powder Sentencing Disparities*, CRS In Focus IF11965

A related line of work expands this focus from defendants to the judges themselves, revealing that disparity varies significantly across districts and individual courtrooms.[6] Scott's study of Massachusetts, for example, illustrates how individual judges within the same district diverged after *Booker* between routine guideline adherence and more aggressive discretionary departures.[7] This district-level focus, however, captures only the origin of a sentence. It misses the crucial appellate pipeline where those sentences are either affirmed, vacated, or resolved in unpublished orders. If trial-level disparities are ultimately sustained or corrected through unequal appellate review, then analyzing the district courts alone tells only half the institutional story.

The appellate-behavior literature demonstrates, first, that criminal appeals are not well described by a simple attitudinal model. Broad studies of appellate voting do find partisan patterning across federal dockets in large modern datasets, including measurable effects of panel

(Congressional Research Service, 2021), https://www.congress.gov/crs-product/IF11965 on persistent statutory disparities in crack-versus-powder cocaine sentencing; Cassia Spohn and Jerry Cederblom, "Race and Disparities in Sentencing: A Test of the Liberation Hypothesis," *Justice Quarterly* 8, no. 3 (1991): 305–27; Darrell Steffensmeier and Stephen Demuth, "Ethnicity and Judges Sentencing Decisions: Hispanic-Black-White Comparisons," *Criminology* 39 (2001): 145–78; Darrell Steffensmeier et al., "The Interaction of Race, Gender, and Age in Criminal Sentencing: The Punishment Cost of Being Young, Black, and Male," *Criminology* 36, no. 4 (1998): 763–98; Cassia Spohn and Lisa L. Sample, "The Dangerous Drug Offender in Federal Court: Intersections of Race, Ethnicity, and Culpability," *Crime & Delinquency* 59, no. 1 (2013): 3–31.

[6] Some scholars emphasize inter-district variation driven by local environments and aggregated demographics. See Richard D. Hartley and Robert Tillyer, "Inter-District Variation and Disparities in Federal Sentencing Outcomes: Case Types, Defendant Characteristics, and Judicial Demography," *Criminology, Criminal Justice, Law & Society* 20, no. 3 (2019): 46–63; Amy Farrell et al., "Intersections of Gender and Race in Federal Sentencing: Examining Court Contexts and the Effects of Representative Court Authorities," *The Journal of Gender, Race & Justice* 14, no. 1 (2010): 85–126; Matthew S. Crow and Natalie Goulette, "Sex, Politics, and U.S. District Court Outcomes: Examining Variation in Judge-Initiated Downward Guideline Departures," *American Journal of Criminal Justice* 48, no. 2 (2023): 295–318, https://doi.org/10.1007/s12103-021-09648-3; Matthew S. Crow and Keith A. Johnson, "Race, Ethnicity, and Habitual-Offender Sentencing: A Multilevel Analysis of Individual and Contextual Threat," *Criminal Justice Policy Review* 19, no. 1 (2008): 63–83; Matthew S. Crow and Natalie Goulette, "Judicial Diversity and Sentencing Disparity Across U.S. District Courts," *Journal of Criminal Justice* 82 (2022): 101973. A crucial strand of research instead focuses on intra-district, judge-level discretion. Cohen and Yang and Yang both find that after the *Booker* decision made the Guidelines advisory, inter-judge variation increased significantly, often splitting along demographic and political lines. See Alma Cohen and Crystal S. Yang, "Judicial Politics and Sentencing Decisions," *American Economic Journal: Economic Policy* 11, no. 1 (2019): 160–91; Crystal S. Yang, "Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker," *New York University Law Review* 89, no. 4 (2014): 1268–342. See also Paul J. Hofer et al., "Effect of the Federal Sentencing Guidelines on Interjudge Sentencing Disparity," *Journal of Criminal Law and Criminology* 90, no. 1 (1999): 239–322; Cassia Spohn and Patricia K. Brennan, "The Joint Effects of Offender Race/Ethnicity and Gender on Substantial Assistance Departures in Federal Courts," *Race and Justice* 1, no. 1 (2011): 49–78; Nicholas Goldrosen et al., "Racial Disparities in Criminal Sentencing Vary Considerably Across Federal Judges," *Journal of Institutional and Theoretical Economics* 179, no. 1 (2023): 92–113.

[7] Ryan W. Scott, "Inter-Judge Sentencing Disparity After Booker: A First Look," *Stanford Law Review* 63, no. 1 (2010): 1–66.

political composition and political alignment with lower-court judges.[8] The canonical panel-effects literature, however, emphasizes that judges rarely vote as isolated partisans; rather, their choices are constrained by doctrine and collegial dynamics. Cross and Tiller's whistleblower account, for example, illustrates how the presence of a dissent-capable colleague from an opposing party forces a panel majority to adhere more strictly to precedent.[9] Kim similarly relocates these effects away from simple partisan sorting and into the mechanics of deliberation and circuit-wide norms.[10] Sunstein's research sharpens this point specifically for the criminal context, where he found that partisan divergence is notably weaker in criminal appeals than in other legal domains, suggesting that the standard Democrat-versus-Republican axis fails to capture how appellate sentencing review actually operates.[11]

A second line of work shows that specific demographic characteristics and panel assignment practices can shape appellate outcomes in certain domains. Kastellec finds that adding a Black judge to a panel dramatically shifts affirmative-action rulings, while Peresie observes a similar voting shift in sex-discrimination cases when a female judge is present.[12] Hinkle extends this insight, showing that panel diversity alters not just who wins, but how the resulting opinions frame facts and articulate precedent.[13] Because these interactions can be consequential, the exact mechanism of panel formation becomes a significant variable. The Seventh Circuit's practitioner handbook explicitly describes its panel assignments as computer-randomized and walled off from

[8] Alma Cohen and Rajeev H. Dehejia, *Judges Judging Judges: Partisanship and Politics in the Federal Circuit Courts of Appeals*, Working Paper 32920 (National Bureau of Economic Research, 2024), https://doi.org/10.3386/w32920; Alma Cohen, "Pervasive Influence of Political Composition on Circuit Court Decisions," *Journal of Legal Analysis* 17, no. 1 (2025): 14–41.

[9] Frank B. Cross and Emerson H. Tiller, "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals," *Yale Law Journal* 107, no. 7 (1998): 2155–76.

[10] Pauline T. Kim, "Deliberation and Strategy on the United States Courts of Appeals: An Empirical Exploration of Panel Effects," *University of Pennsylvania Law Review* 157, no. 5 (2009): 1319–81.

[11] Cass R. Sunstein et al., "Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation," *Virginia Law Review* 90, no. 1 (2004): 301–54; Cass R. Sunstein et al., *Are Judges Political? An Empirical Analysis of the Federal Judiciary* (Brookings Institution Press, 2006).

[12] Jonathan P. Kastellec, "Racial Diversity and Judicial Influence on Appellate Courts," *American Journal of Political Science* 57, no. 1 (2013): 167–83; Jennifer L. Peresie, "Female Judges Matter: Gender and Collegial Decisionmaking in the Federal Appellate Courts," *Yale Law Journal* 114, no. 7 (2005): 1759–90.

[13] Rachael K. Hinkle, "Panel Effects and Opinion Crafting in the U.S. Courts of Appeals," *Journal of Law and Courts* 5, no. 2 (2017): 313–36.

case calendaring to prevent judge-shopping.[14] However, Chilton and Levy's empirical work cautions that formal design and realized assignment are not necessarily equivalent in practice; they demonstrate how related-case rules, scheduling conflicts, and the discretionary authority of court clerks can complicate true randomization.[15] Even so, this rich literature rarely asks a sentencing-specific question, leaving a gap regarding whether judge traits or panel mechanics actually correlate with sentence relief across a circuit's full merits stream.

The mechanics of appellate review further obscure these disparities because the federal courts of appeals do not speak through a single, uniformly visible docket. The decision to withhold an opinion from publication is structured rather than random, creating a secondary, less visible body of appellate law that manages heavy caseloads but receives less doctrinal scrutiny than published precedent.[16] The scale of this phenomenon is substantial, with roughly eighty-seven percent of federal appellate decisions now unpublished.[17] As McAlister's research demonstrates, this lower-visibility docket is not evenly distributed. It disproportionately contains what she calls "bottom-rung" appeals—the lowest tier of judicial triage where high-volume, low-status cases like criminal, prisoner, and *pro se* matters are resolved through thinner procedures and abbreviated reasoning.[18] Those categories matter deeply here because criminal appeals, especially routine sentencing challenges, are especially likely to be processed through these low-visibility, nonprecedential forms of review.

The most recent empirical work pushes that argument further by showing that publication is not merely a neutral formatting choice. Hinkle argues that the decision to publish can be shaped

[14] United States Court of Appeals for the Seventh Circuit, "Practitioner's Handbook for Appeals to the United States Court of Appeals for the Seventh Circuit," 2020, https://www.ca7.uscourts.gov/rules-procedures/Handbook.pdf.
[15] Adam S. Chilton and Marin K. Levy, "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals," *Cornell Law Review* 101 (2015): 1–56; Marin K. Levy, "Panel Assignment in the Federal Courts of Appeals," *Cornell Law Review* 103, no. 1 (2017): 65–116.
[16] Donald R. Songer, "Criteria for Publication of Opinions in the U.S. Courts of Appeals: Formal Rules Versus Empirical Reality," *Judicature* 73 (1990): 307–13; Deborah Jones Merritt and James J. Brudney, "Stalking Secret Law: What Predicts Publication in the United States Courts of Appeals," *Vanderbilt Law Review* 54, no. 1 (2001): 71–121; Donna S. Stroud, "The Bottom of the Iceberg: Unpublished Opinions," *Campbell Law Review* 37 (2015): 333–85.
[17] Rachel Brown et al., "Is Unpublished Unequal? An Empirical Examination of the 87% Nonpublication Rate in Federal Appeals," *Cornell Law Review* 107 (2022): 1–150.
[18] Merritt E. McAlister, "Downright Indifference," *Michigan Law Review* 118, no. 4 (2020): 533–610; Merritt E. McAlister, "Rebuilding the Federal Circuit Courts," *Northwestern University Law Review* 116 (2022): 1137–226; Merritt E. McAlister, "Bottom-Rung Appeals," *Fordham Law Review* 91 (2023): 1355–422.

by ideological and strategic considerations, while Newman and Levy show how unwritten circuit customs help govern when an opinion becomes precedential.[19] Varsava's latest dataset adds that author characteristics correlate with publication status in a large modern appellate sample.[20] Read together with earlier findings on how panel composition alters opinion crafting, this frontier suggests that the "unpublished" label is part of the institutional machinery by which circuits allocate attention, visibility, and lawmaking authority.[21]

Existing scholarship therefore leaves a narrower but critically important question unanswered. We have district-court studies of sentencing disparity, appellate studies of panel effects, and institutional critiques of selective publication, but they remain isolated. What remains missing is a judge-identifying, circuit-specific account of sentencing appeals that integrates these insights by mapping the full merits stream across its various levels of procedural visibility. This thesis asks whether apparent differences in appellate sentencing relief are better explained by stable judge traits or by the institutional organization of review itself. It does so by dividing appellate decisions into three tracks: the Aggregate track (the full universe of merits decisions), the Routine track (the unpublished, nonprecedential dispositions), and the Doctrine track (the published opinions that make binding law). By standardizing for case mix across these three tracks, the thesis tests whether a judge's demographic and professional traits drive sentencing relief, or whether the inherent processes of appellate review explain more of the disparity.

Methodology

This study begins with the Seventh Circuit's public repository of opinions, nonprecedential dispositive orders, and oral-argument materials, from which it collects all criminal appellate orders issued between January 1, 2020, and December 31, 2025. Those records are then narrowed to the merits portion of the sentencing docket: direct appeals and resentencings after remand, excluding

[19] Rachael K. Hinkle, *Selective Publication in the U.S. Courts of Appeals: The Invisible Norm That Perpetuates Inequality* (Oxford University Press, 2024); Jon O. Newman and Marin K. Levy, *Written and Unwritten: The Rules, Internal Procedures, and Customs of the United States Courts of Appeals* (Cambridge University Press, 2024).
[20] Nina Varsava, "Opinion Authorship and Precedential Status," *Washington University Law Review* 101 (2024): 1593–674.
[21] Hinkle, "Panel Effects and Opinion Crafting in the U.S. Courts of Appeals."

collateral attacks, revocations, detention and bail, or other procedural orders.[22] This is a deliberate choice since the thesis is designed to compare like with like within appellate sentencing review, not to pool together procedurally heterogeneous matters that are governed by different standards and remedial stakes.

The data pipeline proceeds in four stages. Source PDFs are first deduplicated into canonical documents using byte, text, and visual checks so that amended versions, duplicate uploads, and overlapping records do not inflate the corpus. Each canonical document is then processed through a segmented, schema-constrained LLM extraction system that codes panel membership, publication status, posture, offense category, issues, and relief outcomes into a uniform structure. Every field is stored with both a substantive value and a status marker indicating whether the information was `present`, `not_mentioned`, `unclear`, or `not_applicable`.[23] The resulting decision-level table is then expanded into a judge-vote table, with one row for each judge participating in each case, because appellate sentencing review is rendered by panels rather than by individual judges sitting alone. Those judge-vote rows are not independent because judges on the same panel share the same case-level outcome, which is why the inferential model later includes a case-level random intercept rather than treating panel votes as separate sentencing events.

