When Do LLMs Need Human Experts? Evidence for Social Science from Jurisprudential Classification

Caroline Cheng [1] **Edward H. Stiglitz** [2] **David Mimno** [1] **Matthew Wilkens** [1]
[1] Cornell University [2] Cornell Law School
{cyc59, js2758, mimno, wilkens}@cornell.edu

Abstract

Social scientists increasingly use large language models (LLMs) to classify text at scale, raising a key question: when can LLMs replace expert human annotation? Prior work found that earlier generative models failed on complex social science tasks while fine-tuned BERT succeeded, but whether current frontier-scale models close this gap remained untested. We investigate this question on a challenging legal reasoning task—classifying paragraphs from U.S. Supreme Court opinions as employing formal, grand, or no reasoning. Testing frontier LLMs including GPT-5.2 and leading open-weight alternatives, we find that even the most capable prompted models consistently underperform fine-tuned BERT. Only when high-parameter-count generative LLMs are fine-tuned on human-annotated training data does performance improve, and fine-tuned BERT remains a cost-effective alternative. Contrary to a common view, our results demonstrate that scaling to frontier-size LLMs does not eliminate the need for expert annotation on tasks requiring deep domain expertise—a finding with important implications for computational social science measurement.

1 Introduction

Social scientists increasingly rely on LLMs to classify text at scale, replacing expensive human coding (Ziems et al., 2024). But for tasks requiring deep domain expertise, can LLMs replace expert annotators? This question has direct implications for measurement validity: if LLM classifications diverge from expert judgments on complex constructs, downstream social science inferences may be biased (Egami et al., 2023).

Recent work finds that zero-shot LLM performance on social science coding tasks can be remarkably low, and that supervised fine-tuning on human-labeled data substantially improves smaller open-weight models (Halterman and Keith, 2025). However, these studies tested early or small models; whether frontier-scale LLMs close the gap with fine-tuned domain models is untested.

We conduct an up-to-date survey of the performance of LLMs against a challenging legal reasoning benchmark established in Thalken et al. (2023): classifying United States Supreme Court opinions as employing "formal" reasoning, "grand" reasoning, or no legal reasoning.[1] This task requires understanding of jurisprudential philosophy, rather than merely common or surface-level legal knowledge, making it a strong test for whether LLMs can match expert human annotators on domain-specific measurement tasks common to social science research. Similar to Halterman and Keith (2025), Thalken et al. (2023) found that generative LLMs perform poorly on their task without human annotation and instead observed strong performance with lightweight fine-tuned, in-domain BERT models.

Yet results as in Thalken et al. (2023) must be viewed as a snapshot, and a long-standing view is that models will outpace humans through scaling laws and greater application of compute (Kaplan et al., 2020). We now use this difficult legal benchmark to re-evaluate the question of when LLMs need human experts using more advanced frontier models, including OpenAI's most recent "reasoning" model.

We find that: (1) more recent in-domain fine-tuned models perform comparably to the original in-domain BERT models on this task; (2) prompted SOTA in-context LLMs continue to under-perform fine-tuned BERT models; (3) only fine-tuned SOTA generative LLMs—trained on human annotated samples—surpass the BERT baseline, with GPT-4.1 achieving the strongest performance.

Our findings illustrate that in a highly complex and specialized domain, scaling to frontier-size rea-

[1]The data in Thalken et al. (2023) later became the foundation of several social science papers, including Stiglitz and Thalken (2026).

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 103–112

soning LLMs does not obviate the need for human annotation, and fine-tuned BERT models remain a competitive and cost-effective alternative for social scientists.[2]

2 Related Work

LLMs for Social Science Measurement. A growing body of work examines whether LLMs can serve as reliable measurement tools for social science. Ziems et al. (2024) find that LLMs generally fail to outperform fine-tuned models on CSS classification tasks. Halterman and Keith (2025) report zero-shot LLM F_1 as low as 0.21 on political science tasks, and show that supervised fine-tuning on human-labeled data substantially improves 7–12B parameter open-weight models. They do not compare against BERT-like encoder baselines or examine frontier reasoning models. Egami et al. (2023) show that modestly inaccurate LLM labels can produce biased downstream statistical inference without correction. Chae and Davidson (2025) find that fine-tuning smaller models is competitive with zero-shot large models. Pangakis and Wolken (2025) test GPT-4 on a variety of CSS tasks and conclude that human annotation is essential. However, on social media classification, Törnberg (2025) find that GPT-4 outperforms expert human coders and supervised classifiers; see also Gilardi et al. (2023). Combined, LLM performance may turn heavily on the degree to which the task requires expert domain knowledge.

Legal NLP and This Task. There has been significant recent progress in legal NLP (Siino et al., 2025), and LLMs show strong performance on both conventional human benchmarks, such as bar exams, as well as curated-task benchmarks (Guha et al., 2023). Classifying legal philosophy, however, requires deeper expertise than many legal tasks; even the human experts in Thalken et al. (2023), which established the benchmark, disagreed sometimes on the correct classification. They found that prompt-based generative models, including GPT-4, were out-performed by lightweight fine-tuned BERT models (Thalken et al., 2023).

