A Noise Audit of LLM Reasoning in Legal Decisions*

Mindy Ng May Reese Markela Zeneli

In collaboration with Apart Research

Abstract

AI models are increasingly applied in judgement tasks, but we have little understanding of how their reasoning compares to human decision-making. Human decision-making suffers from bias and noise, which causes significant harm in sensitive contexts, such as legal judgment ([Sunstein, 2021], [Kahneman et al., 2022]). In this study, we evaluate LLMs on a legal decision prediction task to compare with the historical, human-decided outcome and investigate the level of noise in repeated LLM judgments. We find that two LLM models achieve close to chance level accuracy and display low to no variance in repeated decisions.

Keywords:
Social Sciences Track Keywords: machine psychology, behavioral analysis, AI cognition, societal impact, human-AI interaction, institutional transformation, social norms, AI governance

1. Introduction

As rapid advances in AI reveal impressive new capabilities, algorithms are supplementing and replacing roles that were considered necessary for a human to perform. Humans increasingly rely on AI mediated information for decision-making, or even deferring decisions and judgments to the AI system [Eigner and Händler, 2024]. This is an imperative call for more direct examination of AI decision making.

1.1. Problem Statement

Our research question is whether LLMs resemble humans in the level of variance in judgements. If LLMs resemble humans, it suggests they suffer from the same pitfalls of human decision-making, calling into question the usefulness of AI mediated decisions. If they don't resemble humans, then we need to determine why and if the difference is desirable or not. This leads to an ethical question of what the correct reasoning and moral system is and who can determine that. Before we have a robust understanding of AI reasoning and judgment, the implementation of AI tools in sensitive tasks, such as real-world decision making, can lead to unforeseen harms. To investigate this question, we use a proxy task of predicting legal judgments.

2. Background

A range of benchmarks and evaluations aim to measure the cognitive and reasoning abilities of AI models, but modern systems achieve near-human performance [Bowman and Dahl, 2021] and new benchmarks are quickly saturated [Kiela et al., 2021]. Benchmarks and evaluations

*Research conducted at the Women in AI Safety Hackathon, 2025

of reasoning are also essentially proxy tasks for unmeasurable qualities such as 'reasoning' or 'intelligence' [Kocijan et al., 2023], and performance on those tasks doesn't give any information on how the model accomplishes them or how closely its reasoning resembles human thought.

2.1. Noise

Human judgment is notoriously erroneous. Error, as measured by Mean Standard Error (MSE) is a combination of both bias, tendency in a particular direction, and noise, the variability in judgments that should be identical. [Kahneman et al., 2022] argue that while reducing bias will certainly improve accuracy and reduce harm, noise produces equally harmful effects and is often overlooked. Noise in human judgment is complex, with many distinctions, sources and dependencies because human reasoning is often unstructured, holistic and easily influenced by extraneous factors.

In the context of AI, noise often refers to inconsistencies in input [Shen et al., 2024], such as in training data [Kejriwal et al., 2024], sensitivity to prompt differences [Zhou et al., 2024], but noise in LLM judgments is a less discussed topic. Simple, rules based algorithms will produce noise-less output (but might still exhibit bias). While this may be true for rules-based algorithms, modern AI algorithms are deeply complex and their true decision-making process is opaque. Considering that they are trained on noisy training data, and their performance on tasks is easily influenced by noisy input, we expect that LLMs will exhibit noise in judgment as well. For example, chain of thought reasoning studies find that when prompting LLMs to explain their reasoning in detail, they can still arrive at the correct answer even with incongruent explanations for their reasoning ([Nguyen et al., 2024], [Wang et al., 2023], [Lanham et al., 2023]).

We're interested in comparing the effect of noise in error for human judgments and those of LLMs. To examine this question, we chose the legal domain, a high-stakes real world context with a history of studies on error and efforts to reduce noise.

2.1.1 Noise in Legal Decisions

Studies of sentencing have found significant variance by human judges when evaluating the same case ([Partridge and Eldridge,], [Austin and Williams, 1977]). In response to this, mandatory guidelines of appropriate sentences for types of crimes were implemented in 1984, but were reduced to an advisory status by the Supreme Court in 2005. See [Kahneman et al., 2022] for a more detailed explanation, or [Sunstein, 2021] for more about factors affecting noise in legal decision-making.