High non-present rates in some extracted fields do not automatically indicate model failure. Because each extracted field is stored as a status-value pair, the pipeline records both what an opinion states and what it declines to state. A field marked `present` carries substantive content; fields marked otherwise are not carried forward as observed substantive values and instead remain missing, unknown, or grouped residual categories depending on the variable, which makes missingness legible.[24] A higher non-present rate may partly reflect extraction limits at the margin,

[22] Appendix V's analytical sample-construction audit reports the full corpus and each exclusion step in sequence.

[23] Outputs that fail schema or contract validation are logged and set aside for manual correction, so the final dataset reflects only records that satisfy the pipeline's legal and structural rules. Appendix I reports the extraction schema, validation rules, and fuller model diagnostics in detail. In practical terms, "set aside" means that outputs violating schema or status/value contract (e.g., an empty entry with a "present" label would not pass the check) are excluded from the analytical table until the underlying extraction error is resolved by manual review.

[24] In the current quality audit, `decision_relief` is known in 1,567 of 1,591 decisions (98.5 percent), while offense category is present in 1,191 decisions (74.9 percent). Appendix IV reports the fuller variable-level non-present audit. The key methodological point is that non-`present` fields are not silently treated as observed content: the status-value contract forces those entries to remain null at extraction, after which the preprocessing pipeline carries them forward as

but it more often reflects the selective disclosure practices of appellate opinions themselves.

The final inferential design uses three related universes, labeled Aggregate, Routine, and Doctrine. The final training sample contains 1,375 merits decisions and 4,157 judge-vote rows, of which 728 decisions and 2,180 rows fall in Routine, and 647 decisions and 1,977 rows fall in Doctrine.[25] Holding the model specification constant across all three tracks permits direct comparison between the court's overall output and the two publication regimes that compose it.

To estimate relief across those tracks, the thesis uses a hierarchical Bernoulli-logit generalized linear mixed model fit by Markov Chain Monte Carlo, specifically the No-U-Turn Sampler. Standardized judge rates are then computed against a common reference docket so that judges are compared on the same mixture of case types rather than on whatever mix happened to reach their panels. This design is intentionally conservative. The point is to determine whether apparent inter-judge differences persist once the court's own decisional structure is taken into account.

Several limitations follow from that choice. Because panel compositions are thinly distributed, the model cannot estimate stable panel-specific interaction effects.[26] Because publication is itself a judicial decision, the analysis can distinguish subsets but cannot fully recover the doctrinal or strategic reasons cases are sorted into one subset rather than another. And because per-judge counts remain modest in some strata, partial pooling necessarily compresses small apparent differences by shrinking sparse judge-specific estimates toward the overall mean. Those are real limits, but they are preferable to the more serious distortion that results from treating published opinions alone, or raw relief rates alone, as an adequate account of appellate sentencing review.

Analysis

This chapter proceeds in six parts. It first defines the relevant denominators in each track, then turns to the Aggregate, Doctrine, and Routine subsets. It then turns to judge-attribute compar-

missing, unknown, or grouped residual categories according to the variable's role in the analysis.

[25] The judge-vote counts are not simple triple multiples of the case counts because five en banc cases include 10, 10, 11, 11, and 12 judge-rows rather than the ordinary three-judge panel.

[26] Statistically, 21 judges yield 1,330 distinct three-judge panels ($\binom{21}{3}$), but only 375 are observed across 1,394 unique decisions; 232 of those observed combinations (61.9%) appear three times or fewer.

isons within those tracks, model diagnostics, and the resulting synthesis. The guiding premise is that appellate sentencing disparity cannot be evaluated responsibly unless the analysis distinguishes the relief environment litigants actually face from the publication regimes through which the court structures that environment.

A. Universes and Denominators

Table 1 states the central descriptive fact of this chapter: once publication track is made visible, the Seventh Circuit's merits docket separates into two sharply different tracks.

Table 1: Three-track sentencing universes.

Track	Decisions	Vote Rows	Agg. Share (%)	Relief (%)	Per curiam (%)
Aggregate	1,375	4,157	100.0	11.23	55.3
Routine	728	2,180	52.4	3.44	100.0
Doctrine	647	1,977	47.6	19.83	5.1

Overall, relief is a rare event in the Seventh Circuit. Even without any statistical modeling, the descriptive disparity is immediate. Across the full merits stream, the court grants sentence relief in only 11.23 percent of decisions. But that Aggregate figure pools two very different tracks of publication. Routine contains slightly more than half, 52.4 percent, of the decisions yet only a 3.44 percent relief rate, whereas Doctrine contains 47.6 percent of the decisions yet a 19.83 percent relief rate, nearly six times Routine. Once publication status is introduced, the merits docket no longer looks like one continuous field with modest variation around a common center. It breaks into two tracks with sharply different rates of relief, modes of authorship, and doctrinal visibility.

The per curiam column further reveals how the appellate panels process, write, and present sentencing appeals. All orders in Routine are labeled per curiam throughout, without a single exception in the six-year period, while Doctrine is rarely so labeled, with 94.9 percent of decisions explicitly authored by a judge. Read as a whole, Aggregate can make the court look like a single

merits docket with one average relief rate. Read by track, the same docket looks much more divided by publication form and decisional style than by any simple judge-to-judge pattern. That is why the usual appellate denominators in the literature are not enough here. Whole-docket studies and published-opinion studies both blur the sharp split that appears inside a single circuit's criminal merits stream.

B. Aggregate Lens

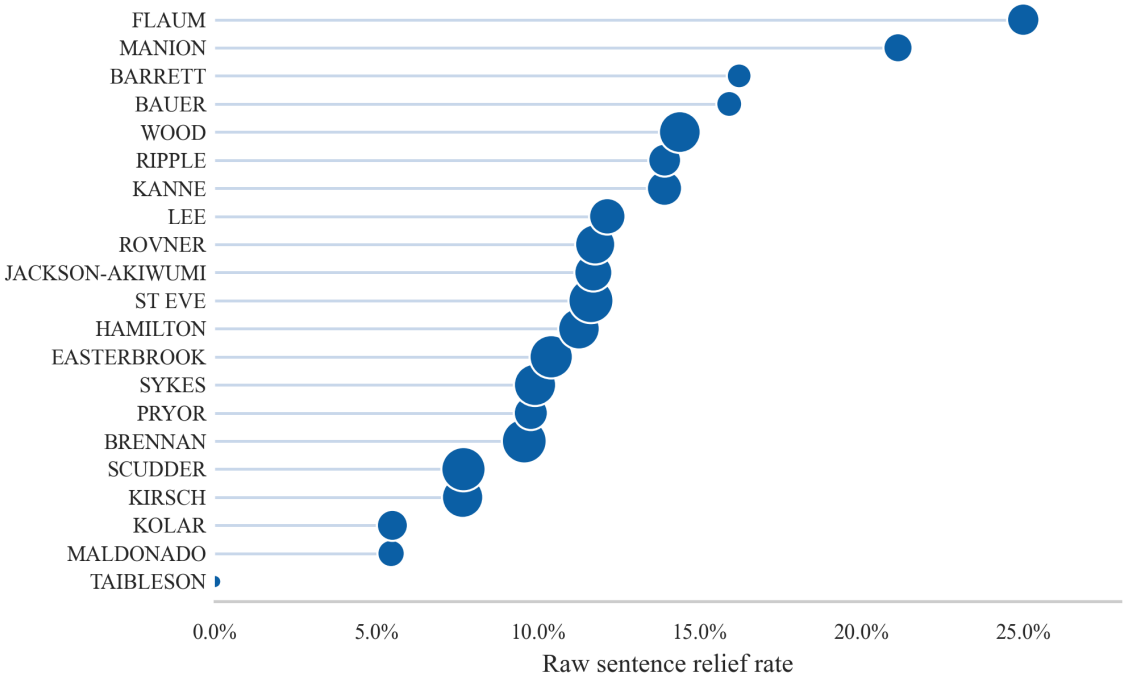

Figure 1: Raw Aggregate sentence-relief rates by judge across all panel judges. Bubble size scales with Aggregate panel votes.

Before any standardization, the Aggregate merits stream reveals a court with large judge-to-judge disparities in sentence relief. Figure 1 plots raw sentence-relief rates for every judge who appears in the Aggregate panel data, with bubble size scaled to panel-vote volume. Across the full judge set, the observed rate runs from 0 to nearly 25 percent. A lawyer scanning outcomes judge by judge could easily conclude that some members of the court are several times more willing than others to grant relief on appeal.

That reaction is understandable, but the raw ranking is only the observed surface of the Aggregate docket, not a clean measure of judicial ideology or stable leniency. Some of the most dramatic rates belong to judges with thinner panel exposure than the court's most frequently sitting members, and even the larger bubbles reflect very different portfolios of sentencing appeals. Some judges may sit on more published cases, more guideline-calculation disputes, or more predicate-offense questions; others may see heavier shares of Routine affirmances and highly deferential reasonableness claims. The immediate task is therefore to show what kinds of cases populate the Aggregate stream before asking what remains once judges are compared on a common docket.

Table 2: Aggregate offense composition. Shares are calculated over Aggregate decisions with known relief. Only the largest categories are shown.

Offense	Share (%)	Relief (%)
Drugs	32.0	11.6
Firearms	19.9	14.5
Fraud	10.5	9.0
Child exploitation	8.2	8.0
Violent	7.1	13.3
Sex offense	5.5	5.3

The offense table comes first because offense category is the most stable descriptive summary of the docket's composition. Drugs and firearms together make up just over half of the Aggregate docket, so they anchor the court's baseline. Fraud, child-exploitation, violent-crime, and sex-offense appeals then shift that baseline around the edges. A judge who sees more firearms and violent-crime appeals will usually face a somewhat more correction-rich docket than one who sees heavier shares of sex-offense or low-relief categories.

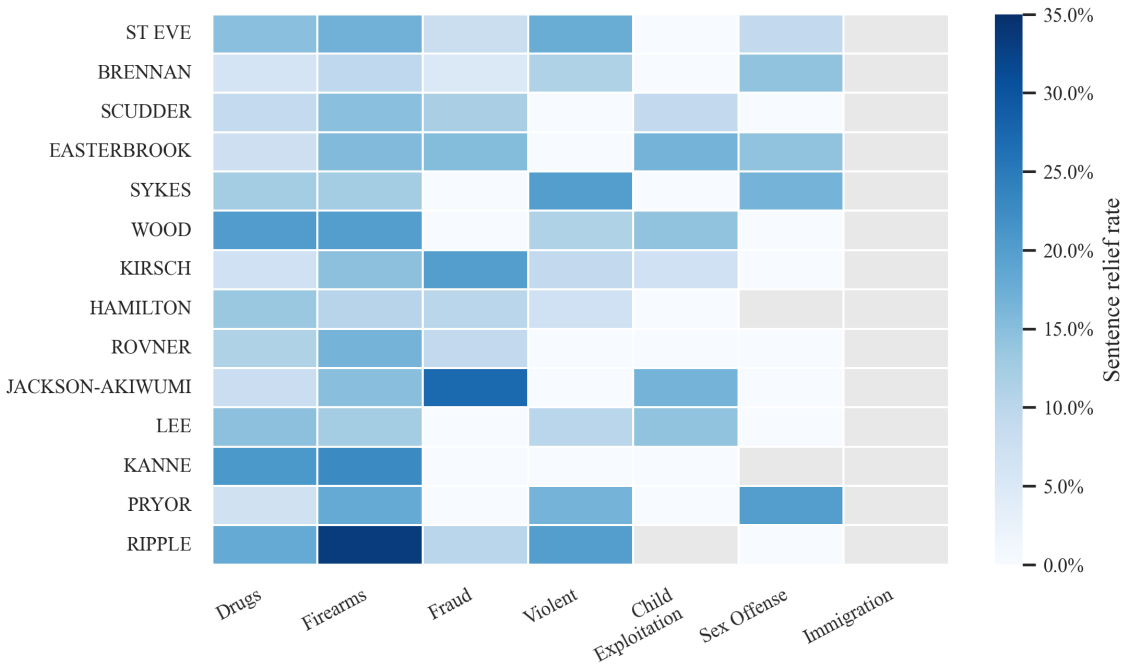

Figure 2: Judge-by-offense raw sentence-relief rates in the Aggregate merits universe. Gray cells indicate sparse judge-offense cells.

The offense heatmap is more muted than the raw judge ranking because offense category is only a coarse proxy for case difficulty. Each offense bucket contains many different sentencing theories, standards of review, and procedural postures. Even so, the same baseline structure still appears, as drugs and firearms dominate the docket and therefore anchor the Aggregate baseline, while lower-relief categories such as sex offenses and immigration pull that baseline downward and more correction-rich categories within the common offense mix pull it upward. The issue-level breakdown that follows sharpens that picture at a finer doctrinal level.

