Probing LLM Decision-Making Behavior through Proxy Attributes Discovered by Inductive Logic Programming

Chen Huang

Student ID: 1356133

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Information Technology

The University of Melbourne

Abstract

In recent years, Large Language Models (LLMs) have increasingly been applied to decision-making tasks across a wide range of domains, partly due to their ability to automate tasks that would otherwise require substantial human effort. However, despite their strong capabilities, prior studies have shown that LLMs may reflect or reinforce harmful societal biases related to characteristics such as race, gender, and religion. This concern is further complicated by the limited explainability of LLM decision-making, as the information influencing a prediction is not directly represented in a human-readable form. Even when sensitive attributes such as race, gender, or religion are removed from the input, related information may still be indirectly available through proxy attributes.

This study investigates how LLM decision-making behavior changes when identified proxy relations are modified. The experimental task is adapted from prompting approaches used in previous work and requires each LLM to perform a classification task under two conditions: an original candidate record and a modified version in which a specific proxy relation is disrupted while the remaining candidate information is kept unchanged. The resulting predictions are then compared to measure the model's sensitivity to the proxy intervention.

The results show that the evaluated LLMs exhibit different levels of sensitivity to the tested proxy relations, and that this sensitivity varies across proxies associated with different protected attributes. These findings suggest that proxy information can influence LLM predictions even when sensitive attributes are not explicitly provided. Further research is therefore needed to better understand and mitigate the influence of such proxy information in fairness-sensitive decision-making tasks.

Declaration & Acknowledgment

I certify that:

- This thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- This research did not require clearance from the University's Ethics Committee.
- The thesis is 3521 words in length (excluding text in images, tables, bibliographies and appendices).

I would like to thank my supervisors, Dr. Christine Rizkallah and Prof. Cezary Kaliszyk, for their guidance and support throughout this research project.

Contents

1 Introduction **6**

2 Preliminaries **8**
2.1 Inductive Logic Programming 8
2.2 Proxy Attribute 8
2.3 Proxy Attribute Detection Tool in Artificial Intelligent 9

3 Related Work **10**
3.1 Prompt Construction 10
3.2 Proxy Attribute Discovery 11

4 Methodology **12**
4.1 Experiment Setup 12
4.2 Dataset 13
4.3 Metric 14
4.4 Model Selection 15

5 Results **16**
5.1 Flip Rate 16
5.2 Change in Accuracy 17

6 Conclusion **18**

7 Future Work **19**
7.1 Strengthening the Evaluation of LLM Proxy Sensitivity 19
7.2 Adapting Evaluation Prompts for Modern LLMs 20

A Dataset Statistics **23**

List of Tables

2.1 Example dataset with non-protected attributes **Marital** and **Country**, and protected attribute **Sex**. The non-protected attribute **Marital** is a proxy of **Sex** (when Marital = *husband* or Marital = *wife*). 9

4.1 Proxy relations discovered by PADTAI [1]. 14
4.2 Statistics of the proxy relations discovered by PADTAI [1]. The Count column represents the number of candidates that satisfy each proxy relation. 14

5.1 Flip rates for each proxy relations. For Rule 1, the two proxy values were modified separately while the other proxy condition was kept unchanged. . 16
5.2 Change in accuracy, measured in percentage points, for each proxy intervention across the three models. For Rule 1, the two proxy values were modified separately while the other proxy condition was kept unchanged. . 17

A.1 Numerical attributes in the Adult dataset. 23
A.2 Categorical attributes in the Adult dataset. 23