Benchmark Validity. On published legal benchmarks such as LegalBench[3] and BigLaw Bench,[4] SOTA generative LLMs approach perfect performance. However, with the capacity for pre-training LLMs unclear, there is concern that performance is increasing due to training on the benchmarks (Ni et al., 2025; Li and Flanigan, 2024). We evaluate on a recent, specialized, expert-annotated dataset less likely to have been included in pre-training corpora.

3 LLMs on Jurisprudential Classification

Dataset. Our dataset, established by Thalken et al. (2023), consists of paragraphs from U.S. Supreme Court opinions annotated by domain experts as containing formal reasoning, grand reasoning, or neither.[5] These categories derive from jurisprudential philosophy (Llewellyn, 1960), requiring annotators to distinguish between modes of legal reasoning rather than surface legal features. Inter-annotator agreement measured by Krippendorff's α reached 0.65 in annotation sessions, reflecting the genuine difficulty of this classification task; the original annotation procedure used an iteratively-developed codebook, a decision chart that systematically improved agreement (Thalken et al., 2023, Figs. 1 and 4). This moderate agreement establishes a human performance ceiling that contextualizes model results. The annotated corpus contains 2,748 paragraphs (329 formal, 551 grand, 1,869 none), with seeds used to sample paragraphs and "none" intentionally oversampled to account for the heterogeneity of paragraphs not engaged in legal reasoning (Thalken et al., 2023).

Several pieces of external evidence support data and construct validity: predictions from a model trained on labels for these randomly selected paragraphs recover the consensus historical periodization in legal scholarship (Thalken et al., 2023; Stiglitz and Thalken, 2026, 2024); the justice-level aggregations of predictions reflect common views about justices' jurisprudence (Stiglitz and Thalken, 2026, 2024); predictions of jurisprudence correlate with the partisanship of the authoring justice in expected ways (Thalken and Stiglitz, 2026); famous historical episodes of changes in jurisprudence of justices show up in the predictions (Stiglitz and Thalken, 2024).

[2] Code is available at: https://github.com/caroline-y-cheng/llms-legal-reasoning. We do not test retrieval-augmented generation, large-scale few-shot regimes, or agentic tool-use approaches, any of which might, in the right configuration, narrow the gap. Cost, latency, and data-governance constraints common in applied social science motivate our focus on lightweight prompting and small-to-mid-scale fine-tuning.

[3] https://www.vals.ai/benchmarks/legal_bench

[4] https://www.harvey.ai/blog/gemini-3-pro-public-preview-early-access-evaluation-results

[5] See Thalken et al. (2023) for full dataset description.

We test LLMs on the task of distinguishing types of legal reasoning between formal and grand (or none) classes.[6] We compare performance between sets of in-domain fine-tuned models, a new set of prompted generative LLMs, and a set of supervised fine-tuned generative LLMs. We chose models based on performance on other legal benchmarks and accessibility (i.e., open-weight models and model size), being cognizant of resources needed for applied researchers to fine-tune and perform inference with extremely large and expensive models. In-domain fine-tuned models include LEGAL-BERT (Chalkidis et al., 2020),[7] LEGAL-RoBERTa,[8] and LEGAL-ModernBERT.[9] Generative LLMs prompted to identify legal reasoning include GPT-OSS (120B) (OpenAI, 2025a), GPT-4.1 (OpenAI, 2023), GPT-5.2 (OpenAI, 2025b), Llama-3.1-Instruct (8B and 70B) (Grattafiori et al., 2024), Qwen3-Instruct-2507 (4B), Qwen3-A3B-Instruct-2507 (30B), Qwen3-Next-A3B-Instruct (80B) (Yang et al., 2025), and DeepSeek-V3.1 (DeepSeek-AI, 2025). A subset of these LLMs was fine-tuned.

We created five stratified splits of the annotated data with 80% of the data in the training set and 20% of the data in the test segment. The in-domain fine-tuned and generative in-context models were evaluated over five splits, while the fine-tuned generative models were evaluated over one split due to model size and cost.

3.1 In-Domain BERT Models

For the task of identifying *types* of legal reasoning in text, the strongest results in Thalken et al. (2023) derived from the procedure of fine-tuned multi-class classification based on hand-labeled annotations. BERT can be fine-tuned on the GPU of a reasonably equipped recent laptop. Following this approach, we again observe similar performance among LEGAL-BERT, LEGAL-RoBERTa, and LEGAL-ModernBERT (Table 1).

[6]We also examine performance on a simpler binary task: classifying whether the passage engages in legal reasoning (of any form, either formal or grand) or not. The results from this exercise tend to support those from the more difficult multi-class problem. We report the results from this binary task in Table B.2.

[7]In Thalken et al. (2023), the strongest results derived from fine-tuning LEGAL-BERT on the annotated dataset.