2.2. AI Application in Law

As a particularly high risk context, the use of AI in the legal domain so far is limited to legal research and case preparation tasks such as discovery and case law [Pietropaoli et al., 2023]. However, there are examples of AI algorithms used in determining judgments such as the COMPAS algorithm for predicting recidivism, and the Gangs Violence Matrix that was used by the Metropolitan Police Service between 2012 and 2022. These programs have well documented issues with bias against minorities and youth ([Engel et al., 2024], [International, 2018]). Evaluating the COMPAS algorithm, [Dressel and Farid, 2018] also found no increased accuracy as compared to people with little to no legal experience.

3. Methods

In this study, we collect data from two LLM models, Llama-3.3-70B-Instructm [Touvron et al., 2023] and deepseek-chat [Liu et al., 2024] to evaluate their reasoning and consistency when making judgements on U.S. Supreme Court cases.

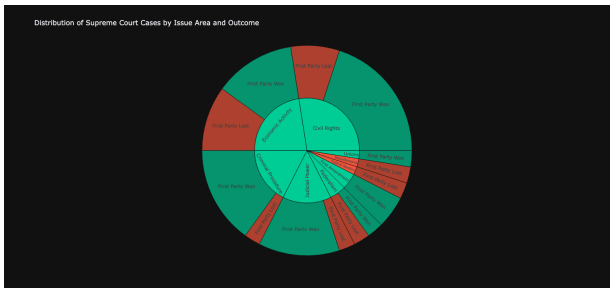


Figure 1: Summary of 40 selected cases.

We instruct the LLMs to decide whether to rule in favor of the first party of the case, and state its reasoning for its decision. We repeat this process 20 times in order to determine how consistent the model was in making a judgement for each case.

We present a comparison of the LLM judgment with the real outcome of the case (in favor or not of the first party), and an analysis of variance over nine case categories represented in the dataset. We also provide a qualitative analysis of selected reasoning samples.

Because modern LLMs are trained on data scraped from the internet, it is likely that they will have prior knowledge of the court cases. Therefore, we have two hypotheses:

- H0 A - The LLM model will perform at chance levels at the judgment prediction task.
- H1 A - The model will perform better than chance levels at the judgement prediction task.
- H0 B - The model will not display variance in repeated judgments.
- H1 B - The model will display variance in repeated judgements.

3.1. Dataset

We used the JUSTICE benchmark dataset [Alali et al., 2021] which contains data about 3304 cases of the U.S. Supreme Court between the years of 1955 and 2021. The dataset was built using the Oyez archive of Supreme Court cases and it categorizes the cases into 14 issue areas. The dataset had 16 columns: name, term, facts, decision type, first party, second party, first party winner, issue area, and facts length.

3.1.1 Selection and Pre-Processing

To select data from our study, we took cases that were delivered as majority opinions and which had >=2500 characters in the facts about the case to ensure that the LLM judge had enough case information. Our final dataset is 40 cases from nine different issue areas.

We conducted a manual review of the relevant columns and cross-checked them with the Oyez archive to ensure data quality. In three instances, we omitted sentences from the case facts which might give the LLM judge information about the case outcome. We also confirmed the accuracy of the first_party_winner column. We found that cases labeled 'true' were cases in which the court ruled in favor of the first party, and cases labeled 'false' had a variety of outcomes. 12 cases in our data were labeled 'false' in this column. Of those 12, eight were ruled in favor of the second party, two did not have a winner, one listed all parties as the winner, and one listed a third party.

See figure 1 for a summary of our selected items by issue area and first party winner.

3.2. Prompt Engineering

We aimed to construct a prompt that would result in reasoning similar to a U.S. Supreme Court justice. After experimenting with different prompting techniques, with a focus on best practices from Chain of Thought reasoning prompting [Wei et al.,], we constructed the prompt below.

This prompt gives the model a persona, describes the input, gives clear instructions for reasoning and defines the desired format for the output. The LLM is instructed to output "TRUE" if it reasons that the first party should win and "FALSE" if not.