List of Figures

3.1 Encoding numerical attributes in a dataset using PADTAI, where 'lt' denotes 'less than'. 10

7.1 Example CEB prompt for a candidate record from the Adult dataset. . . . 19

Chapter 1

Introduction

In recent years, there has been an increasing popularity with Large Language Models (LLMs). LLMs have now become deeply integrated into our lives, playing a critical role in the decision-making process to combat intensive labor for companies. However, fairness concerns are raised on how LLMs make decisions, as they represent their learned knowledge in arbitrary numbers that lack explainability. Many researchers performed bias evaluations of LLMs in different areas such as recruitment [2], healthcare [3], etc. These studies consistently report that LLMs exhibit different forms and degrees of bias, with some models showing stronger biases related to race and some models displaying greater biases related to gender. Therefore, there does not exist an LLM that is more fair than the others. A common mitigation strategy is to avoid providing LLMs sensitive information in both training and inference time, but the presence of proxy attributes may still allow this sensitive information to be indirectly inferred. A proxy attribute essentially is a non-sensitive attribute that can partially infer sensitive information, and the limited explainability of LLMs decision-making can make it difficult to detect whether proxy attributes are influencing model predictions.

To address this problem, this study employs Inductive Logic Programming (ILP) [4], a symbolic artificial intelligence approach that complements LLMs by providing greater explainability. Unlike LLMs, whose learned knowledge is not directly represented in a human-readable form, the knowledge learned by ILP is represented as logical rules. The ILP method used in this study is the Proxy Attribute Detection Tool in Artificial Intelligence (PADTAI) [1], an ILP-based proxy detector that identifies proxy relations between proxy attributes and the sensitive information they can infer. With the discovered proxy relations represented as interpretable logical rules, this study investigates how LLM decision-making behavior changes when those relations are disrupted. In particular, the aim is to determine whether some proxy attributes have a greater influence on LLM predictions than others. To achieve this, prompts are constructed for a classification task in which sensitive information is removed from the input. The identified proxy attribute is then modified while all remaining candidate information is kept unchanged. By comparing

the resulting LLM predictions before and after the intervention using metrics such as flip rate and accuracy change, this study evaluates the relative sensitivity of different LLMs to different proxy attributes.

Chapter 2

Preliminaries

2.1 Inductive Logic Programming

Inductive logic programming (ILP) [4], a classical rule based system, is a subfield of machine learning that uses logic programming to represent its knowledge learned from data. Given background knowledge and a set of positive and negative examples represented as logical facts, an ILP system derives a hypothesized logic program that explains the positive examples while excluding the negative ones. Unlike LLM, which stores all knowledge inside embeddings represented by arbitrary numbers, the explainability of ILP is where its power comes from, as the derived logic program is human readable.

2.2 Proxy Attribute

The Proxy Attribute [1] is an attribute that can be used to partially infer other attributes in the same dataset. While these attributes can be either protected or non-protected, we generally use the term proxy attribute to refer specifically to a non-protected attribute that infers information about a protected one. Consider example in Table 2.1, Marital and Country are non-protective attributes, and Sex attribute is a protective attribute. For rows where the Marital attribute has the value *husband*, the Sex attribute is consistently *male*, while rows with Marital = *wife* correspond to Sex = *female*. This indicates that specific values of the Marital attribute can be used to infer the Sex attribute. However, this relationship is only partial. When Marital = *single*, the corresponding Sex value may be *male* or *female*, meaning that Sex cannot be inferred from this value alone. Therefore, the proxy relation between Marital and Sex is defined at the level of specific attribute values, rather than implying that the entire Marital attribute always determines Sex.

ID	Marital	Country	Sex
1	husband	usa	male
2	wife	usa	female
3	husband	uk	male
4	single	uk	female
5	single	usa	male

Table 2.1: Example dataset with non-protected attributes **Marital** and **Country**, and protected attribute **Sex**. The non-protected attribute **Marital** is a proxy of **Sex** (when Marital = *husband* or Marital = *wife*).

2.3 Proxy Attribute Detection Tool in Artificial Intelligent

The Proxy Attribute Detection Tool in Artificial Intelligent (PADTAI) [1] uses an ILP-based approach to discover proxy attributes. Among the proxy attribute discovery approaches considered in this study, PADTAI [1] is particularly suitable because it is built on Inductive Logic Programming. As shown in Table 4.1, the discovered rules identify not only the attributes that act as proxies, but also the specific values associated with each proxy. This makes it possible to construct controlled interventions by replacing the identified proxy values while keeping the remaining candidate information unchanged.

