LAWMA: THE POWER OF SPECIALIZATION FOR LEGAL ANNOTATION

Ricardo Dominguez-Olmedo[1], **Vedant Nanda**[2], **Rediet Abebe***[1,3], **Stefan Bechtold***[4], **Christoph Engel***[5], **Jens Frankenreiter***[6], **Krishna Gummadi***[2], **Moritz Hardt***[1], **and Michael Livermore***[7]

[1]Max Planck Institute for Intelligent Systems, Tübingen, and Tübingen AI Center [2]Max Planck Institute for Software Systems, Saarbrücken [3]ELLIS Institute, Tübingen [4]ETH Zurich [5]Max Planck Institute for Research on Collective Goods, Bonn [6]Washington University in St. Louis - School of Law [7]University of Virginia School of Law

ABSTRACT

Annotation and classification of legal text are central components of empirical legal research. Traditionally, these tasks are often delegated to trained research assistants. Motivated by the advances in language modeling, empirical legal scholars are increasingly turning to commercial models, hoping that it will alleviate the significant cost of human annotation. In this work, we present a comprehensive analysis of large language models' current abilities to perform legal annotation tasks. To do so, we construct `CaselawQA`, a benchmark comprising 260 legal text classification tasks, nearly all new to the machine learning community. Starting from GPT-4 as a baseline, we show that it has non-trivial but highly varied accuracy, often exhibiting performance that may be insufficient for legal work. We then demonstrate that a lightly fine-tuned Llama 3 8B model vastly outperforms GPT-4 on almost all tasks, typically by double-digit percentage points. A few tens to hundreds of examples suffice to achieve high classification accuracy. Our work points to a viable alternative to the predominant practice of prompting commercial models. For concrete legal tasks with some available labeled data, researchers are better off using a specialized open-source model. Code, datasets, and fine-tuned models are available at https://github.com/socialfoundations/lawma.

1 INTRODUCTION

The legal system generates a staggering volume of complex documents. United States federal courts alone process hundreds of thousands of cases a year, each having substantial case files. Much empirical legal research involves the systematic collection and analysis of such data in order to understand how laws function in practice and what impact they have on society. What limits researchers across the board is the cost of annotating and classifying legal documents. Legal classification tasks vary in complexity, but often require substantial expertise and effort. Employing trained research assistants stretches to a few thousand documents at a time, but is no match for the sheer scale of legal data.

There has long been an interest by empirical legal scholars in NLP tools for feature extraction (i.e., annotation) in lieu of human annotators (Livermore & Rockmore, 2019). Starting from sentiment analysis and topic models, to now large language models. The costs and error of existing methods is the single most important bottleneck in the empirical legal studies pipeline. Yet, the use of large language models to annotate legal text remains a critically understudied area.

Nonetheless, motivated by the rapid advances in language models, law scholars increasingly try out commercial models, such as GPT-4, on a variety of legal tasks, hoping to boost the efficiency of legal research. The underlying assumption is that large models such as GPT-4 provide the best solution to the problem that is currently available. In this work, we critically examine this assumption.

[*] Alphabetical order.

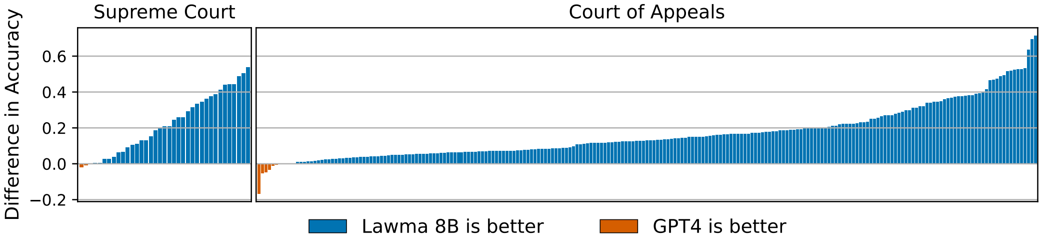

Figure 1: The cost of generality: Difference in accuracy between Lawma 8B and GPT4. Each vertical bar represents the accuracy difference on one task, sorted in ascending order.

1.1 Our contributions

We introduce and study a collection of 260 legal classification tasks, nearly all new to the machine learning community. The tasks we introduce are actual legal annotation tasks based on the U.S. Supreme Court (Spaeth et al., 2023) and Court of Appeals (Songer) databases. These databases offer rich annotations for court cases, which we utilize as labels to create challenging multi-class classification tasks. We aggregate these tasks into an easy-to-use benchmark, which we call `CaselawQA`. We detail in Section 2 the process used to construct this benchmark.

We then evaluate in Section 3 the zero-shot performance of 28 language models, including GPT-4. We find that only a handful of them—notably Llama 3 70B Instruct and GPT-4—perform significantly better than a constant classifier that outputs the majority class. Of these models, GPT-4 delivers the strongest performance. Still, its average performance is poor (**62.0%** accuracy), and there are dozens of tasks where it performs worse than random guessing. Evaluating GPT-4 few-shot does not improve performance. Based on our comprehensive evaluations, we conclude that that the performance of current LLMs is far from sufficient for actual legal annotation work.

Next, we leverage our large corpus of legal classification tasks to fine-tune a single Llama 3 8B Instruct (MetaAI, 2024) model, which we call *Lawma 8B* (Section 4). We show that Lawma 8B achieves vastly superior performance to GPT-4[1] (Figure 1). Specifically, Lawma 8B outperforms GPT-4 by **20.0** accuracy points, attaining in absolute terms **81.9%** accuracy. Although it is expected that fine-tuning improves performance, the strong superiority of fine-tuning an open-weights model at much smaller scale is highly surprising. Our results demonstrate that, for legal classification tasks, researchers are better off using small specialized models rather than large general-purpose LLMs.

Finally, we conduct several additional large-scale fine-tuning experiments that further demonstrate the benefits and practicality of specializing models:

- Larger models respond better to fine-tuning. Across nine different base models, accuracy increases steadily with pretraining compute (Section 4.2, Figure 7). We fine-tune a single Llama 3 70B Instruct model, which we call Lawma 70B, which attains **83.3%** accuracy.
- Fine-tuning is data efficient. A few hundred examples typically suffice to achieve high classification accuracy (Section 4.3, Figure 8). This is crucial, since labeling a few hundred data points is often financially feasible for many legal scholars.
- Fine-tuning generalizes to unseen tasks. Fine-tuning only on Court of Appeals tasks improves its accuracy on Supreme Court tasks by 18.8 accuracy points (Section 4.4, Figure 9).
- We can simultaneously fine-tune on all 260 tasks. There is not a large loss compared with fine-tuning on a single specific task (Appendix D, Figure 11). This is desirable in practice, as it obviates the need to train and maintain a separate model for each task.

Our results speak to the power of specialization for legal classification tasks. The methods described in our paper can radically expand the capacity of legal scholars to engage in quantitative work, empowering legal scholars to apply the "law as data" paradigm to a host of novel research questions.

[1] We evaluate `gpt-4-0613`, which is what at the time of writing `gpt-4` points to in the API. The recently released GPT-4o and o1 models are currently not available for our region via the Azure OpenAI Service.

Annotations of existing datasets can become much more fine-grained. Entire jurisdictions that have hitherto escaped academic attention, such as the many courts of U.S. States, may finally be analyzed.

From a benchmarking perspective, the tasks presented in this work are of independent interest. They are challenging multi-class classification problems that require some amount of legal expertise. The best models achieve non-trivial, but modest zero-shot performance. And even fine-tuned models don't reach intercoder agreement rates (Section C). Our empirical findings suggest that these legal classification tasks are diverse, non-trivial evaluation tasks for future model advances.

Finally, our work challenges the prevailing narrative about the suitability of "generalist" models. In commercial APIs, users are generally limited to prompting generalist models, as fine-tuning is costly for the model provider. But as we show, generalist models are neither sufficiently good nor best possible for many practical tasks. Specializing models to concrete tasks of interests, even using relatively few labeled examples, can provide a simple, practical, and far more accurate solution.

1.2 Related work

Benchmarks for legal tasks. LegalBench (Guha et al., 2023) is a recent multi-task benchmark for natural language understanding in legal domains. As of writing, LegalBench consists of 162 tasks gathered from 40 contributors. LegalBench draws on numerous earlier benchmarking efforts in different legal domains, specifically, inference on contracts (Koreeda & Manning, 2021; Hendrycks et al., 2021), merger agreement understanding (Wang et al., 2023), identifying the legal holding of a case (Zheng et al., 2021), statutory reasoning (Holzenberger & Van Durme, 2021), privacy compliance and policy (Wilson et al., 2016; Zimmeck et al., 2019; Ravichander et al., 2019), and identifying unfair clauses in terms of service (Lippi et al., 2019). Bhambhoria et al. (2024) evaluate the performance of general-purpose models on legal question-answering tasks and advocate for the development of open-source models tailored to the legal domain. We extend and strengthen these valuable efforts to benchmark large language models in legal settings. We focus on core legal classification tasks based on the U.S. Supreme Court Database (Spaeth et al., 2023) and the U.S. Courts of Appeals database (Songer). Our evaluation suite measures the performance of models in annotating court opinions, focusing on tasks that are of interest to the field of empirical legal studies. The tasks we study are complementary to those in LegalBench. We do not evaluate our model on LegalBench, since our model is specialized to the Supreme Court and Appeals Court data.

Large language models for the legal domain. General-purpose language models are likely to be trained on a substantial amount of legal data because much of this data is publicly available on the internet. For example, the FreeLaw dataset includes a large collection of court opinions (Gao et al., 2021). Legal-BERT (Chalkidis et al., 2020) is a BERT-like transformer model that was pretrained on a few hundred thousand legal documents. The more recent SaulLM models (Colombo et al., 2024b;a) adapt the open-weights Mistral (Jiang et al., 2023; 2024) models to the legal domain both by continual pretraining and instruction-tuning on legal text. In contrast to Lawma, we consider SaulLM to be a general-purpose model for the legal domain, not tailored to any specific legal task. Our approach differs significantly; we focus on developing models specialized for annotation tasks of practical interest to empirical legal studies. We demonstrate that specialization is highly effective, with our Lawma models significantly outperforming all other evaluated LLMs. For a discussion on the adoption of large language models in the legal community, refer to Appendix A.

Data annotation and labeling. Hall & Wright (2008) provide an overview of the use of human annotators in empirical legal studies. Student coders have been deployed to extract a wide variety of features from legal data. Although student researchers are much less expensive than private attorneys, the costs can quickly become prohibitive. Depending on the size of the document and the complexity of the task, research assistants can label roughly dozens of examples per hour. Projects involving the labeling of hundreds of documents are financially feasible for many legal scholars, but projects involving many thousands of documents are largely impractical. In an example of a larger annotation effort, Frankenreiter et al. (2021) employed human coders to annotate several thousands of corporate charters. Using ChatGPT for a similar task, Frankenreiter & Talley (2024) estimated that employing human coders would have been approximately ten times more costly.

Data annotation and labeling also play a major role in machine learning benchmarks and applications, see, e.g., Aroyo & Welty (2015); Gray & Suri (2019); Hardt & Recht (2022) for background.

```
What follows is an opinion from the Supreme Court of the United States.
Your task is to identify  whether the opinion effectively says that the
decision in this case  "overruled" one or more of the Court\'s own
precedents. Alteration also extends to language in the majority opinion
that states that a precedent of the Supreme Court  has been "disapproved,"
or is "no longer good law". Note, however, that alteration does not
apply to cases in which the Court "distinguishes" a precedent.

[COURT OPINION]

Question: Did the the decision of the court overrule one or more of the
Court’s own precedents?
A. Yes
B. No
Answer:
```

Figure 2: Example task corresponding to the Supreme Court “precedent alteration” variable.

Dorner & Hardt (2024) give an extended discussion about label quality and annotator disagreement in the context of machine learning benchmarks.

1.3 LIMITATIONS

Fine-tuning increases accuracy to about 80% in our evaluation suite compared with around 60% for non-specialized models. While we are rather certain that 60% accuracy is insufficient for consequential legal work, we emphasize that 80% is still far from perfect. In addition, the variance in accuracy across tasks remains high. Although our work meets the ethical and technical recommendations by Kapoor et al. (2024) for “developers of legal AI”, we maintain caution about the use of large language models for consequential legal tasks. To which extent these models are suitable for use in specific applications requires additional substantive investigation. We add that the legal documents we consider are exclusively from either the U.S. Supreme Court or appellate courts in the United States. We cannot speak to how these results may change for tasks in other legal domains within the United States or legal systems in other countries.

2 CASELAWQA

In this work, we focus on legal classification tasks. Legal classification tasks range in complexity, from extremely simple tasks that require little specialized knowledge, to highly sophisticated tasks that involve specific legal knowledge, familiarity with legal principles or discourse, and the ability to engage in nuanced analogical or conceptual reasoning. For example, labeling the ideological valence of a decision requires the annotator to understand how specific legal issues map onto contemporary political debates. Labeling the standard of review applied by an appellate court requires detailed knowledge of these standards as well as the ability to parse procedural history. Many legal doctrines are quite complicated, involving multipart tests, nuanced exceptions, and balancing inquiries.

Our reasons to study legal classification tasks are both technical and substantive. From a technical machine learning perspective, these tasks provide highly non-trivial classification problems where even the best models leave much room for improvement. From a substantive legal perspective, efficient solutions to such classification problems have rich and important applications in legal research, see Appendix A.1 for a detailed discussion.

