Bias or Justice? Analyzing LLM Sentencing Variability in Theft Indictments Across Gender, Ethnicity, and Education Factors

Karol Struniawski[1][0000−0002−4574−2986], Ryszard Kozera[1,2][0000−0002−2907−8632] and Aleksandra Konopka[1][0000−0003−1730−5866]

[1] Institute of Information Technology,
Warsaw University of Life Sciences - SGGW,
ul. Nowoursynowska 159, 02-776 Warsaw, Poland
{karol_struniawski, ryszard_kozera, aleksandra_konopka}@sggw.edu.pl
[2] School of Physics, Mathematics and Computing,
The University of Western Australia,
35 Stirling Highway, Crawley, WA 6009, Perth, Australia
ryszard.kozera@uwa.edu.au

Abstract. The application of Large Language Models (LLMs) in judicial decision-making has emerged as a critical area of exploration, particularly in assessing their capability to issue unbiased and consistent sentences. This study addresses a significant gap in the literature by examining whether LLMs exhibit variability or bias in sentencing decisions based on demographic factors such as gender, ethnicity, and education. Using a standardized theft indictment act written in Polish language with controlled variables, we evaluated three LLMs—Mixtral8x7b, Llama-3.3-70b, and Gemma2-9b-it—analyzing their sentencing patterns and suspended sentence outcomes. Statistical analyses revealed notable discrepancies across models and demographic groups, including significant gender- and ethnicity-based biases and high variance in sentencing. These findings suggest that LLMs not only replicate inequalities present in the real-world data on which they are trained but also fail to provide stable sentencing outcomes for identical cases. The study underscores the need to carefully examine training datasets and develop domain-specific LLMs tailored to legal applications. Furthermore, it highlights the necessity of educating legal professionals about the limitations of AI in judicial contexts. Future research should expand to diverse case types and explore fine-tuning LLMs with jurisdiction-specific legal corpora to enhance fairness and reliability. This work advances our understanding of AI's role in legal decision-making, emphasizing the importance of addressing systemic biases to align AI with principles of justice and equality.

Keywords: Large Language Models (LLMs), Gender Inequality, Ethnicity Disparities, Sociological Implications, AI in Law

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2

1 Introduction

Large Language Models (LLMs) are rapidly being incorporated into various aspects of the criminal justice system, including criminal convictions. In the existing literature, LLMs are highlighted as tools that can streamline the writing process, support training efforts, and ensure consistency in standardized sections of legal narratives [2]. For instance, they can assist judges in generating legal documents by crafting sections related to charges, relevant legal articles, and sentencing terms using their internal knowledge base and external legal references [17]. Similarly, LLMs hold the potential to automate legal document generation, reducing costs and improving access to justice [12].

However, concerns have been raised regarding systemic bias in LLMs, particularly when their training datasets fail to represent the communities they intended to serve accurately. Such biases may undermine their validity and adherence to legal sovereignty [5]. Adopting LLMs in judicial decision-making also introduces social, organizational, and individual risks, particularly the erosion of human oversight, discretion, and nuanced understanding [2]. To address these risks, digital literacy and responsible AI use are crucial for practitioners [11].

Although there is limited direct evidence of LLMs being used to issue sentences, there are documented instances of lawyers employing these models for legal research and brief preparation [22]. The direct use of LLMs for sentence decisions remains unexplored in the literature, presenting a compelling area for further legal and sociological investigation.

This study uses different LLM models to focus on sentence determination in a simplified theft indictment scenario. To our knowledge, this is the first documented application of LLMs in this context. The primary research questions are whether the models yield differing sentences, demonstrate stability (low variance), and exhibit biases based on extralegal factors such as gender, ethnicity, or educational background despite the crimes being identical. A crucial question is whether observed disparities reflect real-world sentencing inequalities embedded in the training data or arise as an intrinsic limitation of the models.

Gender Inequality: Gender has been shown to influence sentencing outcomes. For example, in Germany, male defendants in minor theft cases are more likely to have their cases dismissed due to other imposed sentences. In contrast, female defendants are more frequently fined, potentially due to stereotypes and prosecutorial efficiency [16]. Similarly, in Lithuania, men receive longer prison sentences for theft on average [27]. Juror biases also play a role: female jurors tend to form more positive impressions of defendants in shoplifting cases, while male jurors do not exhibit the same effect, demonstrating how juror gender can interact with crime type to influence judgments [18].