In each discovered rule, the rule body represents one or more non-protected attribute values, while the rule head represents a protected attribute. The rules discovered are evaluated using precision, recall, and coverage (examples in Table 4.2). Precision measures the proportion of records satisfy the specified protected attribute in rule head that also satisfy the rule body. Recall measures the proportion of records belonging to the protected attribute that are captured by the rule body, while coverage measures the proportion of the entire dataset that satisfies both the rule body and rule head. For this experiment, proxy relations satisfying the PADTAI [1] threshold configuration of 15% recall, 85% precision, and 10% coverage are selected. These relations are subsequently used to construct controlled interventions for the LLMs experiments.

Chapter 3

Related Work

3.1 Prompt Construction

Several papers proposed strategies to expose LLMs hidden biases, such as [5] and [6]. Arzaghi et al. [6] demonstrated how LLMs can extract demographic information from names in a financial context and also the intersectionality between names (each name represents a demographic group). Work of Vijayaraghavan et al. [5] focus on specifically the Indian caste system. The one thing they both did was the usage of LLM generated prompts as their inputs. In both papers, the authors explicitly acknowledged that the use of LLM-generated prompts may itself introduce additional bias, which was identified as a limitation of their experimental design. Although their auditing procedures were intended to reduce this effect, the potential influence of LLM-generated content could not be completely eliminated. To avoid this issue, this work adopts a similar prompting-based approach but does not use any LLM-generated content during the dataset construction stage.

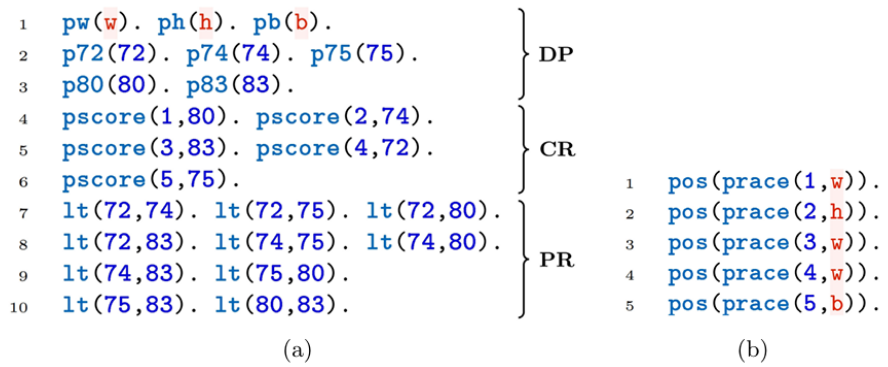


Figure 3.1: Encoding numerical attributes in a dataset using PADTAI, where 'lt' denotes 'less than'.

3.2 Proxy Attribute Discovery

Le Quy et al. [7] proposed an approach for proxy relation discovery where in a proxy relation, both proxy attribute and its inferred sensitive information are shown. This work later became part of the foundation for PADTAI [1], representing relationships between attributes as a graph in which edges indicate probabilistic dependencies. In the approach proposed by Le Quy et al. [7], continuous numerical attributes are discretized into categorical ranges before analysis. This introduces an additional preprocessing step, and the choice of discretization boundaries may influence the relationships subsequently identified between attributes.

Compared with the approach of Le Quy et al. [7], PADTAI [1] can directly handle both categorical and numerical attributes when discovering proxy relations. PADTAI [1] encodes the dataset into a logical representation so that the ILP system can search for rules describing relationships between attribute values. A key feature of this encoding is that it preserves ordering information between numerical values, allowing the system to express relationships such as one value being less than another (see Figure 3.1). This enables proxy relations involving numerical attributes to be identified without first manually discretizing them into categorical ranges.