2.1 DATA SOURCES

Central to our study are the U.S. Supreme Court Database (Spaeth et al., 2023) (SCDB) and the U.S. Courts of Appeals database (Songer) (USCAD). The SCDB compiles comprehensive information on U.S. Supreme Court decisions from 1946 onward, and includes variables such as case outcomes, issue areas, legal provisions, and vote counts. The USCAD contains detailed information about decisions made by the U.S. Courts of Appeals from 1925 to 1988. It includes data on judicial decisions, panel compositions, and case characteristics. Both databases provide essential tools

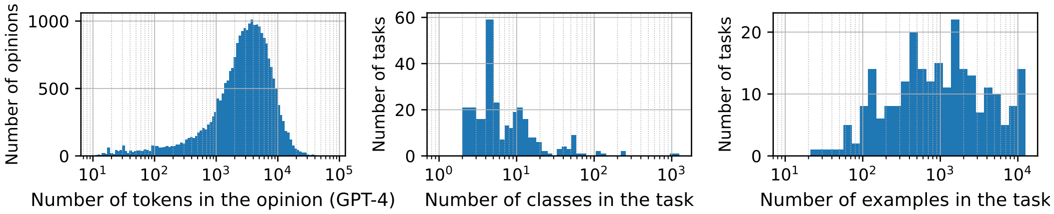

Figure 3: General statistics of the court opinions and legal classification tasks considered.

for scholars conducting quantitative analyses of the judicial system, decision-making, ideological trends, and the impact of various factors on case outcomes.

The SCDB and USCAD have been instrumental in advancing research on judicial decision making within the fields of political science and empirical legal studies (Epstein et al., 2013; Segal & Spaeth, 2002; Martin & Quinn, 2002). These datasets have been used to drive a substantial research program by allowing scholars to systematically analyze large numbers of court cases, uncovering patterns, trends, and factors influencing judicial outcomes. By providing detailed information on case characteristics, judge attributes, and decision outcomes, these databases have enabled researchers to test theories of judicial behavior, examine the impact of ideology on court decisions, and explore the dynamics of judicial decision-making at different levels of the court system. The insights gained from research using these databases have had significant implications for legal practitioners, policymakers, and the broader legal community, contributing to a better understanding of how courts operate and how legal outcomes are shaped.

2.2 Construction of the classification tasks

We use the variables of the USDB and the USCAD to construct a set of classification tasks. We construct a total of 260 distinct classification tasks, 38 of them corresponding to the Supreme Court database and 232 to the U.S. Court of Appeals. The annotations in the USDB and USCAD serve as labels for these classification tasks. For each task, we additionally construct a prompt template consisting of a general description of the task, followed by a multiple choice question containing each of the possible variable codes. We formulate the task description, question, and answer choices by closely following the databases' variable descriptions. See Figure 2 for an example task.

For every case contained in the USDB and the USCAD, we use the provided case citations to search for its corresponding majority opinion of the court on the Caselaw Access Project, a database of digitized court opinions. We match a total of 24,916 court cases, which we divide into a 70%/10%/20% train/validation/test split. That is, models may not train on any of the court cases used for evaluation.

Since many of the classification tasks contain heavily imbalanced classes, we subsample the majority class such that there are at most as many task examples in the majority class as task examples in all other classes combined. As a result, a constant classifier that outputs the majority class label will never achieve more than 50% accuracy on any individual task. This results in a more honest measure of model performance, as models cannot attain high accuracy simply because a task is heavily imbalanced. We report in Appendix E results without subsampling of the majority class.

We plot some statistics of the tasks in Figure 3. First, court opinions tend to be long, with 12% having above 8,000 tokens, the typical maximum context size for current state-of-the-art models, such as Llama 3. Second, some tasks have a large number of classes, with 28% of tasks having more than 10 classes. Third, there is a large variability in terms of the number of task examples, ranging from a couple dozen to 18500 task examples. Our final dataset comprises 718,971 task examples.

To reduce the compute required for evaluating the benchmark, we select at random 5,000 examples from the Supreme Court tasks and 5,000 examples from the Court of Appeals tasks. We include only court cases where the court opinion, including the head matter, contains at least 2,000 characters, ensuring the opinion is at least a few sentences long. These 10,000 task examples comprise the test set of `CaselawQA`. We nonetheless make available all 143,635 task examples corresponding to the test court cases, which we call the *extended test set*. Evaluating on the extended test set is 14x as expensive, but provides much more fine-grained information on models' performance across all 260

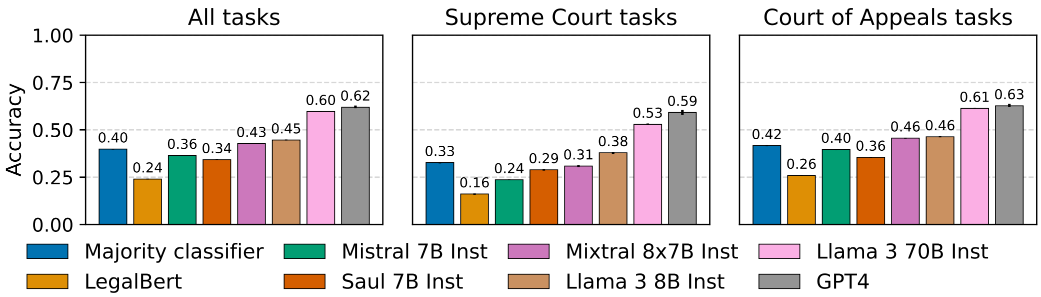

Figure 4: Accuracy of publicly-available LLMs on the extended test set of CaselawQA.

legal classification tasks, rather than simply an aggregate measure of model performance. In this work, we report accuracy on the extended test set, unless otherwise stated.

2.3 Evaluation methodology

We evaluate models using a prompt template identical to the one for the MMLU benchmark (Hendrycks et al., 2020). Since many popular benchmarks are phrased as multiple-choice questions, recent models tend to do well for them (Dominguez-Olmedo et al., 2024). Due to diverse set of models and large number of tasks under consideration, we perform no prompt tuning.[2]

We use accuracy as the evaluation metric. Given that the tasks we consider involve vastly differing numbers of answer choices, accuracy provides an interpretable measure of performance. Additionally, accuracy is the standard metric used in knowledge-testing LLM benchmarks. For completeness, we also report balanced accuracy and macro-averaged F1 scores in Appendix E.

When reporting aggregate performance across multiple tasks (e.g., all Supreme Court tasks), we compute the average accuracy across all task examples. Intuitively, we can visualize the Supreme Court database as a large table with dimensions corresponding to the number of court cases (rows) and the number of tasks (columns). The aggregate accuracy, in this case, represents the fraction of entries in this table that the model correctly predicts. For completeness, we also report mean task accuracy (i.e., macro-averaging rather than micro-averaging across tasks) in Appendix E.

3 Evaluation baselines

We start by evaluating the performance of different language models on the extended test set of CaselawQA. We choose language models that are of particular relevance to the legal domain: LegalBert (Chalkidis et al., 2020), as well as SaulLM 7B (Colombo et al., 2024b), its base model Mistral 7B Instruct (Jiang et al., 2023), and its Mixture-of-Experts variant Mixtral 8x7B (Jiang et al., 2024). We additionally evaluate GPT-4 (Achiam et al., 2023) due to its prevalent use among legal scholars, and the Llama 3 Instruct (MetaAI, 2024) models, which are arguably the best performing open-weights models at present time. We report the evalution results in Figure 4. We include as baseline the majority classifier which simply outputs the majority class of each classification tasks.

Despite the popularity of LegalBERT, we observe that it performs worse than the majority classifier baseline. This is unsurprising, as by current standards it has both a small size (110M parameters) and a small context window (256 tokens). SaulLM 7B, the other legal-domain model, similarly fails to beat the majority classifier baseline. In fact, SaulLM 7B underperforms compared to its base model Mistral 7B Instruct both across all tasks and Court of Appeals tasks. This indicates that broadly adapting models to the legal domain may not prove beneficial for annotation work. Overall, we find that existing LLMs tailored for the legal domain obtain trivial performance in our annotation tasks.

[2] Note that more involved prompting strategies –e.g., chain-of-thought (Wei et al., 2022)– can yield better task performance but are substantially more expensive. For legal tasks specifically, the choice of prompt can have a significant effect in performance (Li et al., 2024). However, prompt tuning requires task-specific domain knowledge and can be reasonably time consuming.

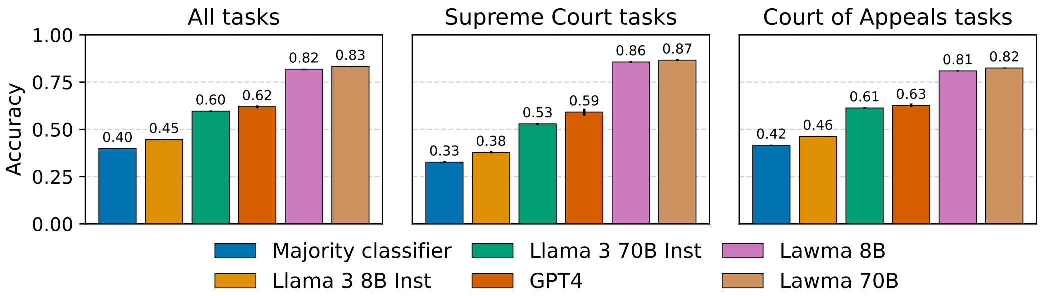

Figure 5: Accuracy of the Lawma models on the extended test set of `CaselawQA`.

In fact, we observe that only the two largest models tested, Llama 3 70B Instruct and GPT-4, perform substantially better than the majority classifier baseline. Still, their performance is modest ($< 65\%$ accuracy), and there are dozens of tasks where both models perform worse than random guessing, see Figure 16 in Appendix E. Our evaluations therefore indicate that, for most tasks, the performance of general-purpose LLMs is clearly insufficient for consequential legal annotation work.

Few-shot evaluation We also consider whether evaluating GPT-4 few-shot leads to any improvements. Whereas the default context window for GPT-4 is 8,000 tokens, a version with 32,000 tokens is available at twice the cost. We evaluated the 32k version with 3-shot prompting, since it is often unfeasible to fit more than 3 task examples within the 32k context window. Labeling each example 3-shot is $3 \times 2 = 6$ times more expensive compared to the zero-shot GPT-4 evaluation. To compensate for the increase in cost, we evaluate the model 3-shot on roughly 5% of the test examples compared those used to evaluate GPT-4. Our evaluation shows that GPT-4's performance with 3-shot prompting does not improve over zero-shot prompting, as detailed in Table 1. This is likely because most legal classification tasks involve more than three classes, meaning that three in-context examples do not cover all possible answer choices. Consequently, the model often responds with one of the three presented examples, leading to a significant drop in performance. Few-shot prompting is therefore not a fruitful strategy to adapt the model to the legal classification tasks at hand.

Table 1: Zero-shot and few-shot accuracy of GPT-4.

Model	**All tasks**	**Supreme Court**	**Court of Appeals**
GPT-4 zero-shot	62.0 ± 0.4	59.2 ± 0.8	62.7 ± 0.5
GPT-4 $32k$ 3-shot	60.4 ± 1.9	50.5 ± 4.3	62.9 ± 2.1

4 Fine-tuning and the power of specialization

In this section, we present a detailed analysis of how models can be specialized for legal classification tasks. We start by fine-tuning Llama 3 8B Inst and Llama 3 70B Inst on all 260 tasks simultaneously, resulting in our Lawma 8B and Lawma 70B models. We then perform additional fine-tuning experiments highlighting different aspects, including the scaling behaviour of fine-tuning, its sample efficiency, and its generalization to unseen tasks and Courts.

4.1 The Lawma models

We first leverage our large corpus of legal classification tasks to fine-tune Llama 3 8B Instruct and Llama 3 70B Instruct on *all tasks* simultaneously. We refer to these fine-tuned models as Lawma 8B and Lawma 70B, respectively. We fine-tune on the 260 classification tasks described in Section 2.2. The fine-tuning dataset contains a total of 503,698 task examples and 1.96B tokens. See Appendix F for additional details regarding the model fine-tuning.

We compare in Figure 5 the task accuracies of Lawma 8B and Lawma 70B to that of their respective base models Llama 3 8B Instruct and Llama 3 70B Instruct, as well as GPT-4. Fine-tuning leads

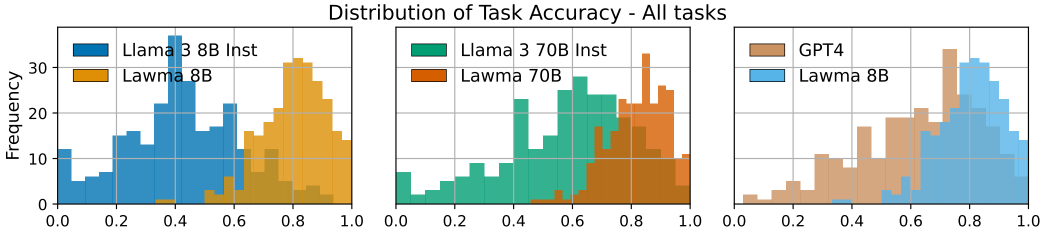

Figure 6: Distribution of task performance across all tasks for Llama 3, GPT4, and Lawma.

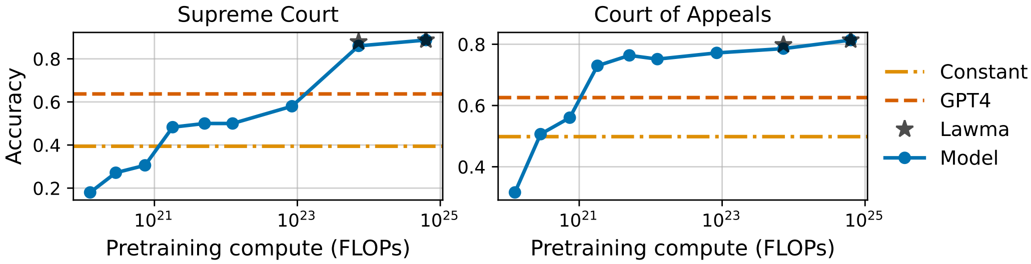

Figure 7: Performance after one epoch of fine-tuning increases monotonically in the amount of pretraining of the base model. Models left to right (blue dots): Pythia 70M, Pythia 160M, Pythia 410M, Pythia 1B, Pythia 2.8B, Pythia 6.9B, Llama 2 7B, Llama 3 8B Inst, Llama 3 70B Inst.

to remarkably large improvements in average task accuracy: Lawma 8B outperforms Llama 3 8B Instruct by **37.2** accuracy points and Lawma 70B outperforms Llama 3 70B Instruct by **21.3** accuracy points. Both Lawma 8B and Lawma 70B outperform GPT-4, Lawma 8B by **19.9** accuracy points and Lawma 70B by **21.3** accuracy points. In fact, both Lawma 8B and Lawma 70B outperform GPT-4 in about 95% of all tasks, see Figure 1. Figure 6 further demonstrates the large effect of fine-tuning by showing the histogram of task accuracies of Lawma in comparison with Llama 3 and GPT-4.

