Journal of Law and Courts (2026), 1–22
doi:10.1017/jlc.2025.10012

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Measuring Jurisprudence

Rosamond Thalken and Edward H. Stiglitz

Cornell University, Ithaca, USA
Corresponding author: Rosamond Thalken; Email: ret85@cornell.edu

(Received 18 April 2025; Revised 16 December 2025; Accepted 19 December 2025)

Abstract

Legal jurisprudence is widely debated but rarely measured. We present the first comprehensive measure of jurisprudence in U.S. Supreme Court opinions from 1870 to 2024. Building on qualitative studies of legal reasoning, we classify court opinions into two contrasting types: "formal" reasoning and anti-formal or "grand" reasoning. The foundation of this measurement dataset is a smaller, hand-annotated dataset created by a team of domain experts. Using this annotated dataset, we fine-tune and evaluate a foundational large language model, which is then employed to predict legal reasoning across all opinions in the full dataset. We demonstrate the potential of this new measure for applications in empirical research, enabling analyses of shifts in jurisprudence over time, the reasoning styles of individual justices, and the relationship between legal reasoning and other judicial features, such as ideology. To support further research, we release the annotated dataset, the fine-tuned model, and the final measures, offering a resource for both studying legal reasoning and judicial behavior and evaluating language models in the legal domain.

Keywords: jurisprudence; legal reasoning; measurement; formalism

Introduction

Jurisprudence provides judges with a theory for what the law is, what it ought to be, and how generally to apply it. Debates over legal jurisprudence fill the volumes of law review articles, play out in the front pages of newspapers, and intersect with social science efforts to pull apart "law" from "politics." Jurisprudence can be sliced in many ways, but perhaps the most prominent cut is between two contrasting schools of thought: formalism, which sees the law as a closed, logical, and objective system for producing legal determinations, and anti-formalism, which sees law as an on-going, innovative, and consequence and workability-oriented system for producing legal determinations that is open to social and economic considerations. Qualitative scholars have examined the sweeping patterns of jurisprudence over time, and full-throated debates center on the merits, demerits, and implications of jurisprudential sub-species – for example, for formalism, textualism, and originalism.

© The Author(s), 2026. Published by Cambridge University Press on behalf of the Law and Courts Organized Section of the American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Yet no quantitative measure of jurisprudence exists. This absence hampers the systematic study of law, from fundamental descriptive questions about the sweep of jurisprudence, to thornier causal questions of whether jurisprudence matters for case outcomes or for society more broadly. In this work, we describe our process for creating a novel quantitative measure of jurisprudence that covers all United States Supreme Court opinions issued between 1870 and 2024. To create this measure, we leverage fine-tuned large language models (LLMs) and legal domain expertise. Although scholars have long used computational methods in the legal domain, jurisprudence is too abstract and complex to be studied using dictionary or more conventional machine learning approaches, such as support vector machines. LLMs offer the necessary algorithmic dexterity to identify abstract concepts such as jurisprudence (Thalken et al. 2023). By assessing language in context, LLMs can identify complex concepts with performance that was impossible until recently (Devlin et al. 2019; Vaswani 2017).

To develop this measure, we started by training a team of domain expert annotators, four upper-year law students at a leading law school, to identify legal reasoning in historical U.S. Supreme Court opinions. Once this team of annotators reached consistent and reasonably high levels of agreement, they labeled legal reasoning in nearly three thousand Court opinion paragraphs, creating the expert-annotated dataset. We then used this annotated dataset to fine-tune and evaluate the ability of a series of large language models to identify legal reasoning, and selected the best-performing model to make predictions across our full dataset, which contains nearly 25,000 cases. These predictions represent our final measure.

This comprehensive measure allows scholars to probe substantive legal and political questions that are otherwise inaccessible. For example, this measure makes it possible to examine systematically the sweeping changes in jurisprudence since Reconstruction (Stiglitz and Thalken 2024). The dataset can also be used to study changes of jurisprudence *within* justices (Stiglitz and Thalken 2026). Do they evolve over time or remain fixed in their commitments? If they evolve, why? Such questions begin to strike at the issue of how we ought to think of jurisprudence and its function in our political-legal system. To take a related open and unresolved question, what is the relationship between jurisprudence and ideology? Should we regard jurisprudential commitments as merely justificatory frameworks for ideological positions? Or is the relationship between these concepts more complicated? Likewise, does jurisprudence exert independent causal effect on case outcomes (or social outcomes more broadly)? Our hope is that scholars will use the measure and related datasets to deepen our knowledge of these and other areas.

Elsewhere, we have begun to apply this measure to some of the substantive questions noted earlier (Stiglitz and Thalken 2024, 2026). Our aim in this paper is to detail strategies underlying the measure of jurisprudence, to make the final measure available to the research community, and to release our final, fine-tuned model and intermediate human-annotated dataset. That intermediate dataset required hundreds of hours of labor to create, as described below, and will be of interest both to social scientists seeking to develop even more advanced measures of jurisprudence and to natural language processing researchers seeking to study the performance of models in a complex domain unlikely to be tainted by model pre-training.

The objective of this paper is to detail our measurement strategy and to release resources to the research community. Through this paper, we introduce the following resources: 1) the measure of jurisprudence in Supreme Court opinions, from 1870 to 2024, 2) our best-performing model with updated weights, and 3) the annotated

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Table 1. Codebook Definition for Each Class of Legal Reasoning

Class	Definition
Formal	A legal decision made according to a rule, often viewing the law as a closed and mechanical system. It screens the decision-maker off from the political, social, and economic choices involved in the decision.
Grand	A legal decision that views the law as an open-ended and on-going enterprise for the production and improvement of decisions that make sense on their face and in light of political, social, and economic factors.

dataset. Using our new measure, empirical legal scholars may continue probing the evolution of legal reasoning and its relationship with other factors in the legal system. Using the model, researchers may generate new measures for years or jurisdictions not covered by our measure, such as those issued before or after our dataset begins and ends or issued by a lower court. And finally, with the annotated dataset, researchers may improve quantitative performance in this task, using different or newer models and innovative training or prompting procedures.

Legal reasoning

Our study focuses on formalist versus anti-formalist legal reasoning. These modes of legal reasoning, called macro-jurisprudence elsewhere, describe the general approach of a judge to law and its interpretation (Stiglitz and Thalken 2024).[1] Macro-jurisprudence is evident in the written opinion of a judicial decision, where the judge explains how they reached their holdings. For this reason, court opinions serve as valuable records of legal reasoning.

Table 1 defines the two categories of legal reasoning we examine: formalism and anti-formalism, the latter of which we refer to using Llewellyn's term of *grand* reasoning (Llewellyn 1960). When using formal reasoning, a judge regards the law as a closed, logical, and objective device for producing legal determinations. Prominent examples of formalism include textualism and originalism. At least in outward aspiration, textualism reaches for the formalist ideal by carefully parsing public, authoritative legal documents and using third-party dictionaries to uncover meaning (Pierce 1995); originalism does the same by finding meaning in the public understanding of authoritative legal documents when they were passed, which in principle might be assessed empirically through contemporary commentary or other documents (Scalia et al. 1997; Kay 2009). Distinctive characteristics of formalism include a deductive, almost "mechanical" (Pound 1908), method of reasoning, with a closed system of what counts as evidence of meaning, requiring the judge to exclude external factors in decision-making (Schauer 1988; Wiecek 2001). Legal meaning tends to be fixed and resistant to change under formalism.