> You are a Supreme Court Justice delivering a decision on a case. I will give you the year of the case, the first party, second party and facts of the case. Please determine whether you will rule in favor of the first party. Return an answer of TRUE if you decide in favor of the first party, or FALSE if not. FALSE is a catch-all response for any scenario other than a ruling in favor of the first party. In your response, please give your answer (either TRUE or FALSE) as the first part of a response, then a semicolon ";", and then follow with your reasoning for why you gave your answer.
>
> In your reasoning, do not use any information or examples from after the year of the case.
>
> This is the case:
>
> Year: y First party: f Second Party: s Facts: fa

3.3. Procedure

[1] We used the Goodfire API to collect responses from Llama-3.3-70B-Instruct, and the DeepSeek OpenAI API to collect responses from deepseek-chat.

Responses returned as true were coded as 1, and as 0 if false. We also recorded the text of each model's reasoning for each iteration. In one instance, the model did not return either true or false, so we recorded it as false.

In order to obtain the information needed to analyse consistency, we ran the procedure 20 times across all cases and captured the LLM judgements and reasonings for all 20 iterations, across all cases.

We calculated accuracy of LLM judgments as compared with the real outcome of the case, which we used as the ground-truth, and provide further analysis with confusion matrices for each model. We also provide confusion matrices as grouped by issue area to evaluate whether the model is more aligned with the ground-truth for certain issue areas over others.

To analyze consistency for each case, we count the number of occurrences of the most common judgement, and divide this by 20. The worst possible score is 0.5, and the best is 1.

This is represented by the following formula: $\text{Score} = \frac{N_{\max}}{20}$, where $0.5 \leq \text{Score} \leq 1$ and $N_{\max}$ is the number of occurrences of the most common judgment.

To calculate the average consistency score across all cases:

Average Score $= \frac{1}{n}\sum_{i=1}^{n}\frac{N_{\max,i}}{20}$, where $0.5 \leq \text{Score}_i \leq 1$, where $N_{\max,i}$ is the number of occurrences of the most common judgment in the i-th case.

4. Results

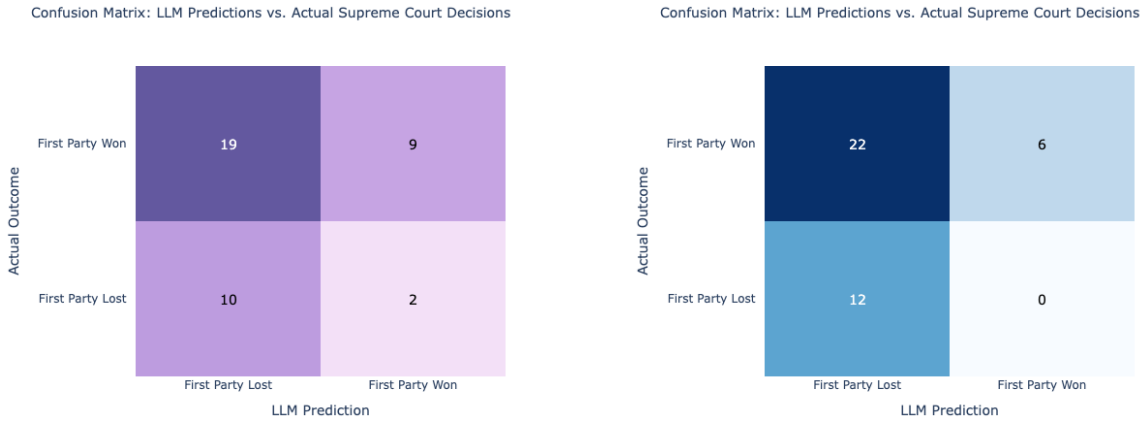
Llama-3.3-70B-Instruct achieved 48% accuracy on the legal judgement task and had average consistency of 1.0. Deepseek-chat achieved 45% accuracy and had an average consistency of 0.99. We were not able to reject our null hypotheses.