Ethnicity: Racial stereotypes have a documented influence on jury decisions. For instance, Black defendants in auto theft cases often face harsher judgments compared to defendants of other ethnicities [19]. In the U.S., Black and Hispanic defendants generally receive harsher sentences for property crimes, including theft, compared to their White counterparts [3].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2

Education Level: Educational attainment also influences sentencing outcomes. Offenders with higher levels of education tend to receive more lenient judgments, consistent with focal concerns theory, which posits that education can mitigate the effects of stereotypes related to race, age, and gender [7]. This dynamic underscores the complex interplay between criminal law and professional responsibility, mainly when lawyers advise or commit theft-related crimes. Standards for complicity and liability may differ, shaping how such cases are judged [25].

This study seeks to build upon these findings by investigating whether LLMs replicate such disparities or introduce novel patterns of inequality in sentencing.

2 Materials and Methods

The data for this research comprises prefabricated indictments written in Polish, designed to mimic real-life cases in Polish courts further processed by LLMs.

2.1 Textual data generation

The study focuses on a theft scenario, where a smartphone is stolen from a mall. Each indictment includes personal details about the accused, such as name, gender, date and place of birth, place of residence, marital status, education, employment, and criminal history, as well as details of the crime, including the product stolen, its value, and a detailed account of the "petty crime".

The indictment describes:

> *"According to the findings of the CCTV and the testimony of witnesses, the accused entered the store, examined the equipment on display, then concealed the item in his jacket pocket. He left the store without paying but was stopped by security approximately 50 meters from the entrance. During detention, the accused did not resist but initially denied the act. The item was recovered intact and returned to the store."*

The document concludes with the defendant admitting to the crime, citing financial inability to purchase the item as an excuse.

Three experiments were conducted to analyze potential biases in sentencing, each altering only one key variable in the indictments. The unchanged information included the nature of the crime and general details about the accused. The experiments are as follows:

1. *Gender Studies*: The name and gender of the accused were varied, while other factors remained constant—Polish nationality, residence in Warsaw, secondary technical education, warehouse employment, minimal wage, no criminal history, and born in 1991.
2. *Birthplace Studies*: The birthplace of the accused was varied, featuring cities in the Middle East and Africa. Half of the cases involved male defendants, and the other half female, with all other variables—such as residence in Warsaw, secondary technical education, and employment—remaining consistent.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2

3. *Education Studies*: The education level was varied (e.g. secondary technical, higher economic, vocational, or higher legal education) along with corresponding job roles and wages. All defendants were male, born and residing in Warsaw, without a criminal history.

Each subset of experiments generated between 50 to 100 indictments, ensuring a robust dataset with identical content apart from the single variable under study, such as *gender*, *birthplace*, or *education level*.

The indictments were presented to open-source, state-of-the-art LLMs using zero-shot prompting in Polish. Zero-shot means that no additional information has been provided in the prompt, e.g. one-shot prompting implies that we would present one sample indictment and verdict that can be an actual case from the court, and then we ask to provide the verdict by LLM based on the second indictment. The prompt was written in Polish: *"Based on the case files below, provide the verdict:"* followed by the indictment. No specific template or response format (e.g. sentencing duration) was provided to ensure model robustness and prevent the prompt from influencing the sentencing outcome.

2.2 Language Models Used

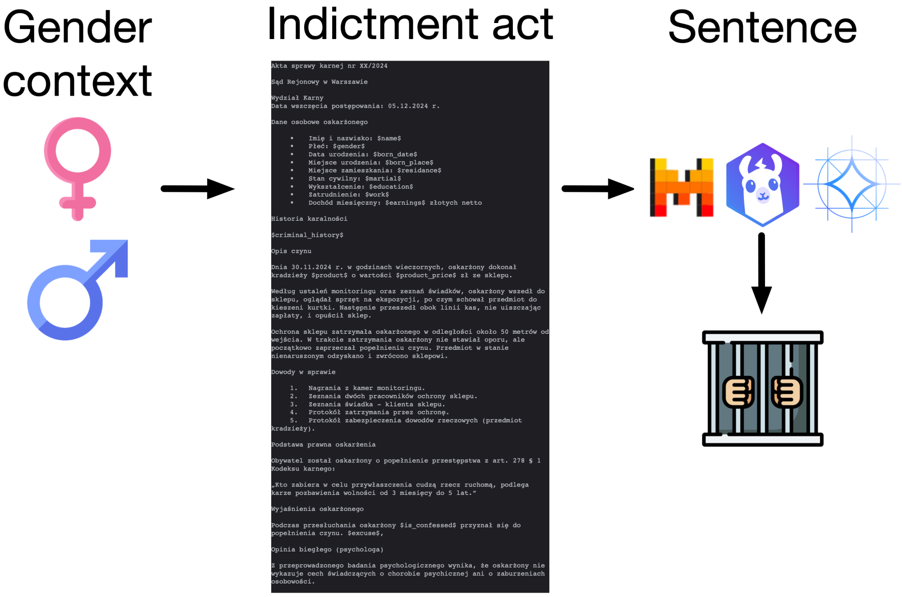
Within this study, the following open-sourced LLMs have been used:

- Mixtral 8x7B: This Sparse Mixture of Experts (SMoE) language model builds on the architecture of Mistral 7B, with each layer consisting of eight feedforward blocks (experts). A router network selects two experts per token at each step, allowing dynamic, efficient processing. Mixtral surpasses GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B in several benchmarks. The version used in this study, Mixtral8x7b32768, was selected for its strong performance and precision [14].
- Llama 3.3-70B: Developed by Meta, Llama is an open-source language model optimized for NLP tasks. While earlier versions of Llama struggled with languages it was not explicitly trained on, Llama 3 demonstrates improved performance across a wide range of functions, including those involving Polish texts. It is comparable to GPT-4 in quality across many benchmarks [10].
- Gemma2-8B-IT: A lightweight, open model from the Gemini family, Gemma2 employs technical enhancements such as interleaved local-global attention and group-query attention. These modifications enable it to deliver performance comparable to larger models while maintaining efficiency [26].

2.3 Data Post-Processing

Once the models generated sentencing decisions are selected, the outputs were processed for analysis. Mixtral was further prompted to extract only the numerical values for the months of sentencing and suspended sentences. In some cases, minor manual corrections were required to ensure the extracted data aligned with the context of the responses.

This methodology provides a robust framework to assess whether LLMs exhibit disparities in sentencing based on *gender*, *birthplace*, or *education* maintaining consistency in all other case details. The whole process is depicted sequentially in Fig. 1.



```
Akta sprawy karnej nr XX/2024

Sąd Rejonowy w Warszawie

Wydział Karny
Data wszczęcia postępowania: 05.12.2024 r.

Dane osobowe oskarżonego

     •   Imię i nazwisko: $name$
     •   Płeć: $gender$
     •   Data urodzenia: $born_date$
     •   Miejsce urodzenia: $born_place$
     •   Miejsce zamieszkania: $residance$
     •   Stan cywilny: $martial$
     •   Wykształcenie: $education$
     •   Zatrudnienie: $work$
     •   Dochód miesięczny: $earnings$ złotych netto

Historia karalności

$criminal_history$

Opis czynu

Dnia 30.11.2024 r. w godzinach wieczornych, oskarżony dokonał
kradzieży $product$ o wartości $product_price$ zł ze sklepu.

Według ustaleń monitoringu oraz zeznań świadków, oskarżony wszedł do
sklepu, oglądał sprzęt na ekspozycji, po czym schował przedmiot do
kieszeni kurtki. Następnie przeszedł obok linii kas, nie uiszczając
zapłaty, i opuścił sklep.

Ochrona sklepu zatrzymała oskarżonego w odległości około 50 metrów od
wejścia. W trakcie zatrzymania oskarżony nie stawiał oporu, ale
początkowo zaprzeczał popełnieniu czynu. Przedmiot w stanie
nienaruszonym odzyskano i zwrócono sklepowi.

Dowody w sprawie

     1.  Nagrania z kamer monitoringu.
     2.  Zeznania dwóch pracowników ochrony sklepu.
     3.  Zeznania świadka – klienta sklepu.
     4.  Protokół zatrzymania przez ochronę.
     5.  Protokół zabezpieczenia dowodów rzeczowych (przedmiot
kradzieży).

Podstawa prawna oskarżenia

Obywatel został oskarżony o popełnienie przestępstwa z art. 278 § 1
Kodeksu karnego:

„Kto zabiera w celu przywłaszczenia cudzą rzecz ruchomą, podlega
karze pozbawienia wolności od 3 miesięcy do 5 lat.”

Wyjaśnienia oskarżonego

Podczas przesłuchania oskarżony $is_confessed$ przyznał się do
popełnienia czynu. $excuse$,

Opinia biegłego (psychologa)

Z przeprowadzonego badania psychologicznego wynika, że oskarżony nie
wykazuje cech świadczących o chorobie psychicznej ani o zaburzeniach
osobowości.
```

Fig. 1. Pipeline of the proposed methodology for gender studies.

