

Crafting Cultural Models: The Vital Role of Domain Expertise in Language Modeling

Rosamond Thalken, April 4
PhD Candidate, Information Science, Cornell University

Roadmap

1. What is the role of document classification in studying cultural heritage?
2. How have changes in machine learning (and the rise of language modeling) changed document classification?
3. Can we use generative models for domain-specific document classification?

What are the practicalities, limitations, and benefits of using generative models to annotate documents?

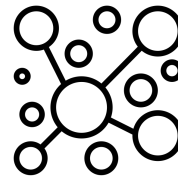
Document classification is a critical step in making use of cultural documents.



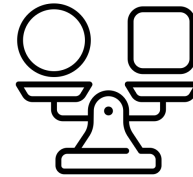
Retrieval



Trends

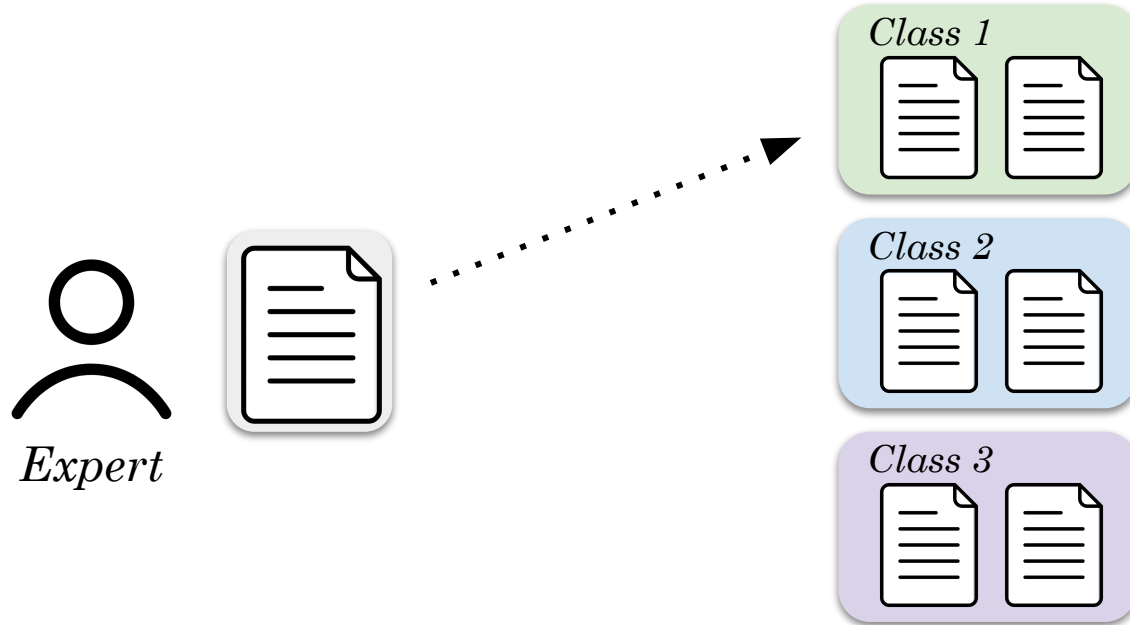


Collections

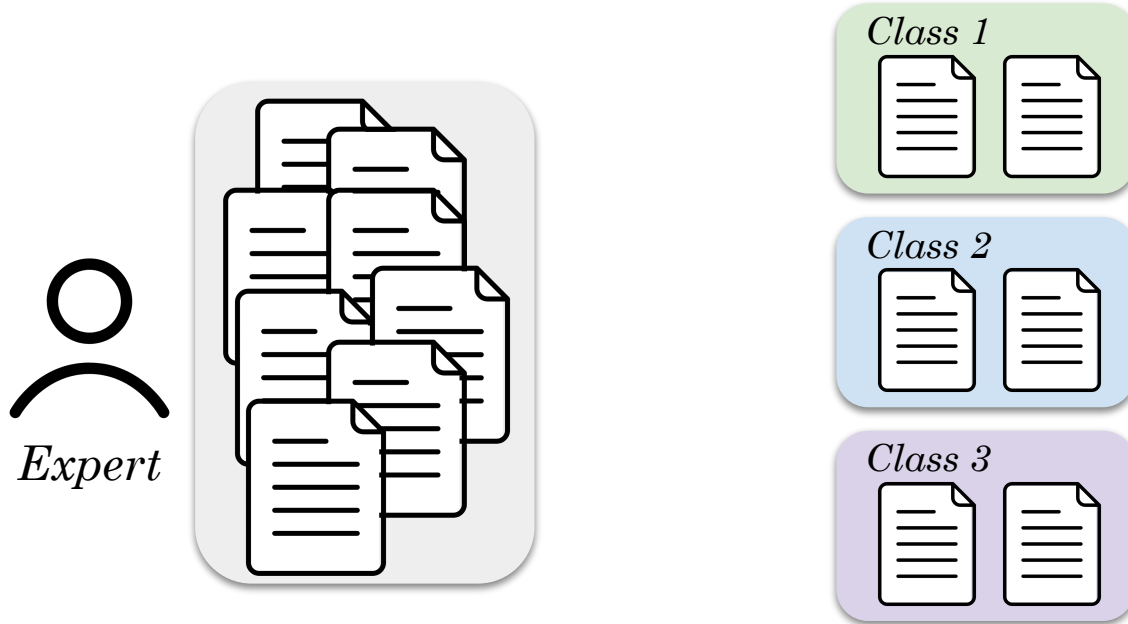


Comparison

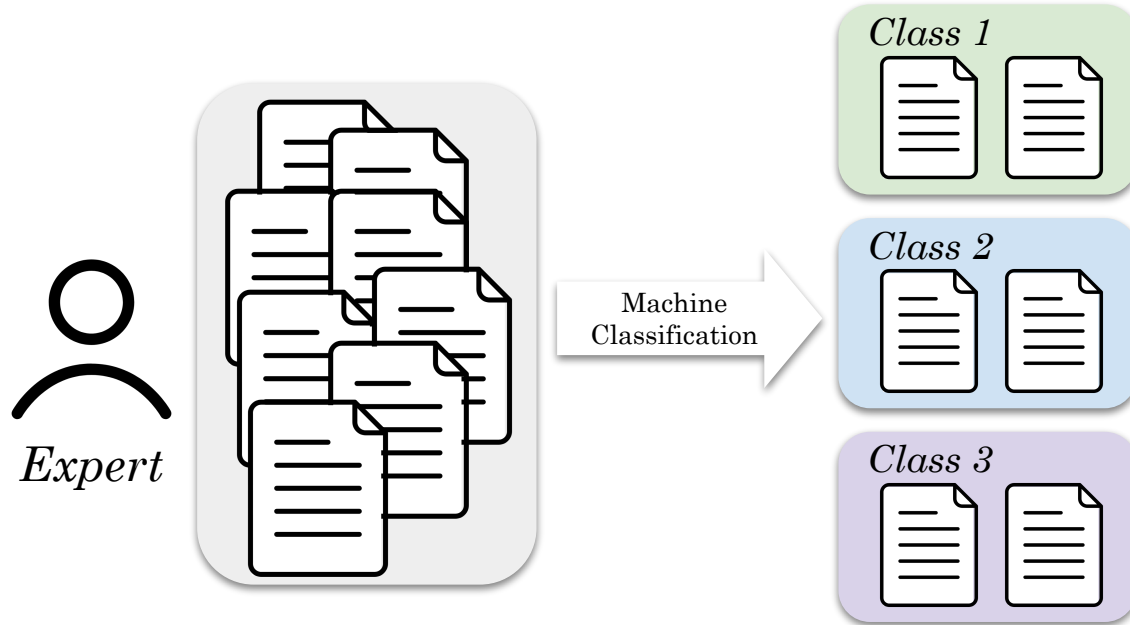
In the best-case scenario, an expert categorizes every document.