[8]https://huggingface.co/Saibo-creator/legal-roberta-base

[9]https://free.law/2025/03/11/semantic-search/

3.2 Prompted Generative LLMs

To establish a baseline for the more recent generative LLMs, we test their performance when simply given instructions equal to those presented to our human annotators. In our multi-class classification task, we explore three prompting strategies: *descriptions* of each legal-reasoning class (zero-shot); *examples*, a 3-shot prompt with one canonical paragraph per class drawn from the codebook (Appendix D); and *chain-of-thought* (CoT), which walks through the annotation decision chart (Appendix C). The three in-context examples are fixed across test items. Full prompts appear in Appendix A.

3.3 Supervised Fine-tuned Generative LLMs

We determined a subset of the new set of generative LLMs to fine-tune based on their prompt-based performance and accessibility. GPT-4.1 is OpenAI's most powerful proprietary model that can be fine-tuned via the OpenAI API. Qwen3-A3B-Instruct-2507 (30B) resulted in higher macro-averaged F_1 scores than Qwen3-Next-A3B-Instruct (80B) in the multi-class task under each of the prompting strategies (Table B.1; Table 1) and was more feasible to fine-tune. DeepSeek-V3.1 was consistently outperformed by other more feasibly trained models on the baseline measurements (Table 1).

For fine-tuning, we prepared one split of the annotated dataset as prompt-completion examples:

- *Prompt*: Each of the three multi-class prompting strategies (Appendix A) followed by a paragraph to classify.
- *Completion*: Domain-expert classification of the paragraph.

We fine-tuned GPT-4.1 with the OpenAI API[10] with OpenAI autoconfigs. The Llama (8B and 70B) and Qwen (4B and 30B) models were fine-tuned using 4-bit quantized models and parameter-efficient techniques (Dettmers et al., 2023). They were fine-tuned in 200 steps, with a 10% warm-up ratio, a maximum learning rate of 3e-5, with a weight decay of 0.01.

4 LLM Performance

First, we show that more recent in-domain fine-tuned BERT models have comparable performance differences on this task to the older LEGAL-BERT model (Table 1).

[10]https://openai.com/api/

Model	Strategy	Macro F1
Fine-Tuned BERT (5-split avg.)		
LEGAL-BERT	–	0.71
LEGAL-RoBERTa	–	0.71
LEGAL-ModernBERT	–	0.70
Prompted (5-split avg., best strategy)		
GPT-5.2	Examples	0.68
GPT-4.1	Examples	0.64
GPT-OSS (120B)	Examples	0.59
Qwen3-A3B (30B)	Examples	0.53
DeepSeek-V3.1	CoT	0.52
Qwen3-Next-A3B (80B)	Examples	0.48
Llama-3.1 (8B)	CoT	0.43
Qwen3 (4B)	Examples	0.40
Llama-3.1 (70B)	Desc.	0.36
Fine-Tuned Generative (single split, best strategy)[†]		
GPT-4.1	CoT	**0.79**
Llama-3.1 (70B)	CoT	0.76
Qwen3-A3B (30B)	CoT	0.70
Qwen3 (4B)	Examples	0.69
Llama-3.1 (8B)	CoT	0.65

Table 1: Multi-class macro F_1 summary. Best prompting strategy shown per model. [†]Single-split results; see Limitations. Full per-class results in Appendix Table B.1.

We then tested the performance of the new set of LLMs with various prompting strategies on the same splits of data and find that without fine-tuning, they all perform worse than the in-domain fine-tuned BERT models.[11] On our multi-class task, the best performing prompted LLM is GPT-5.2 when given examples of the legal reasoning classes, whose macro F_1 score was 0.68 compared to LEGAL-BERT's 0.71. Llama-3.1-Instruct (8B and 70B) without task-specific fine-tuning very rarely predicted the "none" class (Table 1).[12] Our results suggest that without fine-tuning, generative LLMs continue to fall short of alignment with human annotation on highly complex and specialized classification tasks.

The fine-tuned LLMs all performed better than their non-fine-tuned versions within the same prompting strategy. However, SFT gains are scale-dependent: fine-tuned 8B and 4B models still underperform BERT, while only GPT-4.1 and 70B-class models surpass it on the evaluated split. On the multi-class classification task, fine-tuned GPT-4.1 consistently outperformed the in-domain fine-tuned models: the macro F_1 for GPT-4.1 with CoT prompt-completion examples is 0.79, 0.08 higher than the best macro F_1 for the in-domain fine-tuned models (Table 1; Table B.1). Fine-tuned Llama-3.1-Instruct (70B) and Qwen3-A3B-Instruct-2507 (30B) sometimes outperformed the in-domain fine-tuned models (Table B.1). Because notable performance gains require fine-tuned GPT-4.1 or larger open-weight LLMs trained on expert-annotated data, fine-tuned BERT remains a competitive and substantially more cost-effective alternative.