In contrast, anti-formalism or "grand" reasoning views the law as evolving across time and responsive to the social, political, and economic needs of the historical

[1]The prefix "macro" is adopted self-consciously: these scholars believe these categories reflect the most prominent cleavage in the literature, but jurisprudence is a high-dimensional concept and this measure cannot track all the possibly relevant lines. In this respect, there is an analogy to ideology, which is likewise high-dimensional, but still can be usefully discussed in terms of a left-right spectrum. Part of our hope in releasing these data is that future scholars will develop more textured measures of jurisprudence.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

moment (Llewellyn 1960; Horwitz 1992). The law is continuously created and improved, with an emphasis on workability and common sense, and an openness to the consequences of legal decisions. Examples of this form of legal reasoning include open-ended balancing tests and holistic assessments. For example, the rule of reason in anti-trust law requires judges to holistically balance the pros and cons of allegedly anti-competitive behavior. Living constitutionalism, likewise, fits in this class of jurisprudence (D. A. Strauss 2010).

Measurement strategy and procedure

We use a combination of legal data sources to create our comprehensive measure that is based on all opinions issued by the U.S. Supreme Court between 1870 and 2024. Our basic strategy involved domain experts curating a dataset of legal reasoning examples and scaling the insights from that dataset to the universe of Supreme Court opinions using a fine-tuned version of recent transformer-based language models. Without these recent language models, this measure would not be possible, but equally indispensable, and in many ways more challenging, is the integrity of the human-annotated dataset, and accordingly we devote attention both to the model and annotation protocols.

Raw opinion data

First, we collect all United States Supreme Court opinions available in the Harvard Case Law Access Project.[2] We limit this historical court data to opinions issued in 1870 at the earliest, but this dataset only extends until 2014.[3] Court opinions and metadata from 2014 to 2024 are collected through the Supreme Court's official website.[4] Then, using the Supreme Court Database's list of cases (Spaeth et al. 2024),[5] we identify cases that are missing from these sources. We scrape missing cases from Justia using either U.S. Reporter citation or docket identifier.[6][7] The combined dataset includes 24,890 unique cases and 36,430 opinions.

Over time, the number of cases in our dataset decreases, especially since the 1990s, as shown in Figure 1. The number of opinions issued each year has also gradually declined, with its peak in the 1980s, as shown in Figure 2.

Creating the annotated dataset

After collecting the raw opinion data, we created a smaller dataset for hand annotation by sampling from the full dataset. First, we filter to opinions that likely perform

[2]https://case.law/

[3]The dataset thus starts roughly from Reconstruction, a common delineator in historical studies. As scholars note elsewhere, the ideas of formalism and anti-formalism extend to the early Republic, with most observers arguing that jurisprudence in the early Republic was inventive and anti-formalist (Stiglitz and Thalken 2024). However, we have not studied the period before Reconstruction quantitatively, and our hope is that publishing our data and model will facilitate others to study that period.

[4]https://www.supremecourt.gov/xmls/

[5]http://scdb.wustl.edu/

[6]https://www.justia.com/

[7]In total, there are 97 cases that we are not able to identify in any of these sources.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

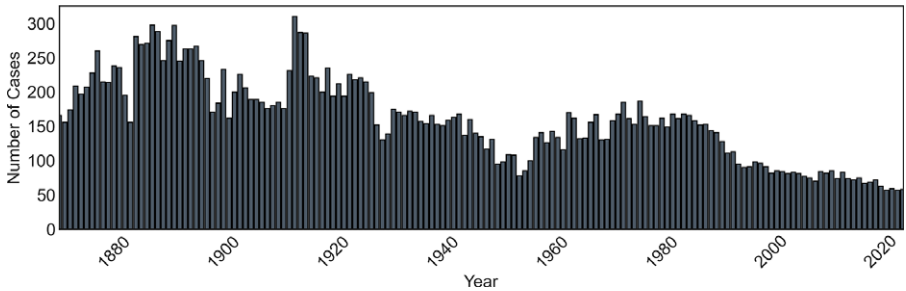


Figure 1. Number of court cases in every year in our dataset.

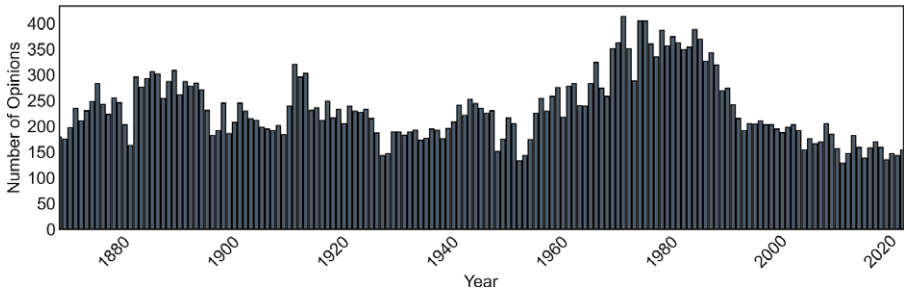


Figure 2. Number of court opinions issued in every year in our dataset.

statutory interpretation. Focusing on statutory interpretation was a strategic research choice. Where possible, we sought to ease the cognitive burden we placed on our annotators, and tracking categories of legal reasoning across different legal domains is taxing. The risk of this approach is that the model orients to statutory interpretation rather than law more generally. One reason we thought that this risk was tolerable is that much of the qualitative historical literature on trends in jurisprudence speaks in general terms that integrates over many different legal domains (Stiglitz and Thalken 2024). As noted later, we evaluate this risk and find that model performance does not drop off in non-statutory domains, suggesting that the orientation for statutory interpretation is aligned with other legal domains (Stiglitz and Thalken 2024).

To do this filtering, we select opinions that include any of the word tokens "statute," "legislation," or "act," within 200 characters of the tokens "mean," "constru" (i.e. construct), "interpret," "reading," or "understand." We choose these words as a simple way to increase the likelihood that statutory interpretation is present in an opinion.

Of the cases that pass the initial statutory interpretation filter, we chunk the opinions by paragraph. Because we anticipate formal and grand reasoning to be relatively infrequent in court opinions, we up-sample paragraphs that are likely to have formal or grand reasoning by identifying the occurrence of seed words that are indications of formal or grand reasoning (Table 2).

This subsampled dataset was gradually created through periodic sampling from the full dataset during the annotation task. In the initial sample, we select 50% of paragraphs with no seed term, 25% of paragraphs with at least one of the grand seed terms, and the remaining 25% with at least one of the formal seed terms. This

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Table 2. Seed Words Used to Up-sample Paragraphs Likely to Exhibit Formal or Grand Legal Reasoning

Formal Seeds	dictionary, dictionarium, liguae britannicae, world book, funk & wagnalls, expressio, expresio, inclusio, noscitur a sociis, noscitur a socis, ejusdem generis, last antecedent, plain language, whole act, whole-act, whole code, whole-code, in pari materia, meaningful variation, consistent usage, surplusage, superfluit, plain meaning, ordinary meaning, word
Grand Seeds	conference report, committee report, senate report, house report, assembly report, senate hearing, house hearing, assembly hearing, committee hearing, conference hearing, floor debate, legislative history, history of the legislation, conference committee, joint committee, senate committee, house committee, assembly committee, legislative purpose, congressional purpose, purpose of congress, purpose of the legislature, social, society

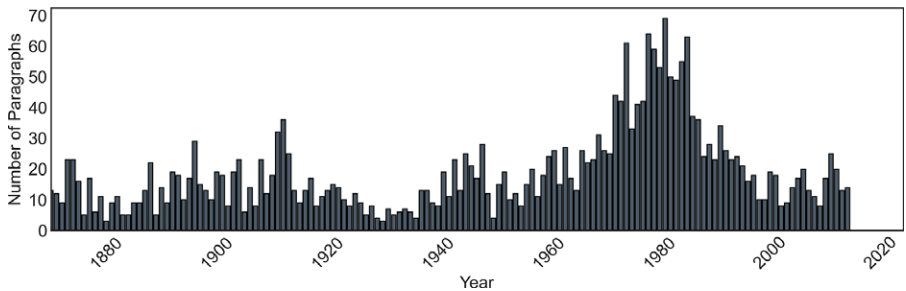


Figure 3. Number of paragraph samples from every year included in our annotated dataset. Paragraphs from the 1980s are over-represented.

sampling procedure remains constant until the final stages of annotation, where we increase the proportion of formal and grand seeds to 40% in both classes (and 20% without any seed term) to increase the number of formal and grand examples in our annotated dataset.

