Bias in Legal Data for Generative AI

Holli Sargeant [1] **Måns Magnusson** [2]

Abstract

As large legal corpora become more abundant, its use in developing generative legal AI is poised to transform the legal sector. However, the use of case law data necessitates a more critical examination of the ethical and legal implications for the development of generative legal AI tools. This research conducts a survey of various types of bias, their sources, and potential impacts.

1. Introduction

As artificial intelligence (AI) and machine learning (ML) revolutionise industries worldwide, the legal sector is poised for a transformative shift. An industry traditionally perceived as conservative and resistant to change now faces a technological revolution that promises both substantial opportunities and risks. Put simply, ML is learning from data (Hastie et al., 2009), and large language models (LLMs) learn from vast amounts of data (Chang et al., 2024). The development of AI in the legal domain hinges substantially on the availability and quality of legal data. Legal text has distinct characteristics compared to generic corpora, as the field uses notoriously complex, domain-specific language (Ruhl, 2008; Katz & Bommarito, 2014; Nazarenko & Wyner, 2017; Dale, 2017; Friedrich, 2021; Glogar, 2023; Trancoso et al., 2024). One of the primary strategies for enhancing the capabilities of legal AI involves pre-training language models on a large corpus of legal text (Katz et al., 2020; Chalkidis et al., 2022; Wang et al., 2023), shown in several recent legal LLMs (Chalkidis et al., 2020; Xiao et al., 2021; Zheng et al., 2021; Song et al., 2022; Huang et al., 2023).

The expansion of legal corpora across various jurisdictions has played a critical role in advancing computational research on legal texts (Poudyal et al., 2020; Hwang et al., 2022; Henderson et al., 2022; Niklaus et al., 2023a;b; Östling et al., 2023; Harvard University, 2024). However, these legal datasets are not without challenges, particularly concerning the prevalence of bias. Bias in legal data will likely impact generative AI models across their various applications. From unfairness or errors in prediction tasks to biased information generation in question-answer tasks, addressing bias in various contexts is crucial for fair and reliable generative legal AI tools. This paper surveys relevant types of bias in legal data, emphasising their implications for the development and deployment of generative AI.

2. Types of Bias in Legal Data

Under-representation bias arises when relevant information is missing from a dataset, often resulting in reduced accuracy (Shahbazi et al., 2023). The selective publication of judicial decisions results in significant gaps and biases in legal datasets. Factors such as case importance, complexity, court hierarchy, and precedent-setting potential all influence this discretionary publication process (Byrom, 2022; Justice UK, 2023). Research suggests that large portions of UK case law is missing from public repositories, limiting access to representative data (Shubber, 2022; Hoadley et al., 2022; House of Commons, 2022; Gisborne et al., 2022). If the data used to train the generative AI model does not reflect the population it is deployed on, it may also be non-representative data. For instance, models trained on US law may be unsuitable for certain generative AI applications for the UK. Additionally, there is a considerable selection bias towards only litigating "edge" or "marginal" cases where the interpretation of a legal issue is unclear in prior precedent or legislation (Priest & Klein, 1984). Under-representation bias can lead to bias in generative AI models that misrepresent legal realities.

Historical bias occurs when data does not reflect current reality (Lattimore et al., 2020). Datasets like the UK Cambridge Law Corpus contains cases from the 16th century onwards (Östling et al., 2023), meaning older case law often mirrors outdated laws and social norms that are unacceptable today. For example, historical case law will include references to laws that allowed what would now be considered unlawful discrimination and will use language that is now considered outdated or offensive. Researchers have identified the risks of demographic disparities in legal

[1]Faculty of Law, University of Cambridge, Cambridge, UK [2]Department of Statistics, Uppsala University, Uppsala, Sweden. Correspondence to: Holli Sargeant <hs775@cam.ac.uk>.