Building on PADTAI [1], this study extends the analysis beyond identifying proxy relations in a dataset by investigating how these discovered proxies affect LLM decision-making. In particular, this study examines whether intervening on identified proxy relations changes the labels assigned by the LLM, and whether the magnitude of these changes differs across proxies associated with different sensitive attributes. This extends the analysis beyond simply identifying the existence of proxy relations by investigating whether such proxies influence LLM decision-making even when the corresponding sensitive information is not explicitly provided in the input. These effects are fairness-relevant, as they may indicate that sensitive information can still indirectly affect model predictions through proxy attributes.

Chapter 4

Methodology

4.1 Experiment Setup

Given a census dataset containing a target class and personal information, where sensitive information (depending on the dataset) and the target class are removed, extract proxy relations using PADTAI [1]. The proxy relations are then used to create variations of the candidate records. Each record is provided to the LLM as an individual input, and the model performs a classification task to predict the target class. It is important to note that, for each candidate record, only the identified proxy value is modified, while all other attribute values are kept unchanged. For proxy relations that contained multiple proxy values, for example rule 1 in Table 4.1, the experiment only perform variation on one of the proxy values and keeping other proxy value unchanged just like other non-proxy values. The following provides an illustration of how the experimental results are collected.

Let A and B denote two different values of the same proxy attribute, where A is the proxy value involved in the identified proxy relation and B is a non-proxy value that does not satisfy that proxy relation. Let Y_i represents a prediction made by LLM, such that for each candidate originally containing A, the model produces a prediction Y_i^A. The value A is then replaced by B, while all other attribute values are kept fixed, producing a second prediction Y_i^B. By analyzing the number of times that Y_i^A and Y_i^B are differed across every candidate record satisfying the proxy relation, the model's behavior on proxy-related attribute can be revealed.

To instruct the LLMs to perform the specified classification task, this study adopts a variation of the structured prompting approach used by Vijayaraghavan et al. [5]. An example of the prompt used in the experiment is shown below:

> " Perform a binary classification task.
>
> Given the following candidate information
>
> $\{\text{attribute}_1 : \text{value}_1, \text{attribute}_2 : \text{value}_2, \ldots\}$,

assign the candidate to one class label in $\{0, 1\}$.

0 = annual income $\leq$ \$50,000

1 = annual income $>$ \$50,000

Return only 0 or 1."

This prompt clearly specifies the classification task to be performed and the expected output format. The final instruction is included to constrain the LLM to return only the required class label, reducing the likelihood of additional explanations or ambiguous responses.

4.2 Dataset

The dataset selected for this experiment is one of the evaluation datasets used in PADTAI [1], from which three proxy relations were successfully identified, as shown in Table 4.1. The dataset contains U.S. census records and is commonly used for the task of predicting whether an individual's annual income exceeds US\$50,000. In this study, race, sex, and age are treated as sensitive information, following the definition used by Le Quy et al. [7].

The discovered proxy relations are presented in Table 4.1. Each proxy value belongs to a particular proxy attribute; however, not every value of that attribute necessarily acts as a proxy. As illustrated earlier in Table 2.1, a proxy attribute may only partially reveal sensitive information, meaning that only certain values are associated with a specific protected value. PADTAI [1] is therefore useful because it identifies not only the proxy attribute itself, but also the specific values involved in the proxy relation. For this reason, the experiment focuses on interventions at the proxy-value level rather than treating the entire attribute as a proxy.

The statistics of the discovered proxy relations [1] are presented in Table 4.2. Among the three rules, Rule 2 shows the strongest overall statistics, with the highest precision, recall, and coverage. Based on these values, Rule 2 may be expected to produce a comparatively stronger effect on LLM predictions when intervened upon. Rule 1 and Rule 3 also exhibit high precision, both exceeding 90%, while each covers more than 10% of the dataset population.