The final annotated dataset includes 2,748 paragraphs. Of these paragraphs, 37% of paragraphs include a formal seed, 36% of paragraphs include a grand seed, and 28% of paragraphs contain no formal or grand seed. In this dataset, paragraphs from court opinions that were issued in the 1980s are overrepresented (Figure 3), yet the 1980s were also the decade with the largest number of opinions issued per year, as shown in Figure 2.

The annotation task

In the annotation task, our team assigned a paragraph one of three labels: *formal*, *grand*, or *none*. Formal and grand reasoning are our categories of interest, yet not every court opinion paragraph performs legal reasoning – many paragraphs recite procedure or facts, for example. For this reason, the "none" label represents an absence of legal reasoning or legal reasoning that is neither formal nor grand.

We created a codebook with definitions and examples of each type of legal reasoning to support the annotation team. Our final codebook, which evolved modestly in the first weeks of the task, is included in full in Appendix. This codebook includes a section for each of the three classes, and a definition and set of core and periphery examples for each class. The core class is intended to represent a canonical

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

ideal of the class; the peripheral examples represent less central examples that still fit within a class.

We used the tool Prodigy as a simplified interface for the annotation task to reduce the cognitive workload of labeling data (Montani and Honnibal). This system allowed us to load the sampled paragraphs into a database system that presented annotators with a single paragraph and the class options (see Figure 4). Annotators were asked to assign a class label to the paragraph without any context (i.e. case title, year, author name, etc., were excluded). Additionally, we replaced any case citation with the token "[CITE]" using the eyecite Python package to identify citations (Cushman, Dahl, and Lissner 2021). Annotators were expected to choose a label using only their expertise in legal reasoning, the codebook, and the paragraph itself.

The annotation interface asks the annotator to consecutively label every highlighted sentence as formal, grand, or none. Then, after labeling every sentence in the paragraph, the annotator must assign a label to the entire paragraph. Though only the paragraph text and labels were used to train and evaluate the models, we included sentence-level labeling during the annotation task to ensure that every component of the paragraph was taken into account before the annotator reached a decision on the paragraph.

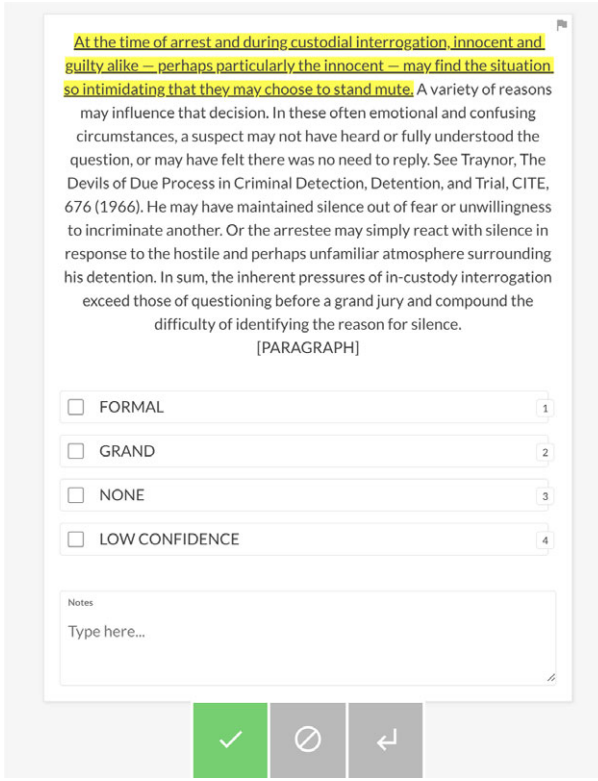


Figure 4. Annotation interface for labeling. An annotator must assign a label to every highlighted sentence before assigning a label to the paragraph as a whole.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Annotation team training

The annotation task took place over the spring 2023 semester. After developing the initial codebook, we recruited four upper-year law students, all of whom had completed coursework relevant to identifying legal reasoning. Annotators were assigned background reading on legal reasoning to review before the initial training session (Schauer 1988), in which they were introduced to the classes and tested the annotation interface. All annotators were then given an overlapping set of paragraphs to label in the next week.

In weekly intervals, after labeling the set of assigned paragraphs, we evaluated the quality of the annotations with inter-rater reliability. We use Krippendorff's *alpha* to account for the number of classes and annotators (Krippendorff 2011). Each week, the group discussed paragraphs with high disagreement to decide the paragraph's best class label and form a better collective understanding of each class. Annotators were typically assigned anywhere from 25 to 200 paragraphs to label in a week. During the first weeks, all annotators labeled the same paragraphs to allow for careful assessment of inter-rater reliability at any point in the annotation task; in later weeks, after reaching sufficient inter-rater reliability, we continued assigning at least 25 paragraphs out of the total set to every annotator.

For the first four weeks, inter-rater reliability was both low and inconsistent from week to week. During this training period, we used disagreements during the weekly discussions to refine the codebook. The most successful codebook change was the introduction of a decision chart, which we believe reduced cognitive load in the labeling task by breaking down the classification of a given paragraph into a set of simple binary questions (Figure 5).

After the introduction of this decision chart, inter-rater reliability increased and stabilized from week to week (Figure 6). Beginning in week five, the week following the introduction of the decision chart, we finalized the codebook and increased the number of paragraphs assigned to all annotators, so as to label as much data as possible.

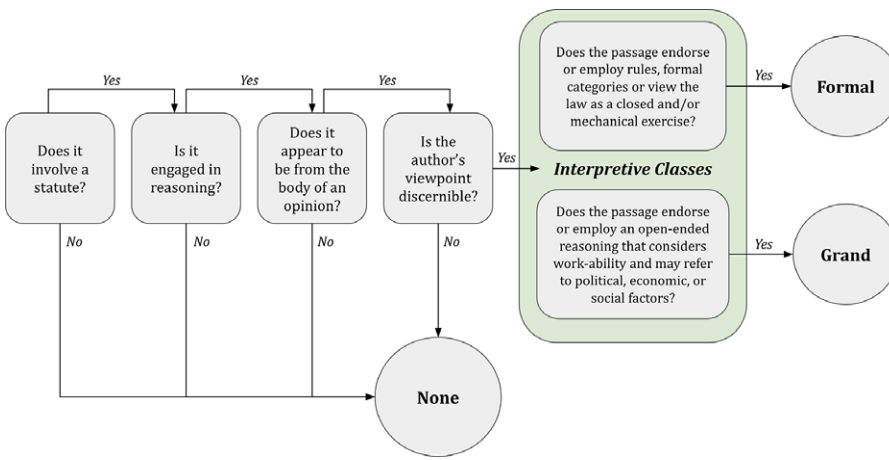


Figure 5. The decision chart provided to the annotation team beginning in week 5.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