We find that Lawma 8B closely matches the performance of Lawma 70B. Specifically, Lawma 70B outperforms Lawma 8B by only 1.0 accuracy points for Supreme Court tasks and by 1.5 accuracy points for Appeals Court tasks (Figure 5). This suggests that, for our fine-tuning dataset, further scaling model size (e.g., fine-tuning GPT-4) is unlikely to yield major improvements. A more promising direction is to instead improve the diversity and quantity of the fine-tuning data. On the flip side, practitioners may choose to use Lawma 8B instead of the 70B model with little cost in performance.

4.2 Performance after fine-tuning scales with pretraining compute

The performance of specialized models tends to scale with pretraining compute (Dominguez-Olmedo et al., 2024). We investigate how performance after fine-tuning scales with the pretraining compute of the base model. We fine-tune the following models for a single epoch: Pythia 70M, Pythia 160M, Pythia 410M, Pythia 1B, Pythia 2.8B, Pythia 6.9B (Biderman et al., 2023), Llama 2 7B (Touvron et al., 2023), Llama 3 8B Instruct and Llama 3 70B Instruct. We fine-tune on all 260 tasks simultaneously. We approximate pretraining compute in FLOPs as $C \approx 6 \cdot N \cdot D$ (Kaplan et al., 2020), where N is model size and D is the number of pretraining tokens.

We find that mean task accuracy after fine-tuning improves with pretraining compute (Figure 7). However, we find signs of diminishing returns. For the Supreme Court tasks, performance increases steadily from 10^{20} to 10^{24} FLOPs, but further scaling to 10^{25} FLOPs only improves performance by an additional 4.0 accuracy points. For Appeals Court tasks, performance sharply increases from 10^{20} to 10^{21} FLOPs (i.e., Pythia 1B – which interestingly already beats GPT-4 zero-shot), but stagnates thereafter, only improving by an additional 8.5 accuracy points when scaling to 10^{25} FLOPs.

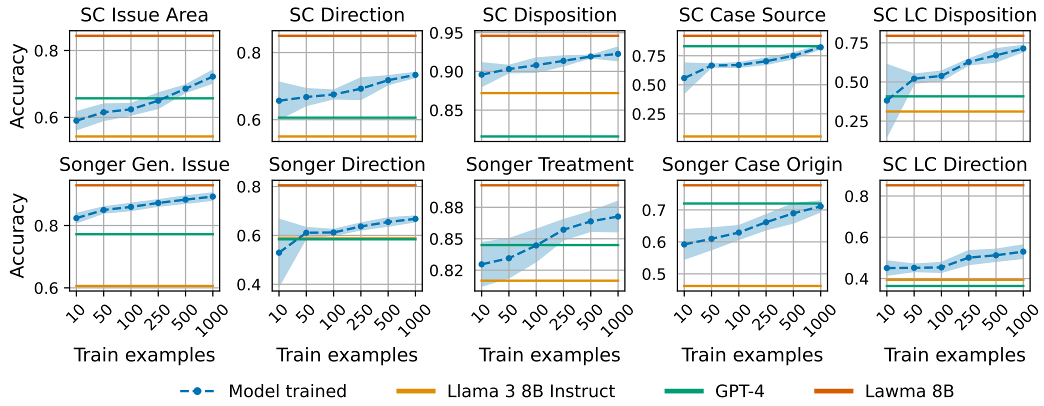

Figure 8: Sample efficiency of fine-tuning Llama 3 8B on a single task. Hundreds of task examples are typically enough to match or beat the zero-shot performance of GPT-4. Dashed blue line indicates the accuracy of Llama 3 8B fine-tuned on a single task as a function of the number of training examples. The shaded area indicates the 95% confidence interval over 5 random seeds.

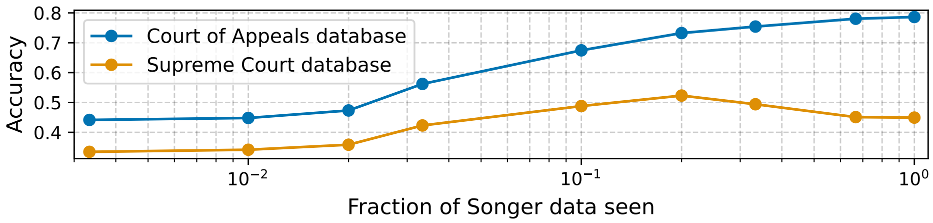

Figure 9: Fine-tuning on the Court of Appeals tasks improves accuracy on Supreme Court tasks.

Our findings suggest that major improvements will likely not come from model scale alone. Rather, future work should focus on obtaining better scaling behavior. One promising direction is to improve the quality, quantity and diversity of the fine-tuning data.

4.3 Sample efficiency

We study how task accuracy scales as models fine-tune on more training examples. We consider the 10 tasks highlighted in Section B. We fine-tune Llama 3 8B Instruct on each task independently, rather than on all tasks simultaneously as in the previous experiments. For each task, we fine-tune on 10, 50, 100, 250, 500, and 1000 task examples. We select task examples uniformly at random, and train 5 different models corresponding to different random seeds on the examples selected for training. We therefore fine-tune and evaluate a total of $10 \cdot 6 \cdot 5 = 300$ models. We fine-tune for a maximum of 20 epochs and early stop when validation loss increases for 3 consecutive epochs.

Figure 8 shows how accuracy improves with the number of training examples. Fifty training examples are enough to match or beat the GPT-4 zero-shot baseline for 6 out of the 10 highlighted tasks, and 250 traning examples are enough to match or beat GPT-4 for 8 out of the 10 highlighted tasks. This is crucial, since labeling a few hundred data points is often financially feasible for many legal scholars (Hall & Wright, 2008). With relative few labelled task examples, fine-tuning reasonably small publicly available models can be competitive with state-of-the-art closed models. Moreover, accuracy continues to improve significantly with additional examples. With one thousand training examples, fine-tuning Llama 3 8B Inst matches or beats the GPT-4 baseline for all highlighted tasks.

4.4 Generalization to unseen databases

We now investigate whether fine-tuning only on the Songer Appeals Court database allows us to generalize to the Supreme Court database. We fine-tune Llama 3 8B Inst for one epoch on all Songer tasks simultaneously. We plot in Figure 9 the mean accuracy for Court of Appeals tasks and Supreme Court tasks at intermediate checkpoints. As expected, performance on Court of Appeals tasks improves monotonically with the number of training examples seen. More interestingly, we observe that mean task accuracy for the Supreme Court also improves substantially, by up to 18.8 accuracy points at 20% of the training steps[3]. Thereafter, performance degrades, seemingly plateauing at 11.3 accuracy points above the baseline non-finetuned performance of Llama 3 8B Inst.

Our findings indicate that, since there is some degree of overlap between Court of Appeal and Supreme Court tasks, fine-tuning on the former transfers to the latter. This suggests that Lawma might be of practical use beyond the Supreme Court and Court of Appeals tasks it was trained on.

Note, however, that fine-tuning only on the Court of Appeals database results in a mean case accuracy of 51.6%, compared to 82.4% for Lawma 8B. That is, not fine-tuning on Supreme Court cases results in a 30.9 accuracy points decrease in performance. These results again highlight the importance of fine-tuning precisely on the target tasks of interest.

5 Discussion

We introduce and study a collection of 260 legal classification tasks, nearly all new to the machine learning community. CaselawQA, our introduced dataset, serves a double purpose: a benchmark to evaluate the ability of LLMs to perform legal annotation work of practical interest to legal scholars, and a fine-tuning dataset to specialize existing models to such legal classification tasks.

As we show, the performance of existing "generalist" LLMs is far from sufficient for consequential legal annotation work. In contrast, we demonstrate the power of specialization: we fine-tune and make available the Lawma 8B and Lawma 70B models, which strongly outperform all other models evaluated, including GPT-4 and two existing legal-domain LLMs.

The CaselawQA dataset, the Lawma models, and more broadly the specialization methodology presented in this work, are all of practical interest to legal research. The cost of human annotators represents a considerable bottleneck for the field of empirical legal studies. The advent of low-cost and flexible tools for data extraction can lead to tremendous boosts in scholarly productivity and knowledge production. For example, the falling cost of genetic sequencing led to a paradigm shift across the biological sciences, as genetic data became increasingly available in fields as disparate as public health and entomology (Köser et al., 2012; Ballare et al., 2019). A flexible automated feature extraction tool for legal texts holds similar potential for empirical legal studies, as a large realm of conceivable but impracticably expensive research projects becomes accessible.

The tasks we introduce are also interesting from a broader LLM benchmarking perspective. The accuracy numbers are neither too low nor too high. The best models achieve non-trivial, but modest zero-shot performance. And even fine-tuned models don't reach intercoder agreement rates. This situation suggests that these legal classification tasks may be good test cases for future model advances. As such, we hope to extend and strengthen existing evaluation efforts.

Lastly, our work challenges the prevailing narrative about the suitability of "generalist" models. The generalist abilities of large language models are vital for commercial APIs, where users are largely restricted to prompting. But as we show, generalist models may be neither sufficiently good nor the best possible solution for many practical tasks.

We show that this is certainly the case for annotation work that arises in empirical legal research. Lightly fine-tuned special purpose models achieve significantly higher accuracy from relatively few labeled examples. Labeling a few hundred cases is often financially feasible. This suggests a simple and practical strategy for solving legal classification tasks: Obtain a few hundred labeled examples, fine-tune an-open weights model, and use the fine-tuned model to annotate the remaining cases.

[3] Note that 20% is the optimal amount of Songer data to train on if the goal is to generalize precisely to Supreme Court cases. If the goal is to generalize to some other dataset, then 20% need not be the optimal amount of Songer data to train on.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

Kimberly M Ballare, Nathaniel S Pope, Antonio R Castilla, Sarah Cusser, Richard P Metz, and Shalene Jha. Utilizing field collected insects for next generation sequencing: Effects of sampling, storage, and dna extraction methods. *Ecology and Evolution*, 9(24):13690–13705, 2019.

Rohan Bhambhoria, Samuel Dahan, Jonathan Li, and Xiaodan Zhu. Evaluating ai for law: Bridging the gap with open-source solutions. *arXiv preprint arXiv:2404.12349*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Caselaw Access Project. Caselaw access project. *URL: https://case.law/*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: the muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

Jonathan H Choi. How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics (Forthcoming)*, 2023.

Jonathan H Choi and Daniel Schwarcz. AI assistance in legal analysis: An empirical study. *Available at SSRN 4539836*, 2023.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. ChatGPT goes to law school. *J. Legal Educ.*, 71:387, 2023.

Axolotl AI Cloud. Axolotl, 2024. URL `https://github.com/axolotl-ai-cloud/axolotl`. Accessed: 2024-07-17.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Sofia Morgado, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. *arXiv preprint arXiv:2407.19584*, 2024a.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024b.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*, 2024.

Florian E Dorner and Moritz Hardt. Don't label twice: Quantity beats quality when comparing binary classifiers on a budget. In *International Conference on Machine Learning*, 2024.

Christoph Engel and Richard H Mcadams. Asking GPT for the ordinary meaning of statutory terms. *MPI Collective Goods Discussion Paper*, (2024/5), 2024.

Lee Epstein, William M Landes, and Richard A Posner. *The behavior of federal judges: a theoretical and empirical study of rational choice*. Harvard University Press, 2013.

Jens Frankenreiter and Eric L Talley. Sticky charters? the surprisingly tepid embrace of officer-protecting waivers in delaware. *The Surprisingly Tepid Embrace of Officer-Protecting Waivers in Delaware (March 19, 2024). European Corporate Governance Institute-Law Working Paper*, (762), 2024.

Jens Frankenreiter, Cathy Hwang, Yaron Nili, and Eric Talley. Cleaning corporate governance. *U. Pa. L. Rev.*, 170:1, 2021.

L Gao, S Biderman, S Black, L Golding, T Hoppe, C Foster, J Phang, H He, A Thite, N Nabeshima, et al. The pile: an 800gb dataset of diverse text for language modeling 2020. *arXiv preprint arXiv:2101.00027*, 2021.

Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. *Philosophical Transactions of the Royal Society A*, 382(2270):20230155, 2024.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

Mark A Hall and Ronald F Wright. Systematic content analysis of judicial opinions. *Calif. L. Rev.*, 96:63, 2008.

Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

Nils Holzenberger and Benjamin Van Durme. Factoring statutory reasoning as language understanding challenges. *arXiv preprint arXiv:2105.07903*, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

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

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint arXiv:2402.01656*, 2024.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. GPT-4 passes the Bar Exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

Yuta Koreeda and Christopher D Manning. ContractNLI: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*, 2021.

Claudio U Köser, Matthew J Ellington, Edward JP Cartwright, Stephen H Gillespie, Nicholas M Brown, Mark Farrington, Matthew TG Holden, Gordon Dougan, Stephen D Bentley, Julian Parkhill, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. 2012.

LexisNexis. LexisNexis Launches Lexis+ AI, a Generative AI Solution with Linked Hallucination-Free Legal Citations. News Release, Oct. 25 2023.

Zehua Li, Julian Nyarko, and Sarath Sanga. The terms of freedom: Black labor contracts in the reconstruction south. 2024.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139, 2019.

Michael A. Livermore and Daniel N. Rockmore (eds.). *Law as Data: Computation, Text, & the Future of Legal Analysis*. Santa Fe Institute Press, 2019.