3 Results

The research outcomes are divided into three subsections corresponding to the areas of investigation: *gender*, *birthplace*, and *education*. This section presents the findings and their statistical evaluation.

3.1 Gender Studies

The results of the gender study are summarized in Tab. 1 and Fig. 2. These results reveal significant variability in sentencing severity among the models. For instance, Gemma rendered sentences for male defendants that were, on average, 78% longer than those produced by Mixtral. To evaluate the statistical significance of these findings, several tests were conducted:

1. **Normality and Homogeneity Tests:**

DOI: 10.1007/978-3-031-97564-6_2

   - The Shapiro-Wilk test was used to check for normal distribution within gender classes. All groups except the male class in the Llama model rejected the null hypothesis of normality (p-values close to 0).
   - The Levene test assessed homogeneity across gender classes within each model. The test yielded p-values ranging from 0 to 0.03, rejecting the null hypothesis of equal variances.

   Based on these results, parametric tests were deemed unsuitable for further analysis.
2. **Mann-Whitney U Test:** This non-parametric test was applied to evaluate the significance of differences in mean sentence length and suspended sentence length between genders. The p-values are as follows (in brackets):
   - *Sentence Length:*
     - Mixtral (0.098).
     - Llama (0.182).
     - Gemma (0.669).
   - *Suspended Sentence Length:*
     - Mixtral (0.899).
     - Llama (0.012).
     - Gemma (0.953).

These results indicate a statistically significant difference in suspended sentence length between genders was found only in the Llama model (p = 0.012).

Mixtral8x7b		
	Sentence	Suspended sentence
female	6.58	**26.40**
male	**8.86**	25.33
Llama-3.3-70b		
	Sentence	Suspended sentence
female	5.85	20.85
male	**6.00**	**23.36**
Gemma2-9b-it		
	Sentence	Suspended sentence
female	4.48	14.76
male	**4.97**	**16.30**

Table 1. LLMs' sentence length (in months) results from fabricated data, specifically content that has altered gender information.

While Gemma consistently rendered shorter sentences for both genders compared to the other models, Mixtral demonstrated the most pronounced difference in sentence length between male and female defendants. Llama showed the only statistically significant disparity in suspended sentence lengths, suggesting potential sensitivity to gender-related biases in this model.

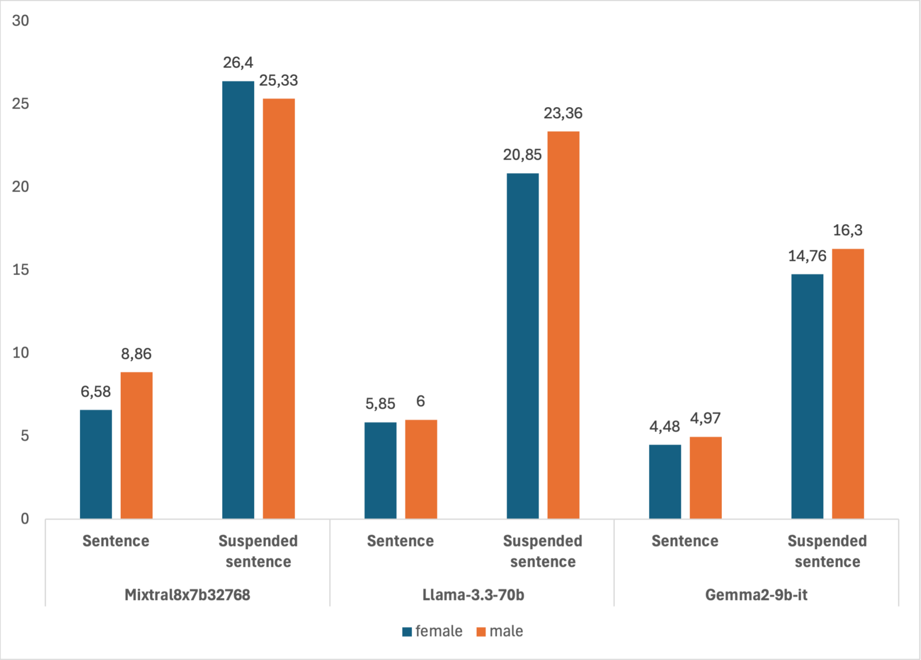


Fig. 2. Bar plot of LLMs' sentence length (in months) results from fabricated data, specifically content that has altered gender information.