Per-class results (Appendix Table B.1) reveal that the models have the most trouble with the Formal and Grand classes; the None class is relatively easy for the models to detect. This may be because the reasoning classes reflect specialized domain-specific usage of common terms (Schauer, 1987). Even where aggregate scores converge, error patterns differ: GPT-5.2 matches BERT's Formal F_1 (0.56) but with lower precision (0.48 vs. 0.57) and higher recall (0.70 vs. 0.56), indicating it over-predicts formal reasoning, likely triggered by surface legal language rather than the underlying concept and method of reasoning. For CSS measurement, such systematic classification errors could bias downstream inference (Egami et al., 2023). Fine-tuning produces the largest gains on these difficult legal reasoning classes: fine-tuned GPT-4.1 CoT improves Formal F_1 to 0.68 (+0.16 over best GPT-4.1 without FT), Grand to 0.77 (+0.20), and None to 0.92 (+0.08), suggesting expert-annotated training data is most valuable where the classification requires the deepest domain knowledge. The models continue to make errors with fine-tuning, but at lower rates, and the errors tend to be more balanced between precision and recall, reducing systematic over- or under-prediction of any single class.

5 Conclusion

Extending Thalken et al. (2023) with frontier-scale models, we find that scaling alone does not solve this task: even GPT-5.2, a cutting edge reasoning model, underperforms fine-tuned BERT when prompted, and only fine-tuned generative LLMs surpass the BERT baseline on the evaluated split. This contrasts with tasks where frontier LLMs match or exceed human coders (Törnberg, 2025), suggesting that the need for human annotation turns

[11] Our reported results employ a user-role and zero temperature for classification.

[12] Some models often return additional text beyond the class label; we extract the class label (if it occurs) from the text and use that as the label for evaluation. Also, just prompting the GPT-OSS and DeepSeek models often generate responses that do not contain a label in our support; we calculate the classification reports based only on the predictions within the set of true labels in our codebook.

on the depth of domain expertise a task requires. Fine-tuned BERT remains a competitive and substantially more cost-effective alternative to fine-tuned frontier LLMs on this task. For CSS researchers working on tasks requiring deep domain expertise, our findings add to growing evidence (Ziems et al., 2024; Halterman and Keith, 2025; Gu et al., 2025) that investing in high-quality human annotation yields greater returns than relying on increasingly capable models alone. We reveal limitations in SOTA LLMs' capabilities to align with human reasoning, emphasizing the continued importance of domain-expert annotation.

Limitations

The prompting strategies explored for LLM baseline performance and supervised fine-tuning followed those of Thalken et al. (2023), which are limited to the scope of the instructions provided to human annotators. We did not evaluate retrieval-augmented generation, dynamic few-shot example selection, self-consistency or majority-vote schemes, or multi-agent or tool-using setups. Any of these could in principle narrow or close the gap to fine-tuned models, and exploring how much expert supervision can be substituted by stronger inference-time procedures is a natural next step.

In addition, this up-to-date survey of LLMs was limited by model size and cost. We prioritize open-weight models, 4-bit quantized models, and models that can be fine-tuned on a single H100 GPU due to resource constraints. Better performance on this task may be possible with emerging models. For the same cost-based reasons, we limit our study to a single split of the fine-tuned generative models; comparative claims between fine-tuned generative and BERT models should be interpreted with this caveat. We also do not vary the labeled-data budget for the same reason of cost. Prior work documents that prompting can substitute for hundreds of fine-tuning examples on some classification tasks (Le Scao and Rush, 2021); scaling the labeled data budget would allow us to identify the threshold at which prompted frontier LLMs become competitive with fine-tuning. This would be a productive next-step in this analysis.

The Krippendorff $\alpha = 0.65$ inter-annotator agreement is moderate and bounds the absolute level of model–human alignment achievable on this task. This ceiling reflects the genuine difficulty of distinguishing modes of jurisprudential reasoning. Though comparisons across models are evaluated against the same labels, so team-specific annotation conventions cannot drive the relative rankings reported here, absolute F_1 values should be read against the ceiling of human agreement.

Our findings are based on the study of a single domain, jurisprudential philosophy in Supreme Court opinions. Other CSS domains may show different patterns.

Ethics Statement

No human subjects were involved in this study beyond the annotation described in Thalken et al. (2023). Our data is in the public domain.

Acknowledgments

Many thanks to Rosamond Thalken and the team for establishing the foundation for the continuation of this work and to Prof. Stiglitz, Prof. Mimno, and Prof. Wilkens for their guidance and support.

References

Youngjin Chae and Thomas Davidson. 2025. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

DeepSeek-AI. 2025. Deepseek-v3 technical report.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. 2023. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*, pages 1–17.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Teven Le Scao and Alexander M Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.

Karl N. Llewellyn. 1960. *The Common Law Tradition: Deciding Appeals*. W.S. Hein, Buffalo, NY.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2025. Training on the benchmark is not all you need. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24948–24956.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2025a. gpt-oss-120b & gpt-oss-20b model card.

OpenAI. 2025b. Openai gpt-5 system card.

Nicholas Pangakis and Samuel Wolken. 2025. Keeping humans in the loop: Human-centered automated annotation with generative AI. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1471–1492.