Issue categories then add the finer-grained legal picture that offense labels cannot capture. Substantive reasonableness appears the most often, but it produces relief in only 4.5 percent of issue-pairs. Guideline miscalculation appears slightly less often, yet it succeeds at roughly three times that rate. Procedural reasonableness sits in between, and guideline interpretation remains materially more relief-rich than the common substantive claims that dominate sentencing appeals. In practical terms, that means two judges can sit on similarly large numbers of sentencing cases yet inherit very different correction opportunities depending on whether their panels receive a heavier

mix of technical guideline and predicate disputes or a heavier mix of deferential reasonableness challenges, which is solely determined by the appellants' legal arguments.

Table 3: Aggregate issue composition. Shares are calculated over unique decision-issue pairs in the known-outcome Aggregate universe. Only the largest categories are shown.

Issue	Share (%)	Relief (%)
Substantive reasonableness	10.5	4.5
Guidelines miscalculation	9.2	13.5
Procedural reasonableness	8.7	11.0
Waiver or jurisdiction	7.9	8.3
Evidentiary	6.1	10.7
Guidelines interpretation	5.8	11.4

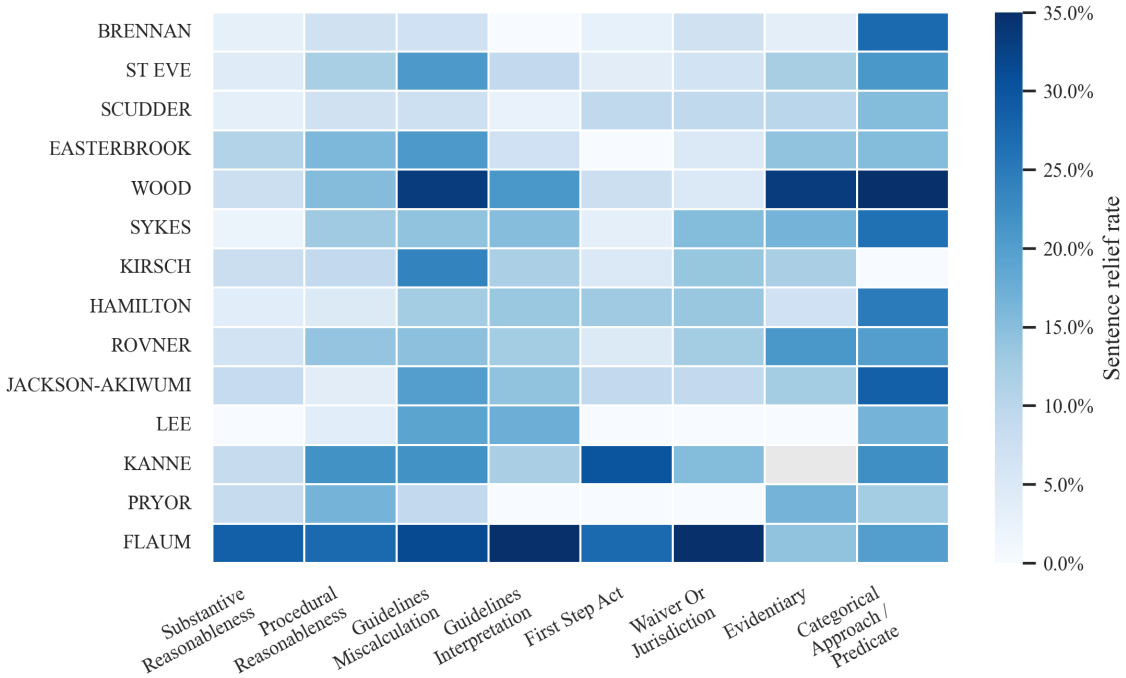

Figure 3: Judge-by-issue raw sentence-relief rates in the Aggregate merits universe. Gray cells indicate sparse judge-issue cells.

Figure 3 makes that pattern visible. The darkest cells cluster around guideline-calculation

and predicate-offense questions, while substantive-reasonableness cells remain comparatively pale across even the most heavily exposed judges. That pattern is legally intuitive because appeals framed in technical or rule-bound terms create more opportunities for appellate correction than claims governed by highly deferential review. A judge who sees more of the former and less of the latter will therefore look more relief-favorable before any adjustment is made, even if the judge is not generally more willing to grant relief across the docket as a whole.

Taken together, the offense and issue displays show why the raw ranking cannot be read in isolation. Judges do not confront the same mix of offenses, issues, or publication environments, and those differences are large enough to alter the apparent rate of relief. The core model therefore standardizes on posture and offense, which give each decision a stable baseline structure. Issue tags stay at the descriptive stage because they are usually multi-label, sparser, and much closer to the doctrinal disputes already being contested on appeal. Using them as core controls would risk turning the model into a partial restatement of those disputes rather than a cleaner baseline for comparison. The next section introduces the standardized rates directly and compares Aggregate, Routine, and Doctrine to show how the apparent spread changes once the merits stream is separated back into its component tracks.

C. Track Separation

Section C applies the standardization described in methodology and asks what remains once judges are compared on a common docket. Figure 4 shows what happens when that adjustment is applied. The standardized distributions rescore each judge on the same reference docket within each track, then partially pool thin-exposure judges back toward the relevant track mean. That makes the procedure deliberately conservative in the very cells where raw rates are most volatile. Once the adjustment is imposed, the pooled Aggregate spread contracts sharply, and the three tracks tighten around very different centers: a low-relief Routine environment and a much higher-relief Doctrine environment.[27]

[27] Separate opinions are rare in the modeled merits universe. Only 54 of the 1,375 modeled merits decisions, about 3.9 percent, contain a concurrence or dissent. That low rate is consistent with broad panel agreement, although it does not

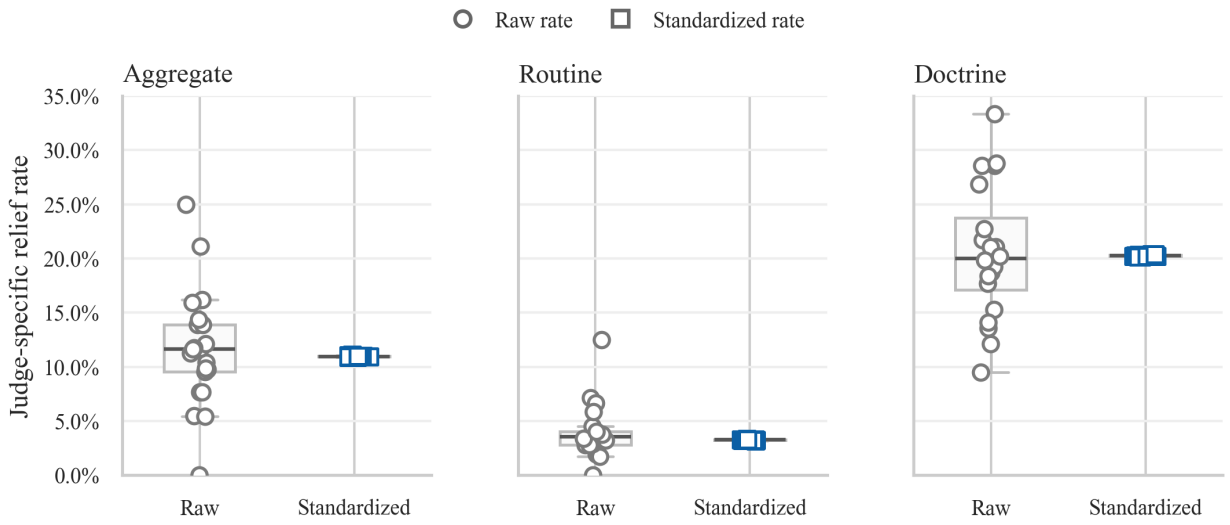

Figure 4: Judge-level raw and standardized sentence-relief-rate distributions across Aggregate, Routine, and Doctrine. Each point is a judge; the boxplots summarize the judge-level distributions within each track.

Table 4 extracts the same shift numerically. The observed relief rate and posterior predictive mean are very close in all three tracks, and the observed rate falls comfortably inside the 95 percent posterior predictive interval each time. That means the fitted model is reproducing the baseline relief environment of each track well before it asks whether any residual judge-level spread remains.

Table 4: Core track diagnostics.

Track	Obs. (%)	PP Mean (%)	95% PP Int.	Dispersion	p
Aggregate	11.23	11.34	[10.85, 11.84]	2.84x	<0.001
Routine	3.44	3.61	[3.03, 4.22]	1.23x	0.212
Doctrine	19.83	19.98	[19.12, 20.89]	1.35x	0.172

The inferential result directly supports the descriptive split. Aggregate still shows excess residual spread (2.84x, $p < 0.001$), but Routine does not (1.23x, $p = 0.212$), and neither does Doctrine (1.35x, $p = 0.172$). Those p-values are statistically not significant, and they do not

capture disagreement that never appears in a separate opinion.

measure the probability that the null is true. Rather, they mean that under a null of no excess residual judge-level dispersion, spreads at least this large would arise about 21 percent of the time in Routine and about 17 percent of the time in Doctrine. That is too often to count as strong evidence of within-track disparity once posture, offense, and shared case-level dependence are taken into account. Because the model shrinks thin cells toward the relevant track mean, it is harder, not easier, for a large within-track spread to survive. What remains after that adjustment is a modest pooled Aggregate residual sitting on top of two very different publication baselines. The distance between Routine and Doctrine is still about sixteen to seventeen percentage points (about six times) in observed relief and nearly as large in the standardized estimates; the residual within-track spread is small.

D. Judge Attributes Within Track

Once the docket is split by track and judges are compared on a common baseline, the attribute figures become much easier to read. The central visual question is whether those differences remain large after standardization. In each figure, the open circle marks the raw category rate, the open square marks the standardized estimate, and the vertical interval shows the uncertainty around that standardized estimate. Across this section, some categories sit slightly above others, but those within-track differences are modest and the much larger divide still runs between Routine and Doctrine.

Figure 5 provides the clearest illustration because party is the attribute most likely to invite an overread. The standardized Democratic estimate is slightly higher than the Republican estimate in all three tracks, but only by a narrow margin: about 1.0 percentage point in Aggregate, 0.4 points in Routine, and 2.1 points in Doctrine. Those gaps are small beside the sixteen-to-seventeen-point distance separating Routine from Doctrine themselves, and the Democratic intervals are somewhat wider because the cohort contains only seven judges compared to fourteen. The more interesting movement is the shift from raw to standardized rates. In Aggregate and Routine, Democratic appointees move upward more than Republicans once they are compared on the

same docket, which suggests that the raw summaries partly reflected a less relief-rich case mix rather than a cleaner partisan divide. Overall, Democrats sit a little higher than Republicans after standardization, but the effect remains modest and heavily overshadowed by track.

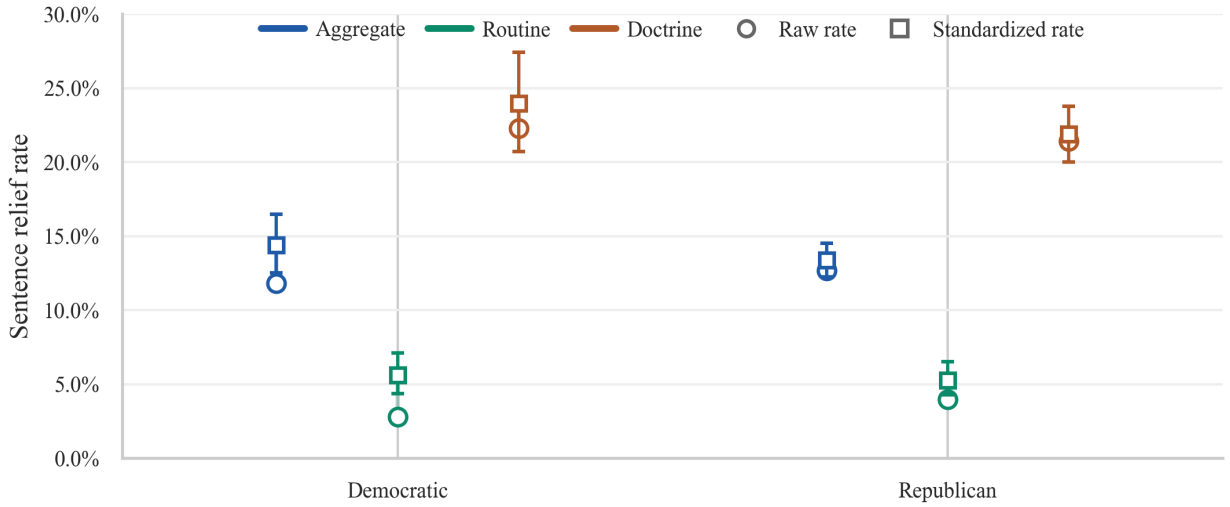

Figure 5: Raw and standardized sentence-relief rates by appointing party on a common axis. Track-colored segments connect raw rates to standardized estimates with 95 percent intervals across Aggregate, Routine, and Doctrine.