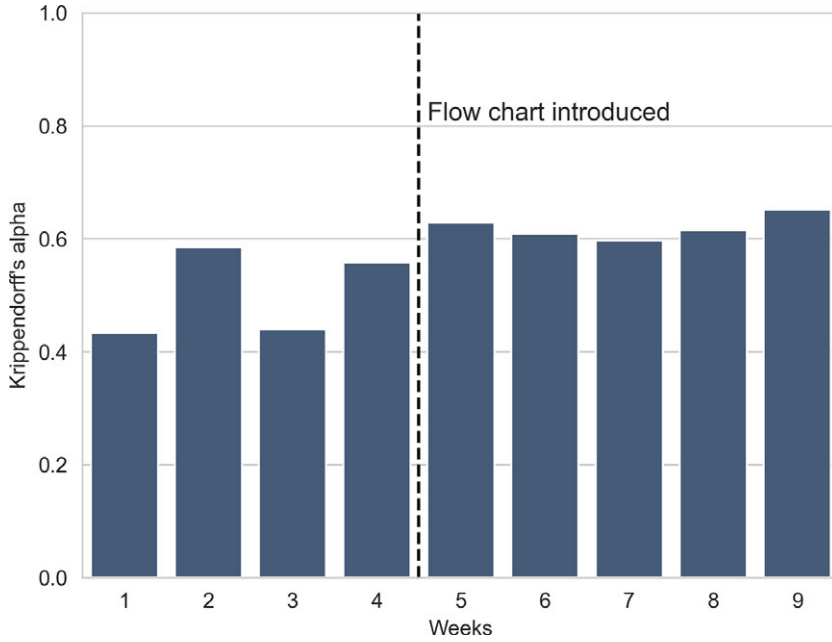


Figure 6. Weekly progression of Krippendorff's *alpha* for paragraphs and sentences assigned the formal, grand, and none labels. The decision chart was added before week 5.

Low confidence annotations

During the annotation process, some paragraphs had such unclear legal reasoning that the annotator could not confidently assign a label without a group discussion of the paragraph. We added the "low confidence" label so that an annotator could signal this lack of certainty; this label could only be added in addition to the three primary classes.

If a paragraph received the low confidence label, the annotation team would deliberate the best label for the paragraph until forming consensus. Often, paragraphs with the low confidence label engage in an internal argument: for example, the paragraph might critique grand reasoning, but for the further improvement of that type of reasoning, not to promote formal reasoning. In these instances, we asked annotators to assign a label based solely on the paragraph text and to avoid making assumptions about the role of the paragraph in the entire opinion. Some of the peripheral examples in the codebook were added after discussing these low-confidence paragraphs (see Appendix).

Ending the annotation task

Annotation tasks are costly in terms of the time and energy of the annotators, as well as their compensation for performing the task. Every annotator in our project was also a law student, so we were careful to end the annotation task when the marginal improvement of additional training data became low. More data would certainly be better, as it would allow for more partitioning and testing of the models, but that must

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

be balanced against the costs of human labor and the risk of diminished annotation quality as the task is prolonged.

After reaching our desired inter-rater reliability, we began evaluating model performance with each additional sample of data to identify the point at which additional labeled data had no empirical effect on model performance. We evaluate the performance of three fine-tuned large language models, BERT, LEGAL-BERT, and DistilBERT, to identify the point at which all three had plateaued on the various classes (Devlin et al. 2019; Chalkidis et al. 2020; Sanh et al. 2020).

Based on the plateaus in Figure 7, we ended data collection after ten total weeks, the last six of which occurred after reaching adequate inter-rater reliability, once we found no substantial improvement in model performance with additional data. We included only data from the last six weeks in our annotated dataset and discarded all annotated data before the inter-rater reliability improvement.

The annotated dataset

In total, there are 2,748 paragraphs in the annotated dataset, with an overall inter-rater reliability (IRR) of 0.63 Krippendorff's *alpha* for annotations included in the final set. Although this IRR score represents an improvement over the annotator training stage, it remains below the ideal level for an annotation task. However, given

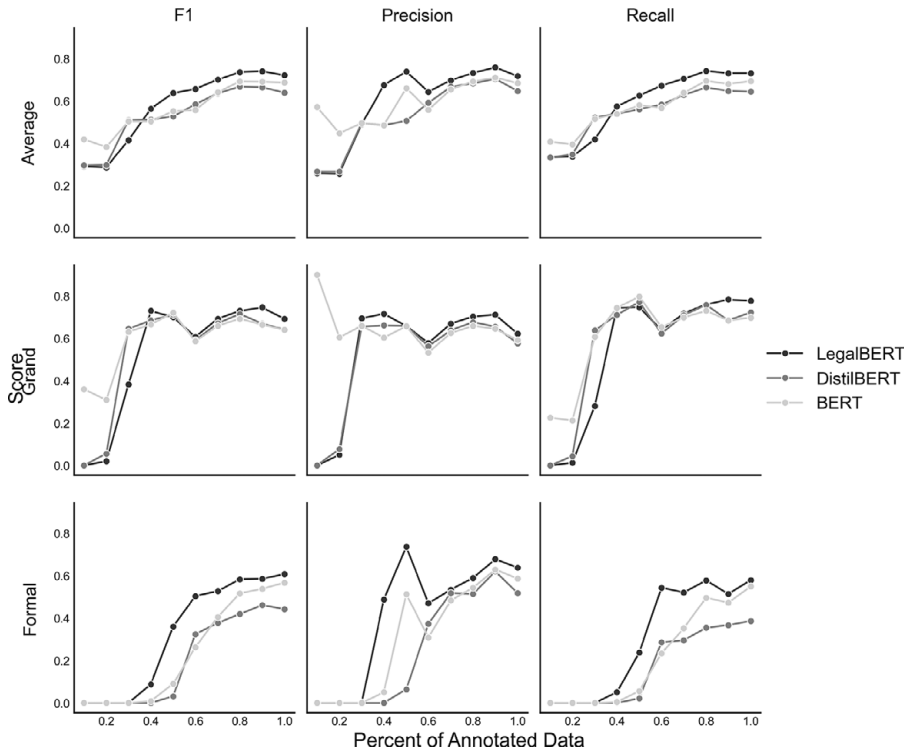


Figure 7. Model performance over 10-fold cross-validation across growing random samples of our annotated dataset, from 10% of the data to the entire dataset.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Table 3. Number of Paragraphs Fitting Each Class Identified by Annotators, and Number Assigned to the Low Confidence (LC) Class by Its Initial Annotator

Class	# Ident'd	Ident'd %	# LC	LC %
Formal	329	11.9	37	11.2%
Grand	551	20.0	33	6.0%
None	1869	68.0	31	1.7%
Total	2748		101	

the inherent complexity of classifying legal reasoning, particularly in cases where extracts may involve ambiguity, we regard this as an acceptable result.[8] We worried, moreover, that further amplifying annotator training or guidelines would reduce the task to a rule-following exercise, at the expense of the nuanced human judgment of our domain experts. Users of this dataset should recognize that the IRR score is lower than desired, but also that this is a task where a degree of expert disagreement is to be expected.

Table 3 shows the number of paragraphs assigned to each class label, as well as how many of those labels were also assigned the low confidence score. Though we upsampled paragraphs likely to use formal or grand reasoning, the "none" class was the most common. Grand reasoning is the second most common class, followed by formal reasoning as the least common class in our annotated dataset.

Evaluating models

Using the annotated dataset, we evaluate the performance of various models to optimize the quality of the predictions on the entire dataset. The objective is to find the model and training procedure that best align with the expert annotations in the annotated dataset. Because our object of interest is legal reasoning, we are most concerned with model performance in the grand and formal classes.

Related earlier work compares the performance of fine-tuned models, which update the model weights using training data, to generative models, such as GPT, prompted to identify formal, grand, and no legal reasoning. Full details of this model evaluation can be found in Thalken et al. 2023.[9]

This work evaluates two fine-tuning procedures: the first procedure tunes a standalone model to classify a paragraph into one of the three classes. The alternative procedure breaks the classification task into two serial steps: first, classifying legal reasoning (i.e. a formal or grand label) and the absence of legal reasoning (i.e. the none label); then, for paragraphs classified as legal reasoning, whether the legal reasoning is formal or grand. The fine-tuned models include BERT-base, LEGAL-BERT-base, DistilBERT, T5-base, and T5-small (Devlin et al. 2019; Chalkidis et al. 2020; Sanh et al. 2020; Raffel et al. 2020).