Frederick Schauer. 1987. Formalism. *Yale Lj*, 97:509.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13:18253–18276.

Edward H Stiglitz and Rosamond Thalken. 2024. Historical trends in macro-jurisprudence: A language model assessment, 1870-2023. *Md. L. Rev.*, 84:46.

Edward H. Stiglitz and Rosamond Thalken. 2026. Undertanding change in jurisprudence. *Journal of Law, Economics, and Organization*.

Rosamond Thalken, Edward Stiglitz, David Mimno, 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*, pages 9252–9265, Singapore. Association for Computational Linguistics.

Rosamond Thalken and Edward H Stiglitz. 2026. Measuring jurisprudence. *Journal of Law and Courts*, pages 1–22.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*.

An Yang et al. 2025. Qwen3 technical report.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Appendices

A Prompts

Our approaches to prompting LLMs to identify legal reasoning and legal reasoning classes in text:

- *In Context, Descriptions*: An in-context prompt that provides the model with descriptions of the legal reasoning classes before asking for inference on a new paragraph.
- *In Context, Examples*: An in-context prompt that provides the model with one example of each of the three legal reasoning classes before asking for inference on a new paragraph.
- *Chain-of-Thought*: A CoT prompt that provides the model with steps of reasoning to follow in order to determine the class of legal reasoning before asking for inference on a new paragraph.

These prompts are included in Figure A.1. Each prompting strategy is derived from our codebook (Appendix D) or decision chart (Appendix C).

B Full Results

Table B.1 contains the full results for the multi-class task (formal reasoning versus grand reasoning versus no interpretation). Table B.2 contains the results for the simpler binary task (interpretation versus no interpretation).