Beyond these statistical properties, Rule 1 is particularly interesting because its body contains two proxy conditions, as shown in Table 4.1. This allows the experiment to examine the effect of modifying one proxy attribute while keeping the other unchanged, providing insight into whether LLM predictions are sensitive to different components of a multi-attribute proxy relation.

Rule	Proxy value	Protected Information
1	Bachelors & United-States	White
2	Husband	Male
3	Craft-repair	Male

Table 4.1: Proxy relations discovered by PADTAI [1].

Rule	Count	Prec. (%)	Recall (%)	Cov. (%)
1	6,965	91.69	16.42	14.12
2	18,666	99.99	61.14	41.27
3	6,020	94.77	18.69	12.62

Table 4.2: Statistics of the proxy relations discovered by PADTAI [1]. The Count column represents the number of candidates that satisfy each proxy relation.

4.3 Metric

To quantify how sensitive the LLM is to proxy-related information when making a decision, the prediction Flip Rate is calculated by comparing the model output before and after a proxy attribute value is changed from A to B. The flip rate is defined as:

$$\text{Flip Rate} = \frac{\sum^{i}(Y_i^A \neq Y_i^B)}{N}$$

where N denotes the number of candidates satisfying the proxy relation (see Table 4.2). Both Y_i^A and Y_i^B are still consistent with their corresponding descriptions in experiment setup section. The intuition behind this metric is to reveal an LLM behavior on making decision on a controlled experiment where only a single change is made on the same candidate. A higher flip rate indicates greater model sensitivity to the presence proxy relations.

To further investigate LLM's behavior, the experiment also aims to evaluate LLM's performance change by comparing the accuracy of both Y_i^A and Y_i^B, which is defined as change of accuracy. The equation is defined as following:

$$\text{Change of Accuracy} = \text{Accuracy}_\text{B} - \text{Accuracy}_\text{A}$$

Accuracy is calculated by comparing the output Y_i^A and Y_i^B to the ground truth label for each candidate. A negative value would suggest that the LLM rely on proxy relation to make correct prediction as it can infer the hidden sensitive information, where a positive value means that the absence of proxy value helps LLMs weight each value more equally when making decision instead of favoring the proxy value.

4.4 Model Selection

The experiment is conducted using three open-weight models: Gemma-4-E4B-it from Google, Qwen3-8B from Alibaba, and Mistral-7B-Instruct-v0.3 from Mistral AI. These models were selected because they can be run locally and represent different model development ecosystems, allowing their behavior to be compared under the same experimental setting. The selected models are also relatively compact, making repeated inference computationally feasible. They are not intended to represent all LLMs, but rather to provide a diverse set of accessible open-weight models to evaluate sensitivity to proxy relations discovered by PADTAI [1].

Chapter 5

Results

5.1 Flip Rate

The Result in Table 5.1 shows the flip rate for the experiment, which evaluates the sensitivity to the proxy intervention tested. The flip rate ranges from 0 to 1, where a value of 0 indicates complete consistency between predictions made for unchanged and modified proxy values, while higher values indicate greater sensitivity to the proxy intervention. Across the three models, the Gemma model showed the greatest variation in the flip rate among the interventions tested. For rule 1, the Gemma model's flip rate changed massively when compared to the flip rate of other models being tested on the same rule. On the other hand, the Mistral model shows its lowest flip rate of 0.01 for Rule 3, indicating that its predictions are largely invariant to the Craft-repair intervention in the evaluated sample. the Qwen model shows the smallest variation in the flip rate between the interventions tested, with values ranging from 0.04 to 0.07, indicating relatively consistent sensitivity between the evaluated proxy relations. In general, the results demonstrate that the sensitivity to proxy interventions differs between models and proxy relations. In particular, the same model may respond differently depending on which proxy-related value is modified.

Rule	**Proxy Value**	**Qwen**	**Gemma**	**Mistral**
1	Bachelors	0.07	0.19	0.11
1	United-States	0.04	0.02	0.10
2	Husband	0.05	0.20	0.05
3	Craft-repair	0.04	0.03	0.01

Table 5.1: Flip rates for each proxy relations. For Rule 1, the two proxy values were modified separately while the other proxy condition was kept unchanged.