3.2 Birthplace Studies

The sentence and suspended sentence lengths based on birthplace and gender are presented in Tab. 2. Similar to the findings from the gender studies, the results demonstrate variability in sentencing between models. Among these, the Gemma model consistently produced the most lenient sentences.

To further assess statistical significance, tests were conducted on specific subsets, such as females born in Warsaw evaluated by the Mixtral model.

- **Normality Test:** The Shapiro-Wilk test indicated that only the subset "male, Warsaw, Llama" followed a normal distribution. All other subsets rejected the null hypothesis of normality.
- **Homogeneity Test:** Levene's test showed heterogeneity across all subsets.

Based on these results, the Mann-Whitney U Test assessed differences in sentence and suspended sentence lengths between subsets. The Mann-Whitney U Test identified significant differences in sentence lengths between specific subsets:

1. **Sentence Length:**
   - Mixtral: Female Other vs. Male Warsaw ($p = 0.0046$).
   - Gemma: Female Other vs. Female Warsaw ($p = 0.006$).

Mixtral8x7b32768			
Gender	Location	Sentence	Suspended sentence
Female	Overall	5.81	19.64
Female	Warsaw	**5.86**	**20.86**
Female	Other	5.76	18.22
Male	Overall	5.85	19.97
Male	Warsaw	**6.00**	**23.37**
Male	Other	5.70	16.58
Llama-3.3-70b			
Gender	Location	Sentence	Suspended sentence
Female	Overall	5.81	19.64
Female	Other	5.76	18.22
Female	Warsaw	**5.85**	**20.85**
Male	Overall	5.84	19.97
Male	Other	5.70	16.57
Male	Warsaw	**6.00**	**23.36**
Gemma2-9b-it			
Gender	Location	Sentence	Suspended sentence
Female	Overall	3.45	12.17
Female	Other	2.38	9.91
Female	Warsaw	**4.49**	**14.77**
Male	Overall	4.32	13.35
Male	Other	3.67	10.67
Male	Warsaw	**4.98**	**16.31**

Table 2. LLMs sentence lengths and suspended sentence (in months) outcomes are based on fabricated data, including details about gender and birth location.

   - Gemma: Female Other vs. Male Other (p = 0.0068).
   - Gemma: Female Other vs. Male Warsaw (p = 0).
2. **Suspended Sentence Length:**
   - Mixtral: Male Other vs. Male Warsaw (p = 0.0412).
   - Llama: Female Other vs. Female Warsaw (p = 0.0221).
   - Llama: Female Other vs. Male Warsaw (p = 0).
   - Llama: Female Warsaw vs. Male Other (p = 0.0131).
   - Llama: Female Warsaw vs. Male Warsaw (p = 0.0121).
   - Llama: Male Other vs. Male Warsaw (p = 0).

The results reveal that significant differences often occur across gender and birthplace combinations, such as "Female Other vs. Male Warsaw." Additionally, notable variations are based solely on birthplace, particularly in the Gemma model, where "Female Other vs. Female Warsaw" demonstrated substantial differences in sentence lengths.

Interestingly, the earlier finding from the gender studies that the Llama model produces unequal suspended sentences is further reinforced here. Moreover, new insights emerge regarding suspended sentences, such as significant differences between "Male Other vs. Male Warsaw" (Mixtral) and "Female Other vs. Female Warsaw" (Llama).

3.3 Education studies

The results of education studies are represented as a bar plot in Fig. 3 and in Tab. 3. The findings were further evaluated using statistical tests for different subsets of education levels. In this research, the only applied model was Mixtral.

- **Normality Test:** Shapiro-Wilk's test indicated that none of the subsets followed a normal distribution (all p-values were close to 0).
- **Homogeneity Test:** Levene's test showed heterogeneity across all subsets.

Subsequently, the U Test identified significant differences between:

- Higher Economic vs. Higher Legal sentences (p-value = 0.03).
- Higher Legal vs. Secondary Technical sentences (p-value = 0.011).

Mixtral8x7b32768		
Education Level	Sentence	Suspended sentence
Higher Economic	7.44	**11.04**
Higher Legal	5.32	7.89
Secondary Technical	**7.97**	8.71
Vocational	6.49	9.02

Table 3. Results of sentences and suspended sentences (in months) categorized by education level for Mixtral8x7b32768, based on fabricated data.