Those fine-tuned models are compared to prompted encoder-decoder models. The authors use three prompting strategies, all of which mirror sections of our codebook (see Appendix). The first prompt uses descriptions of the classes, the

[8]Recall, too, that annotators focused only on the reasoning in isolated paragraphs–we masked internal citations, case metadata, and the content of surrounding paragraphs.

[9]Some central results from those exercises can also be found in the appendix to this paper; see Appendix.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

second uses examples in a few-shot prompting strategy, and the third provides a chain of decisions that the model must make before assigning a label. The models used for prompting are all encoder-decoder models, including GPT-4, FLAN-T5-large, and Llama-2-Chat (7B) (OpenAI et al. 2024; Chung et al. 2022; Touvron et al. 2023).

We evaluate each model using five-fold cross-validation.[10] In each fold, we randomly hold out 25% of the data for evaluation and use the remaining 75% for fine-tuning the fine-tuned models (or omit this data in the case of prompted models). The same 25% subset is used to evaluate all models in a given fold. Final scores are reported as the average across the five folds. This procedure ensures that all models are evaluated on equivalent data, while preventing the fine-tuned models from gaining an advantage by having seen the evaluation data during training. Of all models and training procedures, a fine-tuned LEGAL-BERT model reaches the best performance in this task when fine-tuned to classify a paragraph into one of the three classes (complete results are in the Appendix). In general, fine-tuned models outperform prompted models; no prompted model performs a usable classification on legal reasoning.

Predictions

With this fine-tuned LEGAL-BERT model, we make predictions on all court opinion paragraphs in our entire dataset. The output of this fine-tuned model is a set of logits representing the model's certainty in each of the three classes ("formal," "grand," and "none"); these scores are converted to probabilities across the three classes using the Softmax function. Our primary measurement of legal reasoning, used in Empirical Studies of Legal Reasoning, calculates a formalism score based on predicted class probabilities. At the paragraph level, formalism is weighted by class probabilities: $\overline{p}_i = (1-\pi_{in})\sum_{k\in\{f,g\}} v_k\pi_{ik}$, where class $k \in \{f,g,n\}$ denotes the class of "formal," "grand," and "none" respectively; π_{ik} denotes the probability that the paragraph i is of class k; and v_k is a value assigned to that class for the measure. We set $v_f = 1$ and $v_g = -1$. Thus, a paragraph with the full mass of probability on the formal class receives a score of 1, while one with the full mass on the grand class receives a score of -1. Paragraphs with indeterminate classifications return with scores closer to 0. Finally, for substantive analyses, we typically aggregate paragraphs to the opinion-level and normalize these scores as z-scores.

Analysis of models, in Using and Improving Models, uses the original probabilities derived from the model, given that the object of interest is the model's certainty.

Validation

In this section, we present validation of our jurisprudence measurement. Some of this validation is already present in prior work:

- A yearly analysis of jurisprudence based on our measurement aligns with theories presented by legal scholars (Stiglitz and Thalken 2024). Prior work has explored

[10]The applied models will produce class predictions for each paragraph we feed to them. The encoder-decoder predictions come in the form of a text-based class, which may require standardization using regular expressions. The fine-tuned encoder models produce a determinate class, which reflects the class with the largest probability. The raw logits underlying those class probabilities can also be recovered, and we use those logits for downstream applications, as we note later. For example, the logits allow us to score paragraphs on a spectrum from formal to anti-formal, rather than restricting attention to a single, determinate class.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

this validation exercise and its contribution to debates regarding periods of legal jurisprudence in more detail below (Trends Over Time).
- The classification performance of the best-performing language model remains relatively stable across the early, middle, and late periods of the same dataset (Stiglitz and Thalken 2024).
- Codebook exemplars score as expected by our measurement, with the formal example scoring substantially higher in formalism and the grand examples higher in grand reasoning (Stiglitz and Thalken 2024).
- Although the annotated training data consists solely of excerpts from statutory interpretation cases, we apply the model to all cases in the dataset, regardless of domain. In prior work, two annotators labeled 200 non-statutory examples. We evaluate our model, which was trained on statutory interpretation cases, against this non-statutory set and find that it performs nearly as well as it does on the original statutory examples (Stiglitz and Thalken 2024).

In addition to these existing validation exercises, we present a new validation exercise where we evaluate the language that differentiates grand from formal reasoning. We use the Fightin' Words algorithm to calculate this difference at the word level (Monroe, Colaresi, and Quinn 2008; Hessel 2020). We begin by dropping any words that occur in more than 30% of all paragraphs or in fewer than 300 paragraphs in the full dataset. Then, for every remaining unique word token, we compare its probability in all grand paragraphs to its probability in all formal paragraphs. To do so, we calculate the log-odds ratio for each word, using an informative Dirichlet prior based on word frequencies across the entire dataset. Finally, we calculate z-scores to measure significance for each word. The result suggests how different language usage is between formal and grand paragraphs, at the word level.

Figure 8 shows the top 50 words that distinguish formal from grand, according to our measure. These results further validate our model, as formal words point toward segments of legal documents and the closed interpretation of a document's meaning. In contrast, the grand words emphasize legislative history and legal and political actors beyond the legal text.

Dataset applications

Empirical studies of legal reasoning

Using these paragraph-level measures of legal reasoning, we can empirically analyze longstanding and novel substantive questions related to jurisprudence. In this introductory paper, we highlight three primary hypotheses: first, that the evolution of legal reasoning has now reached a point where formalism is the predominant mode of legal reasoning; second, that conservative ideology is correlated with formal reasoning; and finally, that a case that results in a conservative decision will have been decided with more formal reasoning.

Trends over time

In general, legal historians agree on the basic periods of legal reasoning, as it shifted toward and away from formal or grand reasoning: First, there is scholarly consensus that between the Civil War and roughly World War I, formalism was the predominant mode of legal reasoning (Thomas C Grey 1983; Grossman 2007; Farber 2015).

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

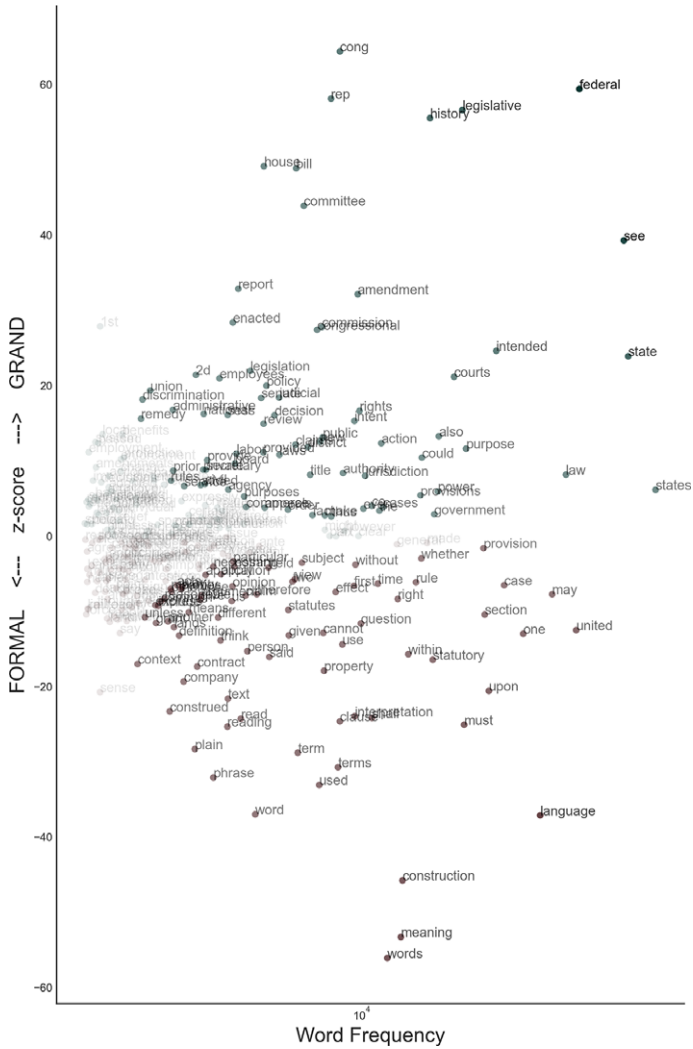


Figure 8. Words that most distinguish paragraphs measured as formal to paragraphs measured as grand, calculated using Fightin' Words.