Workshop on Generative AI and Law (GenLaw) at International Conference on Machine Learning 2024.

texts (Sargent & Weber, 2021). In a recent study, Bozdag et al. (2024) identify that Legal-BERT inherits gender bias most likely from its training data (including case law from the EU and US). Sevim et al. (2023) concluded that legal corpora contained significant gender bias across various countries, which are reflected in NLP models trained on these texts. Addressing historical bias is crucial to prevent the perpetuation of outdated and discriminatory decisions.

Label bias arises when recorded labels in a dataset reflect a disparity across different individuals (Lakkaraju et al., 2017; Jiang & Nachum, 2020). Labelling in legal datasets may be varied by issues of imperfect decisions or human bias.

Imperfect Decisions. Courts are not oracles, and case law should not be understood as absolute truth (Coleman, 1995; Smith, 1985; Heiner, 1986). First, judges have an inevitable position to make decisions under uncertainty with imperfect information (only as presented by the parties and their legal counsel) and may imperfectly use it (Heiner, 1986). Often, even "reasonable minds may differ on the results of given cases" (Smith, 1985). Second, legal corpora will inherently contain cases that have been overturned on appeal to higher courts or in subsequent cases. While this is well understood by lawyers who review the authority of a case, such context will be lost in bulk data if not managed correctly.

Judicial Prejudice. Prejudice within the judiciary further complicates decision-making under uncertainty. In a recent report, the UK Judiciary was identified as "institutionally racist" (Monteith et al., 2022), which builds on previous inquiries into judicial discrimination (Lammy, 2017). Researchers conducted a survey revealing that more than half of the legal professionals witnessed racial biases in action, directed primarily towards black court users, ranging from derogatory remarks to discrimination in judicial decision-making (Monteith et al., 2022). While no computational research has identified biases in UK case law, studies in other jurisdictions highlight the risk of judicial biases embedding in case law. Ash et al. (2024) used NLP to assess gender biases in US State Supreme Courts, creating a gender bias index based on judges' language linking men with careers and women with families. Choi et al. (2022) found that Kenyan judges were 3-5% more likely to allow appeals from co-ethnic individuals, indicating in-group favouritism. Asmat & Kossuth (2021) showed that female judges set lower child support awards than male judges, explained by higher income estimates for fathers. These findings underscore the pervasive impact of both explicit and implicit biases in judicial decision-making.

Base Rates Bias may arise in legal data where it reflects social inequalities and disadvantages that particular groups face that result in differential outcomes in court. The primary concern is that an algorithm treats that information as a general pattern rather than identifying whether a specific person in that protected group has, in reality, a low or high risk. For example, self-represented litigants may be statistically less likely to win cases, however, perpetuating this outcome through predictions is undesirable.

Information leakage occurs when a model uses information from outside the intended dataset (Sarkar & Vafa, 2024). In the legal domain, leakage can happen when using generative AI because it is challenging to separate neutral information about a case from the judgment text. Suppose a model is designed to predict the "case outcome". In that case, it is generally provided with the case judgment text, which will not only contain the verdict but will also reflect the judges' post-hoc knowledge and subjective perspectives that shape their written judgments (Medvedeva & Mcbride, 2023). Judges are often influenced by the performance of counsel and witnesses, such that case judgments cannot be viewed as neutral input text (Smith, 1985). It will therefore be important to understand how generative AI models function so that their responses or predictions can be interpreted transparently (Rudin, 2019). Consequently, generative legal AI pre-trained on such data might inadvertently incorporate subjective perspectives rather than purely legal reasoning, leading to compromised predictions and analyses.

3. Measuring and Mitigating Bias in Legal Data for Generative AI

Context and scope. The implications of bias in legal data for generative AI are highly contingent on the intended scope and application. Researching and evaluating existing law poses fewer risks than using generative AI for predictive tasks, such as predicting judicial decisions or legal recommendations. A nuanced understanding of the model's purpose and the potential downstream effects is crucial in assessing the impact of biased data. There are several avenues for measuring and mitigating bias in legal data for generative AI, including developing contextual bias metrics, techniques for data curation and preprocessing, benchmarking model performance, and minimising hallucinations.