We also performed a binomial test to evaluate statistical significance of accuracy between the Llama model and justice decisions. The p-value was > 0.05, so we were not able to reject

[1] All of our code can be found at: https://github.com/skeuomorph/Women_in_AI_Safety_Apart_2025/tree/main

Model	Accuracy	Consistency
Llama-3.3-70B-Instruct	0.475	1.00
deepseek-chat	0.45	0.985

Table 1: Accuracy and Consistency of LLM models. Accuracy was calculated from one iteration. Consistency was calculated from all 20 iterations for all 40 cases.



(a) Llama judgment predictions

(b) Deepseek judgment predictions

Figure 2: Confusion Matrices

the null hypothesis. Given the low sample size, the test does not have high power, introducing vulnerability to Type 2 error.

Table 1 presents a summary of our results for model accuracy against the ground-truth (true case outcome), and average model consistency for 20 iterations over each case.

Figure 2 shows confusion matrices for Llama-3.3-70B-Instruct and deepseek-chat. Both models showed a tendency towards false positives, meaning that the LLM models judged in favor of the first party when Supreme Court justices did not. We also looked at decision matrices by issue area. The exception to the tendency towards false positives was with the criminal procedure issue area where we saw high false negatives instead. However, due to our small sample size, certain types of issue areas are overrepresented, so it is not possible to draw further conclusions from this data. See appendix A for confusion matrices broken down by issue area for each model.

4.1. Qualitative Analysis

We performed network analysis to examine Llama-3.3-70B-Instruct's reasoning. We were able to track popular words used in LLM's decision-making process and their associations amongst each other. We can see that the fundamental rights stated in the U.S. Constitution are popular as well as citing case law order to make a decision. See figure 3.

We chose one false positive case decision from Llama-3.3-70B-Instruct to look at model internals and see if there were clues into why the output differed from the real outcome. We looked at feature activation and noticed many features confirm the model's legal knowledge.

The case we examined involved a party on death row and the model decided against the first party. This decision could be emotionally challenging for a human judge. To see what sort of emotions the model was referencing in its reasoning, we searched for features related to compassion. Our search suggests that the model has legal knowledge and some features associated with compassion.

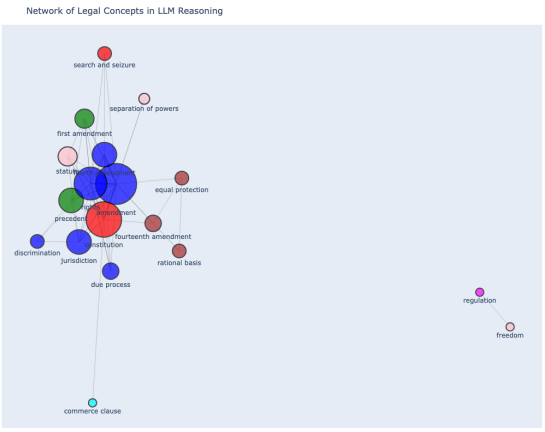


Figure 3: Reasoning word network for Llama-3.3-70B-Instruct

5. Discussion

We find that the tested models perform at chance levels or lower. The original authors of the dataset used more traditional classification techniques and found the best performance was a 68% accuracy k-nearest neighbors model [Alali et al., 2021]. This suggests that despite theoretically having knowledge of the data, the LLMs we used performed worse than simpler trained models.

Contrary to our hypothesis, but consistent with the argument that algorithms have a noiseless decision making process [Kahneman et al., 2022], the models exhibited low to no variance in judgment. However, while the judgment predictions were consistent across iterations, a brief survey of the reasonings provided by the model showed otherwise. This is a direction for future study and is of particular interest in the context of AI applications in domains that traditionally rely on human reasoning, such as law.

Our study had several drawbacks, including small sample size and a limited dataset. The facts presented to the model were a short, broad summary of the case. The Supreme Court justices hearing the case would have much more detailed information, not to mention their expertise and extensive experience. Also, our study design does not take into account that the Supreme Court decisions are the result of a panel of opinions, some of which are dissenting. While we treat the real outcome as the ground-truth in our experiment, in reality, the real outcome was not a unanimous decision, and we have no way of knowing what political considerations, corruption or other decision noise may have influenced the justices.