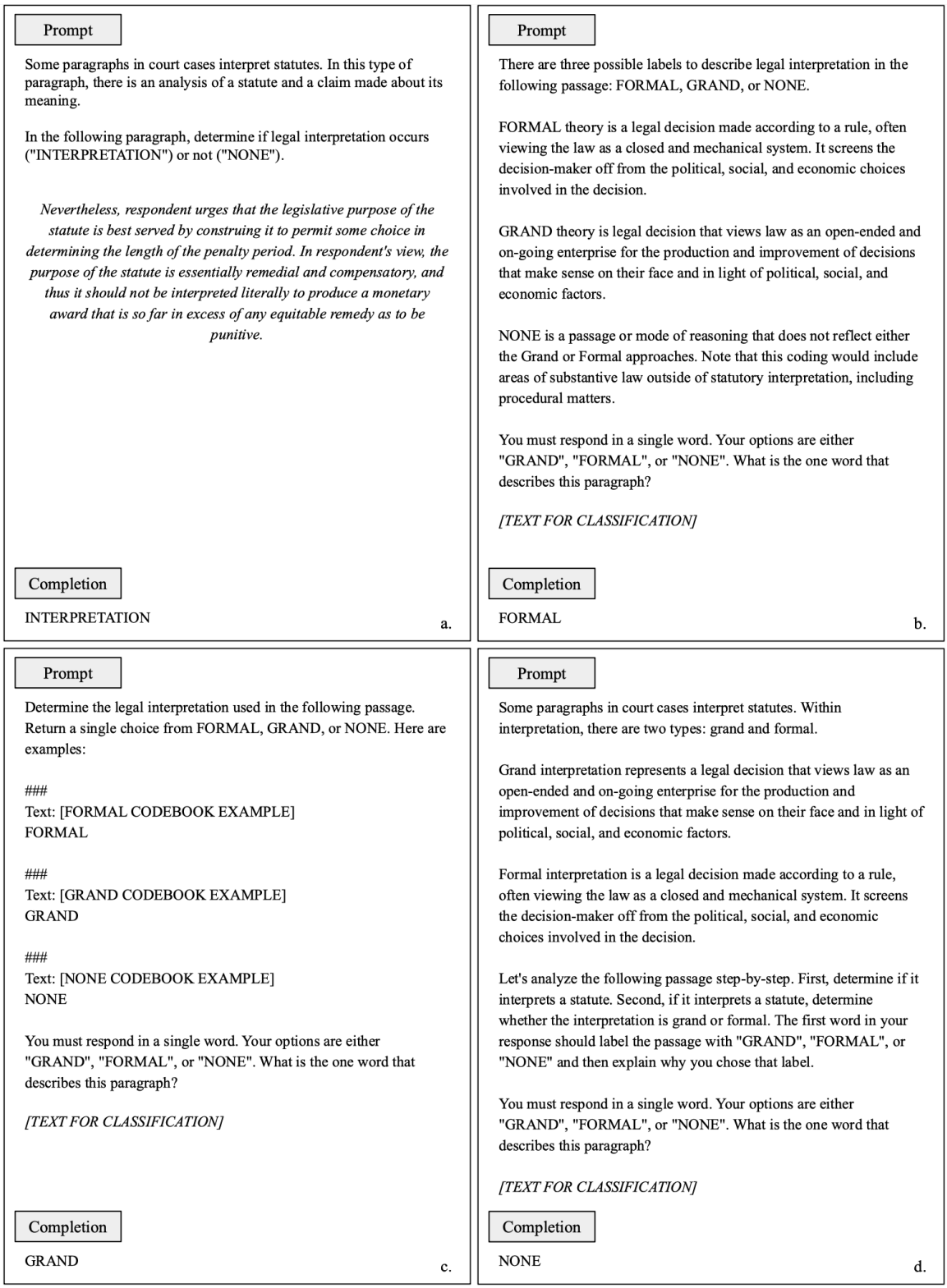

Figure A.1: Prompt a is the prompt used for identifying whether legal interpretation occurs or not. Prompt b is the prompt used for description classification of the classes of legal interpretation. Prompt c is the prompt used for few-shot classification of the classes of legal interpretation. Prompt d is the prompt used for CoT reasoning and the classes of legal interpretation.

C Decision Chart

Figure C.1 presents the decision chart provided to the annotation team. It was the basis of the CoT

Model	*Macro* F1	P	R	*Grand* F1	P	R	*Formal* F1	P	R	*None* F1	P	R
Multi-Class												
LEGAL-BERT	0.71	0.71	0.72	0.69	0.67	0.71	0.56	0.57	0.56	0.88	0.89	0.88
LEGAL-RoBERTa	0.71	0.70	0.72	0.69	0.65	0.73	0.56	0.55	0.58	0.88	0.90	0.86
LEGAL-ModernBERT	0.70	0.72	0.68	0.67	0.70	0.65	0.53	0.58	0.49	0.89	0.87	0.91
In-Context, Descriptions												
GPT-OSS (120B)	0.56	0.57	0.67	0.57	0.46	0.74	0.44	0.31	0.74	0.69	0.93	0.54
GPT-4.1	0.46	0.51	0.58	0.41	0.38	0.44	0.37	0.23	0.87	0.59	0.92	0.43
GPT-5.2	0.49	0.54	0.62	0.43	0.41	0.44	0.40	0.26	0.90	0.66	0.94	0.51
Llama-3.1-Instruct (8B)	0.38	0.43	0.48	0.21	0.28	0.16	0.32	0.20	0.82	0.60	0.82	0.47
Llama-3.1-Instruct (70B)	0.36	0.45	0.50	0.24	0.29	0.21	0.33	0.20	0.92	0.53	0.87	0.38
Qwen3-Instruct-2507 (4B)	0.38	0.46	0.51	0.39	0.27	0.72	0.36	0.27	0.56	0.39	0.85	0.26
Qwen3-A3B-Instruct-2507 (30B)	0.50	0.50	0.58	0.43	0.35	0.55	0.42	0.31	0.67	0.65	0.84	0.53
Qwen3-Next-A3B-Instruct (80B)	0.40	0.50	0.54	0.34	0.36	0.33	0.33	0.20	0.90	0.53	0.93	0.37
DeepSeek-V3.1	0.35	0.48	0.49	0.23	0.36	0.17	0.30	0.18	**0.95**	0.51	0.90	0.36
In-Context, Examples												
GPT-OSS (120B)	0.59	0.59	0.70	0.59	0.45	0.84	0.48	0.36	0.71	0.69	**0.95**	0.55
GPT-4.1	0.64	0.68	0.63	0.57	0.52	0.62	0.52	0.68	0.42	0.84	0.83	0.84
GPT-5.2	0.68	0.66	0.71	0.63	0.61	0.65	0.56	0.48	0.70	0.84	0.88	0.79
Llama-3.1-Instruct (8B)	0.17	0.40	0.30	0.23	0.14	0.58	0.24	0.20	0.30	0.05	0.86	0.02
Llama-3.1-Instruct (70B)	0.18	0.45	0.40	0.32	0.42	0.26	0.22	0.13	0.92	0.00	0.80	0.00
Qwen3-Instruct-2507 (4B)	0.40	0.51	0.50	0.42	0.27	**0.90**	0.32	0.36	0.28	0.48	0.90	0.33
Qwen3-A3B-Instruct-2507 (30B)	0.53	0.55	0.52	0.42	0.45	0.39	0.38	0.44	0.33	0.79	0.76	0.82