Next, at some point in the mid-twentieth century, grand reasoning came to predominate.[11] Legal scholars largely agree that after the 1980s, formalism once again became the most common mode of legal reasoning (Eskridge Jr 1989; Thomas C. Grey 1999; P. L. Strauss 2014). This lean toward formalism continues to the present (Doerfler 2022; Stiglitz and Thalken 2024).

Using the predictions from our model, we can compare these expectations regarding the periodization of legal reasoning to results from our computational

[11]There is also a common view that jurisprudence in the early Republic prior to the Civil War was grand in nature, though our data do not allow us to investigate that thesis.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

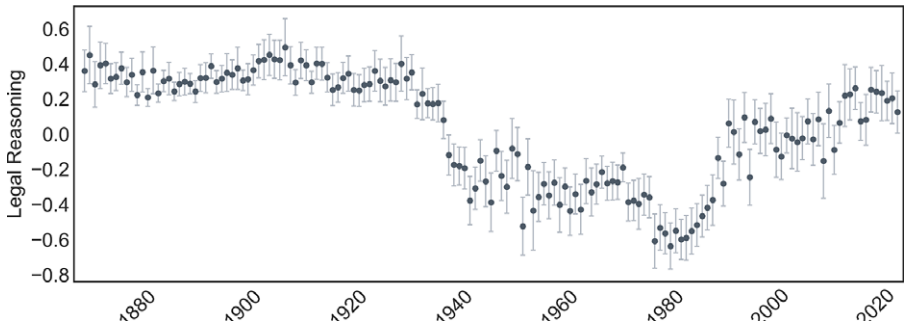


Figure 9. Predicted legal reasoning across years in our dataset. Higher values are formal and lower values are grand.

model. We would expect formal reasoning to be the primary form of reasoning from the beginning of our dataset (1870) and through at least World War I; grand reasoning to rise in popularity during the early-to-mid 20th century, and then formalism to resurge in the 1980s into the present. To assess these historical assumptions, we take the predictions from the model at the opinion level, find the normalized average between formal (1) and grand (-1) reasoning across all paragraphs. We then calculate an average across all opinions issued in a year.

Our results support the core expectations of the legal historians (Figure 9). Formalism does, in fact, dominate from the 1870s and through the 1930s. There is then a rapid shift to grand reasoning following the judicial revolution of 1937, which remains dominant until the 1980s. Formalism surges during the 1980s and remains the primary mode of legal reasoning through all remaining years in our dataset, ending in 2024.

Though these results support prior qualitative assessments regarding the general arc of legal reasoning, legal scholars disagree on the exact turning points when legal reasoning shifted from one form of reasoning to another. For example, legal scholars variously expect the early 20th-century shift from formalism to grand reasoning to have occurred as early as *Lochner v. New York* (Horwitz 1992; Schauer 1988), around World War I (Llewellyn 1960; Gilmore 2015; Kennedy 2006), or decades later, at the "judicial revolution" of 1937 (Wiecek 2001; Tamanaha 2008).

With the granular data provided by our model's predictions, we can identify more exact turning points. In particular, more than any other year, the judicial revolution of 1937 occurred alongside an abrupt change in legal reasoning, from the formal to the grand reasoning. This result is visualized in Figure 9, and prior work has evaluated this turning point in detail and found that it is a statistically significant turning point (Stiglitz and Thalken 2024).

Judicial ideology and reasoning

Common perceptions of legal reasoning align formalism with a conservative judicial ideology (Thomas C. Grey 1999). Through our formalism measure, we can consider whether formalism does in fact correlate with conservatism. By way of preliminary inspection, consider the average formalism scores for recent justices. Figure 10 displays the average legal reasoning score used by every judge since 1973, excluding judges most recently appointed to the Court or those who wrote less than 50

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

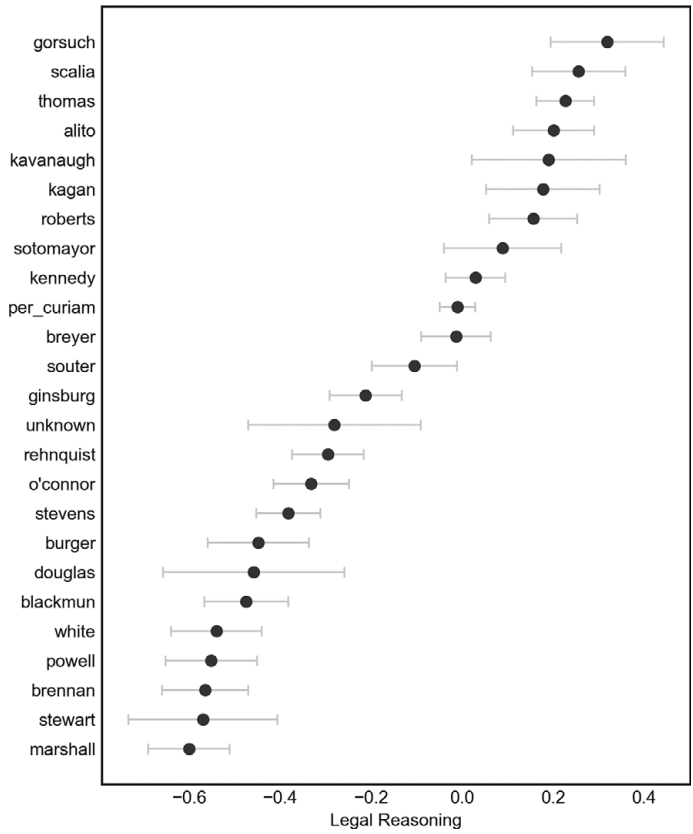


Figure 10. Predicted legal reasoning for every justice since 1973.

opinions between 1973 and 2024.[12] These results support the common categorization of Justice Scalia as a recent formalist (Segall 1993). Justices Gorsuch, Thomas, Alito, and Kavanaugh also lean toward formalism. In contrast, Justices Marshall, Stewart, and Brennan used the most grand reasoning. These judge-level formalism scores further support that we live in a period of high formalism: all currently-serving Roberts Court judges included in our dataset lean more toward formal than grand reasoning.

As indicated in Figure 10, judges commonly associated with more conservative ideology tend to exhibit formalism in our measure, and judges commonly associated with more liberal ideology tend to exhibit anti-formalism. To further evaluate this, we use a simple but common measure of judicial ideology, the binary party of the appointing President, and find that, on average, judges who are appointed by a Republican president use higher levels of formalism (Figure 11).

[12]Closely related figures appear in other studies (Stiglitz and Thalken 2024, 2026).