Legal and ethical considerations. The use of biased legal data in generative AI models raises significant legal implications. From a privacy perspective, the potential for information leakage, particularly concerning personal or sensitive details, poses a risk of violating data protection regulations. Furthermore, the propagation of biases can lead to discriminatory outcomes, contravening anti-discrimination laws and undermining the principles of equal treatment under the law. Crucially, the accuracy and faithfulness of generative AI trained on biased data are called into question, exposing their reliability and trustworthiness in legal applications. Mitigating biases in legal data and ensuring the responsible development of generative legal AI necessitates an interdisciplinary and collaborative effort.

References

Ash, E., Chen, D. L., and Ornaghi, A. Gender Attitudes in the Judiciary: Evidence from US Circuit Courts. *American Economic Journal: Applied Economics*, 16(1):314–350, 2024.

Asmat, R. and Kossuth, L. Gender Differences in Judicial Decisions under Incomplete Information: Evidence from Child Support Cases, 2021. https://ssrn.com/abstract=3964747.

Bozdag, M., Sevim, N., and Koç, A. Measuring and Mitigating Gender Bias in Legal Contextualized Language Models. *ACM Transactions on Knowledge Discovery from Data*, 18(4):79:1–79:26, 2024.

Byrom, N. Oral Evidence: Open justice: court reporting in the digital age, 2022. House of Commons Justice Committee.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pp. 2898–2904. ACL, 2020.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., and Aletras, N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330, Dublin, Ireland, 2022. Association for Computational Linguistics.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 2024.

Choi, D. D., Harris, J. A., and Shen-Bayh, F. Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya. *American Political Science Review*, 116(3):1067–1080, 2022.

Coleman, J. L. Truth and Objectivity in Law. *Legal Theory*, 1(1):33–68, 1995.

Dale, K. Legal Corpus Linguistics: Gambling to Gaming Language Powers and Probabilities Notes. *UNLV Gaming Law Journal*, 8(2):233–252, 2017.

Friedrich, R. Complexity and Entropy in Legal Language. *Frontiers in Physics*, 9, 2021.

Gisborne, J., Patel, R., Paskell, C., and Peto, C. Justice Data Matters: Building a public mandate for court data use, 2022. The Legal Education Foundation, IPSOS.

Glogar, O. The Concept of Legal Language: What Makes Legal Language 'Legal'? *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 36(3):1081–1107, 2023.

Harvard University. Caselaw Access Project, 2024.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Heiner, R. Imperfect Decisions and the Law: On the Evolution of Legal Precedent and Rules. *The Journal of Legal Studies*, 15(2):227–261, 1986.

Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. Pile of Law: Learning Responsible Data Filtering From the Law and a 256gb Open-source Legal Dataset. *Advances in Neural Information Processing Systems*, 35, 2022.

Hoadley, D., Tomlinson, J., Nemsic, E., and Somers-Joce, C. How public is public law? Approximately 55%, 2022. UK Constitutional Law Association.

House of Commons. Open Justice: Court Reporting in the Digital Age, 2022. House of Commons Justice Committee, Fifth Report of Session 2022-23.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., and Feng, Y. Lawyer LLaMA Technical Report, 2023. arXiv:2305.15062.

Hwang, W., Lee, D., Cho, K., Lee, H., and Seo, M. A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551, 2022.

Jiang, H. and Nachum, O. Identifying and Correcting Label Bias in Machine Learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Justice UK. Annual Report and Accounts 2022-2023, 2023. The Supreme Court and Judicial Committee of the Privy Council.

Katz, D. M. and Bommarito, M. J. Measuring the complexity of the law: the United States Code. *Artificial Intelligence and Law*, 22(4):337–374, 2014.

Katz, D. M., Coupette, C., Beckedorf, J., and Hartung, D. Complex societies and the growth of the law. *Scientific Reports*, 10(1), 2020.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, Halifax Canada, 2017. ACM.