Figure 6 shows the most structured raw ordering in the section, but it also shows why the profession results have to be read cautiously. The labels are not mutually exclusive. Of the twenty-one judges in the study, nineteen have private-practice experience, fourteen clerked, twelve served as prosecutors, nine have law-school teaching experience, and only two have a criminal-defense label. That overlap makes the common categories confound each other, while the criminal-defense category is built on the thinnest judge base in the figure. On the raw Aggregate scale, former law-school professors sit visibly above the main cluster, at 15.4 percent compared with roughly 11.0 to 12.6 percent for law clerks, prosecutors, and private-practice lawyers, a raw rate roughly forty percent higher than the lower cluster. After standardization, that ordering still exists but narrows. Law professors remain somewhat higher in Aggregate and Doctrine, while private practice, prosecution, and clerkship cluster tightly together. Criminal-defense judges sometimes appear higher still, but those estimates carry the widest intervals in the figure and rest on only two judges, which makes them too unstable to bear much explanatory weight. Professional background

therefore may track some modest differences at the margins, especially for law professors, but the common profession categories remain far closer to one another than any of them is to the larger divide between Routine and Doctrine.

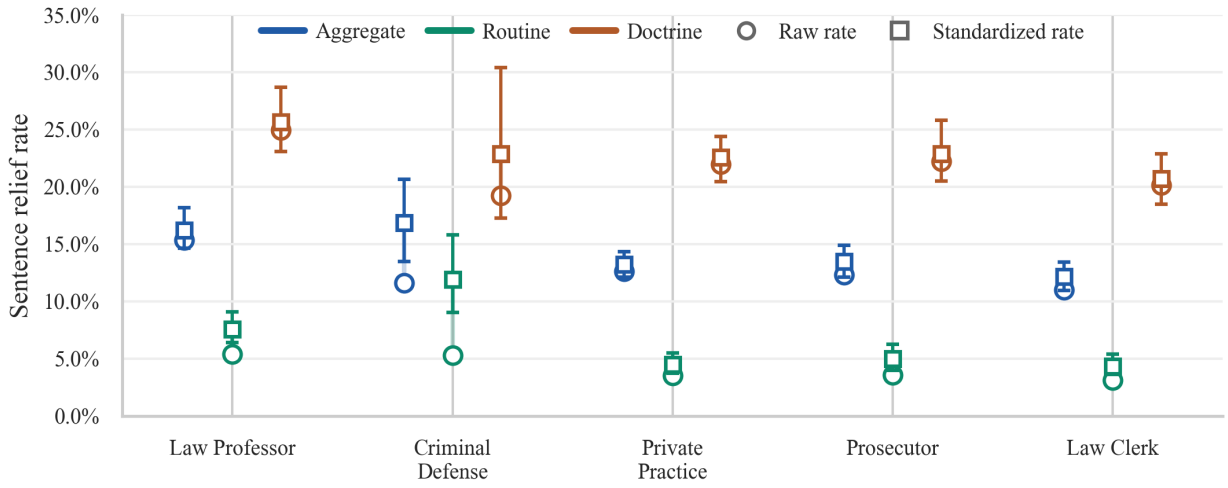

Figure 6: Raw and standardized sentence-relief rates by former profession on a common axis. Track-colored segments connect raw rates to standardized estimates with 95 percent intervals across Aggregate, Routine, and Doctrine.

Figure 7 shows why race is the weakest basis for strong inference in this section. Seventeen judges in the study are coded White, two African, one Asian, and one Hispanic. That means the Asian and Hispanic estimates are effectively single-judge profiles, and the African category is built on only two judges. The resulting intervals are correspondingly wide, especially once the data are split by track. In that setting, raw zeros, sharp standardized jumps, and shifting cross-track orderings are not statistically reliable indicators of group differences. The central point is therefore not any stable ranking among racial categories, but the dataset's inability to support a strong race-based account of residual disparity.

Figure 8 is easier to read because the cohorts are larger and closer in size: nine women and twelve men. Even there, however, the substantive gap is small. Similar to the Democratic versus Republican split, female judges sit slightly above male judges in the standardized estimates across all three tracks, by about 1.0 percentage point in Aggregate, 1.9 points in Routine, and 1.5 points in Doctrine. Those are real directional differences, but they remain minor compared with

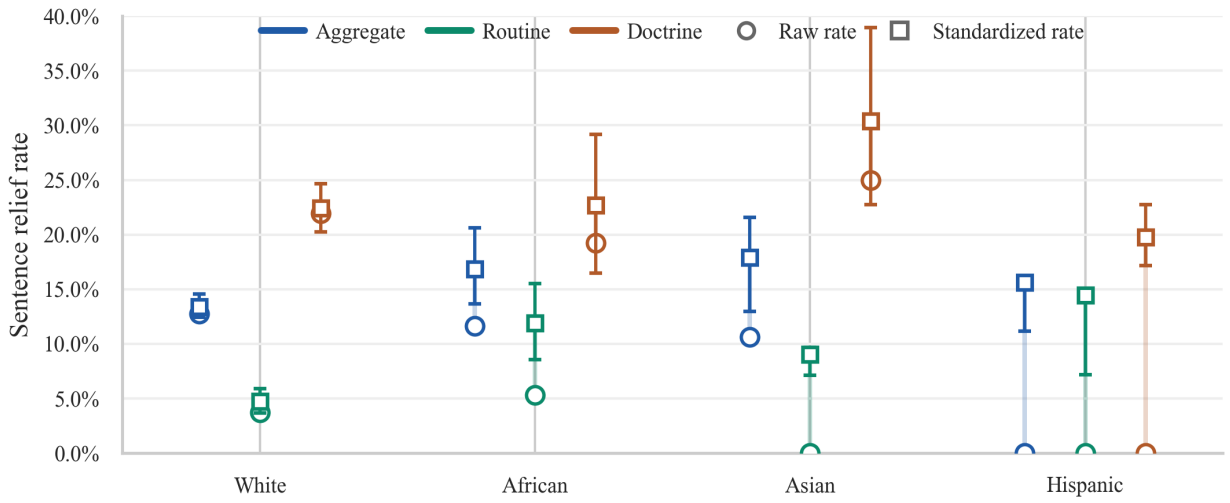

Figure 7: Raw and standardized sentence-relief rates by judge race on a common axis. Track-colored segments connect raw rates to standardized estimates with 95 percent intervals across Aggregate, Routine, and Doctrine.

the much larger change created by moving from the Routine baseline to the Doctrine baseline. The confidence intervals also overlap substantially, and the figure does not show any representative indicator of a clean sex-based division of the court. The most that can be said is that female judges in this sample appear somewhat more relief-favorable than their male colleagues after standardization, but only at the margins and not in a way that rivals publication tracks as an explanation of the observed docket.

Figure 9 and Figure 10 show the most consistent directional trends in the section, and they are best read for shape as much as for level. Across both figures, the three tracks move in roughly parallel directions, with the publication baselines staying far apart, but the curves themselves following similar paths. That means age and tenure may correlate with modest changes in relief, yet they do not reorganize the underlying structure of the docket. The dominant difference is still vertical, between Routine and Doctrine, not horizontal, between younger and older or shorter- and longer-serving judges. Within that narrower frame, however, the age and tenure plots show more consistent patterning than the party, race, gender, or profession figures and therefore merit separate attention.

In Figure 9, tenure shows the clearer of the two gradients. In Aggregate, the standardized

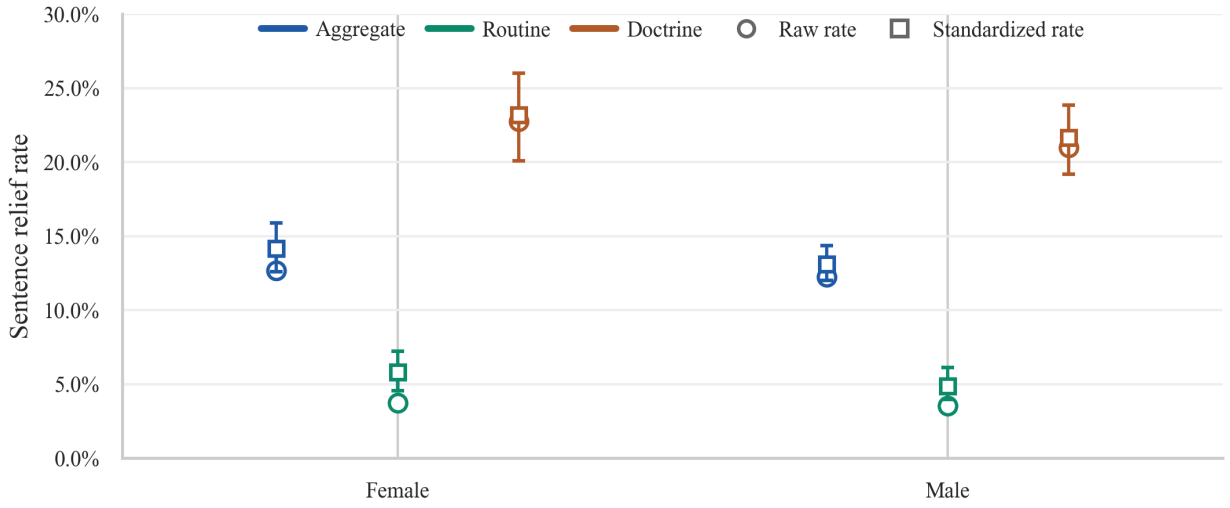

Figure 8: Raw and standardized sentence-relief rates by judge gender on a common axis. Track-colored segments connect raw rates to standardized estimates with 95 percent intervals across Aggregate, Routine, and Doctrine.

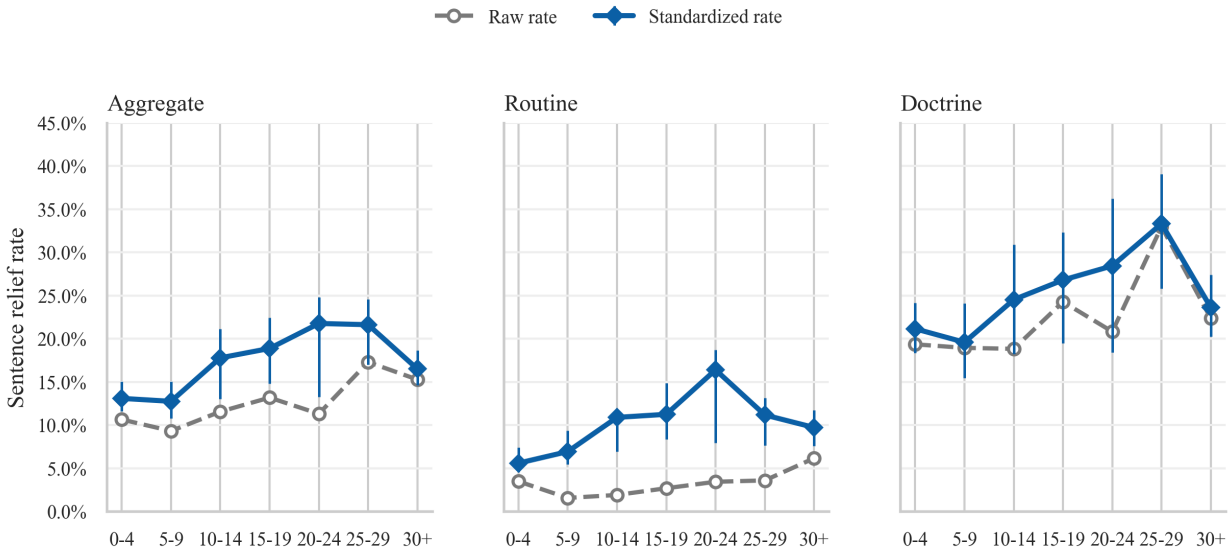

Figure 9: Raw and standardized sentence-relief rates by years on bench across Aggregate, Routine, and Doctrine. Dashed gray lines show raw rates; solid blue lines and vertical bars show standardized rates and 95 percent intervals.

rate rises from about 13.1 percent for judges with 0-4 years on the bench to roughly 21.6 to 21.8 percent in the 20-24 and 25-29 year bins, an increase of about sixty-five percent, before easing back to about 16.5 percent in the 30-plus group. Doctrine follows the same general contour at a higher level, climbing from about 21.2 percent in the newest cohort to 33.3 percent at 25-29 years, an increase of roughly fifty-seven percent, before falling back to 23.6 percent in the longest-serving group. Routine is lower throughout but broadly similar in shape, with a sharper peak at middle-tenure categories above the newest cohort and the oldest cohort dropping slightly but remaining above the early-career baseline. The pattern suggests a bounded seniority effect with the same broad shape across tracks.