In the best-case scenario, an expert categorizes every document.



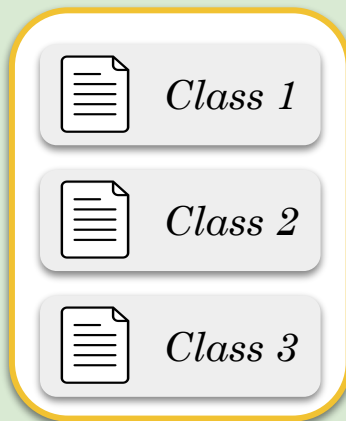
With large datasets, machine learning can replicate hand-labeled annotations at scale.



What training data do we need?

What training data do we need?

#1: Traditional machine learning



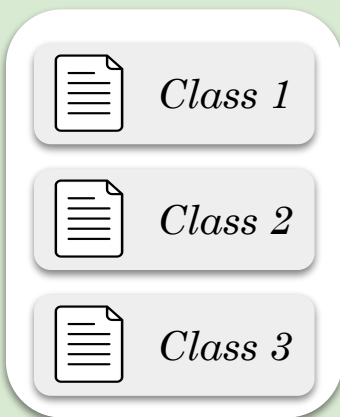
Annotated Dataset

What training data do we need?

#2: Large language models



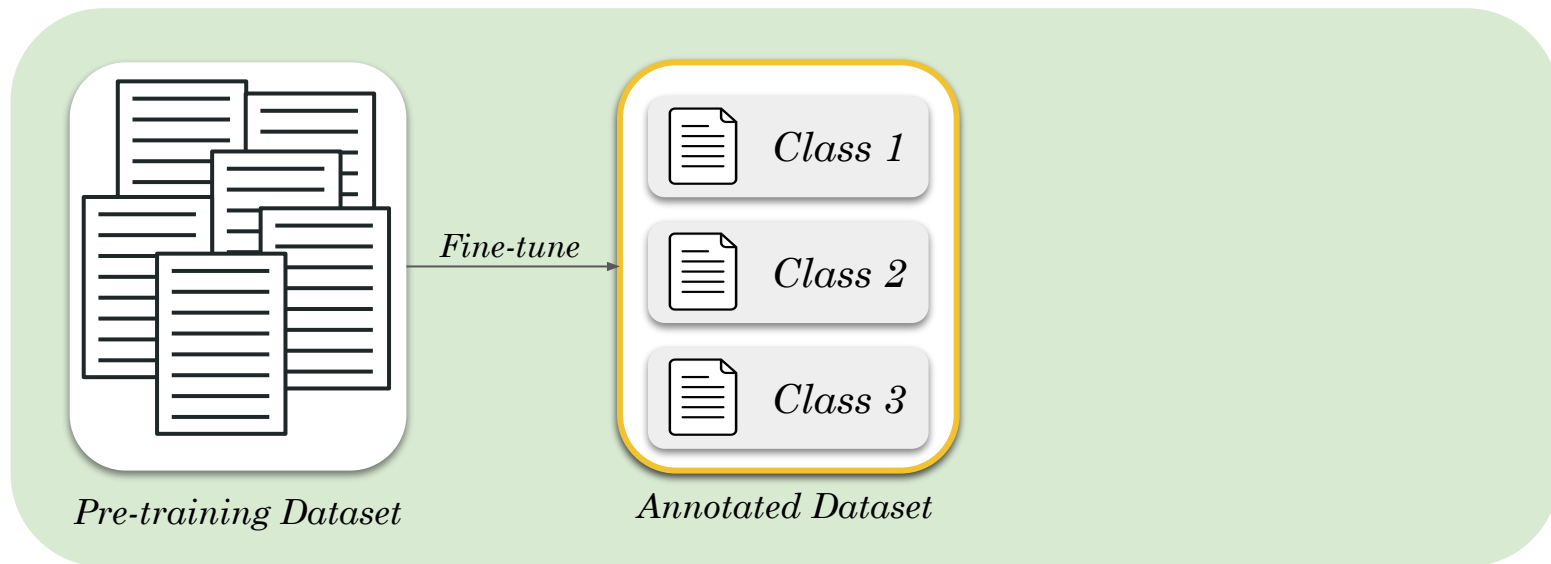
Pre-training Dataset



Annotated Dataset

What training data do we need?

#2: Large language models



What training data do we need?

#3 Generative language models



Pre-training Dataset



Class 1



Class 2



Class 3

Annotated Dataset

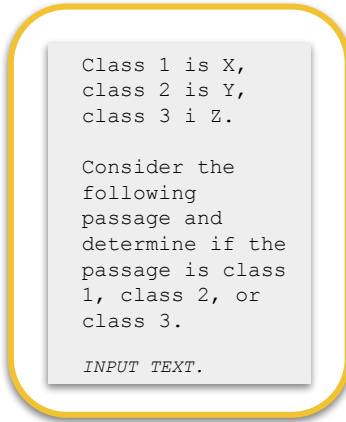
Class 1 is X,
class 2 is Y,
class 3 is Z.

Consider the
following
passage and
determine if the
passage is class
1, class 2, or
class 3.

[INPUT TEXT.]

Prompts

Why replace annotators with generative models?



```
Class 1 is X,  
class 2 is Y,  
class 3 i Z.
```

```
Consider the  
following  
passage and  
determine if the  
passage is class  
1, class 2, or  
class 3.
```

```
INPUT TEXT.
```

Creating a high-quality dataset requires extensive resources.

Annotation tasks entail tedious, repetitive, intensive labor.

Certain annotation tasks can be effectively replicated by generative models.^{1, 2, 3}

[1] He et al. "AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators"

[2] Gilardi et al. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." *PNAS* 2023.

[3] Törnberg. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning."

But what are the **practicalities**, **limitations**, and **benefits** of using generative models for document annotation?