Michael A Livermore, Felix Herron, and Daniel Rockmore. Language model interpretability and empirical legal studies. *Virginia Public Law and Legal Theory Research Paper*, (2023-69), 2023.

Andrew D Martin and Kevin M Quinn. Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political analysis*, 10(2):134–153, 2002.

MetaAI. Llama 3: Advancing open foundation models, 2024. URL `https://ai.meta.com/blog/meta-llama-3/`.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159, 2024.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*, 2019.

Jaromir Savelka and Kevin D Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6, 2023.

Jeffrey A Segal and Harold J Spaeth. *The Supreme Court and the attitudinal model revisited*. Cambridge University Press, 2002.

Donald R. Songer. U.S. Courts of Appeals databases 1925–1996. URL `http://www.songerproject.org/us-courts-of-appeals-databases.html`.

Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2023 supreme court database, version 2023 release 01, 2023. URL `http://supremecourtdatabase.org`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Steven H Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. MAUD: An expert-annotated legal nlp dataset for merger agreement understanding. *arXiv preprint arXiv:2301.00876*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Kyle Wiggers. Harvey, which uses AI to answer legal questions, lands cash from OpenAI. *TechCruch*, Nov. 23 2022.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1330–1340, 2016.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pp. 159–168, 2021.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66, 2019.

A Adoption of large language models in the legal community.

The legal community has moved relatively quickly in adopting GPT models. Several startups have begun using incorporating large language models, including GPT, into legal products (Wiggers, 2022). Lexis Nexis, a major commercial provider of law-related services, has partnered with Open AI and Anthropic to offer legal text generation (LexisNexis, 2023). Legal scholars have evaluated GPT's performance on the bar exam (Katz et al., 2024) as well as law school exam (Choi et al., 2023). Choi & Schwarcz (2023) examined how GPT-4 can improve student performance on law school exams. Nay et al. (2024) examined how LLMs perform on answering multiple choice questions related to tax law. Gray et al. (2024) used GPT models to extract information from cases concerning the factors that predict the constitutionality of police stops. Choi (2023) used GPT-4 to extract information concerning interpretative techniques from U.S. Supreme Court decisions. Livermore et al. (2023) tested the performance of GPT models for categorizing cases by issue areas and in recommending citations based on case similarity. Savelka & Ashley (2023) evaluate the zero-shot performance of GPT-4 on a variety of semantic legal annotation tasks. Engel & Mcadams (2024) ask GPT for the ordinary meaning of statutory terms. In the area of corporate law, Frankenreiter & Talley (2024) use GPT-4 to extract information about the contents of corporate charters.

A.1 Potential applications of efficient solutions to legal classification tasks

More efficient ways to solve legal classification tasks would be tremendously useful in practice. A well functioning system to automatically extract relevant features from legal texts could, in particular, facilitate empirical legal study across a wide range of domains. This research could include not only social scientific study of the causes or consequences of judicial decisions, but also more traditional research modalities based on doctrinal interpretation (Livermore & Rockmore, 2019). There is an almost unlimited variety of features that legal scholars could study, ranging from the factors cited by judges when deciding the outcomes of property law disputes to the relationship between the party affiliation of judges and their use of different interpretative styles. With the digitization of legal texts at the U.S. state level and outside the U.S., low-cost and flexible featurization can also boost efforts to show the geographic diffusion of legal concepts.

B Highlighted tasks

Throughout this paper, as in Figure 4, we provide detailed results for ten tasks. Six of these tasks are from the SCDB, and four are from the USCAD. We selected tasks that we believe are particularly relevant to the legal community and chose tasks with varying levels of complexity, ranging from relatively simple (e.g., determining the issue area) to more complex (e.g., determining the ideological 'direction' of the court decision).

Four tasks from the USCAD and all tasks from the SCDB were selected to form pairs, with each pair consisting of one task from the USCAD and one from the SCDB that capture similar concepts. It is important to note that, despite capturing broadly similar concepts, the precise formulation of the tasks might differ between the USCAD and the SCDB, making them less than perfectly comparable. In addition to the four pairs, we include two tasks from the SCDB that involve determining features of the decision reviewed by the Supreme Court on the basis of the Supreme Court opinion. The following is a description of the task pairs:

- **SC Issue Area / Songer Gen Issue:** These tasks capture the case's issue area, requiring a determination of whether the case belongs to one of several broadly defined categories, such as criminal cases or First Amendment cases. These tasks are expected to be of relatively low complexity.

- **SC Case Source / Songer Case Origin:** These tasks require identifying the court or adjudication body where the case was originally initiated before moving up the judicial hierarchy. Like the previous pair, these tasks are expected to be of relatively low complexity.
- **SC Disposition / Songer Treatment:** These tasks involve determining how the deciding court treated the lower court opinion it reviewed, such as whether it affirmed or reversed the opinion. We consider these tasks to be of relatively low complexity.
- **SC Direction / Songer Direction:** These tasks involve determining the ideological 'direction' of the decision, specifically whether the decision supports a "conservative" or "liberal" outcome. We consider these tasks to be comparably complex.
- **SC LC Disposition / SC LC Direction:** These tasks involve determining the disposition and ideological 'direction' of the decision reviewed by the Supreme Court. As these tasks require analyzing features of another decision based on the text of the Supreme Court decision, we consider these tasks to be comparably complex.

C Intercoder agreement analysis

The Songer Appeals Court database provides intercoder agreement rates for a subset of the variables. These intercoder agreement rates provide valuable context for the performance of our model. Specifically, intercoder agreement gives us information about the inherent label noise in the annotation procedure. In particular, the intercoder agreement rate gives a natural upper bound on model performance, as we cannot expect the model to perform well when the label is uncertain or subject to interpretation.

However, we cannot directly compare intercoder agreement rates with the accuracy numbers we report. The reason is that in each task we subsampled the majority class to be no larger than the union of all other classes. This is a design choice we made to account for class imbalance. In this section, we map our model's accuracy to *adjusted* accuracy numbers that undo the subsampling step. This results in accuracy numbers that are commensurate with the intercoder agreement rate.

Name	**IC Agreement**	**Adj accuracy**	(unadjusted)	Keep
WEIGHTEV (songer_weightev)	**76**	**78.7%**	(77.2%)	28.72%
PROCEDUR (songer_procedur)	**78**	**75.2%**	(73.9%)	83.08%
ORIGIN (songer_origin)	**83.2**	**80.1%**	(77.7%)	53.13%
DIRECT2 (songer_direct2)	**85.6**	**67.5%**	(67.5%)	100.00%
DIRECT1 (songer_direct1)	**94**	**80.5%**	(80.5%)	100.00%
TREAT (songer_treat)	**95.2**	**91.1%**	(90.1%)	71.26%
GENISS (songer_geniss)	**97.6**	**93.2%**	(92.9%)	84.77%
CIRCUIT (songer_circuit)	**100**	**93.2%**	(93.2%)	100.00%
COMMENT (songer_comment)	**100**	**100.0%**	(91.7%)	0.13%

Table 2: Intercoder agreement rates, Lawma accuracies, and fraction of the majority class retained in our sample. Rows are sorted in increasing order of agreement rate.

Table 2 considers several tasks from the Appeals Court database, including the selected ones we highlighted in various figures. Each row corresponds to one task and provides the intercoder agreement rate, adjusted (and unadjusted) accuracy achieved by Lawma 8B, and the fraction of samples we retained in the majority class. A fraction of 100% means that we kept all samples. The smaller the fraction the larger the majority class is relative to the other classes.

The table contains several interesting insights:

- The adjusted accuracy of Lawma 8B is generally within single digit percentage points of the intercoder agreement rate for easy tasks such as general issue classification (GENISS).

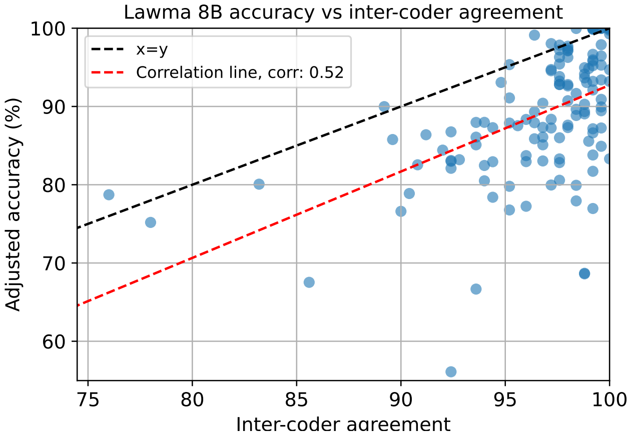

Figure 10: Lawma task accuracy against inter-coder agreement. Lawma

- Lawma 8B is surprisingly close on the two tasks with the lowest intercoder reliability, i.e., WEIGHTEV and PROCEDUR. This shows that high intercoder reliability is no prerequesite for the model to perform well, i.e., close to the agreement rate.
- On harder tasks, like identifying the ideological valence of a decision (DIRECT1 and DIRECT2), Lawma 8B is below the agreement rate by double digit percentage points.
- Tasks with very high agreement rate (e.g., CIRCUIT and COMMENT) are not all alike. Some of them (e.g., COMMENT) correspond to a task with extreme class imbalance. Here, the model reaches the agreement rate. Other tasks (e.g., CIRCUIT) have perfect agreement rate, no class imbalance, and yet Lawma is far from the agreement rate.

These findings speak to the task heterogeneity and the non-trivial nature of the task suite as a classification benchmark.

D Specializing for single tasks

We now study how much accuracy we stand to gain by fine-tuning on a *single* task. We specialize models for each of the 10 tasks highlighted in Section B. We specialize the follow models: Llama 3 8B Inst, Llama 3 8B Inst fine-tuned for one epoch on all tasks, and Lawma 8B (i.e., Llama 3 8B Inst fine-tuned for three epochs on all tasks). For each task, we fine-tune for a maximum of 20 epochs and early stop when validation loss increases for 3 consecutive evaluation stpes, each corresponding to one tenth of an epoch.

Figure 11 shows the results of specialization to single tasks. First, we observe that, for 7 out of 10 tasks, Llama 3 8B Inst fine-tuned on all tasks for one epoch (yellow) outperforms Llama 3 8B Inst specialized for a single task (blue). That is, there is value to fine-tuning on our entire dataset rather than overspecializing for a single task. One explanation is that there is substantial cross-task overlap, and fine-tuning on the entire dataset amounts training on many more examples –even if on average these examples are less relevant.

Secondly, we observe that after fine-tuning on *all* 260 tasks for 1 epoch (yellow), further specializing for a single task (green) improves performance on all cases. Importantly, the latter outperforms the specialized Llama 3 8B Inst (blue) in all tasks. That is, a model that is fine-tuned on everything provides a “better” foundation from which to then “overspecializing” for a single task.

Thirdly fine-tuning on everything for three epochs (i.e., Lawma 8B, in red) again improves over the specialized models (i.e., green). Lastly, “overspecializing” Lawma 8B for a single task results in small single digit improvements for 3 out of the 10 tasks. However, we observe no benefits from

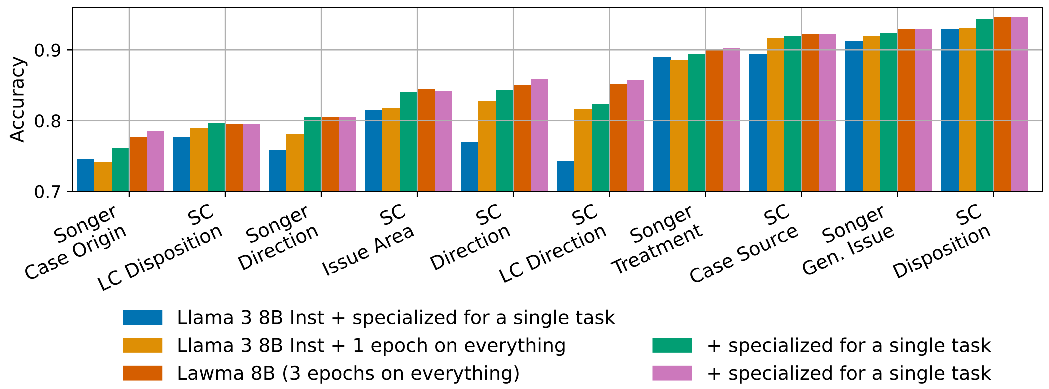

Figure 11: Specializing Lawma 8B to individual tasks can yield small improvements in accuracy.

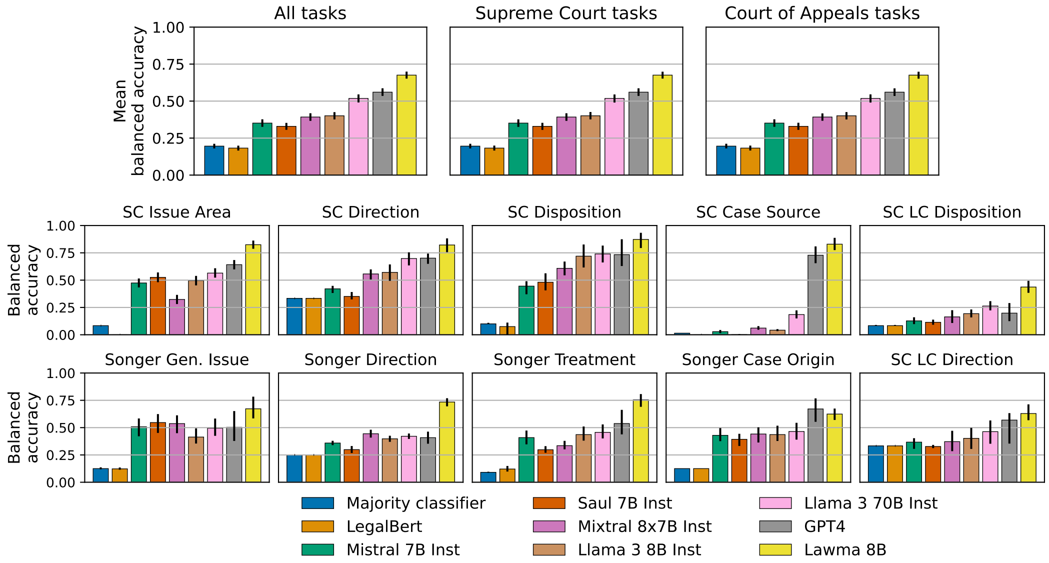

Figure 12: Evaluation results using balanced accuracy as the evaluation metric.

specializing Lawma 8B for most (7/10) of the tasks.[4] These results show that we don't leave much accuracy on the table by fine-tuning a single model for all tasks. This is practically quite appealing, since it obviates the need to maintain a separate model for each task. A single model suffices.