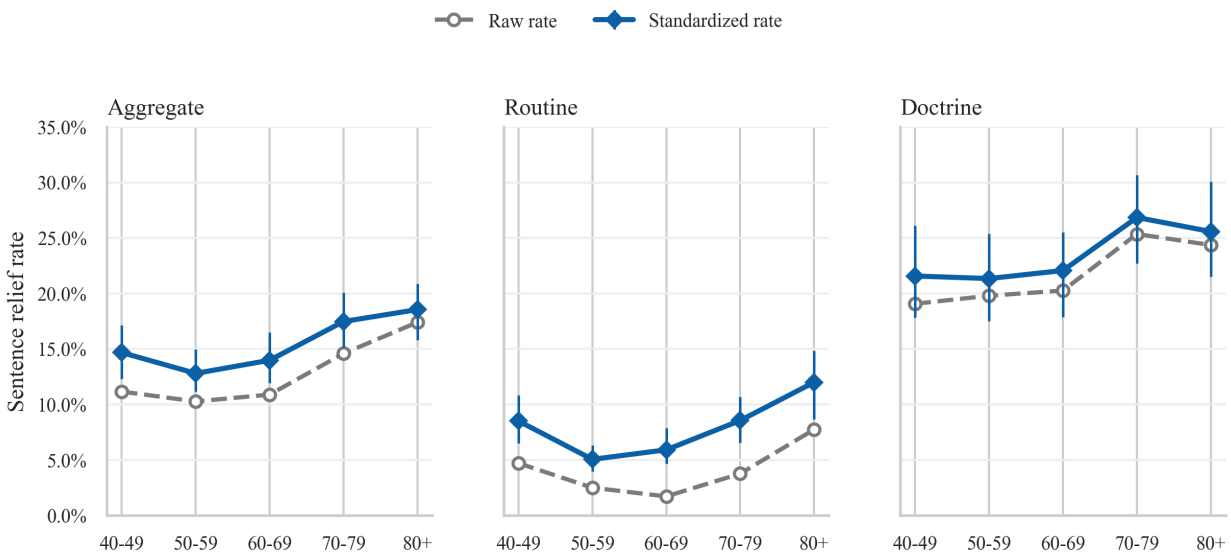

Figure 10: Raw and standardized sentence-relief rates by judge age across Aggregate, Routine, and Doctrine. Dashed gray lines show raw rates; solid blue lines and vertical bars show standardized rates and 95 percent intervals.

In Figure 10, age shows a related, slightly smoother pattern. The youngest age bin does not sit at the bottom of every track, but the second-youngest cohort is consistently lower, after which the standardized rates rise steadily into the older groups. In Aggregate, the standardized rate moves from 14.7 percent for ages 40-49 down to 12.8 percent for ages 50-59, then climbs through 14.0 and 17.5 to 18.5 percent in the 80-plus group, about a forty-five percent increase from the low point to the oldest cohort. Doctrine again follows the same shape at a higher baseline, while

Routine follows it at a lower one. Taken together, the age and tenure plots support a restrained reading that more senior judges are somewhat more relief-favorable, and the pattern is more pronounced than the earlier party, race, gender, or profession contrasts, but it is still gradual and smaller than the split created by publication track itself.

This pattern is likely entangled with judicial biography. In this sample, older and longer-serving judges are disproportionately Republican appointees from Reagan and Bush administrations and are more likely to have assumed senior status. The profession data also point in the same direction, with the law-professor group being older and longer-serving on average than the very small criminal-defense cohort. Age and tenure therefore deserve more attention, but they should be read as correlated variables, not as completely independent causal mechanisms.

E. Robustness

A pooled sensitivity sequence reinforces the chapter's central claim. Table 5 shows what happens to the pooled model as successive blocks of covariates are added. `Tail Loss` is the variational fit criterion used here, so lower values indicate a closer fit; `Delta` records the gain from each added block; and `Judge Sigma` with its `95% CrI` shows how much residual judge-level dispersion remains after each step. On those terms, posture and offense barely move the pooled model at all. Publication does significantly. Once publication status enters the sequence, tail loss falls by `128.6`, dwarfing every other change in the table, while a coarse standard-of-review grouping yields only a modest further gain of `7.2`. The pooled sequence therefore points in the same direction as the three-track analysis itself. The largest omitted distinction inside the criminal merits stream is publication track, not the baseline offense or posture controls.

Table 5: Pooled VI sensitivity sequence.

Model	Added	Tail Loss	Delta	Judge Sigma	95% CrI
M0	Baseline	1507.0	0.0	1.012	[0.560, 1.586]
M1	Posture	1502.5	4.5	1.011	[0.550, 1.606]

Model	Added	Tail Loss	Delta	Judge Sigma	95% CrI
M2	Offense	1508.2	-5.7	1.060	[0.581, 1.664]
M3	Publication	1379.6	128.6	1.135	[0.637, 1.772]
M4	Std. review	1372.4	7.2	1.183	[0.641, 1.831]

The sigma-prior check addresses a narrower robustness concern. The judge-effect prior controls how strongly the model regularizes residual judge heterogeneity, so varying it tests whether the chapter's rankings depend heavily on one statistical prior setting. They do not. In Aggregate, changing the prior from `\sigma = 0.5` to `1.0` to `1.5` moves standardized judge rates by at most about 0.04 percentage points, with a median shift near 0.02 points; the median rank shift is 0, and the maximum shift is only three places. Routine and Doctrine move just as little in rate terms, never exceeding about 0.06 percentage points across the tested priors.

Figure 11 shows the same point visually. Most Aggregate ranks do not move at all as the prior changes. The judges who do move are concentrated in the middle of the ordering, where standardized rates are already tightly clustered, while the top and bottom clusters remain largely fixed. The prior can nudge near-tied middle positions, but it does not reorder the judges who sit clearly at the top or bottom of the Aggregate distribution, which demonstrates the stability and robustness of the model.

Table 6 shows why judge biography is a weak rival explanation for the remaining spread. The table reports posterior means, which are the model's best estimates of direction and magnitude, together with 95% credible intervals, which show the range of values compatible with the data. Those estimates cluster close to zero, and every interval spans both positive and negative values. Section D already showed that some groups sit slightly above others in the descriptive and standardized figures; the pooled supplemental trait model places those differences in a small and uncertain range once the core case controls are held constant. The largest positive means belong to age and former defense experience, but both intervals remain wide enough to accommodate substantial movement in either direction. The same pattern holds for party, gender, race, tenure,

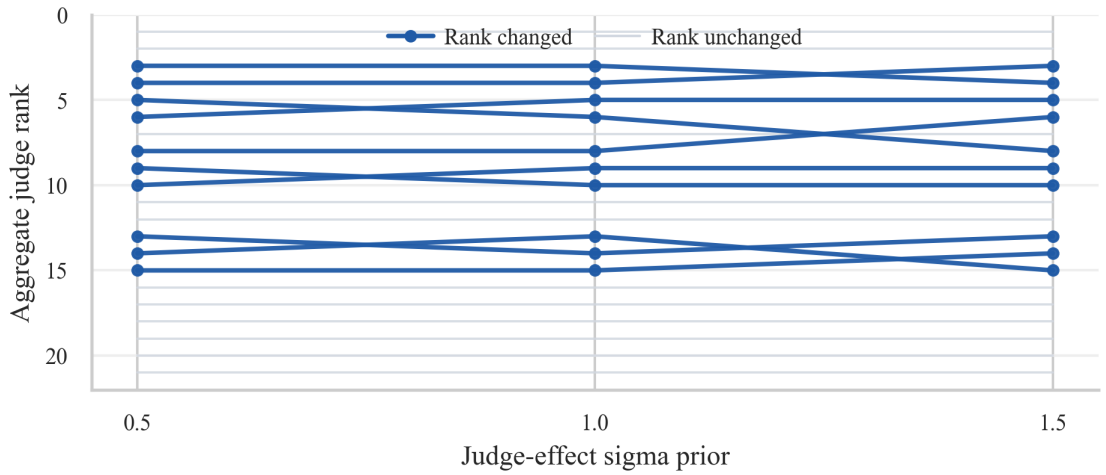

Figure 11: Aggregate judge-rank stability under alternative judge-effect sigma priors. Blue lines mark judges whose Aggregate rank changes as the sigma prior moves from 0.5 to 1.0 to 1.5; gray lines remain fixed.

and prosecutorial background. No biography variable emerges here as a reliable, docket-wide predictor of relief.

Table 6: Supplemental trait uncertainty.

Term	Post. Mean	95% CrI
Republican appointee	-0.035	[-0.845, 0.761]
Male	-0.056	[-0.748, 0.650]
Asian	0.119	[-1.442, 1.548]
Hispanic	-0.164	[-1.869, 1.542]
White	-0.062	[-1.373, 1.288]
Age (scaled)	0.155	[-0.689, 1.018]
Years on bench (scaled)	0.051	[-0.800, 0.907]
Former prosecutor	0.063	[-0.649, 0.756]
Former defender	0.115	[-1.282, 1.544]

F. Synthesis

The Seventh Circuit's criminal sentencing docket does not operate as one uniform merits stream. When viewed in the aggregate, the court appears to exhibit measurable inter-judge disparity, but much of that spread comes from pooling two distinct publication tracks. Once the docket is separated by publication status and judges are compared on a standardized baseline, the apparent spread inside each track narrows sharply. The clearest observed division in appellate outcomes is therefore not between judges' individual traits, but between the low-relief, nonprecedential Routine track and the higher-relief, precedential Doctrine track.

This separation also explains the limited role of judicial background in ordinary criminal appeals. Demographic and professional factors such as party, gender, race, profession, age, and tenure do produce some variation, but they do not organize the docket with the same force as publication track. Age and tenure show the clearest gradients, yet the pooled trait model still places those biography coefficients close to zero and within broad intervals. The directional differences seen in the figures are therefore small and unstable once case composition and sparse exposure are properly modeled. Judges operating inside the same track remain broadly similar on the standardized docket.

The implication is that published sentencing opinions remain indispensable for identifying the rules of appellate review, but they do not fully describe the ordinary environment in which sentencing appeals are resolved. Most appellants encounter a lower-visibility form of adjudication that is more compressed, less correction-oriented, and less visibly differentiated by judge-specific characteristics than published doctrine suggests. A reader who studies precedent alone will therefore capture the court's most visible lawmaking work, but only a partial picture of how sentencing appeals are ordinarily decided, and may misidentify where the clearest observed disparity appears to reside within the docket.

Conclusion

The central finding of this thesis is that apparent inter-judge disparity in Seventh Circuit sentencing appeals is best understood as a product of institutional design rather than stable judicial blocs. In plain terms, not all opinions are made equal. The three-track framework shows that Aggregate mixes two publication regimes with very different baseline relief environments. In the high-volume Routine subset, relief is rare and tightly compressed. In Doctrine, relief is much more common, but that higher baseline is associated with the types of cases that enter the published merits subset, not with durable partisan, demographic, or professional camps acting on a common docket. The point is not that ideology never matters. It is that, in this dataset, ideology is not the best explanation for variation in the ordinary merits stream, and any ideological effects are more plausibly visible in the smaller precedential subset than in the high-volume Routine docket.

That result matters because scholarship and doctrine often infer too much from published opinions alone. A court's precedential output may describe its lawmaking moments (Doctrine) accurately while still misrepresenting the ordinary appellate environment most litigants face (Aggregate). The broader contribution of this thesis is therefore both substantive and methodological: it offers a judge-identifying, track-specific account of sentencing review and a reproducible way to study that hidden docket at scale. At the same time, the design remains limited. It cannot directly identify the mechanisms by which cases are sorted into Doctrine, estimate panel-specific interaction effects, or test how raising different issues alter judicial behavior. Those questions, together with cross-circuit comparisons and longer time spans, are the next step.

Appendices

I. Expanded Methodology Section

Data acquisition and preprocessing. The primary dataset was constructed through an automated script to download the Seventh Circuit's electronic records at https://media.ca7.uscourts.gov/opinion.html. Using Python's BeautifulSoup package, the script reads an HTML page containing the search results of all opinions, without any filtering, published by the court in the six-year period between 01/01/2020 and 12/31/2025, and then performs direct downloads of all PDF documents with Case Type column = "criminal" referenced in that HTML page.[28] The script also extracts case metadata, applies optional date bounds, and writes a manifest spreadsheet (CSV file) that functions as a canonical index of the raw corpus, later reused by the deduplication scripts.[29]

However, circuit court data contain multiple versions of the same decision, including amended opinions, duplicate uploads, and separate writings that point to the same underlying PDF. The pipeline therefore performs both semantic and visual deduplication. Semantic dedupe compares normalized text, while visual dedupe compares rendered page images in a way that approximates what a reader would actually see. Each downloaded PDF is hashed with SHA-256 at the byte level, using 1 MB chunks to accommodate large files while keeping memory use stable. That byte-level hash functions both as an integrity check and as a fallback deduplication key.

Semantic deduplication converts each PDF to text using pdftotext (UTF-8), then normalizes the result to minimize spurious differences by (1) removing zero-width characters, soft hyphens, and ligatures, (2) normalizing hyphenated line breaks, (3) detecting and removing repeated header and footer lines across pages, including page numbers, and (4) lowercasing and reducing the text to

[28] Other case types are excluded from download. The initial dataset, retrieved from the Seventh Circuit's electronic repository, comprised a total of 6,011 raw records. An audit of the primary case-type classifications within this corpus revealed that "Civil" cases represented the largest plurality with 2,714 records (45.2%), followed by "Criminal" cases with 1,853 records (30.8%), and "Prisoner" petitions with 1,130 records (18.8%). The remaining volume was composed of "Agency" reviews (202 records, 3.4%), "Bankruptcy" appeals (69 records, 1.2%), "Tax" cases (31 records, 0.5%), and "Miscellaneous" or unclassified original proceedings (12 records, 0.2%).