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

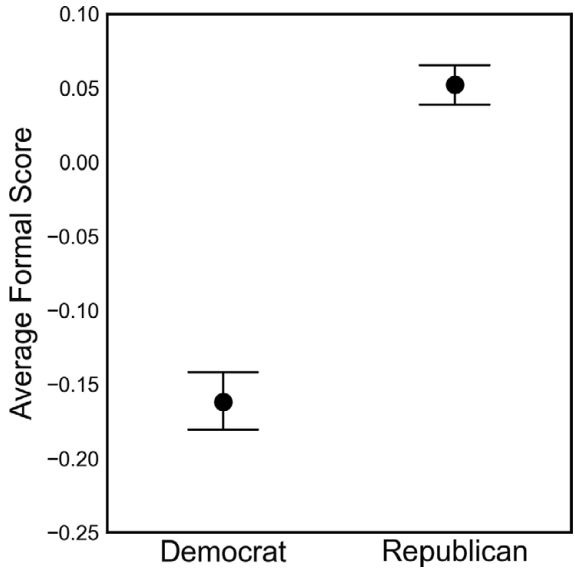


Figure 11. Average formalism scores with 95% confidence intervals for judges appointed by a Democratic versus a Republican President.

Legal reasoning and outcomes

An important margin of interest is the relationship between jurisprudence and *outcomes*. Most proximately, does formalism promote conservative judicial decisions? More distally, does formalism affect the distribution of income, public health outcomes, or other social outcomes we might care about?

Formalism plausibly bears a relationship with both sets of outcomes. More formalist approaches to reasoning, such as textualism, may, for example, promote conservative judicial decisions by narrowing the window of what counts toward legal meaning, eliding broader policy objectives or social considerations. In aggregate, and propagated throughout the judicial system, more conservative judicial decisions may influence broader socioeconomic outcomes.

We note that these would be difficult questions to study. Does formalism drive case outcomes, for example, or is it instead judicial ideology? Are socioeconomic changes co-incident with the rise of formalism driven by changes in jurisprudence, or instead do they relate to broader changes in the political environment that feed through legislative and executive politics? Parsing out other, viable interpretations would be a focus in this line of research.

Despite thorny causal questions, the attention that observers devote to jurisprudence, and the heat surrounding debates over the appropriate method of legal reasoning, suggest that closer attention to the relationship between jurisprudence and outcomes is warranted. The measure offers an opportunity to develop this line of inquiry.

Using and improving models

So far, the applications discussed were based on the final measure relating to Supreme Court opinions after Reconstruction. In addition to this measure, we also publish our annotated dataset and our best-performing model with adjusted weights.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

That best-performing model can be used to extend the study of legal reasoning to other time periods or jurisdictions. For example, scholars may wish to study whether legal reasoning remains formal in future years. Or, they may want to understand the relationship between trends in legal reasoning on the Supreme Court and the lower courts. The model can be fed new data – for example, lower court opinions – and it will produce a measure of jurisprudence relevant to those opinions.

The annotated dataset, in turn, may be used for a variety of purposes. First, as new, still more sophisticated foundational models emerge, they can be fine-tuned on the annotated dataset. The measure of jurisprudence based on these newer models may outperform our best model. A superior measure may be able to address more fine-grained legal questions. These new measures derived from the annotation data will be of interest to scholars interested in substantive legal questions.

Second, the annotated dataset may be useful for those interested in language models or in the technical aspects of measuring legal reasoning. For example, the annotated dataset could be used for benchmarking model performance and to support the evaluation and comparison of both legal-domain and generalized models. We see our annotated dataset as a unique and robust benchmark for several reasons: the annotated dataset was assembled carefully, with extensive documentation requiring hundreds of hours of codebook refinement, annotation, and team conversations; the task itself is a challenging one, requiring both abstract reasoning and domain knowledge; the specific task would not be part of standard pre-training material for a language model.[13]

Like other researchers (Orr and Kang 2024; Walsh, Preus, and Antoniak 2024), we caution against the use of this dataset without attention to its domain-specificity and the known ambiguity of some labeled examples. Throughout the annotation process, we came across examples where two experts could arrive at different labels using valid human judgment. Precisely because this is a challenging task, this primes the annotated dataset for use in probing the limits of increasingly capable models. Indeed, though the ambiguous nature of many legal questions calls for caution, at core we see the ambiguity of the domain as an opportunity for the technical modeling of knowledge uncertainty.

To gesture in this direction, our initial analysis demonstrates that there is alignment between uncertainty on the part of our annotators (measured through the hand-labeled 'low confidence' class) and uncertainty on the part of the model, which can reasonably be measured through the probabilities assigned to each class. Using only predictions for paragraphs in the annotated dataset, we compare the distribution of class probabilities assigned by the fine-tuned LEGAL-BERT model to paragraphs with the low confidence (LC) label to non-low confidence (NLC) paragraphs. We expect the model to be more "confused" about the correct class on the paragraphs our team marked low confidence. This confusion will manifest in dispersion over predicted class probabilities.

We measure model uncertainty using the entropy of the class probability distribution. Entropy captures how evenly the model spreads its prediction probabilities across the possible classes. When the model is confident, it assigns a high probability to one class and low probabilities to the others, resulting in low entropy. When the

[13]Without fine-tuning, other researchers have seen that models have poor performance at identifying legal reasoning (Thalken et al. 2023), suggesting this is a task that reflects a gap in current model knowledge.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

model is uncertain, it assigns more even probabilities across the classes, resulting in higher entropy.

For this exercise, we begin with the model's raw logit predictions for each class ("formal," "grand," and "none"). We convert these logits to class probabilities using the softmax function, which normalizes the outputs so they sum to one. Next, we compute the entropy of the resulting probability distribution for each paragraph. We then calculate the average entropy for LC (low confidence) paragraphs (0.789) and for NLC (non-low confidence) paragraphs (0.539), yielding an observed difference of 0.250. As expected, the LC paragraphs have higher average entropy, indicating that the model is more uncertain on these examples.

To assess the significance of this difference, we conduct a permutation test over 10,000 trials. In each trial, we randomly shuffle the LC labels, recompute the average entropy for the permuted LC and NLC groups, and calculate the difference between them. We then count how many times the permuted difference exceeds the observed difference in magnitude. Dividing this count by 10,000 gives the p-value. As visualized in Figure 12, the observed difference is statistically significant ($p < 0.0001$), providing strong evidence that model uncertainty is systematically higher on LC data.

Exercises such as this one may be exploited for a variety of questions. For example, technically this exercise draws a line between human uncertainty and model uncertainty. One concern with language models is that they can express false certainty, notably "hallucinating" about even simple facts or references (Magesh et al. 2024). It is possible to ask language models to report their level of certainty, but the meaning of the reported quantity is unclear. The exercise in this section shows that it may be possible to recover meaningful assessments of model uncertainty.

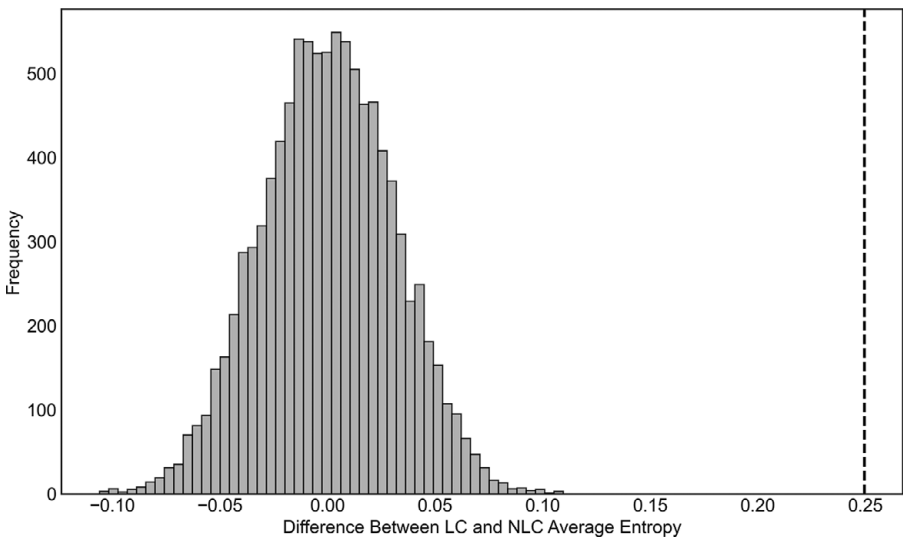


Figure 12. Results from the permutation test. The dashed line represents the observed difference between the average entropy of LC paragraph probabilities and the average entropy of NLC paragraph probabilities in the annotated dataset. The histogram shows the distribution of randomized differences generated through the permutation test.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Conclusion