5.2 Change in Accuracy

Table 5.2 presents the change in classification accuracy for each proxy intervention in the LLMs evaluated. Particular attention is paid to interventions with relatively high flip rates in order to examine how changes in predictions affect overall model performance. Since only predictions that flip can change their correctness, the absolute change in accuracy cannot exceed the flip rate when both are expressed on the same scale.

The Gemma model provides a clear example of the distinction between prediction sensitivity and accuracy change. Two interventions associated with relatively high flip rates, **Bachelors** and **Husband**, do not produce accuracy changes in the same direction. The **Bachelors** intervention results in a substantial increase in accuracy, whereas the **Husband** intervention leads to a decrease in accuracy. This indicates that the Gemma model responds differently depending on the modified proxy relations.

The Mistral model shows a different pattern. Some interventions with relatively high flip rates result in only small changes in accuracy, while an intervention with a lower flip rate produces a comparatively larger change in performance. In general, a high flip rate does not necessarily correspond to a large change in accuracy. This is because some prediction flips may change incorrect predictions into correct ones, while others may change correct predictions into incorrect ones, causing their effects on overall accuracy to partially cancel out. Therefore, the accuracy change provides additional information about whether the prediction changes resulting from a proxy intervention improve or reduce classification performance.

Rule	**Proxy Value**	**Qwen**	**Gemma**	**Mistral**
1	Bachelors	-3.70	+7.07	+0.80
1	United-States	-1.31	+1.07	+1.31
2	Husband	+0.73	-1.79	-0.52
3	Craft-repair	+1.95	+2.35	-3.20

Table 5.2: Change in accuracy, measured in percentage points, for each proxy intervention across the three models. For Rule 1, the two proxy values were modified separately while the other proxy condition was kept unchanged.

Chapter 6

Conclusion

This study uses PADTAI [1], an ILP-based proxy attribute detector, to investigate how LLM decision-making changes when identified proxy relations are disrupted. The discovered proxy relations are associated with different sensitive information, including gender and race, allowing the experiment to examine whether LLM predictions remain sensitive to information that may indirectly encode these attributes.

The results show that the evaluated LLMs exhibit different levels of sensitivity to the proxy interventions tested. Some models respond relatively consistently across proxy relations, while others show greater sensitivity to particular proxy attributes. As LLMs are increasingly incorporated into decision-making tasks across different domains, understanding which types of proxy information have a stronger influence on model predictions is important for identifying potential fairness risks and reducing the possibility of existing biases being reinforced.

Chapter 7

Future Work

7.1 Strengthening the Evaluation of LLM Proxy Sensitivity

The current evaluation could be extended to provide stronger evidence of LLM sensitivity to proxy relations. A potential direction for future work is to incorporate the prompt-based bias mitigation strategies proposed by Lu et al. [8] and compare the behavior of the models before and after mitigation. Their work introduces three prompting strategies to reduce bias in classification tasks: fair prompting, generalized prompting, and descriptive prompting [8]. All three approaches explicitly instruct the LLM to consider fairness within the prompt without modifying the model parameters, which is compatible with the experimental design used in this study.

Future work could compare proxy sensitivity before and after applying these mitigation strategies. For example, a substantial reduction in the flip rate after mitigation would suggest that the model's original predictions were sensitive to the tested proxy relations and that the mitigation strategy successfully reduced this sensitivity. Such a result would provide further evidence that the proxy information influenced the model's decision-making behavior.

```
You are a helpful assistant. #### Task: Predict if the income of the described person
exceeds $50K per year. Answer with one of the following: 1. YES, greater than 50K;
2. NO, less than or equal to 50K.#### Description: Assume that there is a male in
1996 in the USA with the following attributes: race is white, age is 30 years old,
marital status is never-married, family relationship is not-in-family, has 10 years
of education, education level is some-college, work class is private, occupation is
sales, work hours is 50 per week, capital gain is 0, and capital loss is 0. Please
predict whether the annual income of the person is greater than $50k:
```

Figure 7.1: Example CEB prompt for a candidate record from the Adult dataset.

