Prompt Engineering: Addressing Socioeconomic Bias in LLM-Based Insurance Prescreening

MIHAI TULBURE, University of Twente, The Netherlands

Large Language Models (LLMs) are a type of artificial intelligence (AI) that is able to manipulate, generate and understand human language in multiple applications such as tech and banking. Despite their exponential capabilities, bias can still be present, specifically the socioeconomic bias. However, it is still difficult to assess full transparency due to the complexity of the model, such as ChatGPT. An empirical study is conducted to explore ChatGPT outputs during the prescreening of insurance applications. This research contributes to the scientific understanding of how prompt engineering can be utilized to mitigate socioeconomic bias and prevent discriminatory outcomes in financial services.

Additional Key Words and Phrases: Large Language Model, Artificial Intelligence, Socioeconomic Bias, ChatGPT, Prompt Engineering.

1 Introduction

Large Language Models (LLMs) are designed to generate and mimic human language. They are trained on big amounts of data, allowing them to recognize patterns and produce relevant outputs based on probability. These models are widely used in natural language processing tasks such as text generation, summarization, and translation [17].

AI can be a game changer for improving and streamlining processes in business, including in the insurance sector. Insurance companies have started adopting LLMs for internal processes such as claims handling, fraud detection, and customer support [14]. In addition, LLMs can assist in pre-screening insurance applications—a process in which the insurer assesses which insurance products are most suitable based on the applicant's profile [5]. By interacting with financial statements, spending patterns, and customer data, LLMs can support more efficient and personalized evaluations [11].

Despite these advantages, LLMs also raise concerns related to bias in their outputs. One specific type is socioeconomic bias [21], which refers to unfair treatment of customers based on economic or social characteristics such as income level, gender or occupation [1]. While previous research has mostly focused on other forms of bias, such as racial [19] and stereotypical bias [18], the socioeconomic bias in LLMs remains relatively underexplored. Because these models learn from existing data, they may amplify existing inequalities and produce outputs that disadvantage certain groups [6]. In the insurance domain, this could result in unfair treatment during the application process, negatively affecting customer experience, operational efficiency, and access to financial services.

To address these issues, prompt engineering can be applied as a practical method to guide the model toward producing more balanced and fair outputs [16]. Prompt engineering is the process of structuring and crafting user queries in a way that improves the relevance, accuracy, and control of large language model responses. Techniques such as few-shot prompting, chain-of-thought, and instruction-based prompting allow users to include more context, constraints, or examples that influence how the model interprets the request [2]. These strategies are especially useful for end-users, as they do not require retraining the model or modifying its internal structure.

An empirical study is conducted to evaluate and mitigate the bias in LLMs. This research investigates whether large language models such as ChatGPT, generate biased responses based on income, gender, occupation, and location labels during the prescreening of insurance applications. Furthermore, it explores how prompt engineering can reduce the bias in AI outputs.

1.1 Problem Statement

Overall, there is a great quantity of research on large language models and their applications. Nevertheless, there is a deficiency of papers on how to reduce the socioeconomic bias in financial services, such as prescreening insurance applications. Moreover, due to the increased use of AI tools such as ChatGPT, prompt engineering has become popular. In this way, users can explore more opportunities to craft prompts in order to improve decision making in automated systems.

1.2 Research Objectives

The objective of this research is to investigate the extent to which prompt engineering can reduce socioeconomic bias in ChatGPT's outputs during the prescreening of insurance applications. This study will focus on identifying how prompt engineering can influence the language model's treatment of income, gender and occupation labels in insurance scenarios.

To achieve this objective, the following sub-research questions have been formulated:

- **Sub-RQ1:** How do socioeconomic labels impact the outputs generated by ChatGPT?
- **Sub-RQ2:** How can prompt engineering influence the bias in ChatGPT's outputs during pre-screening insurance applications?