Case study: *Classifying legal reasoning with large language models*.^{4, 5}

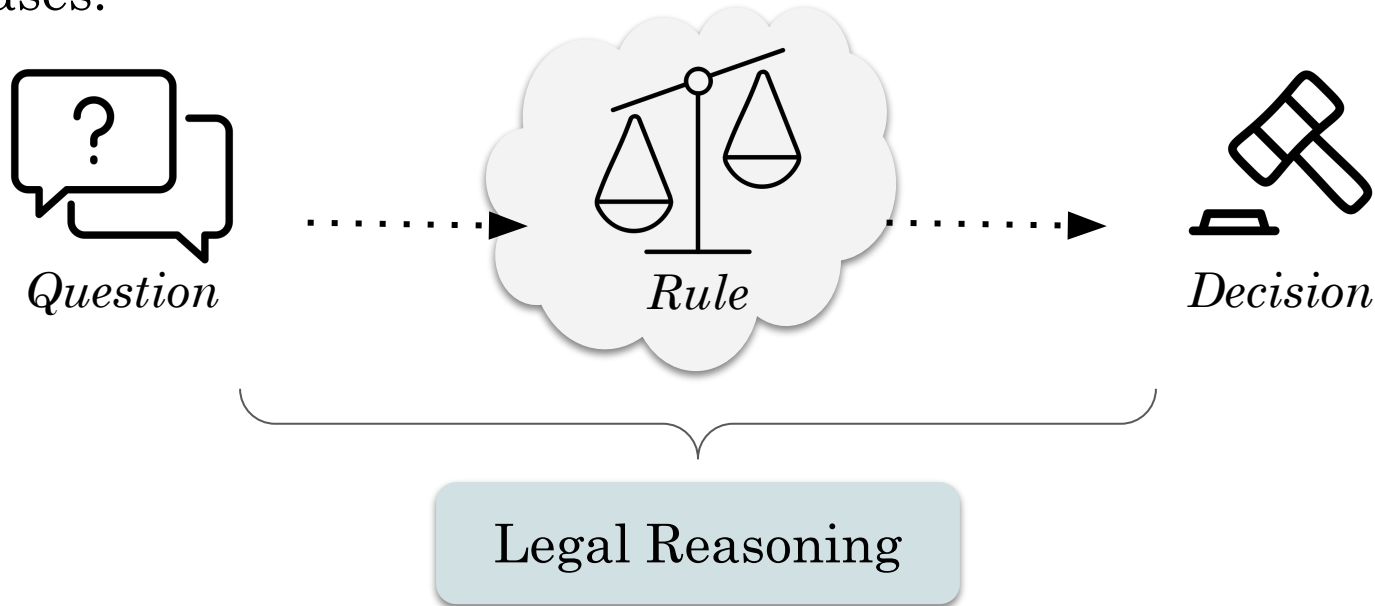
LLMs can lower barriers to document classification, but researchers should take caution when using them.

[4] Thalken, Stiglitz, Wilkens, and Mimno. "Modeling Legal Reasoning: LM Annotation at the Edge of Human Agreement." *EMNLP*, 2023.

[5] Stiglitz and Thalken. "Historical Trends In Macro-Jurisprudence: A Language Model Assessment, 1870-2023. *Maryland Law Review*, forthcoming.

Judges use **legal reasoning** to guide their decision making in specific court cases.

Judges use **legal reasoning** to guide their decision making in specific court cases.



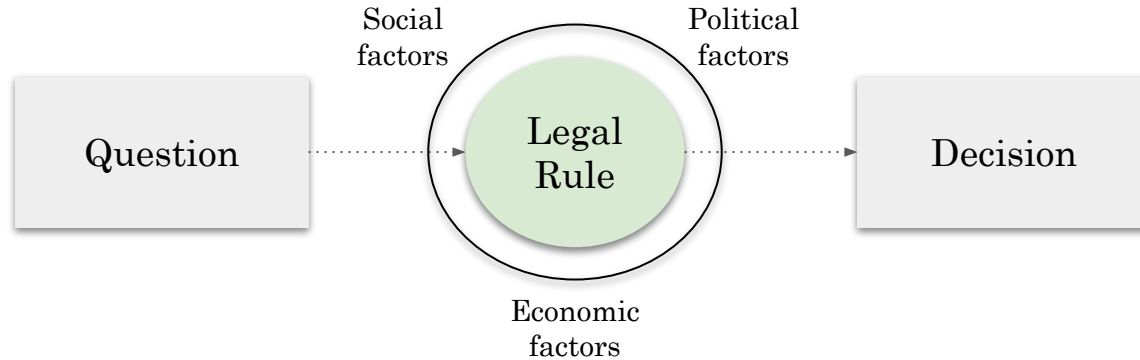
Formal Reasoning

There are two opposing
types of legal reasoning.

Grand Reasoning

Formal Reasoning

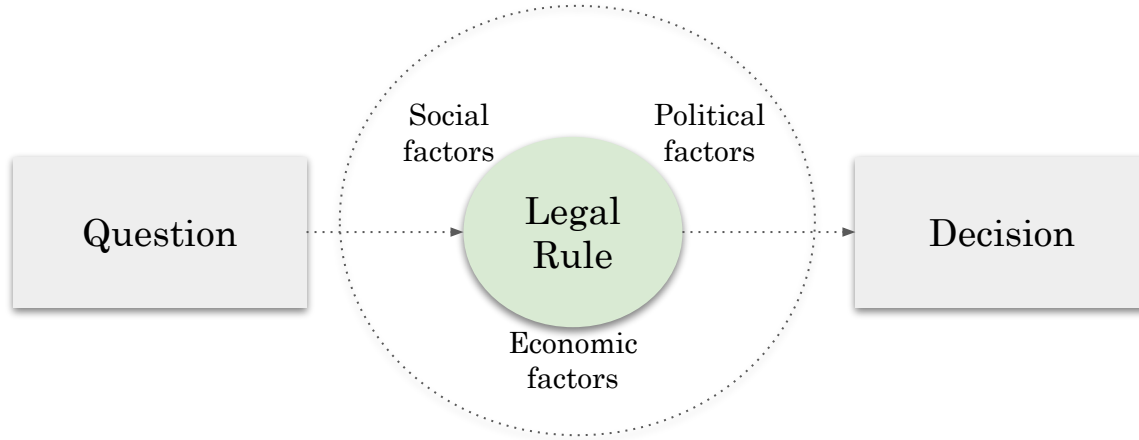
A legal decision made according to a rule, often viewing the law as a **closed and mechanical** system.



*Karl Llewellyn, "The Common Law Tradition: Deciding Appeals"

Grand Reasoning

A legal decision that looks at the law as an **open-ended and ongoing** enterprise.



[illegible]

Legal reasoning is also a tool for writing persuasive court opinions.

Formal Reasoning

- Mechanical and rule-based logic
- References to dictionaries
- Law is understood as a closed system; only the text matters

Grand Reasoning

- Emphasis on legal intent
- Social, political, economic factors taken into consideration
- Law is seen as an open system and changing enterprise

Formal Reasoning

Accepting this point, too, for argument's sake, the question becomes: **What did "discriminate" mean in 1964?** As it turns out, **it meant then roughly what it means today:** "To make a difference in treatment or favor (of one as compared with others)." **Webster's New International Dictionary** 745 (2d ed. 1954). To "discriminate against" a person, then, would seem to mean treating that individual worse than others who are similarly situated. [CITE]. In so-called "disparate treatment" cases like today's, this Court has also held that the difference in treatment based on sex must be intentional. See, e.g., [CITE]. So, taken together, an employer who intentionally treats a person worse because of sex—such as by firing the person for actions or attributes it would tolerate in an individual of another sex—discriminates against that person in violation of Title VII. *Bostock v. Clayton County*

Respondent's argument is not without force. But **it overlooks the significance of the fact** that the Kaiser-USWA plan is an affirmative action plan voluntarily adopted by private parties to eliminate traditional patterns of racial segregation. **In this context** respondent's reliance upon a literal construction of §§ 703 (a) and (d) and upon McDonald is misplaced. See [CITE]. It is a "familiar rule, that a thing may be within the letter of the statute and yet not within the statute, because not within its spirit, nor within the intention of its makers." [CITE]. **The prohibition against racial discrimination in §§ 703 (a) and (d) of Title VII must therefore be read against the background of the legislative history of Title VII and the historical context from which the Act arose.** *Steelworkers v. Weber*.