Lammy, D. The Lammy Review: An Independent Review into the Treatment of, and Outcomes for, Black, Asian and Minority Ethnic Individuals in the Criminal Justice System, 2017. UK Government.

Lattimore, F., O'Callaghan, S., Paleologos, Z., Reid, A., Santow, E., Sargeant, H., and Andrew Thomsen. Using Artificial Intelligence to Make Decisions: Addressing the Problem of Algorithmic Bias, 2020. Australian Human Rights Commission.

Medvedeva, M. and Mcbride, P. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pp. 73–84, Singapore, 2023. ACL.

Monteith, K., Quinn, E., Dennis, A., Joseph-Salisbury, R., Kane, E., Addo, F., and McGourlay, C. Diversity and racial bias in the judiciary | Centre on the Dynamics of Ethnicity | The University of Manchester, 2022.

Nazarenko, A. and Wyner, A. Legal NLP Introduction. *Traitement Automatique des Langues*, 58(2):7–19, 2017.

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. In *Findings of the Association for Computational Linguistics (EMNLP 2023)*, pp. 3016–3054, Singapore, 2023a. ACL.

Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., and Ho, D. E. MultiLegalPile: A 689GB Multilingual Legal Corpus. In *ICML Workshop on Data-centric Machine Learning Research*, 2023b.

Poudyal, P., Savelka, J., Ieven, A., Moens, M. F., Goncalves, T., and Quaresma, P. ECHR: Legal Corpus for Argument Mining. In *Proceedings of the 7th Workshop on Argument Mining*, pp. 67–75. ACL, 2020.

Priest, G. and Klein, B. The Selection of Disputes for Litigation. *Journal of Legal Studies*, 13(1):1–55, 1984.

Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Ruhl, J. Law's Complexity: A Primer. *Georgia State University Law Review*, 24:885, 2008.

Sargent, J. and Weber, M. Identifying Biases in Legal Data: An Algorithmic Fairness Perspective, 2021. arXiv:2109.09946.

Sarkar, S. K. and Vafa, K. Lookahead Bias in Pretrained Language Models, 2024. https://ssrn.com/abstract=4754678.

Sevim, N., Şahinuç, F., and Koç, A. Gender bias in legal corpora and debiasing it. *Natural Language Engineering*, 29(2):449–482, 2023.

Shahbazi, N., Lin, Y., Asudeh, A., and Jagadish, H. V. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Computing Surveys*, 55(13): 1–39, 2023.

Shubber, K. Let in the light to the murky reaches of the English legal system, 2022. https://on.ft.com/4bWQIwZ.

Smith, B. F. Of Truth and Certainty in the Law: Reflections on the Legal Method. *American Journal of Jurisprudence*, 30:97–120, 1985.

Song, D., Gao, S., He, B., and Schilder, F. On the Effectiveness of Pre-Trained Language Models for Legal Natural Language Processing: An Empirical Study. *IEEE Access*, 10:75835–75858, 2022.

Trancoso, I., Mamede, N., Martins, B., Pinto, H. S., and Ribeiro, R. The Impact of Language Technologies in the Legal Domain. In Sousa Antunes, H., Freitas, P. M., Oliveira, A. L., Martins Pereira, C., Vaz De Sequeira, E., and Barreto Xavier, L. (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, pp. 25–46. Springer, Cham, 2024.

Wang, H., Li, J., Wu, H., Hovy, E., and Sun, Y. Pre-Trained Language Models and Their Applications. *Engineering*, 25:51–65, 2023.

Xiao, C., Hu, X., Liu, Z., Tu, C., and Sun, M. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84, 2021.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. 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. ACM, 2021.

Östling, A., Sargeant, H., Xie, H., Bull, L., Terenin, A., Jonsson, L., Magnusson, M., and Steffek, F. The Cambridge Law Corpus: A Dataset for Legal AI Research. *Advances in Neural Information Processing Systems*, 36, 2023. doi: 10.48550/arxiv.2309.12269.