E Additional performance results

E.1 Balanced accuracy and macro-F1

See Figure 12 and Figure 13 for evaluation results using mean balanced accuracy and mean macro-F1 as the evaluation metric, respectively.

E.2 Results without subsampling the majority class

Figure 14 presents the evaluation results when not subsampling the majority class. Models achieve very hight accuracy on many tasks simply because they correctly identify the majority class.

[4] There is a small decrease in performance for SC Issue Area. This is because early stopping is performed with respect to loss on the validation set, but models are evaluated for accuracy on the test set.

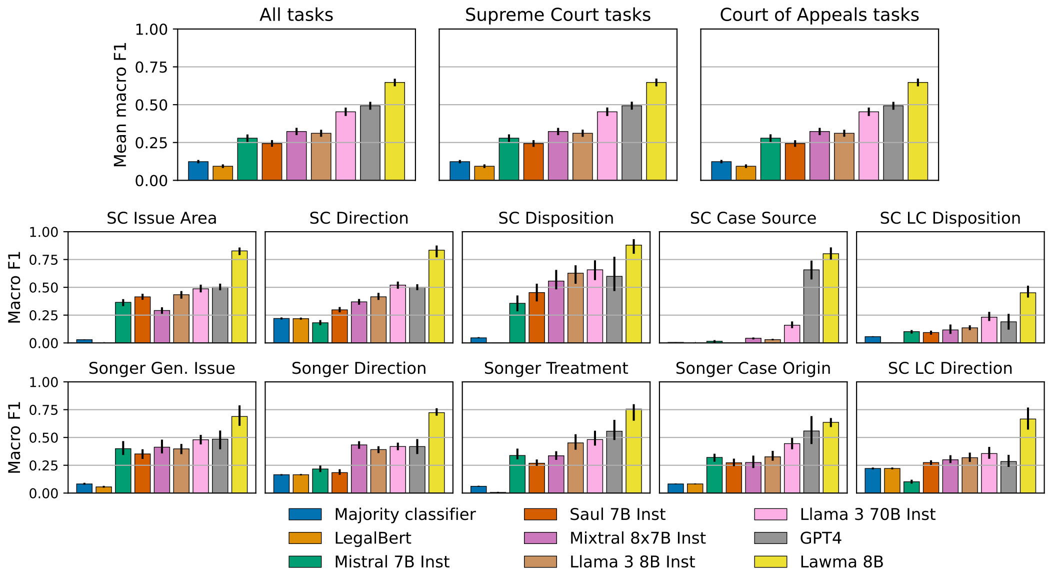

Figure 13: Evaluation results using mean macro-F1 as the evaluation metric.

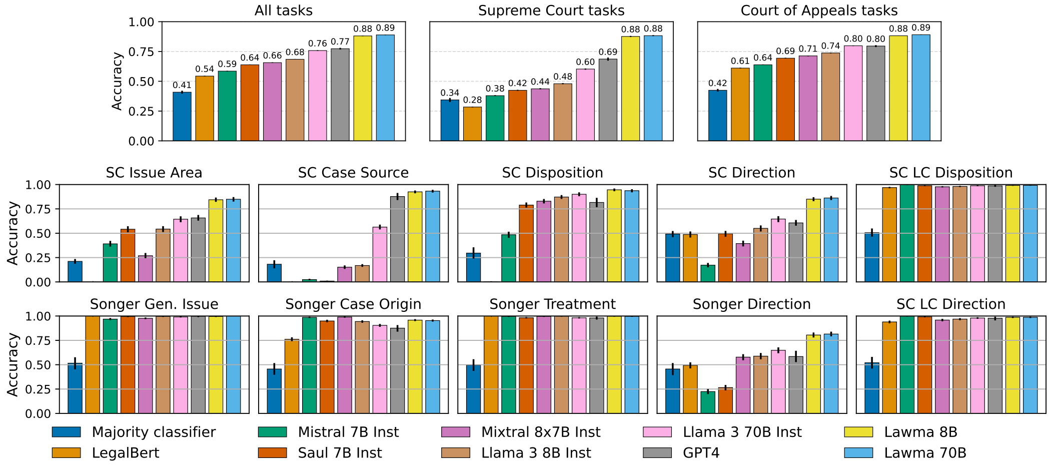

Figure 14: Evaluation results without subsampling the majority class.

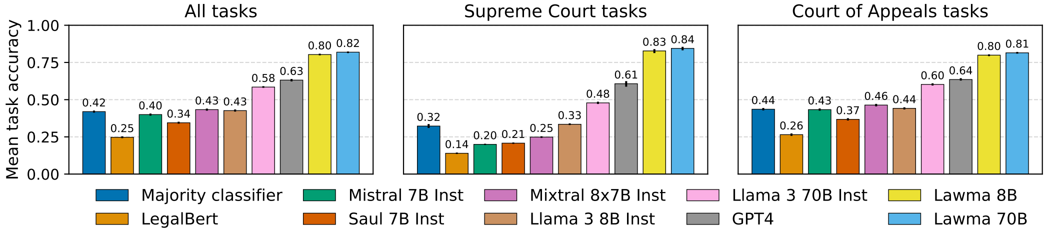

Figure 15: Evaluation results when using mean task accuracy across tasks as the evaluation metric.

E.3 Average task accuracy results

Figure 15 presents the results when using mean task accuracy across tasks as the evaluation metric.

E.4 Comparing Llama 70B Instruct and GPT-4 to the constant classifier

Figure 16 illustrates the difference in performance across tasks between GPT-4 and Llama 3 70B Instruct, and the majority class classifier. GPT-4 and Llama 3 70B Instruct perform worse than the constant classifier for dozens of tasks.

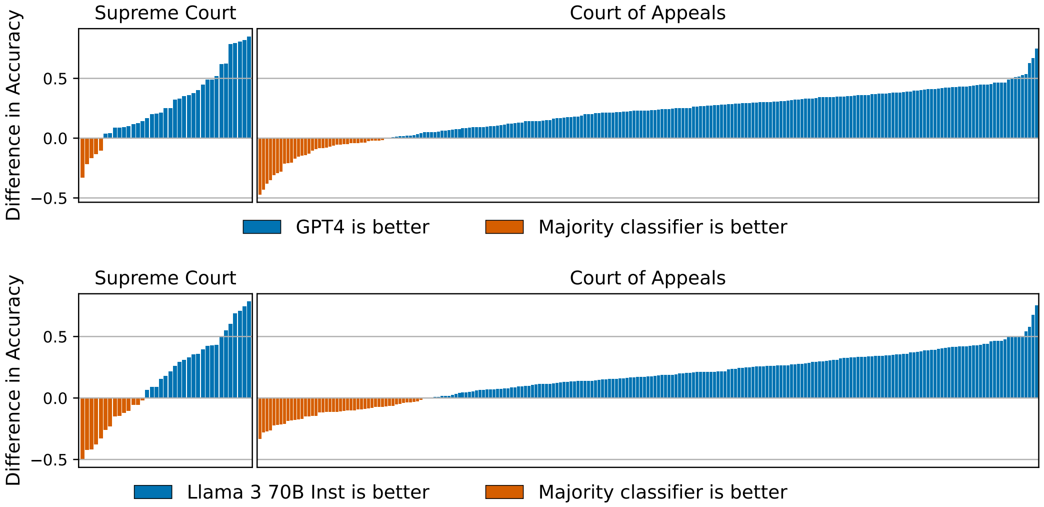

Figure 16: Difference in zero-shot accuracy between GPT4, Llama 3 70B Instruct, and the majority classifier. Each vertical bar represents the accuracy difference on one task, sorted in ascending order.

E.5 Chain of Thought evaluation

We follow the standard methodology of eliciting CoT by appending to the prompt “Let’s think step by step.” Since CoT requires two orders of magnitude more compute for evaluation than the standard QA approach, we only evaluate Llama 3 8B Instruct and Llama 3 70B Instruct. This required over 500 H100 GPU hours. We observe that CoT leads to modest improvements of performance for both the 8B and 70B model, on average of 2 to 3 accuracy points, see Figure 17. Nonetheless, Lawma 8B still strongly outperforms Llama 3 70B, by over 20 accuracy points.

F Fine-tuning details

Compute requirements. We fine-tune on a cluster consisting of NVIDIA H100 GPUs. Fine-tuning on all tasks simultaneously required approximately 600 H100 hours for the 8B model and

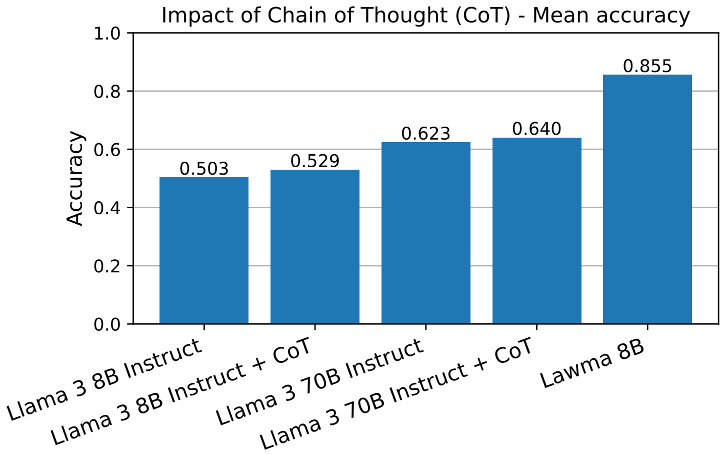

Figure 17: Performance improvements using Chain of Thought (CoT).

1600 GPU hours for the 70B model. In total, the experiments presented in the paper required approximately 8000 H100 GPU hours.

F.1 Lawma

We fine-tuning with a maximum sequence length of 8192 tokens. We use the AdamW optimizer with full precision, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$. We use a peak learning rate of $2 \cdot 10^{-6}$. We use a cosine learning rate schedule, with 180 warm-up steps (approx. 4% of a full epoch) and decay to 10% of the peak learning rate. We use a weight decay of 0.1. We clip gradient to 1.0 max norm. We pack samples using the axolotl library (Cloud, 2024), which improves training efficiency by approximately 40%. For Lawma 8B, we fine-tune Llama 3 8B Instruct for 3 epochs. We train on a node of 7 H100s using DeepSpeed Zero 2, with a global batch size of 56. For Lawma 70B, we fine-tune Llama 3 70B Instruct for 1 epoch. We train on 8 nodes of 8 H100s each using DeepSpeed Zero 3, with a global batch size of 64. We find that additional epochs hurt average task performance, although performance continues to improve for some of the tasks.

F.2 Additional fine-tuning experiments

The hyperparameters are identical to those used for Lawma unless otherwise specified.

Scaling experiments. We fine-tune the Pythia and Llama 2 models with a peak learning rate of $2 \cdot 10^{-5}$, which we find to be result in higher performance than a peak learning rate of $2 \cdot 10^{-6}$. For the Llama 3 models, we use a learning rate of $2 \cdot 10^{-6}$, which we find to be perform better than $2 \cdot 10^{-5}$. We fine-tune for a single epoch. We use a batch size 64. We fine-tune models with their pretraining max sequence length, that is, 2k tokens for Pythia, 4k tokens for Llama 2, and 8k tokens for Llama 3. We use a warm up ratio of 0.03. Due to the costs associated with training the 70B model, we simply take Lawma 70B rather than re-training the model with these slightly different training hyperparameters.

Sample efficiency and specialization We fine-tune for up to 20 epochs. We evaluate the loss on a separate validation set and early stop if the loss increases for 3 consecutive evaluation steps. For the sample efficiency experiments, we evaluate at the end of every epoch. For the specialization experiments, we evaluate every 0.1 epochs. We decay the learning rate to 10% of the peak learning rate over the 20 epochs. We fine-tune with a batch size of 64. For the specialization experiments, we train models both with and without learning rate warm up, and report the accuracy of the best model. We use the AdamW BitsAndBytes 8-bit optimizer, allowing us to fine-tune the models in a single H100 GPU.

Generalization We fine-tune only on the Songer Court of Appeals tasks. We fine-tune with batch size 64. We fine-tune for one epoch and we checkpoint models at 10, 30, 60, 100, 300, 600, 1000, 2000, and 3000 training steps. A full epoch on the Songer Court of Appeal tasks corresponds to 3096 training steps.