Grand Reasoning

Qualitative legal scholars have identified historical periods of legal reasoning based on reading cases:

Civil War to WWI → **Formal**

WWI to 1980s → **Grand**

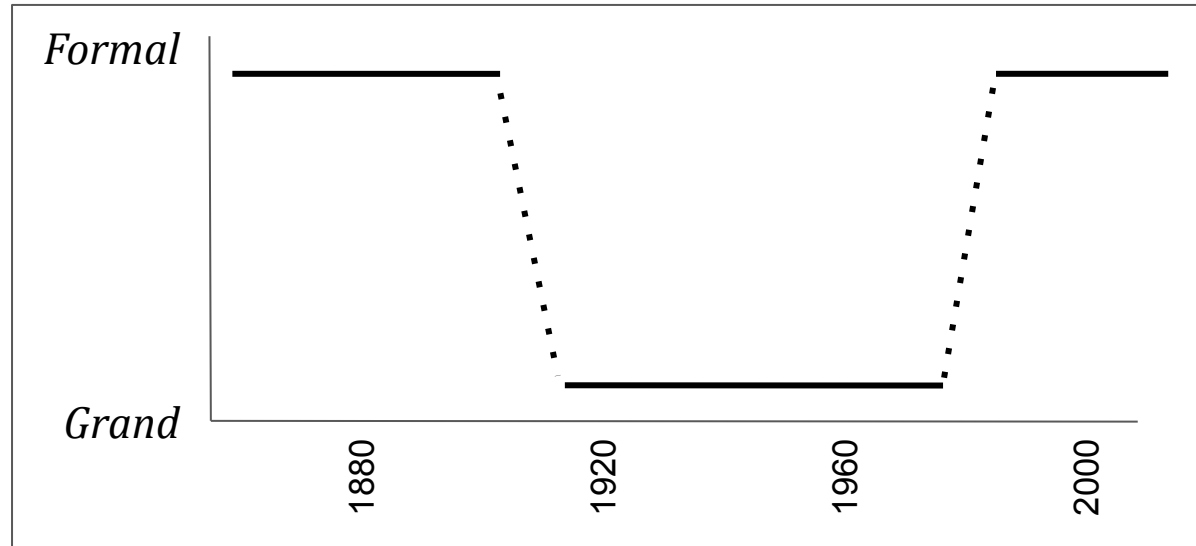
1980s to today → **Formal**

Qualitative legal scholars have identified historical periods of legal reasoning based on reading cases:

Civil War to WWI → **Formal**

WWI to 1980s → **Grand**

1980s to today → **Formal**

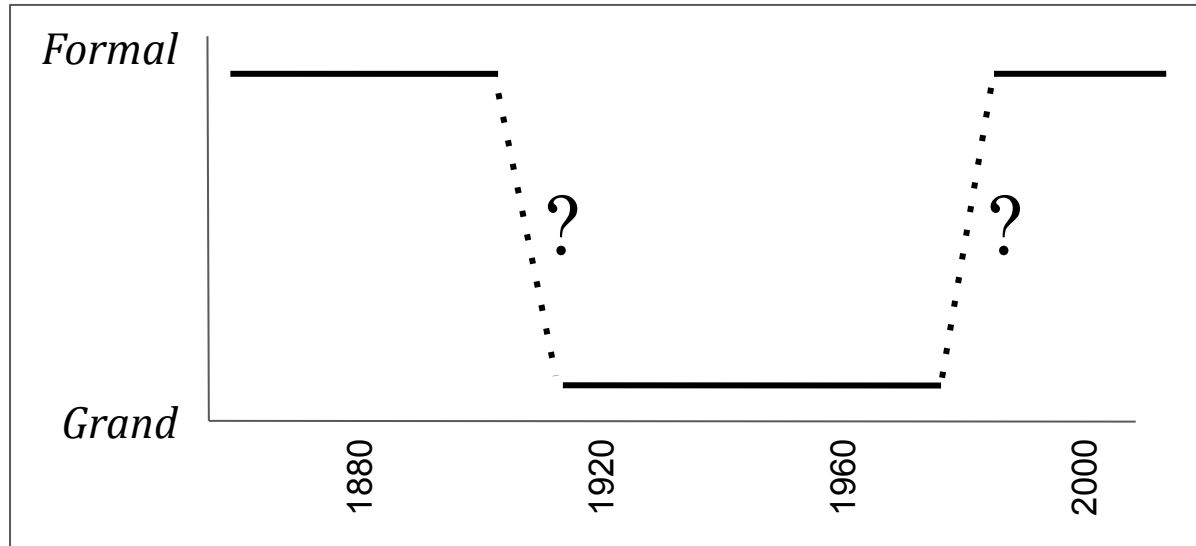


Qualitative legal scholars have identified historical periods of legal reasoning based on reading cases:

Civil War to WWI → **Formal**

WWI to 1980s → **Grand**

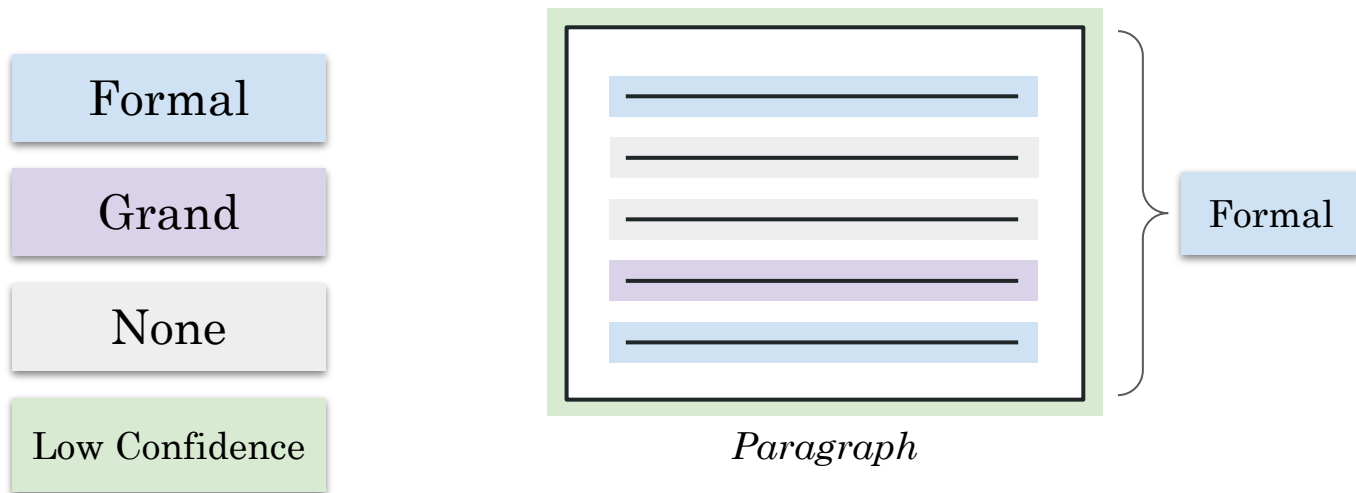
1980s to today → **Formal**



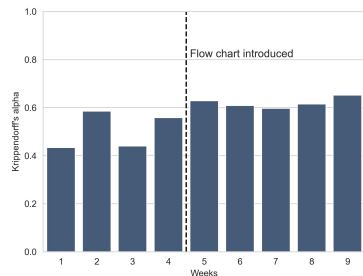
To identify the legal reasoning used in every modern United States Supreme Court opinion (1870-2023) we:

1. **Annotate** a representative sample of court opinion paragraphs for legal reasoning
2. **Fine-tune** or **prompt** language models to identify legal reasoning
3. **Compare** fine-tuned and prompted model performance
4. **Predict** the legal reasoning of *all* modern Supreme Court opinions