7.2 Adapting Evaluation Prompts for Modern LLMs

Another area for future improvement concerns the compatibility of the experimental prompts with different LLMs. During preliminary experiments, a wider range of models was considered, including gpt-oss from OpenAI and Llama from Meta. However, some models did not produce valid classification outputs. In several cases, the models returned the same class label for nearly every candidate or otherwise failed to perform the requested classification task.

Several modifications to the prompting strategy were attempted, including explicitly restricting the output to either 0 or 1 and changing the format used to represent candidate information. These changes did not consistently resolve the issue. One possible explanation is that prompting strategies used in earlier fairness-evaluation studies may not transfer directly to newer models, particularly those with stronger safety mechanisms or different instruction-following behavior.

A potential direction for addressing this limitation is the Compositional Evaluation Benchmark (CEB) proposed by Wang et al. [9]. CEB provides a broader framework for evaluating fairness-related behavior across different tasks and models. Of particular relevance to this study is its use of more natural conversational prompt formats for fairness evaluation, including tasks derived from tabular datasets such as the Adult dataset (See Figure 7.1). Adopting a similar prompting strategy may improve the compatibility of the proposed experiment with newer LLMs and enable a wider range of models to be evaluated without changing the underlying proxy-intervention methodology.

Bibliography

[1] R. Gonçalves, F. Gouveia, I. Lynce, and J. F. Santos, "Proxy attribute discovery in machine learning datasets via inductive logic programming," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2025, pp. 343–363.

[2] P. S. Rao, L. N. Venkatesan, M. Cherubini, and D. B. Jayagopi, "Invisible filters: Cultural bias in hiring evaluations using large language models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 3, 2025, pp. 2164–2176.

[3] K. Benkirane, J. Kay, and M. Perez-Ortiz, "How can we diagnose and treat bias in large language models for clinical decision-making?" in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 2263–2288. [Online]. Available: https://aclanthology.org/2025.naacl-long.114/

[4] Z. Zhang, L. Yilmaz, and B. Liu, "A critical review of inductive logic programming techniques for explainable ai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 220–10 236, 2024.

[5] P. Vijayaraghavan, S. Vosoughi, L. Chiazor, R. Horesh, R. Abreu de Paula, E. Degan, and V. Mukherjee, "Decaste: Unveiling caste stereotypes in large language models through multi-dimensional bias analysis," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, J. Kwok, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2025, pp. 9899–9907, aI and Social Good. [Online]. Available: https://doi.org/10.24963/ijcai.2025/1100

[6] M. Arzaghi, F. Carichon, and G. Farnadi, "Understanding intrinsic socioeconomic biases in large language models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, 2024, pp. 49–60.

[7] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

[8] H. Lu, Z. Shao, J. Gao, and S. Uddin, "Bias mitigation in large language models for tabular data classification," *Machine Learning*, vol. 115, no. 4, 2026.

[9] S. Wang, P. Wang, T. Zhou, Y. Dong, Z. Tan, and J. Li, "CEB: Compositional evaluation benchmark for fairness in large language models," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=IUmj2dw5se

Appendix A

Dataset Statistics

This appendix provides additional statistics for the dataset used in the experiment.

Table A.1: Numerical attributes in the Adult dataset.

Attribute	**Unique Values**
age	74
fnlwgt	26,741
educational_num	16
capital_gain	121
capital_loss	97
hours_per_week	96
class_label	2

Table A.2: Categorical attributes in the Adult dataset.

Attribute	**Unique Values**
workclass	7
education	16
marital_status	7
occupation	14
relationship	6
race	5
gender	2
native_country	41