Qwen3-Next-A3B-Instruct (80B)	0.48	0.53	0.60	0.51	0.46	0.56	0.35	0.23	0.80	0.59	0.91	0.44
DeepSeek-V3.1	0.50	0.62	0.48	0.24	0.53	0.15	0.44	0.58	0.36	0.82	0.73	0.93
Chain-of-Thought												
GPT-OSS (120B)	0.56	0.56	0.66	0.56	0.48	0.68	0.42	0.30	0.71	0.70	0.91	0.58
GPT-4.1	0.51	0.52	0.59	0.43	0.39	0.47	0.41	0.28	0.75	0.69	0.88	0.56
GPT-5.2	0.49	0.55	0.60	0.31	0.47	0.23	0.40	0.26	0.91	0.77	0.93	0.65
Llama-3.1-Instruct (8B)	0.43	0.44	0.46	0.36	0.27	0.51	0.25	0.22	0.30	0.67	0.84	0.56
Llama-3.1-Instruct (70B)	0.33	0.47	0.48	0.24	0.31	0.20	0.30	0.18	0.94	0.46	0.91	0.31
Qwen3-Instruct-2507 (4B)	0.34	0.45	0.48	0.36	0.26	0.61	0.31	0.21	0.62	0.36	0.90	0.22
Qwen3-A3B-Instruct-2507 (30B)	0.42	0.49	0.54	0.44	0.32	0.69	0.33	0.23	0.62	0.48	0.92	0.33
Qwen3-Next-A3B-Instruct (80B)	0.41	0.49	0.53	0.40	0.33	0.52	0.32	0.21	0.74	0.50	0.92	0.34
DeepSeek-V3.1	0.52	0.56	0.59	0.28	0.50	0.19	0.49	0.36	0.79	0.80	0.82	0.78
Fine-Tuned, Descriptions												
GPT-4.1	0.73	0.74	0.74	0.75	0.71	0.80	0.55	0.59	0.51	0.90	0.91	0.90
Llama-3.1-Instruct (8B)	0.61	0.71	0.59	0.61	0.58	0.64	0.35	0.71	0.23	0.87	0.84	0.91
Llama-3.1-Instruct (70B)	0.66	0.81	0.60	0.60	0.79	0.49	0.48	**0.85**	0.34	0.88	0.80	**0.97**
Qwen3-Instruct-2507 (4B)	0.64	0.66	0.62	0.59	0.61	0.56	0.49	0.54	0.45	0.85	0.82	0.87
Qwen3-A3B-Instruct-2507 (30B)	0.68	0.72	0.66	0.70	0.71	0.69	0.48	0.59	0.40	0.88	0.85	0.91
Fine-Tuned, Examples												
GPT-4.1	0.76	0.78	0.75	0.76	0.75	0.77	0.62	0.69	0.55	0.91	0.90	0.93
Llama-3.1-Instruct (8B)	0.59	0.74	0.55	0.55	0.68	0.46	0.37	0.76	0.25	0.86	0.78	0.95
Llama-3.1-Instruct (70B)	0.74	0.76	0.72	0.75	0.75	0.74	0.57	0.65	0.51	0.90	0.88	0.92
Qwen3-Instruct-2507 (4B)	0.69	0.71	0.67	0.67	0.67	0.67	0.54	0.63	0.48	0.86	0.84	0.87
Qwen3-A3B-Instruct-2507 (30B)	0.73	0.77	0.70	0.70	0.74	0.68	0.60	0.72	0.51	0.88	0.85	0.92
Fine-Tuned, Chain-of-Thought												
GPT-4.1	**0.79**	**0.81**	**0.77**	**0.77**	**0.79**	0.75	**0.68**	0.73	0.63	**0.92**	0.90	0.94
Llama-3.1-Instruct (8B)	0.65	0.69	0.62	0.59	0.71	0.51	0.47	0.54	0.42	0.88	0.83	0.93
Llama-3.1-Instruct (70B)	0.76	0.77	0.76	0.75	0.76	0.74	0.64	0.65	0.63	0.90	0.90	0.91
Qwen3-Instruct-2507 (4B)	0.67	0.70	0.65	0.66	0.68	0.64	0.49	0.60	0.42	0.87	0.84	0.90
Qwen3-A3B-Instruct-2507 (30B)	0.70	0.73	0.68	0.68	0.71	0.65	0.56	0.64	0.49	0.88	0.85	0.91