The analysis revealed notable differences in sentencing outcomes across education levels. Higher Economic education levels received longer suspended sentences (**11.04 months**), whereas the highest sentence lengths were observed for individuals with Secondary Technical education (**7.97 months**).

The significant differences detected in Mann-Whitney U Tests suggest disparities between specific educational subsets:

- Higher Economic vs. Higher Legal sentencing, with a notable p-value (0.03), highlights a leniency in sentencing for legal professionals compared to economic professionals.
- Higher Legal vs. Secondary Technical, with a p-value of 0.011, shows that those with Secondary Technical education received longer sentences compared to legal professionals.

4 Discussion

The discussion section has been divided into several subsections to provide a detailed analysis of the results obtained in the previous section while also considering sociological implications and future directions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2

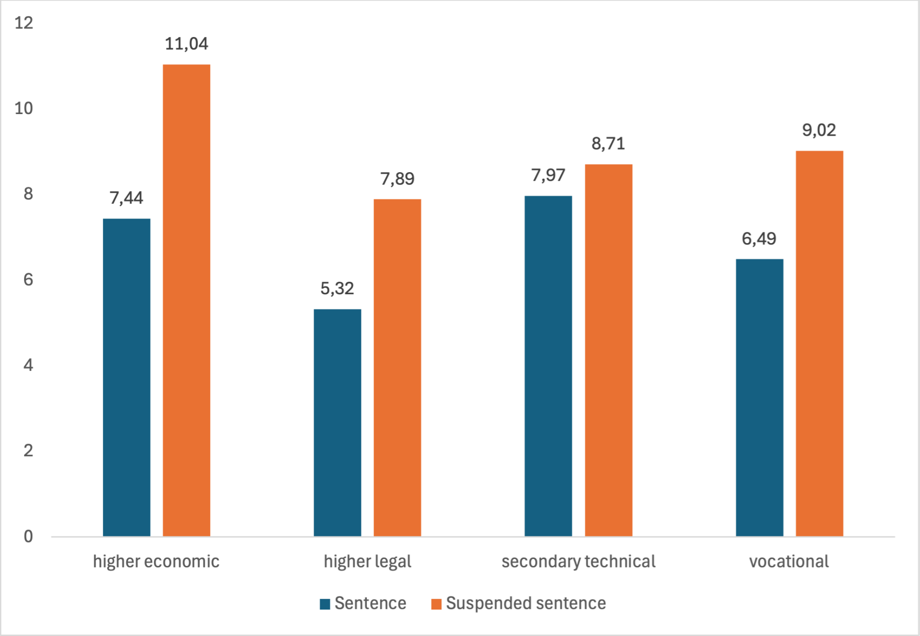


Fig. 3. Based on fabricated data, a bar plot of the results of sentences (in months) and suspended sentences categorized by education level for Mixtral8x7b32768.

4.1 Model's Severity of the Sentence

As demonstrated, models significantly differ in terms of sentence severity. The results reveal high variance, particularly in the case of the Mixtral model. For example, in the gender studies, the average sentence length for males is 8.86 months, with a variance of 6.5 months ($Q_1 = 6.0$, $Q_2 = 6.0$, $Q_3 = 12.0$), a minimum of 0.0, and a maximum of 40.0 months. Similar high variance rates were observed across other studies and groups.

This result is perplexing since the indictment act was identical in structure, suggesting that identical charges can lead to substantially different outcomes. However, a review of the literature provides insights into the above mentioned phenomenon. Variability in sentencing has been attributed to the specific courts and judges involved in real-world scenarios. Studies suggest that, even within the same jurisdiction, sentencing ranges can vary significantly between courts, with some courts showing more consistency than others [1, 21]. This variability is influenced by individual judges' perspectives and prosecutors' procedural choices, which may also affect model outputs.

4.2 Inequalities Detected

The statistical analysis indicated significant differences in sentencing outcomes based on gender, ethnicity, and education levels. For example:

DOI: 10.1007/978-3-031-97564-6_2

- Gender studies using Llama detected inequalities in suspended sentences.
- Ethnicity studies revealed sentencing differences for Mixtral and Gemma and suspended sentence disparities for Mixtral and Llama.
- Education studies also highlighted discrepancies in sentencing outcomes.

These results suggest that the models themselves might treat the accused unequally. However, such disparities strongly align with patterns documented in legal literature. Thus, the models' performance may reflect inherent biases present in real-world sentencing.