Legal Reasoning Annotation



Data Collection



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

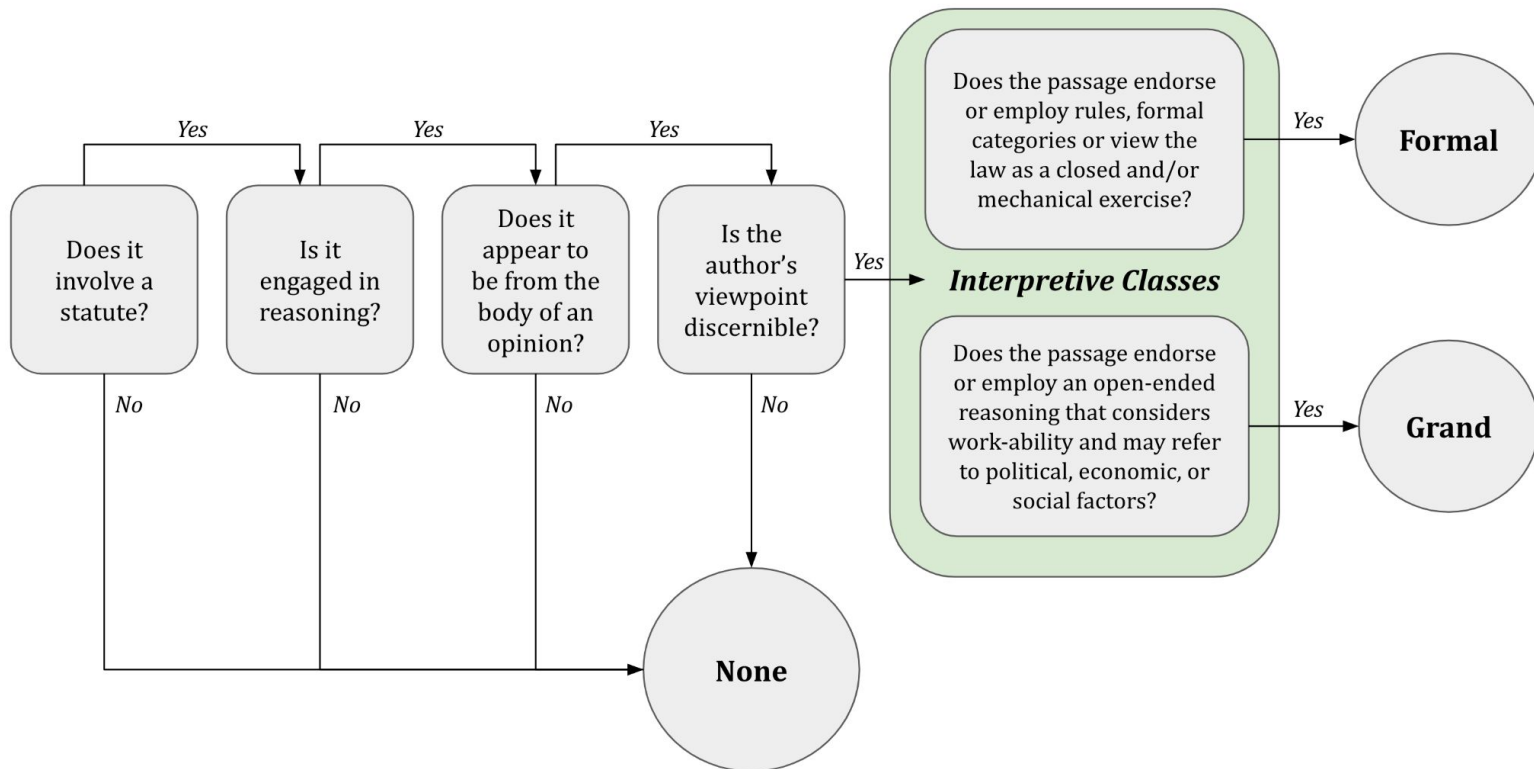
Table 2: Number of paragraphs fitting each class identified by annotators, and number assigned the low confidence (LC) class by its initial annotator.

15,680 court opinions (1870-2023)

2,748 paragraphs were labeled as **formal**, **grand**, or **none** (i.e. no reasoning)

Paragraphs labeled with “low confidence” were deliberated until reaching a majority decision

Introduction of a decision chart improved IRR (Krippendorff’s alpha 0.63).



Models

Prompting {
Descriptions
Examples
Chain-of-Thought

GPT-4

FLAN-T5-large

Llama-2-Chat (7B)

Fine-tuning {
Multi-Class Classification
Nested Binary Classification

BERT-base

DistilBERT

Legal-BERT

T5-small

T5-base

Prompts

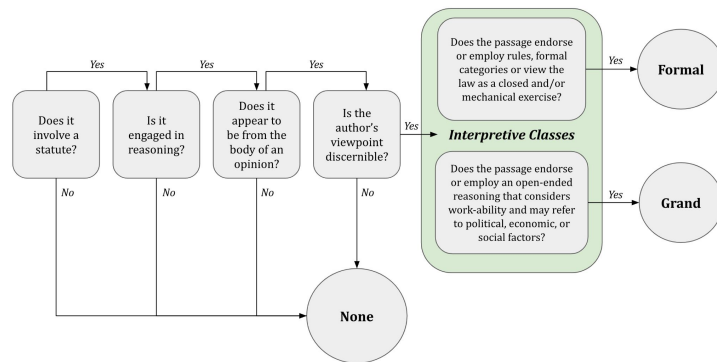
*It's virtually impossible to test all possible prompts. Here, we test only prompts that are **reasonable and helpful for human readers.***

Prompts

*It's virtually impossible to test all possible prompts. Here, we test only prompts that are **reasonable and helpful for human readers**.*

All prompts are directly drawn from our codebook for the legal expert annotators.

1. Prompt with class descriptions → codebook descriptions
2. Prompt with class examples → codebook exemplars
3. Chain-of-thought prompt → decision chart



Prompts

Model Input

Some paragraphs in court cases interpret statutes. Within interpretation, there are two types: GRAND and FORMAL.

FORMAL theory is 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 theory is a legal decision that views 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.

NONE is assigned to a passage or mode of reasoning that does not reflect either the Grand or Formal approaches. Note that this coding would include areas of substantive law outside of statutory interpretation, including procedural matters.