In conclusion, this project advances the study of legal jurisprudence by providing a new measure of legal reasoning in U.S. Supreme Court opinions from 1870 to 2024. By leveraging large language models and domain-specific expertise, we provide both data and methods for systematically analyzing the evolution of legal reasoning. This measure offers empirical legal scholars a valuable resource for exploring the historical trajectories and influences shaping legal reasoning. Additionally, we release our best-performing model for identifying legal reasoning, enabling scholars to expand the dataset or enhance model performance. While this work focuses on the Supreme Court, its implications extend to other state and federal courts, prompting future research into whether legal reasoning trends continue toward formalism, stabilize, or shift back toward grand reasoning as new Supreme Court terms unfold.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/jlc.2025.10012.

Data availability statement. Replication data and code can be found in Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QNC3ZL or in our project GitHub repositories: https://github.com/rosthalken/supreme-court-data.

Acknowledgments. We are grateful for feedback from thoughtful reviewers and for the opportunity to present parts of this work at EMNLP, CELS, OWCAL, PEPL, ICAIL, ALEA, and Cornell University.

Funding support. This research was partially supported by NSF #FMiTF-2019313 and NSF #1652536.

Competing interests. None.

Ethical standard. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. *Legal-BERT: the muppets straight out of law school.* arXiv: 2010.02559 [cs.CL].

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, *et al.* 2022. *Scaling instruction-finetuned language models.* https://doi.org/10.48550/ARXIV.2210.11416. https://arxiv.org/abs/2210.11416.

Cushman, Jack, Matthew Dahl, and Michael Lissner. 2021. Eyecite: a tool for parsing legal citations. *Journal of Open Source Software* 6 (66): 3617. https://doi.org/10.21105/joss.03617.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, volume *1* (long and short papers), edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Doerfler, Ryan D. 2022. Late-stage textualism. *The Supreme Court Review* 2021 (1): 267–313.

Eskridge Jr, William N. 1989. The new textualism. *UCLA l. Rev.* 37:621.

Farber, Daniel A. 2015. Back to the future: legal scholarship in the progressive era and today. *Minn. L. Rev.* 100:1.

Gilmore, Grant. 2015. *The ages of American law*. Yale University Press.

Grey, Thomas C. 1983. Langdell's orthodoxy. *U. PiTT. l. rev.* 45:1.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Grey, Thomas C. 1999. *The new formalism*. Technical report Working Paper No. 4. Available at SSRN:https://ssrn.com/abstract=200732 or https://doi.org/10.2139/ssrn.200732. Stanford Law School, Public Law and Legal Theory Working Paper Series.

Grossman, Lewis A. 2007. Langdell upside-down: James Coolidge Carter and the anticlassical jurisprudence of anticodification. *Yale JL & Human.* 19:149.

Hessel, J. M. 2020. *Fightingwords*. https://github.com/jmhessel/FightingWords. Accessed: 2025-02-06.

Horwitz, Morton J. 1992. *The transformation of American law, 1870-1960: the crisis of legal orthodoxy*. Oxford University Press.

Kay, Richard. 2009. Original intention and public meaning in constitutional interpretation. *Faculty Articles and Papers*, no. 5, https://digitalcommons.lib.uconn.edu/law_papers/5.

Kennedy, Duncan. 2006. *The rise & fall of classical legal thought*. Beard Books.

Krippendorff, Klaus. 2011. Computing Krippendorff's alpha-reliability. https://api.semanticscholar.org/CorpusID:59901023.

Llewellyn, Karl N. 1960. *The common law tradition: deciding appeals*. Buffalo, NY: W.S. Hein.

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading AI legal research tools. *arXiv preprint arXiv:2405.20362*.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16 (4): 372–403.

Montani, Ines, and Matthew Honnibal. *Prodigy: a modern and scriptable annotation tool for creating training data for machine learning models.* https://prodi.gy/.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, *et al.* 2024. *GPT-4 technical report.* arXiv: 2303.08774 [cs.CL]. https://arxiv.org/abs/2303.08774.

Orr, Will, and Edward B Kang. 2024. AI as a sport: on the competitive epistemologies of benchmarking. In *The 2024 ACM conference on fairness, accountability, and transparency*, 1875–1884.

Pierce, Jr., Richard J. 1995. The supreme court's new hypertextualism: an invitation to cacophony and incoherence in the administrative state. *Columbia Law Review* 95:749–781.

Pound, Roscoe. 1908. *Mechanical jurisprudence*. Columbia University Press.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21 (140): 1–67. http://jmlr.org/papers/v21/20-074.html.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. *DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter.* arXiv: 1910.01108 [cs.CL].

Scalia, Antonin, *et al.* 1997. Common-law courts in a civil-law system: the role of united states federal courts in interpreting the constitution and laws. In *A matter of interpretation: federal courts and the law*, STU-Student edition, edited by Amy Gutmann, 3–48. Accessed 4 Nov. 2024. Princeton University Press. http://www.jstor.org/stable/j.ctt7t4dg.4.

Schauer, Frederick. 1988. Formalism. *Yale Law Journal* 97 (4): 509–548.

Segall, Eric J. 1993. Justice Scalia, critical legal studies, and the rule of law. *George Washington Law Review* 62: 991.

Spaeth, Harold J., Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2024. *2024 supreme court database, version 2024 release 01.* Accessed: 2025-04-07. http://supremecourtdatabase.org.

Stiglitz, Edward H., and Rosamond Thalken. 2024. Historical trends in macro-jurisprudence: a language model assessment, 1870–2023. *Maryland Law Review* (November).

Stiglitz, Edward H., and Rosamond Thalken. 2026. Understanding change in jurisprudence, 66–83.

Strauss, David A. 2010. *The living constitution*. Oxford University Press.

Strauss, Peter L. 2014. *Legal methods: understanding and using cases and statutes*. Foundation Press.

Tamanaha, Brian Z. 2008. Understanding legal realism. *Tex. L. Rev.* 87:731.

Thalken, Rosamond, EdwardStiglitz, DavidMimno and Matthew Wilkens. 2023. Modeling legal reasoning: LM annotation at the edge of human agreement. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, edited by Houda Bouamor, Juan Pino, and Kalika Bali, 9252–9265. Singapore: Association for Computational Linguistics, December. https://doi.org/10.18653/v1/2023.emnlp-main.575. https://aclanthology.org/2023.emnlp-main.575.

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, *et al.* 2023. *Llama 2: open foundation and fine-tuned chat models.* arXiv: 2307.09288 [cs.CL].
Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems.*
Walsh, Melanie, Anna Preus, and Maria Antoniak. 2024. *Sonnet or not, bot? poetry evaluation for large models and datasets.* arXiv: 2406.18906 [cs.CL]. https://arxiv.org/abs/2406.18906.
Wiecek, William M. 2001. *The lost world of classical legal thought: law and ideology in America, 1886-1937.* Oxford University Press, USA.

Cite this article: Thalken, Rosamond, and Edward H. Stiglitz. 2026. "Measuring Jurisprudence." *Journal of Law and Courts*, 1–22, doi:10.1017/jlc.2025.10012

https://doi.org/10.1017/jlc.2025.10012 Published online by Cambridge University Press