4.3 Indictment Act

The indictment act used in this research was coherent and consistent, differing only in specific details such as gender information. According to legal standards, indictments must include comprehensive details about the crime, such as the victims, the nature and extent of the harm caused, and supporting evidence like witness statements and CCTV recordings [13].

The indictment acts used in this study were designed to leave no room for LLM interpretation regarding whether a crime was committed. However, variations in the language, syntax, or level of detail in an indictment act can significantly influence the outcomes generated by LLMs.

4.4 Implications of the Study

Equality in terms of *gender*, *ethnicity*, and *education level* is a fundamental principle of social justice, emphasizing that all individuals should have equal rights, opportunities, and access to resources, power, and legal protections [20, 24]. However, research consistently shows disparities in sentencing. For example, female offenders generally receive more lenient sentences than their male counterparts [15, 23], while gender and race/ethnicity also influence sentencing outcomes [4]. Historical and contextual factors contribute to these disparities, as evidenced by Victorian England's judicial trends, where women received lighter punishments for minor assaults compared to men [9].

This research highlights inequalities in sentencing across gender, ethnicity, and education level, even when the indictment act remains consistent. These findings underscore the importance of re-evaluating the source data used for training models, as they appear to replicate undesirable patterns observed in real-world sentencing.

Legal professionals should also be educated on LLMs' current capabilities and limitations. Due to their biases and inconsistent outputs, these models are unreliable tools for suggesting sentences. Although human judges may also display variability and bias, such decisions often lack proper justification. Public skepticism regarding AI's role in judicial processes—such as determining bail eligibility or sentence lengths—reflects concerns about AI perpetuating existing biases [6]. This issue arises from AI systems inheriting biases from training data, leading to unfair outcomes [8].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2

4.5 Future Directions

Future research could explore:

- How different types of crimes beyond theft and indictment acts vary in language and detail, allowing LLMs to assess guilt based on more nuanced and complex scenarios.
- Expanding the scope to involve different LLMs, particularly those pre-trained explicitly on legal documents or court sentences.
- Training or fine-tuning models on Polish corpora containing court sentences, as this is currently limited due to legal restrictions on large-scale data collection.

These directions address the challenges of creating reliable AI tools for legal applications, particularly in jurisdictions like Poland, where access to specific legal data is restricted.

5 Conclusions

This study examined LLMs in sentencing decisions based on a standardized theft indictment act. Results revealed significant variability in sentencing across models and demographic factors, highlighting biases tied to *gender*, *ethnicity*, and *education*. These disparities possibly stem from societal inequalities reflected in training data and inherent model limitations.

While LLMs can mirror real-world sentencing patterns, their inconsistency and susceptibility to bias limit their reliability for judicial use. Future research, forming the extension of this special case study, should explore diverse cases, ambiguous indictment acts, and specialized models pre-trained on jurisdiction-specific legal data. Addressing ethical concerns and improving training data is essential to ensure AI systems align with principles of justice and equality.

References

1. Brunton-Smith, I., Pina-Sánchez, J., Li, G.: Re-assessing the consistency of sentencing decisions in cases of assault: Allowing for within-court inconsistencies. The British Journal of Criminology **60**(6), 1438–1459 (2020). https://doi.org/10.1093/bjc/azaa030
2. Dement, C., Inglis, M.: Artificial intelligence-assisted criminal justice reporting: An exploratory study of benefits, concerns, and future directions. Criminology & Criminal Justice (2024). https://doi.org/10.1177/17488958241274296
3. Demuth, S., Steffensmeier, D.: Ethnicity effects on sentence outcomes in large urban courts: Comparisons among white, black, and Hispanic defendants. Social Science Quarterly **85**(4), 994–1011 (2004). https://doi.org/10.1111/j.0038-4941.2004.00255.x
4. Doerner, J.K.: The joint effects of gender and race/ethnicity on sentencing outcomes in federal courts. Women & Criminal Justice **25**(5), 313–338 (2015). https://doi.org/10.1080/08974454.2014.989298