Table B.1: Model performance averaged over five train test splits for fine-tuned and generative in-context models. Model performance on 1 train test split for fine-tuned generative models. Macro averages represent averages unweighted by class.

prompt for generative LLMs.

D Codebook

Table D.1 presents the codebook with definitions and core examples of each class, which guided annotators and was the basis of the in-context descriptions and in-context examples prompts for generative LLMs.

Model	*Macro* F1	P	R	*Interpretation* F1	P	R	*None* F1	P	R
Fine-Tuned									
LEGAL-BERT	0.82	0.81	0.83	0.76	0.72	0.80	0.88	0.90	0.86
LEGAL-RoBERTa	0.82	0.82	0.83	0.77	0.73	0.80	0.88	0.90	0.86
LEGAL-ModernBERT	0.82	0.82	0.81	0.75	0.77	0.73	0.89	0.88	0.90
Generative In-Context									
GPT-OSS (120B)	0.66	0.69	0.71	0.62	0.49	0.85	0.70	0.89	0.57
GPT-4.1	0.68	0.69	0.72	0.63	0.51	0.80	0.74	0.87	0.64
GPT-5.2	0.71	0.71	0.74	0.65	0.54	0.80	0.77	0.88	0.68
Llama-3.1-Instruct (8B)	0.60	0.64	0.65	0.56	0.44	0.78	0.64	0.84	0.53
Llama-3.1-Instruct (70B)	0.69	0.70	0.73	0.63	0.53	0.79	0.75	0.87	0.66
Qwen3-Instruct-2507 (4B)	0.58	0.65	0.65	0.57	0.42	0.86	0.59	0.87	0.45
Qwen3-A3B-Instruct-2507 (30B)	0.54	0.67	0.65	0.57	0.41	**0.94**	0.52	**0.93**	0.36
Qwen3-Next-A3B-Instruct (80B)	0.62	0.67	0.69	0.59	0.46	0.85	0.66	0.88	0.52
DeepSeek-V3.1	0.64	0.67	0.69	0.59	0.47	0.81	0.68	0.86	0.56
Generative Fine-Tuned									
GPT-4.1	**0.86**	**0.86**	**0.86**	**0.81**	0.82	0.80	**0.91**	0.91	0.92
Llama-3.1-Instruct (8B)	0.78	0.80	0.76	0.68	0.75	0.63	0.87	0.84	0.90
Llama-3.1-Instruct (70B)	0.78	0.83	0.76	0.68	**0.85**	0.56	0.88	0.82	**0.95**
Qwen3-Instruct-2507 (4B)	0.79	0.79	0.79	0.72	0.70	0.73	0.86	0.87	0.85
Qwen3-A3B-Instruct-2507 (30B)	0.81	0.81	0.81	0.75	0.75	0.74	0.88	0.88	0.88

Table B.2: Model performance for binary interpretation averaged over 5 train test splits for fine-tuned and generative in-context models. Model performance for binary interpretation on 1 train test split for fine-tuned generative models. Macro averages represent averages unweighted by class.

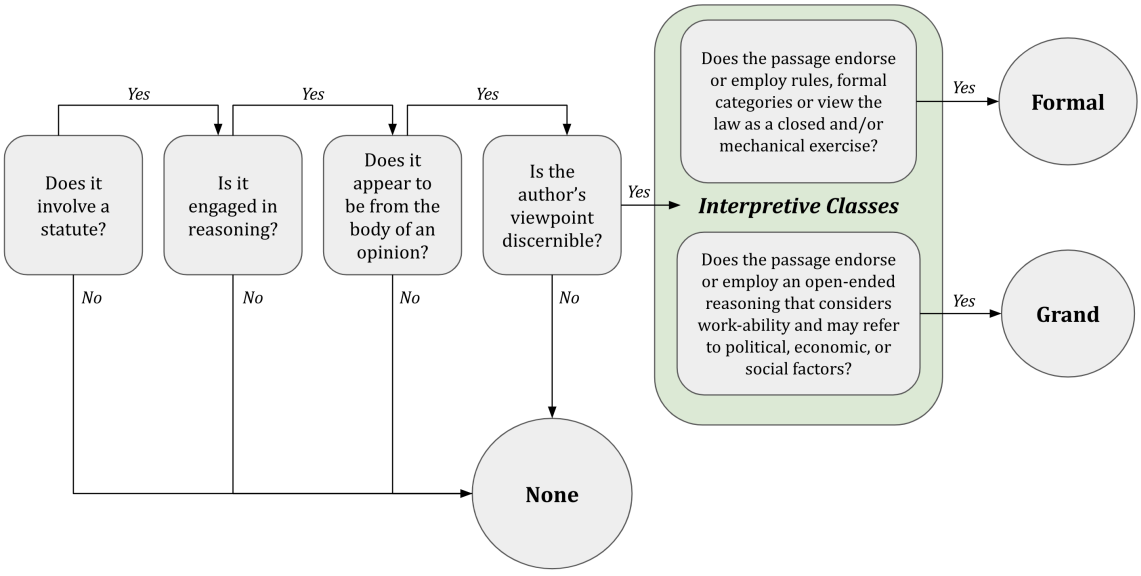

Figure C.1: The decision chart provided to annotators.

Class	Definition	Example
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.	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
Grand	A legal decision that views the law as an open-ended and ongoing enterprise for the production and improvement of decisions that make sense on their face and in light of political, social, and economic factors.	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. See [CITE]. Examination of those sources makes clear that an interpretation of the sections that forbade all race-conscious affirmative action would "bring about an end completely at variance with the purpose of the statute" and must be rejected. [CITE]. See [CITE]. Steelworkers v. Weber
None	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.	The questions are, What is the form of an assignment, and how must it be evidenced? There is no precise form. It may be. by delivery. Briggs v. Dorr, CITE, citing numerous cases; Onion v. Paul, 1 Har. & Johns. 114; Dunn v. Snell, CITE; Titcomb v. Thomas, 5 Greenl. 282. True, it is said it must be on a valuable consideration, with intent to transfer it. But these last are requisites in all assignments, or transfers of securities, negotiable or not. It may be by writing under seal, by writing without seal, by oral declarations, accompanied in all cases by delivery, and on a just consideration. The evidence may be by proof of handwriting and proof of. possession. It may be proved by proving the signature of the payee or obligee on the back, and possession by a third person. 3 Gill & Johns. 218.

Table D.1: Codebook definition and examples of each of the interpretive classes.