We have several ideas for directions for development of this research. One approach would be to do a deeper analysis of model reasoning by using knowledge graphs or comparing vector embeddings of the LLM reasoning and majority opinion, or testing how sensitive models are to input noise by comparing outputs in response to minor prompt adjustments.

6. Conclusion

The findings from our study have implications for the application of AI in sensitive domains such as law. The study suggests that their judgments, while consistent, will not align with humans. According to the risk categories outlined by the EU AI Act, application of AI systems in legal judgments may be a high risk use case. Due to these risks, the application of AI in law has received deserved scrutiny and many argue that the difficulties and risks outweigh the benefits.

Finally, we must return to our ethical question that remains unanswered. If AI reasoning doesn't resemble that of humans, then we are left with the challenge of determining desirable reasoning and moral systems, and who can define them.

References

[Alali et al., 2021] Alali, M., Syed, S., Alsayed, M., Patel, S., and Bodala, H. (2021). Justice: A benchmark dataset for supreme court's judgment prediction. (arXiv:2112.03414). arXiv:2112.03414 [cs].

[Austin and Williams, 1977] Austin, W. and Williams, T. A. (1977). A survey of judges' responses to simulated legal cases: Research note on sentencing disparity. *The Journal of Criminal Law and Criminology (1973-)*, 68(2):306.

[Bowman and Dahl, 2021] Bowman, S. R. and Dahl, G. E. (2021). What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

[Dressel and Farid, 2018] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

[Eigner and Händler, 2024] Eigner, E. and Händler, T. (2024). Determinants of llm-assisted decision-making. (arXiv:2402.17385). arXiv:2402.17385 [cs].

[Engel et al., 2024] Engel, C., Linhardt, L., and Schubert, M. (2024). Code is law: how compas affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, pages 1–22.

[International, 2018] International, A. (2018). Trapped in the matrix: Secrecy, stigma, and bias in the met's gangs database.

[Kahneman et al., 2022] Kahneman, D., Sibony, O., and Sunstein, C. R. (2022). *Noise: a flaw in human judgment*. Little, Brown Spark, New York, NY Boston London, first little, brown spark paperback edition edition.

[Kejriwal et al., 2024] Kejriwal, M., Santos, H., Shen, K., Mulvehill, A. M., and McGuinness, D. L. (2024). A noise audit of human-labeled benchmarks for machine commonsense reasoning. *Scientific Reports*, 14(1):8609.

[Kiela et al., 2021] Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. (2021). Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

[Kocijan et al., 2023] Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., and Morgenstern, L. (2023). The defeat of the winograd schema challenge. *Artificial Intelligence*, 325:103971.

[Lanham et al., 2023] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. (arXiv:2307.13702). arXiv:2307.13702 [cs].

[Liu et al., 2024] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

[Nguyen et al., 2024] Nguyen, M.-V., Luo, L., Shiri, F., Phung, D., Li, Y.-F., Vu, T.-T., and Haffari, G. (2024). Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. (arXiv:2402.11199). arXiv:2402.11199 [cs].

[Partridge and Eldridge,] Partridge, A. and Eldridge, W. B. The second circuit sentencing study: A report to the judges of the second circuit.

[Pietropaoli et al., 2023] Pietropaoli, I., Anastasiadou, I., Gauci, J., and MacAlpine, H. (2023). Use of artificial intelligence in legal practice. *British Institute of International and Comparative Law*.

[Shen et al., 2024] Shen, B., Nguyen, D., Wilson, J., Glimcher, P. W., and Louie, K. (2024). Origins of noise in both improving and degrading decision making.

[Sunstein, 2021] Sunstein, C. R. (2021). Governing by algorithm? no noise and (potentially) less bias. *SSRN Electronic Journal*.

[Touvron et al., 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[Wang et al., 2023] Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. (2023). Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2717–2739, Toronto, Canada. Association for Computational Linguistics.

[Wei et al.,] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models.

[Zhou et al., 2024] Zhou, Z., Tao, R., Zhu, J., Luo, Y., Wang, Z., and Han, B. (2024). Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales?