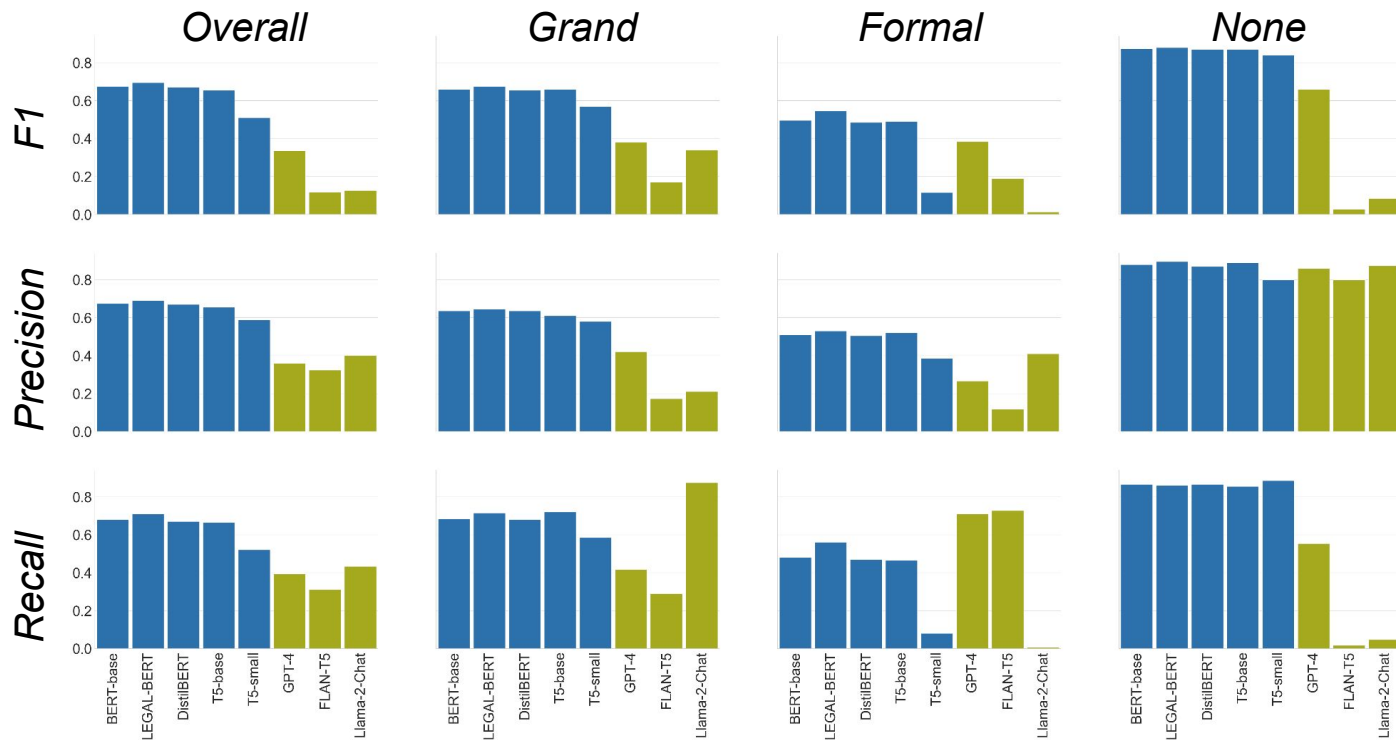
Determine the type of legal interpretation in the following passage. Return a single choice from GRAND, FORMAL, or NONE.

[TEXT FOR INFERENCE]

Model Output

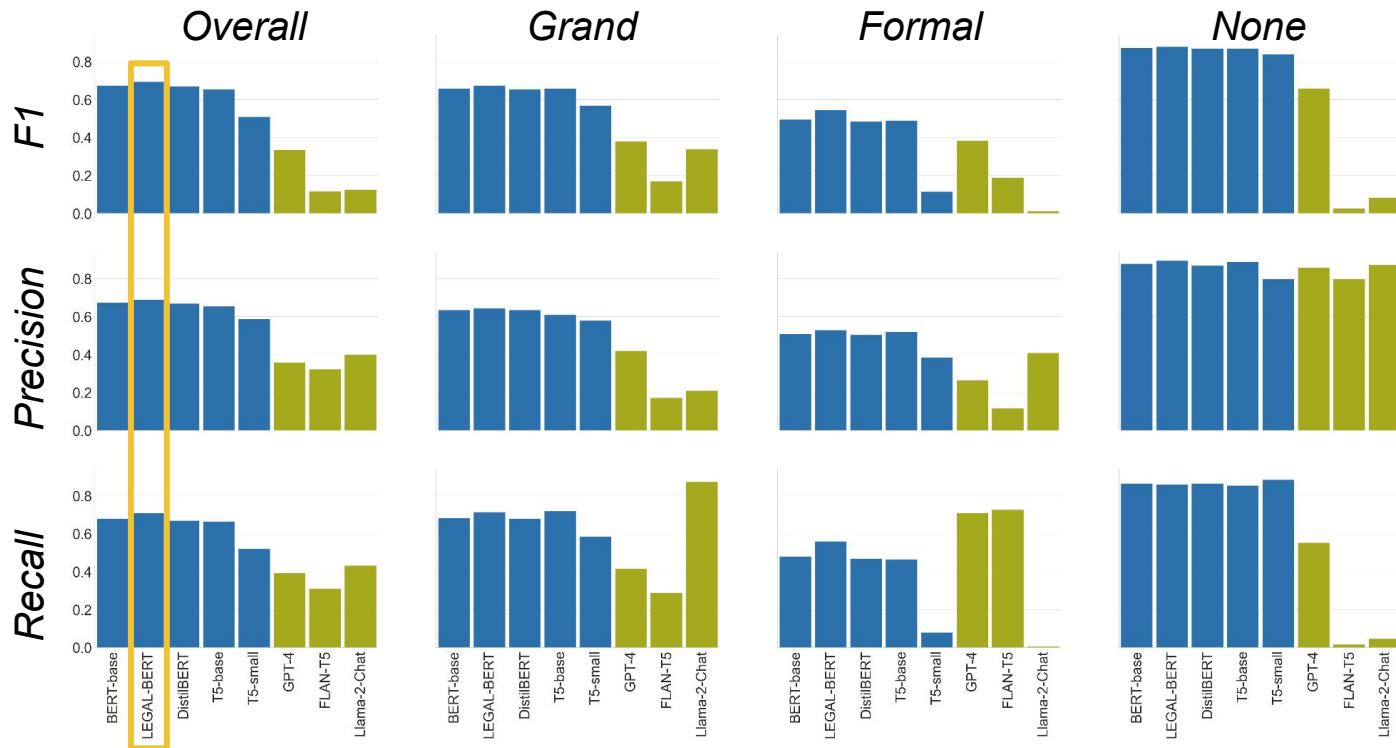
FORMAL

Without fine-tuning, LMs are virtually unusable for this task



A smaller domain-specific model,
Legal-BERT, performs best.

Fine-tuned
Prompted

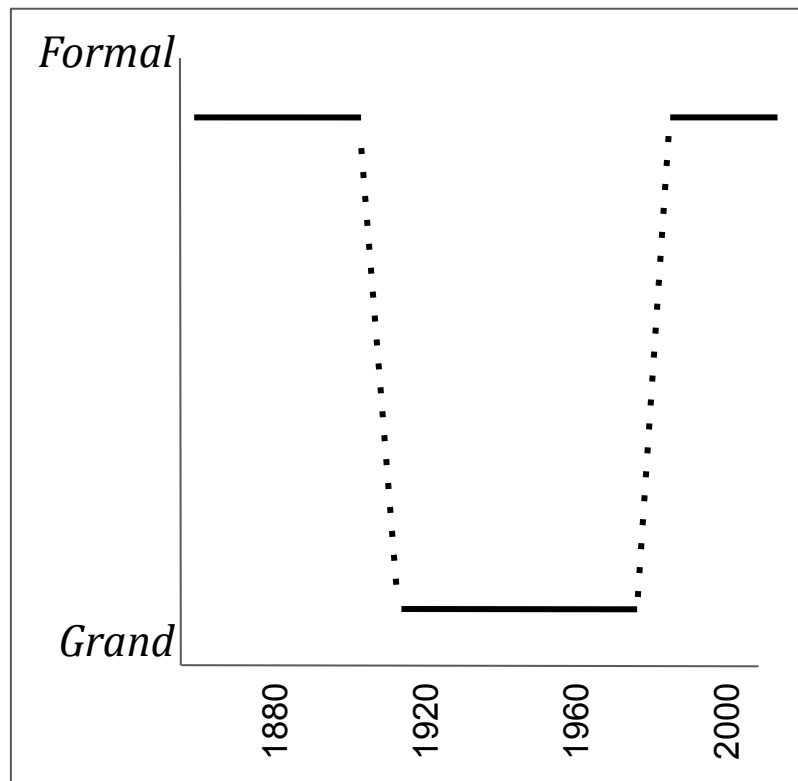


Why is identifying legal reasoning hard?