[29] From each row, the scraper extracts and records: case number, caption, case type, filed date, document label, author (as listed by the court), PDF URL, local filename (generated deterministically with date + case number + label).

alphanumeric tokens. After normalization, the pipeline computes a semantic SHA-256 hash. The script uses a minimum normalized-text threshold of 800 characters to decide whether a document is long enough for semantic hashing. For shorter documents, it falls back to the byte-level SHA-256.

Visual deduplication renders each page to PNG at 150 DPI via pdftoppm and computes a 64-bit difference hash (dHash, 8 x 8) for each page. The page-level hashes are then concatenated and re-hashed with SHA-256 to produce a stable visual signature. In both branches of the pipeline, exact hash equality is required. The design deliberately avoids fuzzy matching so that the deduplication stage cannot create false merges.

For each dedupe group, the pipeline chooses a canonical file according to a specified policy, with earliest filed date as the default. It then writes a unique-pdfs directory containing one canonical PDF per group. Those canonical files, together with the manifest, form the input to the extraction stage. The result is a reduction from 1,853 raw criminal downloads to 1,591 unique criminal decisions. The deduped criminal corpus contains 3,784,061 words in total, with a mean of 2,378 words per decision, a median of 1,560, and a range from 107 to 26,314.

LLM-based structured extraction. The extraction schema is expressed as JSON Schema and designed to enforce strict missing-value semantics. Rather than storing a single flat value for each concept, the schema uses a status-value structure. Each field records both a status and, where appropriate, a value. In practice, that means the pipeline can distinguish information that is present from information that is not mentioned, unclear, or not applicable. That distinction matters because the thesis relies on missingness that is interpretable rather than silent.

The extraction pipeline uses the `gpt-5-mini` model with reasoning set to high through the OpenAI Responses API and requires schema-constrained JSON output.[30] The prompting architecture was designed to discipline the model into extraction rather than synthesis. A system

[30] Model choice followed a manual comparison against a gold-standard sample rather than a purely abstract preference for a larger model class. In that audit, the `gpt-5-mini-high` and `gpt-5.1-high` configurations achieved the strongest field-level agreement and materially lower omission rates on core sentencing variables than smaller alternatives, which is why the appendix describes the selected extractor as a validation-based choice rather than a convenience choice.

prompt imposed the non-inference rule, a short user prompt enforced JSON-only output and schema compliance, and docket metadata were supplied as contextual background rather than as substitute evidence. Since appellate opinions are often lengthy, unevenly organized, and selective in the facts they state, a constrained prompt structure reduces omissions and improves schema compliance in a way that is especially important for legal extraction.

The extraction itself was divided into five thematic passes covering panel composition and outcomes, issue identification, offense and sentencing variables, judicial rationale, and separate opinions. The point of that segmentation was to isolate related tasks and reduce long-context dilution, not to manufacture substantive categories after the fact. Each response was then subjected to both formal JSON-schema validation and a secondary status-value contract. If a field was marked `present`, a non-null value was required; if it was marked `not_mentioned`, the corresponding value had to be null. Invalid outputs were logged and quarantined for manual review. In manual validation against a gold-standard sample, the `gpt-5-mini-high` and `gpt-5.1-high` configurations exceeded 95 percent agreement across 330 audited fields, while smaller models produced materially higher omission rates for core sentencing variables such as `sentencing_at_issue` and `sentence_relief`.

Following extraction, the nested JSON outputs were flattened into a tabular structure. For each variable, the system stores both the substantive content and a corresponding `status` indicator, preserving the extraction state for every observation. This structure allowed for the deterministic derivation of the study's publication-based track architecture. Publication status was derived from document labels using fixed text rules that classify nonprecedential dispositions against all other precedential opinions, a step made necessary by the endogenous role of publication in circuit practice. The pipeline also derives two distinct relief outcomes: `sentence_relief`, which isolates vacaturs tied specifically to sentencing-at-issue challenges, and `decision_relief`, a broader indicator capturing any reversal, vacatur, or remand. By deriving these outcomes from structured disposition fields rather than raw opinion labels, the pipeline ensures that the dependent variables rest on consistent and reproducible legal logic.

The final analytical dataset was constructed at the individual judicial vote level, yielding a discrete observation for each judge-case interaction across the merits docket. This corpus was purposefully restricted to primary merits adjudications, specifically direct appeals and resentencings after remand, and was further refined to include only precedential opinions and nonprecedential dispositions.[31] To mitigate sparsity and ensure the statistical stability of the Bayesian estimator, categorical features for case posture and primary offense were normalized, with infrequent strata (defined by a threshold of $N < 25$) consolidated into a residual "other" category. Judges were integrated into the case-level records through a robust alias-mapping system, ensuring longitudinal consistency across the sample. The primary dependent variable, relief_vote, was operationalized as a binary indicator mapped directly from structured outcome fields. Seniority statuses such as active, senior, and visiting were not used as controls because most senior judges on the bench at the study's end switched their status during the six-year period. For example, Judge Sykes assumed senior status in 2025, making such differentiation unreliable.

Statistical modeling strategy. Circuit-court data depart materially from the assumptions that underlie standard district-court models. At the appellate level, each case is resolved by a three-judge panel, so votes are inherently interdependent; the docket is dominated by per curiam and nonprecedential dispositions with near-zero relief; and the subset of signed, precedential opinions is comparatively sparse. These features render many conventional approaches inappropriate, both because they assume independent observations and because they rely on fine-grained stratification that is infeasible in sparse appellate samples.

Given these constraints, the analysis requires a model that (i) handles binary outcomes, (ii) accommodates shared case-level dependence, (iii) permits partial pooling across judges to stabilize estimates with limited per-judge counts, and (iv) enables explicit standardization to a

[31] Within the final deduped analytical sample (N = 1,591), document-label auditing yields the following categories: Nonprecedential Disposition (894; 56.19%), Amended Nonprecedential Disposition (2; 0.13%), Final Opinion (674; 42.36%), Opinion (7; 0.44%), Amended Opinion (6; 0.38%), Opinion Denying Rehearing (1; 0.06%), Order Correcting Opinion (2; 0.13%), and Rehearing Denial Order Correcting Opinion (2; 0.13%). Three residual records carry anomalous date-like metadata labels and are therefore treated as unknown publication status. Aggregating these labels produces 896 nonprecedential records (56.32%), 692 precedential records (43.49%), and 3 unknown-status records (0.19%), which motivates the publication-based track architecture later reported as Aggregate, Routine, and Doctrine.

reference docket. A Bayesian generalized linear mixed model (GLMM) with random intercepts for both judge and case satisfies these requirements and is better suited to appellate data than several common alternatives. For example, linear probability models violate the [0,1] bounds and misstate uncertainty under heteroskedasticity; fixed-effects logistic models become unstable under sparse panels and are subject to the incidental-parameter problem at the case level; simple shrinkage or Wilson intervals ignore case difficulty and interdependence; and IRT-style models are poorly identified when each case has only three votes and the judge-case graph is thin. The chosen GLMM therefore provides a compromise between oversmoothing and overfitting by explicitly modeling latent case difficulty while partially pooling judge effects.[32]

Publication status is not merely a nuisance covariate in appellate courts; it is a judicial decision that shapes which cases become visible and which remain in the high-volume nonprecedential stream. To avoid confounding publication practices with adjudicative tendencies, the pipeline reports a three-track architecture. The "Aggregate" model uses the full merits docket to capture the bottom-line probability of relief that litigants encounter across the court's sentencing merits stream. The "Routine" model isolates the nonprecedential subset, while the "Doctrine" model restricts the sample to precedential opinions only. This design preserves the full litigant-facing universe while also separating case-management behavior from doctrinal adjudication, which is a central structural feature of appellate courts and far less pronounced in district-court data.

Formal model specification. Let y_{jc} denote the binary relief outcome for judge j in case c. The study models y_{jc} as a Bernoulli random variable, $y_{jc} \sim Bernoulli\left(p_{jc}\right)$. In implementation, y_{jc} is the case-level relief indicator carried onto each judge-case row in the panel-linked data, which is why the case random intercept is necessary to absorb the shared dependence induced by the common case outcome. The log-odds of relief are specified as a linear function of observed case-level covariates and two random intercepts:

$$\text{logit}\left(p_{jc}\right) = \alpha + X_c\beta + u_j + v_c$$

[32] The judge effect (u_j) is a latent, judge-specific deviation in the log-odds of relief after controlling for observed case characteristics and case difficulty. In other words, it captures judge-specific deviation from the overall relief rate, conditional on observed case covariates and case difficulty. It represents residual heterogeneity attributable to unobserved judge-level factors and is not interpreted as a causal or purely extralegal effect.

In this expression, α is the global intercept (baseline relief rate), X_c is the vector of observed case-level factors (posture group and offense group), β is the corresponding coefficient vector, u_j is the judge-level random intercept representing latent leniency, and v_c is the case-level random intercept representing latent case difficulty.

The random effects are modeled hierarchically. Judge effects follow $u_j \sim \mathcal{N}\left(0, \sigma_{\text{judge}}\right)$, while case effects follow $v_c \sim \mathcal{N}\left(0, \sigma_{\text{case}}\right)$. Together, these terms capture systematic between-judge heterogeneity and shared case-level difficulty. In the current implementation, weakly informative priors are used for regularization. The intercept is assigned $Normal(0,\ 1.5)$; fixed-effect coefficients are assigned $Normal(0,\ 1)$; and the standard deviations of the random effects are assigned $HalfNormal(1)$. These priors are designed to yield conservative estimates in sparse judge-level strata while still allowing the data to determine the magnitude of judge and case variability.

Standardization and docket-weighted prediction. The purpose of standardization is to produce relief rates that are comparable across judges by evaluating each judge on the same reference docket. In practice, the study computes a "standard docket" from the empirical distribution of posture and offense categories over the most recent five years (the default window). Each combination of posture and offense defines a docket cell s, and its frequency in that five-year window is used as its weight w_s. These weights sum to 1 and represent the baseline case mix against which all judges are evaluated.

Because the model uses a logit link, the study cannot simply set the case random effect to zero without biasing the expected probability. Instead, it marginalizes over the case-difficulty distribution. For judge j, the standardized relief rate is defined as

$$\hat{p}_j = \sum_s w_s \cdot E_v \left[\text{logit}^{-1}\left(\alpha + X_s\beta + u_j + v\right)\right]$$

where $v \sim \mathcal{N}\left(0, \sigma_{\text{case}}\right)$. In implementation, this expectation is approximated by Monte Carlo integration. Specifically, the study draws M standard normal variates $z_m \sim \mathcal{N}(0,1)$, scales

them by σ_{case}, and computes

$$\hat{p}_j \approx \sum_s w_s \cdot \frac{1}{M} \sum_{m=1}^{M} \text{logit}^{-1}\left(\alpha + X_s\beta + u_j + z_m\sigma_{\text{case}}\right)$$

This procedure directly addresses Jensen's inequality, since $E\left[\text{logit}^{-1}(X)\right] \neq \text{logit}^{-1}\left(E[X]\right)$ for nonlinear links.

Model diagnostics and disparity testing. Model training uses the No-U-Turn Sampler (NUTS), a Hamiltonian Monte Carlo algorithm that adaptively tunes trajectory length to approximate the posterior without random-walk behavior. The default configuration runs 4 chains, each with 1,000 warm-up iterations for step-size and mass-matrix adaptation and 2,000 retained draws, yielding 8,000 posterior samples in total. Convergence is assessed using the potential scale reduction statistic $\widehat{R} \approx 1$ and effective sample size (ESS). For each scalar parameter θ, $\widehat{R} \approx 1$ indicates that within-chain and between-chain variances are statistically indistinguishable, while ESS provides a Monte Carlo precision estimate; low ESS would indicate high autocorrelation and unreliable posterior summaries. These diagnostics are computed via ArviZ and are checked for all key parameters, including the judge-level variance σ_{judge} and the case-level variance σ_{case}.

To detect judge-level disparity beyond random fluctuation in case assignment, the pipeline implements a stratified permutation test. Let r_j be judge j's observed relief rate within a given track (Aggregate, Routine, or Doctrine), and let $\text{Var}(r_j)$ be the cross-judge variance of these rates among judges with at least twenty votes in that track. Under the null hypothesis of exchangeability, the joint distribution of relief votes is invariant to permutation within strata once case mix is fixed. The study therefore defines strata by posture group and offense group and repeatedly permutes relief votes within each stratum, recomputing the judge-level variance after each shuffle. If $V^{(b)}$ denotes the variance from permutation b, the empirical p-value is computed as

$$p = \frac{1}{B} \sum_{b=1}^{B} I\left[V^{(b)} \geq V^{\text{obs}}\right]$$

with $B = 1000$ by default. This produces a nonparametric test of whether observed judge dispersion exceeds the dispersion expected under the null of exchangeable voting behavior within each case-type stratum. The test is executed separately for the Aggregate, Routine, and Doctrine tracks to determine whether disparity is concentrated in the full docket or in one of the

two publication regimes.