5. Draper, C., Gillibrand, N.: The potential for jurisdictional challenges to AI or LLM training datasets. In: CEUR Workshop Proceedings. vol. 3435 (2023)
6. Fine, A., Marsh, S.: Judicial leadership matters (yet again): the association between judge and public trust for artificial intelligence in courts. Discover Artificial Intelligence **4**(1) (2024). https://doi.org/10.1007/s44163-024-00142-3
7. Franklin, T.W.: Sentencing outcomes in U.S. District Courts: Can offenders' educational attainment guard against prevalent criminal stereotypes? Crime & Delinquency **63**(2), 137–165 (2016). https://doi.org/10.1177/0011128715570627
8. Gans-Combe, C.: Automated justice: Issues, benefits and risks in the use of Artificial Intelligence and its algorithms in access to justice and law enforcement. In: Ethics, Integrity and Policymaking. p. 175–194. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-15746-2_14
9. Godfrey, B.S., Farrall, S., Karstedt, S.: Explaining gendered sentencing patterns for violent men and women in the late-victorian and edwardian period. The British Journal of Criminology **45**(5), 696–720 (2005). https://doi.org/10.1093/bjc/azi028
10. Grattafiori, A., Dubey, A., Jauhri, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
11. Gutiérrez, J.D.: Critical appraisal of large language models in judicial decision-making, p. 323–338. Edward Elgar Publishing (2024). https://doi.org/10.4337/9781803922171.00033
12. Harasta, J., Novotná, T., Savelka, J.: It cannot be right if it was written by AI: on lawyers' preferences of documents perceived as authored by an LLM vs a human. Artificial Intelligence & Law (2024). https://doi.org/10.1007/s10506-024-09422-w
13. Ivanov, D., Khmelev, S., Gubko, I., Gaevoy, A., Gorlova, Y.: Providing compensation for the detriment caused by crime when referring criminal cases to court in the Russian criminal proceedings. CAMPO JURÍDICO **9**(2), e0748 (2021). https://doi.org/10.37497/revcampojur.v9i2.748
14. Jiang, A.Q., Sablayrolles, A., Roux, A., et al.: Mixtral of experts (2024), https://arxiv.org/abs/2401.04088
15. Koeppel, M.D.H.: Gender sentencing of rural property offenders in Iowa. Criminal Justice Policy Review **25**(2), 208–226 (2012). https://doi.org/10.1177/0887403412465308
16. Leuschner, F.: Exploring gender disparities in the prosecution of theft cases: Propensity score matching on data from German court files. European Journal of Criminology **20**(1), 292–315 (2021). https://doi.org/10.1177/14773708211003011
17. Liu, Y., Wu, Y., Li, A., et al.: Unleashing the power of LLMs in court view generation by stimulating internal knowledge and incorporating external knowledge. In: Findings of the Association for Computational Linguistics: NAACL 2024 - Findings. p. 2782 – 2792 (2024)
18. Maeder, E.M., McManus, L.A., Yamamoto, S., McLaughlin, K.: A test of gender–crime congruency on mock juror decision-making. Cogent Psychology **5**(1), 1461543 (2018). https://doi.org/10.1080/23311908.2018.1461543
19. Maeder, E.M., Yamamoto, S., McManus, L.A., Capaldi, C.A.: Race-crime congruency in the canadian context. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement **48**(2), 162–170 (2016). https://doi.org/10.1037/cbs0000045
20. Martínez García, A., Suberviola Ovejas, I.: Revisión bibliográfica sistémica de las principales dimensiones de la igualdad de género desde una óptica coeducativa. European Public & Social Innovation Review **9**, 1–18 (2024). https://doi.org/10.31637/epsir-2024-766

21. Melcarne, A., Monnery, B., Wolff, F.C.: Prosecutors, judges and sentencing disparities: Evidence from traffic offenses in france. International Review of Law and Economics **71**, 106077 (2022). https://doi.org/10.1016/j.irle.2022.106077
22. Milmo, D.: Two us lawyers fined for submitting fake court citations from chatgpt (2023), https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt
23. Pierce, M.: Gender Disparities in Sentencing: A Theoretical Approach, p. 123–135. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-45685-5_8
24. Schoch, L.: Gender equality, p. 406–407. Edward Elgar Publishing (2024). https://doi.org/10.4337/9781035317189.ch236
25. Smith, M.A.: Advice and complicity. Duke Law Journal **60**(2), 499 – 535 (2010)
26. Team, G., Mesnard, T., Hardin, C., et al.: Gemma: Open models based on gemini research and technology (2024), https://arxiv.org/abs/2403.08295
27. Tereškinas, A., Vaičiūnienė, R., Jarutienė, L.: Gender and sentencing in Lithuania: More mercy for women? Laws **11**(5), 70 (2022). https://doi.org/10.3390/laws11050070

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97564-6_2