- It requires extensive legal **domain knowledge**
- Reasoning is **abstract** and is not based on specific keywords
- A case's legal reasoning can be **ambiguous** (even to experts)
 - E.g. a judge might critique the quality of another case's use of grand reasoning because they want to improve grand reasoning
- Court paragraphs are **long**



These traits make identifying legal reasoning difficult for humans *and* computational models.



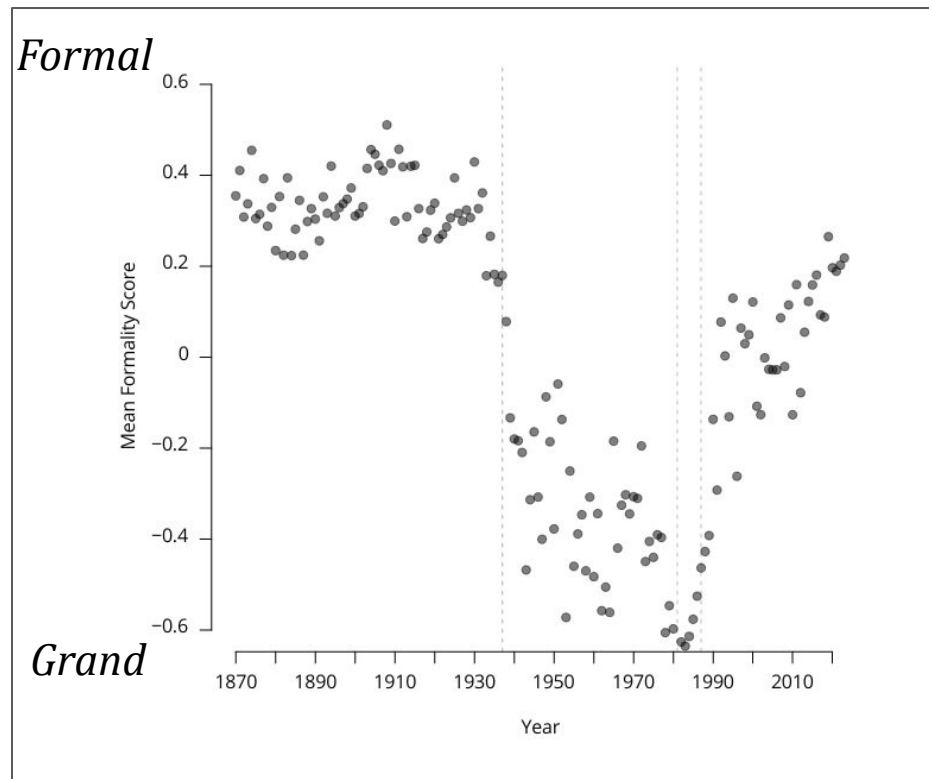
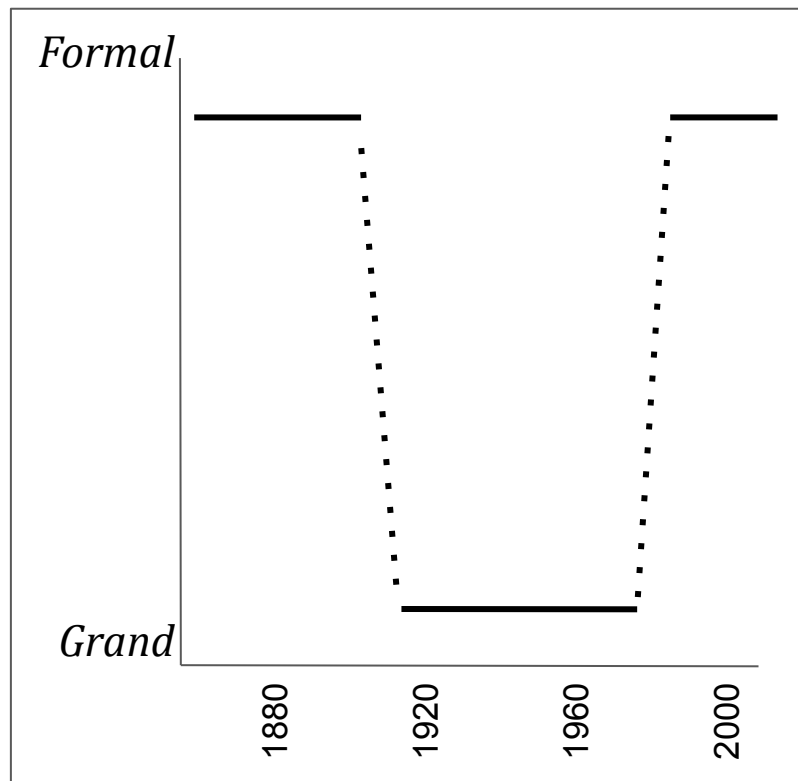
Historical expectations:

Civil War to WWI → **Formal**

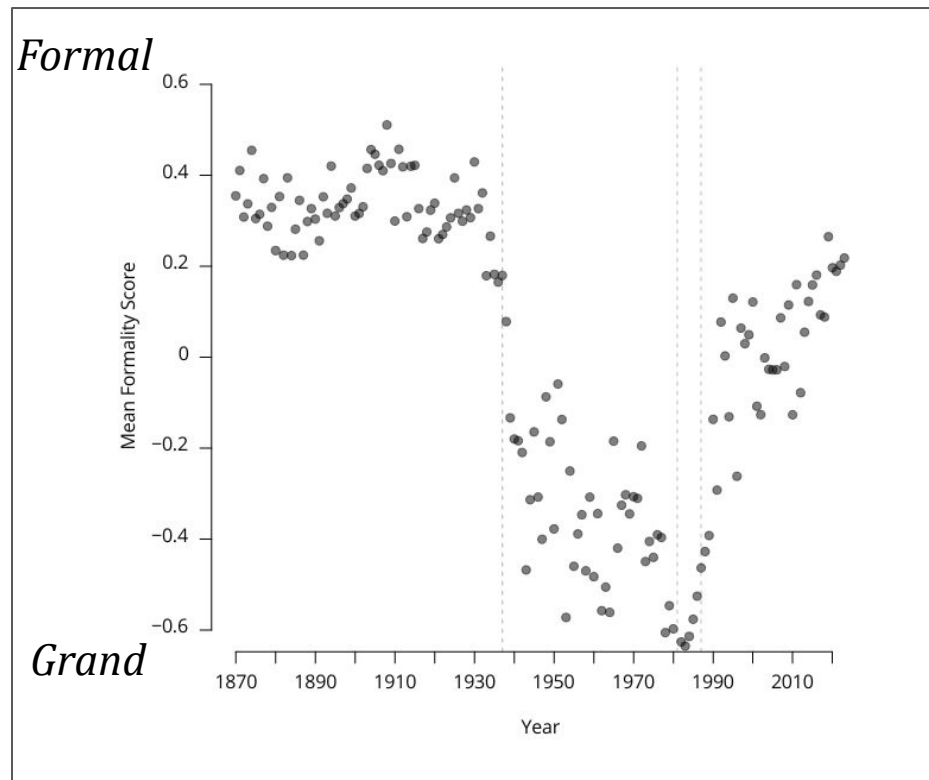
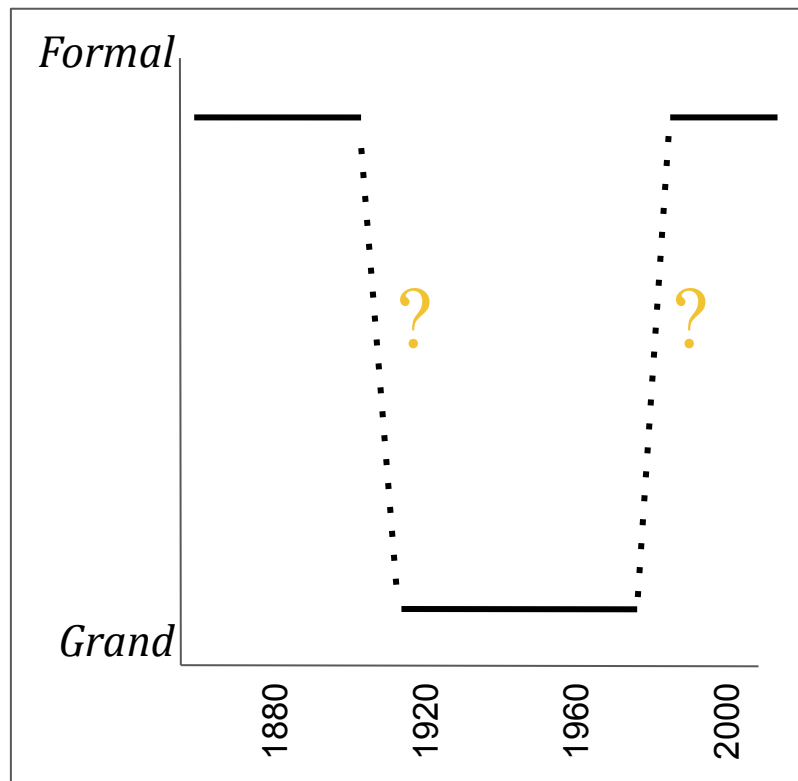
WWI to 1980s → **Grand**