Key assumptions and justifications. The modeling strategy relies on several explicit assumptions that align with the hierarchical logit structure. First, conditional independence is assumed given the linear predictor

$$\eta_{jc} = \alpha + X_c\beta + u_j + v_c$$

the votes y_{jc} are independent across judges and cases. This is the standard exchangeability assumption for hierarchical generalized linear models, and it implies that residual dependence is absorbed by the random intercepts. Second, the judge-level and case-level random effects are assumed to be Gaussian:

$$u_j \sim \mathcal{N}\left(0, \sigma_{\text{judge}}\right)$$
$$v_c \sim \mathcal{N}\left(0, \sigma_{\text{case}}\right)$$

This normality assumption provides analytic regularization and supports shrinkage, ensuring stable estimation under sparse per-judge data. Third, the status-value design makes missingness explicit rather than silent. In the primary GLMM, the modeled outcome is `relief_vote` built from `decision_relief`, a variable with only 1.5 percent missingness, while the grouped posture and offense covariates are carried forward through deterministic preprocessing. Broader extraction sparsity in other fields is therefore treated as a measurement limitation rather than assumed away by the fitted model. Fourth, the standard docket is assumed to be stationary over the chosen five-year window, so that docket weights w_s approximate a stable reference distribution rather than a transient fluctuation. While strict stability is unlikely given exogenous shocks such as COVID-19, the five-year rolling window provides a smoothed institutional average that is more robust than single-year snapshots. Fifth, explicit panel-composition effects are not modeled; instead, panel-level interaction effects are absorbed into the case-level intercept and the remaining uncertainty. This approximation is justified by the fact that each case is heard by exactly one panel, so a separate panel random effect would be weakly identified in a sample of this size.

The complete, versioned pipeline and analysis code are available in the public repository: https://github.com/richardzhux/sentencing/tree/main. The pipeline is implemented in Python and

draws on a standard scientific-computing stack. Data handling and transformation are performed with pandas and NumPy. Bayesian modeling is implemented through Bambi (a high-level formula interface) backed by PyMC, and posterior diagnostics are computed with ArviZ and xarray. Data acquisition uses requests and BeautifulSoup for HTTP and HTML parsing. PDF deduplication relies on pdftotext and pdftoppm for text and image rendering, with Pillow for image processing. Visualization is generated with Plotly for the interactive dashboard, with Matplotlib used for static exports.

II. Representative Status-Value Schema Conversions

Table 7 illustrates how the thesis's status-value schema converts extracted fields into analytically distinct categories for downstream coding.

Table 7: Representative status-value schema conversions.

Field	Original	Status-Value Form
panel_judges	array[string]	object{status,value[array]}
authoring_judge	string|null	object{status,value[string]}
posture	string|null(enum)	object{status,value[enum]}
plea_or_trial	string|null(enum)	object{status,value[enum]}
appeal_waiver_enforced	string|null(enum)	object{status,value[enum]}
issues_raised	array	object{status,value[array]}
outcomes.overall_disposition	string|null(enum)	object{status,value[enum]}
outcomes.sentence_disposition	string|null(enum)	object{status,value[enum]}
outcomes.conviction_disposition	string|null(enum)	object{status,value[enum]}
outcomes.remand_scope	string|null(enum)	object{status,value[enum]}
offense.offense_category	string|null(enum)	object{status,value[enum]}
sentencing.sentence_type	string|null(enum)	object{status,value[enum]}

Continued on next page

Table 7: Representative status-value schema conversions (continued).

Field	Original	Status-Value Form
sentencing.variance_or_departure.direction	string|null(enum)	object{status,value[enum]}
sentencing.variance_or_departure.type	string|null(enum)	object{status,value[enum]}

III. FY2020-2024 CA7 Offense Category and District Composition Data Check

This comparison functions as an external plausibility check, not as a claim of one-to-one equivalence. The LLM corpus is organized at the appellate-decision level, so consolidated appeals and multi-defendant decisions do not map perfectly onto the Sentencing Commission's defendant-level reporting. The time windows are also imperfectly aligned, where fiscal year 2020 in the USSC data begins on October 1, 2019, whereas this corpus begins on January 1, 2020. And some offense labels require translation across different coding conventions, such as treating robbery within a broader violent-crime bucket or combining child-pornography and child-exploitation categories.

Table 8: USSC-LLM offense and district composition check.

Metric	USSC Dataset	LLM Extraction Finding
Drug trafficking	35.8%	40.7%
Firearms	24.7%	24.1%
Fraud, theft, embezzlement	10.8%	11.0%
Robbery	5.5%	6.6% (violent-crime aggregate)
Sexual abuse	4.6%	4.0%
Child pornography	4.0%	10.7% (child-sex-crime aggregate)
Immigration	3.5%	1.3%

Continued on next page

Table 8: USSC-LLM offense and district composition check (continued).

Metric	USSC Dataset	LLM Extraction Finding
N.D. Ill.	2,948 (26.2%)	370 (28.0%)
S.D. Ind.	2,210 (19.6%)	191 (14.4%)
N.D. Ind.	1,483 (13.2%)	158 (12.0%)
E.D. Wis.	1,479 (13.1%)	85 (6.4%)
C.D. Ill.	1,246 (11.1%)	243 (18.4%)
S.D. Ill.	1,201 (10.7%)	168 (12.7%)
W.D. Wis.	689 (6.1%)	107 (8.1%)

IV. Audit of Non-Present Variables in LLM Extraction

The missingness audit is intended to identify which variables are structurally sparse in the source material and which variables were consistently recoverable across the corpus; it does not imply that every non-present field reflects extraction failure. In appellate opinions, many concepts are genuinely absent because the issue never arose, because the variable was inapplicable to the case, or because a short nonprecedential disposition had no reason to discuss it. Table 9 therefore groups nulls, empty strings, empty lists, and explicit unknown or not-applicable markers into a single non-present column. The counts below are ordered from highest to lowest non-present rate. Variables near the top of the table tend to correspond to concurrences and dissents or narrow sentencing subissues. Core docket descriptors and principal outcome variables cluster much lower.

Table 9: Field-level non-present audit.

No.	Field	Non-Present	Rate (%)	Non-Missing
1	pipeline_meta.prompt_cache_retention	1591	100.0	0
2	merge_reasons	1588	99.8	3

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
3	concurring_judges_meta	1557	97.9	34
4	dissenting_judges_meta	1557	97.9	34
5	concurring_judges	1556	97.8	35
6	dissenting_judges	1553	97.6	38
7	separate_opinion_source	1531	96.2	60
8	separate_opinions	1531	96.2	60
9	sentencing.variance_or_departure.type	1407	88.4	184
10	appeal_waiver_enforced	1306	82.1	285
11	sentencing.mandatory_minimum_months	1257	79.0	334
12	sentencing.variance_or_departure.direction	1046	65.7	545
13	authoring_judge_source	964	60.6	627
14	authoring_judge_final	964	60.6	627
15	authoring_judge_meta	960	60.3	631
16	authoring_judge_norm	934	58.7	657
17	authoring_judge_meta_candidates	933	58.6	658
18	sentencing.district_court_guidelines_range_high	912	57.3	679
19	sentencing.district_court_guidelines_range_low	884	55.6	707
20	district_division	757	47.6	834
21	sentence_relief	735	46.2	856
22	offense_category_statute_primary	532	33.4	1059
23	sentencing.rationale.themes	447	28.1	1144
24	offense_categories_primary_statutes	445	28.0	1146
25	sentencing_at_issue	437	27.5	1154
26	sentencing.sentence_at_issue	437	27.5	1154

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
27	offense.offense_category	400	25.1	1191
28	offense_category	400	25.1	1191
29	outcomes.conviction_disposition	311	19.6	1280
30	conviction_disposition	311	19.6	1280
31	sentencing.rationale.notes	305	19.2	1286
32	offense_categories_statutes	300	18.9	1291
33	sentencing.rationale.evidence	300	18.9	1291
34	offense.primary_statutes	269	16.9	1322
35	offense_category_final	245	15.4	1346
36	offense_category_source	245	15.4	1346
37	standard_of_review	231	14.5	1360
38	outcomes.sentence_disposition	220	13.8	1371
39	sentence_disposition	220	13.8	1371
40	sentencing.imprisonment_months	218	13.7	1373
41	plea_or_trial	197	12.4	1394
42	sentencing.sentence_type	158	9.9	1433
43	offense_categories_final	130	8.2	1461
44	offense.statutes	124	7.8	1467
45	extraction_summary	64	4.0	1527
46	decision_relief	24	1.5	1567
47	remand_scope	20	1.3	1571
48	outcomes.remand_scope	20	1.3	1571
49	outcomes.overall_disposition	18	1.1	1573
50	overall_disposition	18	1.1	1573

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
51	panel_key_norm	10	0.6	1581
52	panel_judges_norm	10	0.6	1581
53	panel_size	10	0.6	1581
54	panel_key	10	0.6	1581
55	panel_judges	10	0.6	1581
56	issues_raised	5	0.3	1586
57	publication_status	3	0.2	1588
58	posture	2	0.1	1589
59	district_canonical	2	0.1	1589
60	district	2	0.1	1589
61	district_judge_norm	2	0.1	1589
62	district_judge	2	0.1	1589
63	district_code	2	0.1	1589
64	sentencing.imprisonment_months_status	0	0.0	1591
65	outcomes.sentence_disposition_status	0	0.0	1591
66	sentencing.sentence_type_status	0	0.0	1591
67	outcomes.remand_scope_status	0	0.0	1591
68	sentencing.sentence_at_issue_status	0	0.0	1591
69	outcomes.conviction_disposition_status	0	0.0	1591
70	offense.offense_category_status	0	0.0	1591
71	offense.primary_statutes_status	0	0.0	1591
72	offense.statutes_status	0	0.0	1591
73	sentencing.district_court_guidelines_range_low_status	0	0.0	1591
74	pipeline_meta.run_id	0	0.0	1591

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
75	sentencing.district_court_guidelines_range_high_status	0	0.0	1591
76	pipeline_meta.pass_ids	0	0.0	1591
77	pipeline_meta.model	0	0.0	1591
78	pipeline_meta.prompt_cache_key	0	0.0	1591
79	pipeline_meta.prompt_hash	0	0.0	1591
80	pipeline_meta.prompt_version	0	0.0	1591
81	pipeline_meta.schema_hash	0	0.0	1591
82	pipeline_meta.schema_version	0	0.0	1591
83	extraction_summary_status	0	0.0	1591
84	outcomes.overall_disposition_status	0	0.0	1591
85	pipeline_meta.created_at	0	0.0	1591
86	pipeline_meta.reasoning_effort	0	0.0	1591
87	issue_count	0	0.0	1591
88	sentencing.rationale.evidence_status	0	0.0	1591
89	sentencing.rationale.notes_status	0	0.0	1591
90	sentencing.rationale.themes_status	0	0.0	1591
91	pipeline_meta.multi_pass	0	0.0	1591
92	sentencing.mandatory_minimum_months_status	0	0.0	1591
93	pipeline_meta.pass_specs_path	0	0.0	1591
94	sentencing.variance_or_departure.type_status	0	0.0	1591
95	source_path	0	0.0	1591
96	sentencing.variance_or_departure.direction_status	0	0.0	1591
97	separate_opinions_status	0	0.0	1591
98	decision_id	0	0.0	1591

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
99	district_judge_status	0	0.0	1591
100	concurrence_present	0	0.0	1591
101	visual_hash	0	0.0	1591
102	dedupe_key	0	0.0	1591
103	merged_visual_hashes	0	0.0	1591
104	merged_dedupe_keys	0	0.0	1591
105	canonical_sha256	0	0.0	1591
106	merged_from_groups	0	0.0	1591
107	statute_unparsed_count	0	0.0	1591
108	statute_parsed_count	0	0.0	1591
109	primary_statute_count	0	0.0	1591
110	statute_count	0	0.0	1591
111	offense_category_conflict	0	0.0	1591
112	concurrence_count	0	0.0	1591
113	dissent_count	0	0.0	1591
114	district_status	0	0.0	1591
115	dissent_present	0	0.0	1591
116	separate_opinion_present	0	0.0	1591
117	authorship_type	0	0.0	1591
118	per_curiam_meta	0	0.0	1591
119	doc_label_key	0	0.0	1591
120	doc_label_group	0	0.0	1591
121	doc_label_primary	0	0.0	1591
122	doc_labels	0	0.0	1591

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
123	decision_year	0	0.0	1591
124	decision_dates_list	0	0.0	1591
125	decision_date	0	0.0	1591
126	canonical_filename	0	0.0	1591
127	page_count	0	0.0	1591
128	text_chars	0	0.0	1591
129	word_count	0	0.0	1591
130	filesize	0	0.0	1591
131	issues_raised_status	0	0.0	1591
132	appeal_waiver_enforced_status	0	0.0	1591
133	plea_or_trial_status	0	0.0	1591
134	posture_status	0	0.0	1591
135	per_curiam	0	0.0	1591
136	per_curiam_status	0	0.0	1591
137	authoring_judge	0	0.0	1591
138	authoring_judge_status	0	0.0	1591
139	panel_judges_status	0	0.0	1591
140	source_filenames_list	0	0.0	1591
141	authors_list	0	0.0	1591
142	doc_labels_list	0	0.0	1591
143	filed_dates_list	0	0.0	1591
144	captions_list	0	0.0	1591
145	case_nos_list	0	0.0	1591
146	group_id	0	0.0	1591

Continued on next page

Table 9: Field-level non-present audit (continued).