G LIST OF ALL TASKS

Variable	**Question**	**Sample answer choices**
sc_adminaction	What is the agency involved in the administrative action?	Army and Air Force Exchange Service, Atomic Energy Commission, Secretary or administrative unit or personnel of the U.S. Air Force
sc_adminaction_is	Did administrative action occur in the context of the case?	No, Yes
sc_adminactionstate	What is the state of the state agency associated with the administrative action?	Alabama, Alaska, American Samoa
sc_authoritydecision	What is the basis of the Supreme Court's decision?	judicial review (national level), judicial review (state level), Supreme Court supervision of lower federal or state courts or original jurisdiction
sc_casedisposition	What is the disposition of the case, that is, the treatment the Supreme Court accorded the court whose decision it reviewed?	stay, petition, or motion granted, affirmed (includes modified), reversed
sc_caseorigin	What is the court in which the case originated?	U.S. Court of Customs and Patent Appeals, U.S. Court of International Trade, U.S. Court of Claims, Court of Federal Claims
sc_caseoriginstate	What is the state of the court in which the case originated?	Alabama, Alaska, American Samoa
sc_casesource	What is the court whose decision the Supreme Court reviewed?	U.S. Court of Customs and Patent Appeals, U.S. Court of International Trade, U.S. Court of Claims, Court of Federal Claims
sc_casesourcestate	What is the state of the court whose decision the Supreme Court reviewed?	Alabama, Alaska, American Samoa
sc_certreason	What reason, if any, does the court give for granting the petition for certiorari?	case did not arise on cert or cert not granted, federal court conflict, federal court conflict and to resolve important or significant question
sc_decisiondirection	What is the ideological direction of the decision?	Conservative, Liberal, Unspecifiable
sc_decisiontype	What type of decision did the court make?	opinion of the court (orally argued), per curiam (no oral argument), decrees
sc_declarationuncon	Did the Court declare unconstitutional an act of Congress; a state or territorial statute, regulation, or constitutional provision; or a municipal or other local ordinance?	No declaration of unconstitutionality, Act of Congress declared unconstitutional, State or territorial law, regulation, or constitutional provision unconstitutional
sc_issue_1	What is the issue of the decision?	subconstitutional fair procedure: fugitive from justice, self-incrimination, immunity from prosecution, cruel and unusual punishment, death penalty (cf. extra legal jury influence, death penalty)

sc_issue_10	What is the issue of the decision?	federal pre-emption of state legislation or regulation. cf. state regulation of business. rarely involves union activity. Does not involve constitutional interpretation unless the Court says it does., federal pre-emption of state legislation or regulation. cf. state regulation of business. rarely involves union activity. Does not involve constitutional interpretation unless the Court says it does., national supremacy: public utilities (cf. federal public utilities regulation)
sc_issue_11	What is the issue of the decision?	non-real property dispute between states, non-real property dispute between states, boundary dispute between states
sc_issue_12	What is the issue of the decision?	federal taxation, typically under provisions of the Internal Revenue Code, federal taxation, typically under provisions of the Internal Revenue Code, federal taxation of gifts, personal, business, or professional expenses
sc_issue_2	What is the issue of the decision?	sex discrimination (excluding sex discrimination in employment), Voting Rights Act of 1965, plus amendments, juveniles (cf. rights of illegitimates)
sc_issue_3	What is the issue of the decision?	libel, privacy: true and false light invasions of privacy, parochiaid: government aid to religious schools, or religious requirements in public schools, First Amendment, miscellaneous (cf. comity: First Amendment)
sc_issue_4	What is the issue of the decision?	due process: takings clause, or other non-constitutional governmental taking of property, due process: miscellaneous (cf. loyalty oath), the residual code, due process: miscellaneous (cf. loyalty oath), the residual code
sc_issue_5	What is the issue of the decision?	Freedom of Information Act and related federal or state statutes or regulations, abortion: including contraceptives, abortion: including contraceptives
sc_issue_6	What is the issue of the decision?	attorneys' and governmental employees' or officials' fees or compensation or licenses, commercial speech, attorneys (cf. commercial speech), attorneys' and governmental employees' or officials' fees or compensation or licenses
sc_issue_7	What is the issue of the decision?	labor-management disputes: right to organize, union-union member dispute (except as pertains to union or closed shop), labor-management disputes: employee discharge
sc_issue_8	What is the issue of the decision?	natural resources - environmental protection (cf. national supremacy: natural resources, national supremacy: pollution), Employee Retirement Income Security Act (cf. union trust funds), election of remedies: legal remedies available to injured persons or things

sc_issue_9	What is the issue of the decision?	standing to sue: private or implied cause of action, judicial administration: review of non-final order, judicial administration: jurisdiction or authority of federal district courts or territorial courts
sc_issuearea	What is the issue area of the decision?	Criminal Procedure, Civil Rights, First Amendment
sc_jurisdiction	What is the manner in which the Court took jurisdiction?	cert, appeal, bail
sc_lcdisagreement	Does the court opinion mention that one or more of the members of the court whose decision the Supreme Court reviewed dissented?	Yes, No
sc_lcdisposition	What treatment did the court whose decision the Supreme Court reviewed accorded the decision of the court it reviewed?	stay, petition, or motion granted, affirmed, reversed
sc_lcdispositiondirection	What is the ideological direction of the decision reviewed by the Supreme Court?	Conservative, Liberal, Unspecifiable
sc_partywinning	Consider that the petitioning party lost if the Supreme Court affirmed or dismissed the case, or denied the petition. Consider that the petitioning party won in part or in full if the Supreme Court reversed, reversed and remanded, vacated and remanded, affirmed and reversed in part, affirmed and reversed in part and remanded, or vacated the case. Did the petitioning win the case?	Yes, No
sc_petitioner	Who is the petitioner of the case?	attorney general of the United States, or his office, specified state board or department of education, city, town, township, village, or borough government or governmental unit
sc_petitionerstate	What state is associated with the petitioner?	Alabama, Alaska, American Samoa
sc_precedentalteration	Did the the decision of the court overrule one or more of the Court's own precedents?	Yes, No
sc_respondent	Who is the respondent of the case?	attorney general of the United States, or his office, specified state board or department of education, city, town, township, village, or borough government or governmental unit
sc_respondentstate	What state is associated with the respondent?	Alabama, Alaska, American Samoa
sc_threejudgefdc	Was the case heard by a three-judge federal district court?	Yes, No
songer_abusedis	Did the court conclude that it should defer to agency discretion? For example, if the action was committed to agency discretion.	No, Yes, Mixed answer
songer_adminrev	What federal agency's decision was reviewed by the court of appeals?	Benefits Review Board, Civil Aeronautics Board, Civil Service Commission

songer_agen_acq	Did the court rule for the government in an issue related to agency acquisition of information (e.g. physical inspections, searches, subpoenas, records, etc)?	No, Yes, Mixed answer
songer_alj	Did the court support the decision of an administrative law judge?	No, Yes, Mixed answer
songer_altdisp	Did the court's ruling on an issue arising out of an alternative dispute resolution process (ADR, settlement conference, role of mediator or arbitrator, etc.) favor the appellant?	No, Yes, Mixed answer
songer_amicus	Was there any amicus participation before the court of appeals?	no amicus participation on either side, 1 separate amicus brief was filed, 2 separate amicus briefs were filed
songer_app_stid	What is the state of the first listed state or local government agency that is an appellant?	not, Alabama, Alaska
songer_appbus	What is the total number of appellants in the case that fall into the category "private business and its executives"? Answer with a number.	N/A
songer_appel1_1_2	This question concerns the first listed appellant. The nature of this litigant falls into the category "private business (including criminal enterprises)". What is the scope of this business?	local, neither local nor national, national or multi-national
songer_appel1_1_3	This question concerns the first listed appellant. The nature of this litigant falls into the category "private business (including criminal enterprises)". What category of business best describes the area of activity of this litigant which is involved in this case?	agriculture, mining, construction
songer_appel1_1_4	This question concerns the first listed appellant. The nature of this litigant falls into the category "private business (including criminal enterprises)", specifically "agriculture". What subcategory of business best describes this litigant?	single family farm, commercial farm, agri-business, farm - other
songer_appel1_2_2	This question concerns the first listed appellant. The nature of this litigant falls into the category "private organization or association". What category of private associations best describes this litigant?	business, trade, professional, or union (BTPU), other
songer_appel1_2_3	This question concerns the first listed appellant. The nature of this litigant falls into the category "private organization or association", specifically "business, trade, professional, or union (BTPU)". What subcategory of private association best describes this litigant?	Business or trade association, utilities co-ops, Professional association - other than law or medicine

songer_appel1_3_2	This question concerns the first listed appellant. The nature of this litigant falls into the category ”federal government (including DC)”. Which category of federal government agencies and activities best describes this litigant?	cabinet level department, courts or legislative, agency whose first word is ”federal”
songer_appel1_3_3	This question concerns the first listed appellant. The nature of this litigant falls into the category ”federal government (including DC)”, specifically ”cabinet level department”. Which specific federal government agency best describes this litigant?	Department of Agriculture, Department of Commerce, Department of Defense (includes War Department and Navy Department)
songer_appel1_4_2	This question concerns the first listed appellant. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”. Which category of substate government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_appel1_4_3	This question concerns the first listed appellant. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”, specifically ”legislative”. Which specific substate government agency best describes this litigant?	City/county council, School Board, board of trustees for college or junior college, Other legislative body
songer_appel1_5_2	This question concerns the first listed appellant. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”. Which category of state government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_appel1_5_3	This question concerns the first listed appellant. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”, specifically ”legislative”. Which specific state government agency best describes this litigant?	Legislature or separate house as an organization, Legislative Committee or Commission, Other Legislative Unit
songer_appel1_7_2	This question concerns the first listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the gender of this litigant?Use names to classify the party’s sex only if there is little ambiguity.	not ascertained, male - indication in opinion (e.g., use of masculine pronoun), male - assumed because of name

songer_appel1_7_3	This question concerns the first listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the race or ethnic identity of this litigant as identified in the opinion?	not ascertained, caucasian - specific indication in opinion, black - specific indication in opinion
songer_appel1_7_4	This question concerns the first listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the citizenship of this litigant as indicated in the opinion?	not ascertained, US citizen, alien
songer_appel1_7_5	This question concerns the first listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. Which of these categories best describes the income of the litigant?	not ascertained, poor + wards of state, presumed poor
songer_appel1_8_2	This question concerns the first listed appellant. The nature of this litigant falls into the category ”miscellaneous”. Which of the following categories best describes the litigant?	fiduciary, executor, or trustee, other, nature of the litigant not ascertained
songer_appel1_8_3	This question concerns the first listed appellant. The nature of this litigant falls into the category ”miscellaneous”, specifically ”fiduciary, executor, or trustee”. Which of the following specific subcategories best describes the litigant?	trustee in bankruptcy - institution, trustee in bankruptcy - individual, executor or administrator of estate - institution
songer_appel2_1_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What is the scope of this business?	local, neither local nor national, national or multi-national
songer_appel2_1_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What category of business best describes the area of activity of this litigant which is involved in this case?	agriculture, mining, construction

songer_appel2_1_4	This question concerns the second listed appellant. The nature of this litigant falls into the category ”private business (including criminal enterprises)”, specifically ”agriculture”. What subcategory of business best describes this litigant?	single family farm, commercial farm, agri-business, farm - other
songer_appel2_2_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”private organization or association”. What category of private associations best describes this litigant?	business, trade, professional, or union (BTPU), other
songer_appel2_2_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”private organization or association”, specifically ”business, trade, professional, or union (BTPU)”. What subcategory of private association best describes this litigant?	Business or trade association, utilities co-ops, Professional association - other than law or medicine
songer_appel2_3_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”federal government (including DC)”. Which category of federal government agencies and activities best describes this litigant?	cabinet level department, courts or legislative, agency whose first word is ”federal”
songer_appel2_3_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”federal government (including DC)”, specifically ”cabinet level department”. Which specific federal government agency best describes this litigant?	Department of Agriculture, Department of Commerce, Department of Defense (includes War Department and Navy Department)
songer_appel2_4_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”. Which category of substate government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_appel2_4_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”, specifically ”legislative”. Which specific substate government agency best describes this litigant?	City/county council, School Board, board of trustees for college or junior college, Other legislative body

songer_appel2_5_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”. Which category of state government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_appel2_5_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”, specifically ”legislative”. Which specific state government agency best describes this litigant?	Legislature or separate house as an organization, Legislative Committee or Commission, Other Legislative Unit
songer_appel2_7_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the gender of this litigant?Use names to classify the party’s sex only if there is little ambiguity.	not ascertained, male - indication in opinion (e.g., use of masculine pronoun), male - assumed because of name
songer_appel2_7_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the race or ethnic identity of this litigant as identified in the opinion?	not ascertained, caucasian - specific indication in opinion, black - specific indication in opinion
songer_appel2_7_4	This question concerns the second listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the citizenship of this litigant as indicated in the opinion?	not ascertained, US citizen, alien
songer_appel2_7_5	This question concerns the second listed appellant. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. Which of these categories best describes the income of the litigant?	not ascertained, poor + wards of state, presumed poor
songer_appel2_8_2	This question concerns the second listed appellant. The nature of this litigant falls into the category ”miscellaneous”. Which of the following categories best describes the litigant?	fiduciary, executor, or trustee, other, nature of the litigant not ascertained

songer_appel2_8_3	This question concerns the second listed appellant. The nature of this litigant falls into the category ”miscellaneous”, specifically ”fiduciary, executor, or trustee”. Which of the following specific subcategories best describes the litigant?	trustee in bankruptcy - institution, trustee in bankruptcy - individual, executor or administrator of estate - institution
songer_appfed	What is the total number of appellants in the case that fall into the category ”the federal government, its agencies, and officialss”? Answer with a number.	N/A
songer_appfiduc	What is the total number of appellants in the case that fall into the category ”fiduciaries”? Answer with a number.	N/A
songer_applfrom	What is the type of district court decision or judgment appealed from (i.e., the nature of the decision below in the district court)?	Trial (either jury or bench trial), Injunction or denial of injunction or stay of injunction, Summary judgment or denial of summary judgment
songer_appnatpr	What is the total number of appellants in the case that fall into the category ”natural persons”? Answer with a number.	N/A
songer_appnonp	What is the total number of appellants in the case that fall into the category ”groups and associations”? Answer with a number.	N/A
songer_appstate	What is the total number of appellants in the case that fall into the category ”state governments, their agencies, and officials”? Answer with a number.	N/A
songer_appsubst	What is the total number of appellants in the case that fall into the category ”sub-state governments, their agencies, and officials”? Answer with a number.	N/A
songer_attyfee	Did the court’s ruling on attorneys’ fees favor the appellant?	No, Yes, Mixed answer
songer_bank_app1	Is the first listed appellant bankrupt?	Yes, No
songer_bank_app2	Is the second listed appellant bankrupt?	Yes, No
songer_bank_r1	Is the first listed respondent bankrupt?	Yes, No
songer_bank_r2	Is the second listed respondent bankrupt?	Yes, No
songer_capric	Did the courts’s use or interpretation of the arbitrary and capricious standard support the government? Note that APA allows courts to overturn agency actions deemed to be arbitrary or capricious, an abuse of discretion, or otherwise not in accordance with law. Overton Park emphasized this is a narrow standard, and one must prove that agency’s action is without a rational basis. This also includes the ”substantial justification” doctrine.	No, Yes, Mixed answer