1980s to today → **Formal**

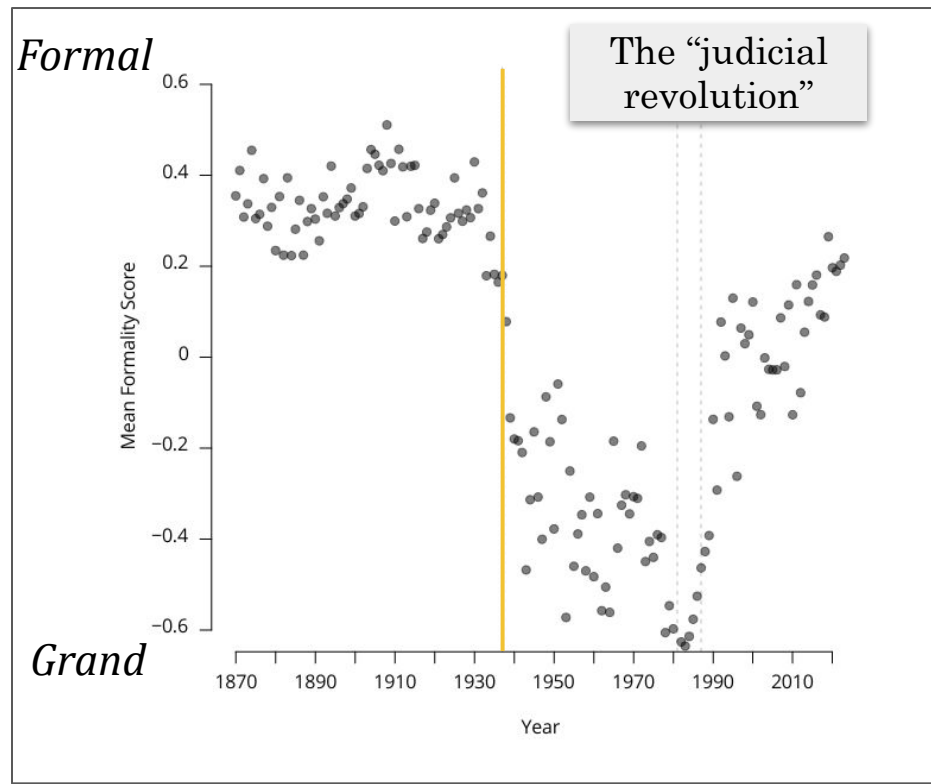
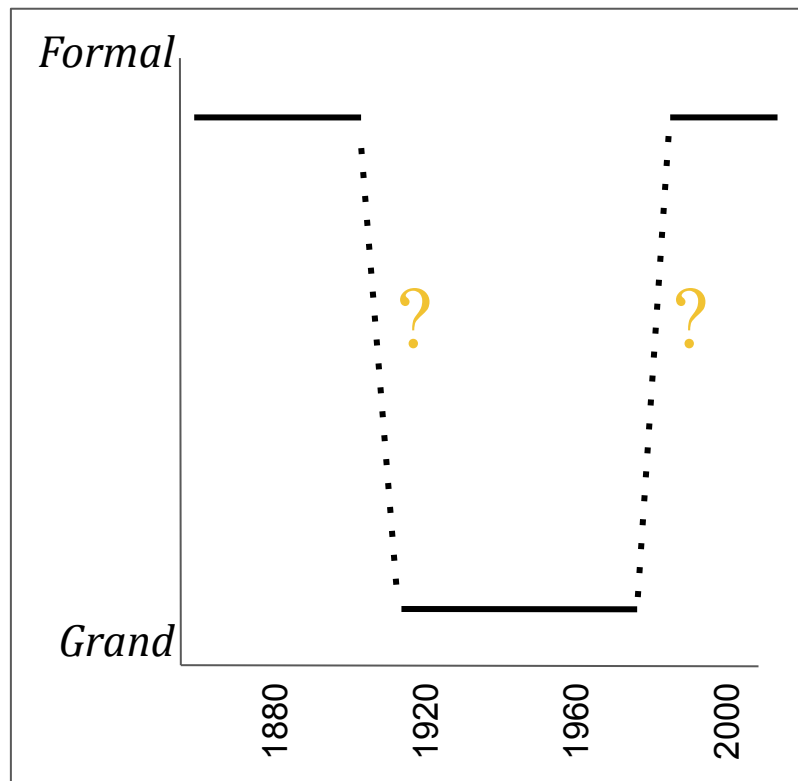
Our model's predictions align with legal scholars on general periods of legal reasoning.



Our model's predictions align with legal scholars on general periods of legal reasoning.



We find clearer turning points in the predominant legal reasoning.



LMs allow us to perform complex document classification tasks, at scale.

While handing annotation over to generative models may be appealing, researchers should proceed with caution.

Identifying legal reasoning represents a best-case language modeling scenario:

- Language models are overwhelmingly built for English
- United States court opinions are included in most pre-training data
- Natural legal language processing is a thriving field
 - There is a domain-specific model built just for English language law!
- Three exclusive classes for prediction

Takeaways

- Qualitative + quantitative methods align on periods of legal reasoning; quantitative methods add detail about turning points.
- Bigger models are not always better.
- What is helpful to human annotators is not always helpful to language models.
- Take caution when using LLMs without fine-tuning.
- Expert annotations remain a critical component of document classification.

Takeaways

- Qualitative + quantitative methods align on periods of legal reasoning; quantitative methods add detail about turning points.
- Bigger models are not always better.
- What is helpful to human annotators is not always helpful to language models.
- Take caution when using LLMs without fine-tuning.
- Expert annotations remain a critical component of document classification.

Thank you!!!