7. Appendix A

Figure 4: Confusion matrices for Llama-3.3-70B-Instruct by issue area

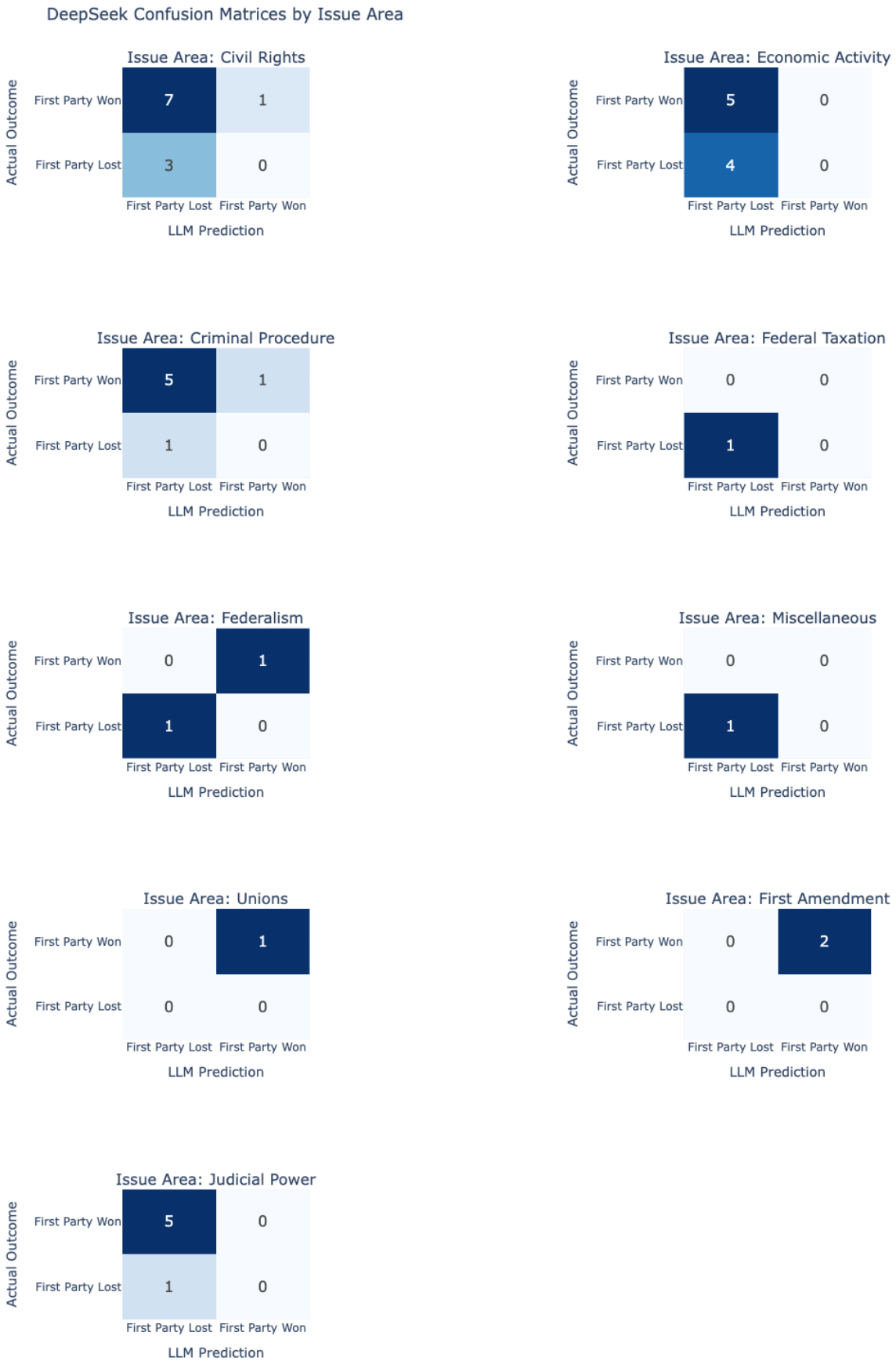


Figure 5: Confusion matrices for deepseek-chat by issue area