songer_casetyp1_1-2	What is the specific issue in the case within the general category of ”issue”?	federal offense, state offense, not determined whether state or federal offense
songer_casetyp1_1-3-1	What is the specific issue in the case within the general category of ”issue”?	murder, rape, arson
songer_casetyp1_1-3-2	What is the specific issue in the case within the general category of ”issue”?	murder, rape, arson
songer_casetyp1_1-3-3	What is the specific issue in the case within the general category of ”issue”?	murder, rape, arson
songer_casetyp1_2-2	What is the specific issue in the case within the general category of ”issue”?	civil rights claims by prisoners and those accused of crimes, voting rights, race discrimination, sex discrimination, other civil rights
songer_casetyp1_2-3-1	What is the specific issue in the case within the general category of ”issue”?	suit for damages for false arrest or false confinement, cruel and unusual punishment, due process rights in prison
songer_casetyp1_2-3-2	What is the specific issue in the case within the general category of ”issue”?	voting rights - reapportionment & districting, participation rights - rights of candidates or groups to fully participate in the political process; access to ballot, voting rights - other (includes race discrimination in voting)
songer_casetyp1_2-3-3	What is the specific issue in the case within the general category of ”issue”?	alien petitions - (includes disputes over attempts at deportation), indian rights and law, juveniles
songer_casetyp1_3-2	What is the specific issue in the case within the general category of ”issue”?	religion, press, commercial, speech and other expression
songer_casetyp1_3-3-1	What is the specific issue in the case within the general category of ”issue”?	commercial speech, libel, slander, defamation, free exercise of religion
songer_casetyp1_3-3-2	What is the specific issue in the case within the general category of ”issue”?	obscenity, association, federal internal security and communist control acts, loyalty oaths, security risks
songer_casetyp1_4-3	What is the specific issue in the case within the general category of ”issue”?	denial of fair hearing or notice - government employees (includes claims of terminated government workers), denial of hearing or notice in non-employment context, taking clause (i.e., denial of due process under the ”taking” clause of the 5th or 14th Amendments)
songer_casetyp1_5-3	What is the specific issue in the case within the general category of ”issue”?	abortion rights, homosexual rights where privacy claim raised, contraception and other privacy claims related to marital relations or sexual behavior (not in 501 or 502)
songer_casetyp1_6-3	What is the specific issue in the case within the general category of ”issue”?	union organizing, unfair labor practices, Fair Labor Standards Act issues
songer_casetyp1_7-2	What is the specific issue in the case within the general category of ”issue”?	taxes, patents, copyright, torts, commercial disputes
songer_casetyp1_7-3-1	What is the specific issue in the case within the general category of ”issue”?	state or local tax, federal taxation - individual income tax (includes taxes of individuals, fiduciaries, & estates), federal tax - business income tax (includes corporate and parnership)

songer_casetyp1_7-3-2	What is the specific issue in the case within the general category of ”issue”?	motor vehicle, airplane, product liability
songer_casetyp1_7-3-3	What is the specific issue in the case within the general category of ”issue”?	contract disputes-general (private parties) (includes breach of contract, disputes over meaning of contracts, suits for specific performance, disputes over whether contract fulfilled, claims that money owed on contract) (Note: this category is not used when the dispute fits one of the more specific categories below), disputes over government contracts, insurance disputes
songer_casetyp1_7-3-4	What is the specific issue in the case within the general category of ”issue”?	bankruptcy - private individual (e.g., chapter 7), bankruptcy - business reorganization (e.g., chapter 11), other bankruptcy
songer_casetyp1_7-3-5	What is the specific issue in the case within the general category of ”issue”?	social security benefits (including SS disability payments), other government benefit programs (e.g., welfare, RR retirement, veterans benefits, war risk insurance, food stamps), state or local economic regulation
songer_casetyp1_7-3-6	What is the specific issue in the case within the general category of ”issue”?	disputes over real property (private), eminent domain and disputes with government over real property, landlord - tenant disputes
songer_casetyp1_9-3	What is the specific issue in the case within the general category of ”issue”?	miscellaneous interstate conflict, other federalism issue (only code as issue if opinion explicitly discusses federalism as an important issue - or if opinion explicity discusses conflict of state power vs federal power), attorneys (disbarment; etc)
songer_casetyp2_geniss	What is the second general issue in the case, other than mainissue?	criminal, civil rights, First Amendment
songer_circuit	What is the circuit of the court that decided the case?	First Circuit, Second Circuit, Third Circuit
songer_civproc1	What is the most frequently cited federal rule of civil procedure in the headnotes to this case? Answer with a number.	N/A
songer_civproc2	What is the second most frequently cited federal rule of civil procedure in the headnotes to this case? Answer with a number.	N/A
songer_classact	Is the case described in the opinion as a class action suit?	No, Yes
songer_comment	Did the agency give proper opportunity to comment?	No, Yes, Mixed answer
songer_concur	What is the number of judges who concurred in the result but not in the opinion of the court?	0, 1, 2
songer_confess	Did the court conclude that a confession or an incriminating statement was improperly admitted? Consider only incriminating statements made by the defendant.	No, Yes, Yes, but error was harmless

songer_const1	What is the most frequently cited provision of the U.S. Constitution in the headnotes to this case? If it is one of the original articles of the constitution, code the number of the article preceeded by two zeros. If it is an amendment to the constitution, code the number of the amendment (zero filled to two places) preceeded by a ”1”. Examples: 001 = Article 1 of the original constitution, 101 = 1st Amendment, 114 = 14th Amendment.	N/A
songer_const2	What is the second most frequently cited provision of the U.S. Constitution in the headnotes to this case? If it is one of the original articles of the constitution, code the number of the article preceeded by two zeros. If it is an amendment to the constitution, code the number of the amendment (zero filled to two places) preceeded by a ”1”. Examples: 001 = Article 1 of the original constitution, 101 = 1st Amendment, 114 = 14th Amendment.	N/A
songer_constit	Did the court's conclusion about the constitutionality of a law or administrative action favor the appellant?	Issue not discussed, The issue was discussed in the opinion and the resolution of the issue by the court favored the respondent, The issue was discussed in the opinion and the resolution of the issue by the court favored the appellant
songer_counsel	Did the court rule that the defendant had inadequate counsel?	No, Yes, Yes, but error was harmless
songer_counsel1	What is the nature of the counsel for the appellant?	none (pro se), court appointed, legal aid or public defender
songer_counsel2	What is the nature of the counsel for the respondent?	none (pro se), court appointed, legal aid or public defender
songer_crmproc1	What is the most frequently cited federal rule of criminal procedure in the headnotes to this case? Answer with a number.	N/A
songer_crmproc2	What is the second most frequently cited federal rule of criminal procedure in the headnotes to this case? Answer with a number.	N/A
songer_crossapp	Were there cross appeals from the decision below to the court of appeals that were consolidated in the present case?	No, Yes, Not ascertained
songer_deathpen	Did the court conclude that the death penalty was improperly imposed? Consider only the validity of the sentence, rather than whether or not the conviction was proper.	No, Yes, Yes, but error was harmless

songer_decuncon	Did the court declare any statute or administrative action unconstitutional?	no declarations of unconstitutionality, act of Congress declared unconstitutional (facial invalidity), interpretation/application of federal law invalid
songer_denovo	Did the court's use of the standard of review, ”de novo on facts” support the government? The courts generally recognize that de novo review is impractical for the bulk of agency decisions so the substantial evidence standard helps provide a middle course. Consider the de novo review of administrative action, not de novo review of trial court by appeals court.	No, Yes, Mixed answer
songer_direct1	What is the ideological directionality of the court of appeals decision?	conservative, liberal, mixed
songer_direct2	What is the ideological directionality of the court of appeals decision?	conservative, liberal, mixed
songer_discover	Did the court's interpretation of rules relating to discovery or other issues related to obtaining evidence favor the appellant?	No, Yes, Mixed answer
songer_dissent	What is the number of judges who dissented from the majority?	0, 1, 2
songer_district	From which district in the state was this case appealed?	Not applicable, Eastern, Western
songer_diverse	Did the court conclude that the parties were truly diverse?	No, Yes, Mixed answer
songer_dueproc	Did the interpretation of the requirements of due process by the court favor the appellant?	No, Yes, Mixed answer
songer_entrap	Did the court rule that the defendant was the victim of illegal entrapment?	No, Yes, Yes, but error was harmless
songer_erron	Did the court's use of the clearly erroneous standard support the government? That is, a somewhat narrower standard than substantial evidence, or ignoring usual agency standards.	No, Yes, Mixed answer
songer_execord	Did the interpretation of executive order or administrative regulation by the court favor the appellant? This does include whether or not an executive order was lawful.	No, Yes, Mixed answer
songer_exhaust	Did the court determine that it would not hear the appeal for one of the following reasons: a) administrative remedies had not been exhausted; or b) the issue was not ripe for judicial action?	No, Yes, Mixed answer
songer_fedlaw	Did the interpretation of federal statute by the court favor the appellant?	No, Yes, Mixed answer

songer_fedvst	Did the court rule that federal law should take precedence over state or local laws in a case involving the conflict of laws (i.e, which laws or rules apply)?	No, Yes, Mixed answer
songer_foreign	Did the court rule that domestic law (federal, state or local) should take precedence over foreign law in a case involving the conflict of laws (i.e., which laws or rules apply- foreign country vs federal, state, or local)?	No, Yes, Mixed answer
songer_freeinfo	Did the court rule in favor of the government when the administrative action in question related to the agency's providing information to those who request it? For example, Freedom of Information, issues of governmental confidentiality, or "government in the sunshine".	No, Yes, Mixed answer
songer_frivapp	Did the court conclude that it could not reach the merits of the case because the motion or appeal was frivolous or raised only trivial issues and was therefore not suitable for appellate review?	No, Yes, Mixed answer
songer_frivol	Did the court conclude that either the original case was frivolous or raised only trivial issues and therefore was not suitable for actions on the merits?	No, Yes, Mixed answer
songer_genapel1	What is the nature of the first listed appellant?	private business (including criminal enterprises), private organization or association, federal government (including DC)
songer_genapel2	What is the nature of the second listed appellant whose detailed code is not identical to the code for the first listed appellant?	private business (including criminal enterprises), private organization or association, federal government (including DC)
songer_geniss	What is the general issue in the case?	criminal, civil rights, First Amendment
songer_genresp1	What is the nature of the first listed respondent?	private business (including criminal enterprises), private organization or association, federal government (including DC)
songer_genresp2	What is the nature of the second listed respondent whose detailed code is not identical to the code for the first listed respondent?	private business (including criminal enterprises), private organization or association, federal government (including DC)
songer_genstand	Did the agency articulate the appropriate general standard? This question includes whether the agency interpreted the statute "correctly". The courts often refer here to the rational basis test, plain meaning, reasonable construction of the statute, congressional intent, etc. This issue also includes question of which law applies or whether amended law vs law before amendment applies.	No, Yes, Mixed answer

songer_habeas	Was the case an appeal of a decision by the district court on a petition for habeas corpus?	no, yes, state habeas corpus (criminal), yes, federal habeas corpus (criminal)
songer_immunity	Did the court refuse to reach the merits of the appeal because it concluded that the defendant had immunity?	No, Yes, Mixed answer
songer_improper	Did the court conclude that there was improper influence on the jury? For example, include jury tampering or failure to shield jury from prejudicial media accounts. Exclude prejudicial conduct by the prosecutor.	No, Yes, Yes, but error was harmless
songer_indict	Did the court rule that the indictment was defective?	No, Yes, Yes, but error was harmless
songer_indigent	Did the court rule that the defendant's rights as an indigent were violated?	No, Yes, Yes, but error was harmless
songer_initiate	What party initiated the appeal?	Original plaintiff, Original defendant, Federal agency representing plaintiff
songer_injunct	Did the court's ruling on the validity of an injunction or the denial of an injunction or a stay of injunction favor the appellant?	No, Yes, Mixed answer
songer_insane	Did the court below err in not permitting an insanity defense?	No, Yes, Yes, but error was harmless
songer_int_law	Did the court rule in favor of the appellant on an issue related to the interpretation of a treaty or international law?	No, Yes, Mixed answer
songer_interven	Did one or more individuals or groups seek to formally intervene in the appeals court consideration of the case?	no intervenor in case, intervenor = appellant, intervenor = respondent
songer_judgdisc	Did the court's ruling on the abuse of discretion by the trial judge favor the appellant? This includes the issue of whether the judge actually had the authority for the action taken, but does not include questions of discretion of administrative law judges.	No, Yes, Mixed answer
songer_judrev	Did the court conclude the decision was subject to judicial review? While questions of fact are subject to limited review, questions of law are subject to full review. The problem becomes determining which are clear questions of law or fact as they are often ”mixed”.	No, Yes, Mixed answer
songer_jurisdiction	Did the court determine that it had jurisdiction to hear this case?	No, Yes, Mixed answer
songer_juryinst	Did the court conclude that the jury instructions were improper?	No, Yes, Yes, but error was harmless
songer_late	Did the court refuse to decide the appeal because the appellant failed to comply with some rule relating to timeliness of the appeal?	No, Yes, Mixed answer