No.	Field	Non-Present	Rate (%)	Non-Missing
147	doc_urls	0	0.0	1591
148	authors	0	0.0	1591
149	filed_dates	0	0.0	1591
150	captions	0	0.0	1591
151	case_nos	0	0.0	1591
152	case_types	0	0.0	1591
153	case_count	0	0.0	1591
154	source_row_count	0	0.0	1591
155	canonical_row_index	0	0.0	1591
156	source_filenames	0	0.0	1591

V. Sample Construction and Model Diagnostics

This section consolidates the denominator transitions and model-audit outputs in one place. Table 10 records the sequence from the 6,011-case scrape to the final 1,375-decision Aggregate universe, with Routine and Doctrine as the two disjoint publication tracks within that training set. Its `Prior (%)` column measures retention relative to the immediately preceding stage, and its `Criminal (%)` column measures retention relative to the 1,853 criminal downloads. Table 11 reports observed relief, the posterior predictive mean, the central 95 percent posterior predictive interval, the disparity ratio, and the stratified permutation p-value. Table 12 reports maximum R-hat, minimum bulk and tail effective sample sizes, and post-warmup Hamiltonian divergences.

Table 10: Analytical sample construction audit.

Stage	Decisions	Vote Rows	Prior (%)	Criminal (%)
Full scrape (all case types)	6,011	–	–	–
Criminal raw downloads	1,853	–	30.8	100.0
Deduped criminal decisions	1,591	–	85.9	85.9
Merits-eligible sentencing pool	1,401	–	88.1	75.6
Panel-linked merits decisions	1,394	4,213	99.5	75.2
Aggregate training universe	1,375	4,157	98.6	74.2
Routine track	728	2,180	52.9	39.3
Doctrine track	647	1,977	47.1	34.9

Table 11: Track-level fit and disparity audit.

Track	Dec.	Rows	Obs. (%)	PP Mean (%)	95% PP Int.	Ratio	p
Aggregate	1,375	4,157	11.23	11.34	[10.85, 11.84]	2.84x	<0.001
Routine	728	2,180	3.44	3.61	[3.03, 4.22]	1.23x	0.212
Doctrine	647	1,977	19.83	19.98	[19.12, 20.89]	1.35x	0.172

Table 12: Sampler convergence audit.

Track	Max R-hat	Min Bulk ESS	Min Tail ESS	Div.
Aggregate	1.000	3928	3649	0
Routine	1.000	4167	3470	0
Doctrine	1.000	5846	3710	0

Bibliography

Primary Sources

Administrative Office of the U.S. Courts. "Judicial Business of the United States Courts." 2025. https://www.uscourts.gov/statistics-reports/analysis-reports/judicial-business-united-states-courts.

Anastasoff v. United States, 223 F.3d 898 (8th Cir. 2000), vacated as moot, 235 F.3d 1054 (8th Cir. 2000). 2000.

Federal Judicial Center. "Biographical Directory of Article III Federal Judges, 1789–Present." 2025. https://www.fjc.gov/history/judges.

Federal Rules of Appellate Procedure. "Rule 32.1. Citing Judicial Dispositions." 2006. https://www.law.cornell.edu/rules/frap/rule_32.1.

Gall v. United States, 552 U.S. 38 (2007). 2007.

Hart v. Massanari, 266 F.3d 1155 (9th Cir. 2001). 2001.

Kimbrough v. United States, 552 U.S. 85 (2007). 2007.

Plumley v. Austin, 574 U.S. 1127 (2015). 2015.

Rita v. United States, 551 U.S. 338 (2007). 2007.

United States Court of Appeals for the Seventh Circuit. "Federal Rules of Appellate Procedure and Circuit Rules of the United States Court of Appeals for the Seventh Circuit." 2024.

United States Court of Appeals for the Seventh Circuit. "Opinions, Nonprecedential Dispositive Orders and Oral Arguments." 2025. https://media.ca7.uscourts.gov/.

United States Court of Appeals for the Seventh Circuit. "Practitioner's Handbook for Appeals to the United States Court of Appeals for the Seventh Circuit." 2020. https://www.ca7.uscourts.gov/rules-procedures/Handbook.pdf.

United States Sentencing Commission. *2023 Demographic Differences in Federal Sentencing*. U.S. Sentencing Commission, 2023. https://www.ussc.gov/research/research-reports/2023-demographic-differences-federal-sentencing.

United States Sentencing Commission. *The Influence of the Guidelines on Federal Sentencing: Federal Sentencing Outcomes, 2005–2017*. U.S. Sentencing Commission, 2020.

United States v. Booker, 543 U.S. 220 (2005). 2005.

Secondary Sources

Brown, Rachel, Jade Ford, Sahrula Kubie, Katrin Marquez, Bennett Ostdiek, and Abbe R. Gluck. "Is Unpublished Unequal? An Empirical Examination of the 87% Nonpublication Rate in Federal Appeals." *Cornell Law Review* 107 (2022): 1–150.

Chilton, Adam S., and Marin K. Levy. "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals." *Cornell Law Review* 101 (2015): 1–56.

Cohen, Alma. "Pervasive Influence of Political Composition on Circuit Court Decisions." *Journal of Legal Analysis* 17, no. 1 (2025): 14–41.

Cohen, Alma, and Rajeev H. Dehejia. *Judges Judging Judges: Partisanship and Politics in the Federal Circuit Courts of Appeals*. Working Paper No. 32920. National Bureau of Economic Research, 2024. https://doi.org/10.3386/w32920.

Cohen, Alma, and Crystal S. Yang. "Judicial Politics and Sentencing Decisions." *American Economic Journal: Economic Policy* 11, no. 1 (2019): 160–91.

Cross, Frank B. *Decision Making in the U.S. Courts of Appeals*. Stanford University Press, 2007.

Cross, Frank B., and Emerson H. Tiller. "Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals." *Yale Law Journal* 107, no. 7 (1998): 2155–76.

Crow, Matthew S., and Natalie Goulette. "Judicial Diversity and Sentencing Disparity Across U.S. District Courts." *Journal of Criminal Justice* 82 (2022): 101973.

Crow, Matthew S., and Natalie Goulette. "Sex, Politics, and U.S. District Court Outcomes: Examining Variation in Judge-Initiated Downward Guideline Departures." *American Journal of Criminal Justice* 48, no. 2 (2023): 295–318. https://doi.org/10.1007/s12103-021-09648-3.

Crow, Matthew S., and Keith A. Johnson. "Race, Ethnicity, and Habitual-Offender Sentencing: A Multilevel Analysis of Individual and Contextual Threat." *Criminal Justice Policy Review* 19,

no. 1 (2008): 63–83.

Farrell, Amy, Geoff Ward, and Danielle Rousseau. "Intersections of Gender and Race in Federal Sentencing: Examining Court Contexts and the Effects of Representative Court Authorities." *The Journal of Gender, Race & Justice* 14, no. 1 (2010): 85–126.

Finklea, Kristin, and Lisa N. Sacco. *Cocaine: Crack and Powder Sentencing Disparities*. CRS In Focus No. IF11965. Congressional Research Service, 2021. https://www.congress.gov/crs-product/IF11965.

Goldrosen, Nicholas, Christian Michael Smith, Maria-Veronica Ciocanel, et al. "Racial Disparities in Criminal Sentencing Vary Considerably Across Federal Judges." *Journal of Institutional and Theoretical Economics* 179, no. 1 (2023): 92–113.

Haire, Susan B., Laura P. Moyer, and Shawn Treier. "Diversity, Deliberation, and Judicial Opinion Writing." *Journal of Law and Courts* 1, no. 2 (2013): 303–30. https://doi.org/10.1086/670724.

Hartley, Richard D., and Robert Tillyer. "Inter-District Variation and Disparities in Federal Sentencing Outcomes: Case Types, Defendant Characteristics, and Judicial Demography." *Criminology, Criminal Justice, Law & Society* 20, no. 3 (2019): 46–63.

Hinkle, Rachael K. "Panel Effects and Opinion Crafting in the U.S. Courts of Appeals." *Journal of Law and Courts* 5, no. 2 (2017): 313–36.

Hinkle, Rachael K. *Selective Publication in the U.S. Courts of Appeals: The Invisible Norm That Perpetuates Inequality*. Oxford University Press, 2024.

Hofer, Paul J., Kevin R. Blackwell, and R. Barry Ruback. "Effect of the Federal Sentencing Guidelines on Interjudge Sentencing Disparity." *Journal of Criminal Law and Criminology* 90, no. 1 (1999): 239–322.

Howard, J. Woodford, Jr. *Courts of Appeals in the Federal Judicial System: A Study of the Second, Fifth, and District of Columbia Circuits*. Princeton University Press, 1981.

Kastellec, Jonathan P. "Racial Diversity and Judicial Influence on Appellate Courts." *American Journal of Political Science* 57, no. 1 (2013): 167–83.

Kim, Pauline T. "Deliberation and Strategy on the United States Courts of Appeals: An Empirical

Exploration of Panel Effects." *University of Pennsylvania Law Review* 157, no. 5 (2009): 1319–81.

Levy, Marin K. "Panel Assignment in the Federal Courts of Appeals." *Cornell Law Review* 103, no. 1 (2017): 65–116.

Mason, Caleb, and David Bjerk. "Inter-Judge Sentencing Disparity on the Federal Bench: An Examination of Drug Smuggling Cases in the Southern District of California." *Federal Sentencing Reporter* 25, no. 3 (2013): 190–96.

McAlister, Merritt E. "Bottom-Rung Appeals." *Fordham Law Review* 91 (2023): 1355–422.

McAlister, Merritt E. "Downright Indifference." *Michigan Law Review* 118, no. 4 (2020): 533–610.

McAlister, Merritt E. "Rebuilding the Federal Circuit Courts." *Northwestern University Law Review* 116 (2022): 1137–226.

Merritt, Deborah Jones, and James J. Brudney. "Stalking Secret Law: What Predicts Publication in the United States Courts of Appeals." *Vanderbilt Law Review* 54, no. 1 (2001): 71–121.

Newman, Jon O., and Marin K. Levy. *Written and Unwritten: The Rules, Internal Procedures, and Customs of the United States Courts of Appeals*. Cambridge University Press, 2024.

Peresie, Jennifer L. "Female Judges Matter: Gender and Collegial Decisionmaking in the Federal Appellate Courts." *Yale Law Journal* 114, no. 7 (2005): 1759–90.

Posner, Richard A. *The Federal Courts: Challenge and Reform*. Harvard University Press, 1996.

Schoenholtz, Andrew I. "Refugee Roulette: Disparities in Asylum Adjudication." *Stanford Law Review* 60, no. 2 (2007): 295–412.

Scott, Ryan W. "Inter-Judge Sentencing Disparity After Booker: A First Look." *Stanford Law Review* 63, no. 1 (2010): 1–66.

Shaughnessy, Joan M. "Commentary: Unpublication and the Judicial Concept of Audience." *Washington and Lee Law Review* 62 (2005): 1597.

Songer, Donald R. "Criteria for Publication of Opinions in the U.S. Courts of Appeals: Formal Rules Versus Empirical Reality." *Judicature* 73 (1990): 307–13.

Spohn, Cassia. "The Effects of the Offender's Race, Ethnicity, and Sex on Federal Sentencing Outcomes in the Guidelines Era." *Law and Contemporary Problems* 76, no. 1 (2013): 75–104.

Spohn, Cassia, and Patricia K. Brennan. "The Joint Effects of Offender Race/Ethnicity and Gender on Substantial Assistance Departures in Federal Courts." *Race and Justice* 1, no. 1 (2011): 49–78.

Spohn, Cassia, and Jerry Cederblom. "Race and Disparities in Sentencing: A Test of the Liberation Hypothesis." *Justice Quarterly* 8, no. 3 (1991): 305–27.

Spohn, Cassia, and Lisa L. Sample. "The Dangerous Drug Offender in Federal Court: Intersections of Race, Ethnicity, and Culpability." *Crime & Delinquency* 59, no. 1 (2013): 3–31.

Steffensmeier, Darrell, and Stephen Demuth. "Ethnicity and Judges Sentencing Decisions: Hispanic-Black-White Comparisons." *Criminology* 39 (2001): 145–78.

Steffensmeier, Darrell, Jeffery Ulmer, and John Kramer. "The Interaction of Race, Gender, and Age in Criminal Sentencing: The Punishment Cost of Being Young, Black, and Male." *Criminology* 36, no. 4 (1998): 763–98.

Stroud, Donna S. "The Bottom of the Iceberg: Unpublished Opinions." *Campbell Law Review* 37 (2015): 333–85.

Sunstein, Cass R., David Schkade, and Lisa Michelle Ellman. "Ideological Voting on Federal Courts of Appeals: A Preliminary Investigation." *Virginia Law Review* 90, no. 1 (2004): 301–54.

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

Varsava, Nina. "Opinion Authorship and Precedential Status." *Washington University Law Review* 101 (2024): 1593–674.

Yang, Crystal S. "Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker." *New York University Law Review* 89, no. 4 (2014): 1268–342.