songer_majvotes	What is the number of judges who voted in favor of the disposition favored by the majority?	0, 1, 2
songer_method	What is the nature of the proceeding in the court of appeals for this case?	decided by panel for first time (no indication of re-hearing or remand), decided by panel after re-hearing (second time this case has been heard by this same panel), decided by panel after remand from Supreme Court
songer_mootness	Did the court conclude that an issue was moot?	No, Yes, Mixed answer
songer_notice	Decisions that affect life, liberty, or property must be preceded by adequate notice and an opportunity for a fair hearing. Did the agency give proper notice?	No, Yes, Mixed answer
songer_numappel	What is the total number of appellants in the case? Answer with a number.	N/A
songer_numresp	What is the total number of respondents in the case? Answer with a number.	N/A
songer_opinstat	Is the opinion writer identified in the opinion, or was the opinion per curiam?	Signed, with reasons, Per curiam, with reasons, Not ascertained
songer_origin	What type of court made the original decision?	Federal district court (single judge), 3 judge district court, State court
songer_othadmis	Did the court rule that some evidence, other than a confession made by the defendant or illegal search and seizure, was inadmissibile (or did ruling on appropriateness of evidentary hearing benefit the defendant)?	No, Yes, Yes, but error was harmless
songer_othappth	Did the court refuse to rule on the merits of the appeal because of some threshhold issue other than timeliness or frivolousness that was relevant on appeal but not at the original trial?	No, Yes, Mixed answer
songer_othcrim	Did the court rule for the defendant on grounds other than procedural grounds? For example, right to speedy trial, double jeopardy, confrontation, retroactivity, self defense. This includes the question of whether the defendant waived the right to raise some claim.	No, Yes, Yes, but error was harmless
songer_othjury	Did the court conclude that the jury composition or selection was invalid or that the jury was biased or tampered with?	No, Yes, Yes, but error was harmless
songer_oththres	Did the court refuse to rule on the merits of the appeal because of a threshhold issue other than lack of jurisdiction, standing, mootness, failure to state a claim, exhaustion, timeliness, immunity, frivolousness, or nonjusticiable political question?	No, Yes, Mixed answer

songer_plea	Did the court rule for the defendant on an issue related to plea bargaining? Plea bargain includes all challenges to plea.	No, Yes, Yes, but error was harmless
songer_polquest	Did the court refuse to rule on the merits of the case because it was considered to be a nonjusticiable ”political question”?	No, Yes, Mixed answer
songer_post_trl	Did the court's ruling on some post-trial procedure or motion (e.g., allocating court costs or post award relief) favor the appellant? This doe not include attorneys' fees, but does include motions to set aside a jury verdict.	No, Yes, Mixed answer
songer_prejud	Was there prejudicial conduct by prosecution?	No, Yes, Yes, but error was harmless
songer_pretrial	Did the court's rulings on pretrial procedure favor the appellant? This includes whether or not there is a right to jury trial, whether the case should be certified as a class action, or whether a prospective party has a right to intervene in the case, but does not include rulings on motions for summary judgment.	No, Yes, Mixed answer
songer_procdis	Did the court uphold the dismissal by district court on procedural grounds?	No, Yes, Yes, but error was harmless
songer_procedur	Did the interpretation of federal rule of procedures, judicial doctrine, or case law by the court favor the appellant?	No, Yes, Mixed answer
songer_r_bus	What is the total number of respondents in the case that fall into the category ”private business and its executives”? Answer with a number.	N/A
songer_r_fed	What is the total number of respondents in the case that fall into the category ”the federal government, its agencies, and officialss”? Answer with a number.	N/A
songer_r_fiduc	What is the total number of respondents in the case that fall into the category ”fiduciaries”? Answer with a number.	N/A
songer_r_natpr	What is the total number of respondents in the case that fall into the category ”natural persons”? Answer with a number.	N/A
songer_r_nonp	What is the total number of respondents in the case that fall into the category ”groups and associations”? Answer with a number.	N/A
songer_r_state	What is the total number of respondents in the case that fall into the category ”state governments, their agencies, and officials”? Answer with a number.	N/A

songer_r_stid	What is the state of the first listed state or local government agency that is a respondent?	not, Alabama, Alaska
songer_r_subst	What is the total number of respondents in the case that fall into the category ”sub-state governments, their agencies, and officials”? Answer with a number.	N/A
songer_realapp	Are the formally listed appellants in the case the ”real parties”, that is, are they the parties whose real interests are most directly at stake?	both 1st and 2nd listed appellants are real parties (or only one appellant, and that appellant is a real party), the 1st appellant is not a real party, the 2nd appellant is not a real party
songer_realresp	Are the formally listed respondents in the case the ”real parties”, that is, are they the parties whose real interests are most directly at stake?	both 1st and 2nd listed respondents are real parties (or only one respondent, and that respondent is a real party), the 1st respondent is not a real party, the 2nd respondent is not a real party
songer_record	Did the agency fail to develop an adequate record? For example, if the court was unable to determine what doctrine was used for the decision or unable to determine the basis of the decision.	No, Yes, Mixed answer
songer_respond1_1_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What is the scope of this business?	local, neither local nor national, national or multi-national
songer_respond1_1_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What category of business best describes the area of activity of this litigant which is involved in this case?	agriculture, mining, construction
songer_respond1_1_4	This question concerns the first listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”, specifically ”agriculture”. What subcategory of business best describes this litigant?	single family farm, commercial farm, agri-business, farm - other
songer_respond1_2_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”private organization or association”. What category of private associations best describes this litigant?	business, trade, professional, or union (BTPU), other
songer_respond1_2_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”private organization or association”, specifically ”business, trade, professional, or union (BTPU)”. What subcategory of private association best describes this litigant?	Business or trade association, utilities co-ops, Professional association - other than law or medicine

songer_respond1_3_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”federal government (including DC)”. Which category of federal government agencies and activities best describes this litigant?	cabinet level department, courts or legislative, agency whose first word is ”federal”
songer_respond1_3_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”federal government (including DC)”, specifically ”cabinet level department”. Which specific federal government agency best describes this litigant?	Department of Agriculture, Department of Commerce, Department of Defense (includes War Department and Navy Department)
songer_respond1_4_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”. Which category of substate government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_respond1_4_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”, specifically ”legislative”. Which specific substate government agency best describes this litigant?	City/county council, School Board, board of trustees for college or junior college, Other legislative body
songer_respond1_5_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”. Which category of state government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_respond1_5_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”, specifically ”legislative”. Which specific state government agency best describes this litigant?	Legislature or separate house as an organization, Legislative Committee or Commission, Other Legislative Unit
songer_respond1_7_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the gender of this litigant?Use names to classify the party’s sex only if there is little ambiguity.	not ascertained, male - indication in opinion (e.g., use of masculine pronoun), male - assumed because of name

songer_respond1_7_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the race or ethnic identity of this litigant as identified in the opinion?	not ascertained, caucasian - specific indication in opinion, black - specific indication in opinion
songer_respond1_7_4	This question concerns the first listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the citizenship of this litigant as indicated in the opinion?	not ascertained, US citizen, alien
songer_respond1_7_5	This question concerns the first listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. Which of these categories best describes the income of the litigant?	not ascertained, poor + wards of state, presumed poor
songer_respond1_8_2	This question concerns the first listed respondent. The nature of this litigant falls into the category ”miscellaneous”. Which of the following categories best describes the litigant?	fiduciary, executor, or trustee, other, nature of the litigant not ascertained
songer_respond1_8_3	This question concerns the first listed respondent. The nature of this litigant falls into the category ”miscellaneous”, specifically ”fiduciary, executor, or trustee”. Which of the following specific subcategories best describes the litigant?	trustee in bankruptcy - institution, trustee in bankruptcy - individual, executor or administrator of estate - institution
songer_respond2_1_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What is the scope of this business?	local, neither local nor national, national or multi-national
songer_respond2_1_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”. What category of business best describes the area of activity of this litigant which is involved in this case?	agriculture, mining, construction

songer_respond2_1_4	This question concerns the second listed respondent. The nature of this litigant falls into the category ”private business (including criminal enterprises)”, specifically ”agriculture”. What subcategory of business best describes this litigant?	single family farm, commercial farm, agri-business, farm - other
songer_respond2_2_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”private organization or association”. What category of private associations best describes this litigant?	business, trade, professional, or union (BTPU), other
songer_respond2_2_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”private organization or association”, specifically ”business, trade, professional, or union (BTPU)”. What subcategory of private association best describes this litigant?	Business or trade association, utilities co-ops, Professional association - other than law or medicine
songer_respond2_3_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”federal government (including DC)”. Which category of federal government agencies and activities best describes this litigant?	cabinet level department, courts or legislative, agency whose first word is ”federal”
songer_respond2_3_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”federal government (including DC)”, specifically ”cabinet level department”. Which specific federal government agency best describes this litigant?	Department of Agriculture, Department of Commerce, Department of Defense (includes War Department and Navy Department)
songer_respond2_4_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”. Which category of substate government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_respond2_4_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”sub-state government (e.g., county, local, special district)”, specifically ”legislative”. Which specific substate government agency best describes this litigant?	City/county council, School Board, board of trustees for college or junior college, Other legislative body

songer_respond2_5_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”. Which category of state government best describes this litigant?	legislative, executive/administrative, bureaucracy providing services
songer_respond2_5_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”state government (includes territories & commonwealths)”, specifically ”legislative”. Which specific state government agency best describes this litigant?	Legislature or separate house as an organization, Legislative Committee or Commission, Other Legislative Unit
songer_respond2_7_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the gender of this litigant?Use names to classify the party’s sex only if there is little ambiguity.	not ascertained, male - indication in opinion (e.g., use of masculine pronoun), male - assumed because of name
songer_respond2_7_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the race or ethnic identity of this litigant as identified in the opinion?	not ascertained, caucasian - specific indication in opinion, black - specific indication in opinion
songer_respond2_7_4	This question concerns the second listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. What is the citizenship of this litigant as indicated in the opinion?	not ascertained, US citizen, alien
songer_respond2_7_5	This question concerns the second listed respondent. The nature of this litigant falls into the category ”natural person (excludes persons named in their official capacity or who appear because of a role in a private organization)”. Which of these categories best describes the income of the litigant?	not ascertained, poor + wards of state, presumed poor
songer_respond2_8_2	This question concerns the second listed respondent. The nature of this litigant falls into the category ”miscellaneous”. Which of the following categories best describes the litigant?	fiduciary, executor, or trustee, other, nature of the litigant not ascertained

songer_respond2_8_3	This question concerns the second listed respondent. The nature of this litigant falls into the category ”miscellaneous”, specifically ”fiduciary, executor, or trustee”. Which of the following specific subcategories best describes the litigant?	trustee in bankruptcy - institution, trustee in bankruptcy - individual, executor or administrator of estate - institution
songer_rtcouns	Did the court rule that the defendant’s right to counsel was violated (for some reason other than inadequate counsel)?	No, Yes, Yes, but error was harmless
songer_search	Did the court below improperly rule for the prosecution on an issue related to an alleged illegal search and seizure?	No, Yes, Yes, but error was harmless
songer_sentence	Did the court conclude that some penalty, excluding the death penalty, was improperly imposed?	No, Yes, Yes, but error was harmless
songer_source	What forum heard this case immediately before the case came to the court of appeals?	Federal district court (single judge), 3 judge district court, State court
songer_st_v_st	Did the court rule in favor of the appellant on the issue of a conflict of laws (which laws or rules apply) other than federal v state or foreign v domestic (e.g., one state vs second state)?	No, Yes, Mixed answer
songer_standing	Did the court determine that the parties had standing?	No, Yes, Mixed answer
songer_state	In what state or territory was the case first heard?	not, Alabama, Alaska
songer_stateclaim	Did the court dismiss the case because of the failure of the plaintiff to state a claim upon which relief could be granted?	No, Yes, Mixed answer
songer_stpolicy	Did the interpretation of state or local law, executive order, administrative regulation, doctrine, or rule of procedure by the court favor the appellant?	No, Yes, Mixed answer
songer_subevid	Did the court’s interpretation of the substantial evidence rule support the government? For example, ”such evidence as a reasonable mind might accept as adequate to support a conclusion” or ”more than a mere scintilla”. This issue is present only when the court indicates that it is using this doctrine, rather than when the court is merely discussing the evidence to determine whether the evidence supports the position of the appellant or respondent.	No, Yes, Mixed answer
songer_suffic	Did the court rule that there was insufficient evidence for conviction?	No, Yes, Yes, but error was harmless

songer_summary	Did the court's ruling on the appropriateness of summary judgment or the denial of summary judgment favor the appellant?	No, Yes, Mixed answer
songer_timely	Did the court conclude that it could not reach the merits of the case because the litigants had not complied with some rule relating to timeliness, a filing fee, or because a statute of limitations had expired?	No, Yes, Mixed answer
songer_treat	What is the disposition by the court of appeals of the decision of the court or agency below?	stay, petition, or motion granted, affirmed; or affirmed and petition denied, reversed (include reversed & vacated)
songer_trialpro	Did the court's ruling on procedure at trial favor the appellant? This includes jury instructions and motions for directed verdicts made during trial.	No, Yes, Mixed answer
songer_two_issues	Are there two issues in the case?	no, yes
songer_typeiss	What is the general category of issues discussed in the opinion of the court?	criminal and prisoner petitions, civil - government, diversity of citizenship
songer_usc1	What is the most frequently cited title of the U.S. Code in the headnotes to this case? Answer with a number.	N/A
songer_usc1sect	What is the number of the section from the title of the most frequently cited title of the U.S. Code in the headnotes to this case, that is, title usc1? Answer with a number.	N/A
songer_usc2	The most frequently cited title of the U.S. Code in the headnotes to this case is usc1. What is the second most frequently cited title of this U.S. Code in the headnotes to this case? Answer with a number.	N/A
songer_usc2sect	What is the number of the section from the title of the second most frequently cited title of the U.S. Code in the headnotes to this case, that is, title usc2? Answer with a number.	N/A
songer_weightev	Did the factual interpretation by the court or its conclusions (e.g., regarding the weight of evidence or the sufficiency of evidence) favor the appellant?	No, Yes, Mixed answer
songer_whlaws	Did the court's discussion of which state's laws should control their ruling in the case support the position taken by the appellant?	No, Yes, Mixed answer