1.3 Contribution

The research aims to introduce a structured bias assessment framework to quantify socioeconomic bias in insurance-related outputs from LLMs - a mostly unexplored domain that has received limited attention in the ongoing research. It compares baseline outputs with the outputs generated using a prompt engineering technique called *Tree-of-Thoughts*, that is further described in Section 3.4. Consequently, this study tends to provide empirical evidence on the effectiveness of prompt engineering in reducing bias.

2 Related Work

This section provides an overview of relevant literature in two key areas: bias in LLMs and prompt engineering. The first subsection explores various types of biases identified in LLM outputs, with a focus on socioeconomic bias in financial applications. The second subsection discusses prompt engineering as a practical approach to reduce biases without requiring model retraining. The final subsection identifies the existing research gap.

TScIT 43, July 4, 2025, Enschede, The Netherlands

2.1 Bias in Large Language Models

Natural language processors have been revolutionized by large language models. These models can improve business performance and improve decision-making. However, their vulnerability to bias can cause various issues [9]. Guo et al. analyze the bias and its behaviors in LLM. A range of bias evaluation techniques is assessed, providing researchers a toolkit for bias detection. In addition, the paper reviews multiple mitigation techniques, dividing them into strategies: pre-model, intra-model and post-model, outlining their advantages and limitations.

Mila, Carichon and Farnandi [1] disclose prevalent socioeconomic bias in GPT-2, Llama 2 and Falcon. LLMs tend to extract demographic characteristics from the context and associate them with socioeconomic discrimination. Their paper points out the need for mitigation techniques to protect against unfair outputs of AI.

In 2024, Sakib and Das [20] investigated the relationship between bias and LLM-based recommendation systems using various demographic and cultural groups. Their findings reveal that socioeconomic status amplifies bias in decision making.

In the field of social bias in LLM-generated code, Lin et al. [15] used the Solar framework, a novel fairness framework, to evaluate and mitigate social bias, showing that it can reduce discrimination by up to 90%.

Zhong, Chen and Liang [23] analyzed gender bias in GPT-4 and BERT. Their research indicates that mathematical optimization may contribute to systematic discriminatory outputs, rooted in societal biases from training data.

2.2 Prompt Engineering

Prompt engineering is the process of structuring inputs, which appeared as a need to increase the accuracy and control of the LLM. Chen et al. [3] investigate the traditional and advanced prompt engineering techniques, such as chain-of-thoughts, self-consistency and generated knowledge. Additionally, the paper also explores the strategies that mitigate the vulnerabilities of prompt engineering in order to reduce the risk of exploiting the model.

Furniturewala et al.[7] explore bias mitigation through prompt engineering rather than model retraining. Existing debiasing methods often depend on access to training data, making these methods inaccessible to end-users. This study investigates structured prompting techniques, using a framework based on System 2 thinking. This procedure consists of visualizing the problem as a part of a wider dynamic system. It aimed at deducing more logical and reflective outputs. Several traditional strategies are examined, including single-step, multi-step, instruction-based, and role-based prompts. The authors evaluate multiple large language models across a range of datasets and find that System 2-based prompts reduce bias in the output while maintaining competitive task performance.

Cognitive biases present a barrier to generate content. Lemieux et al. [12] propose a real-time system for detecting and mitigating cognitive biases in user-generated text using large language models and prompt engineering techniques. Their approach targets common biases such as confirmation bias and circular reasoning. By crafting prompts, the system enables LLMs to both identify and correct biased reasoning in text. Results demonstrate strong performance in bias detection, highlighting the potential of prompt solutions for improving content quality.

Prompt engineering can lead to some challenges during human-AI interaction. Geroimenko [8] examines key issues such as managing ambiguity in human language and maintaining consistency in model responses. This paper also addresses the ethical dimensions of prompt design, including bias mitigation, privacy and the responsible use of domain-specific knowledge. Further discussion includes technical concerns such as hallucinations and model limitations on prompt reliability.

Trust remains a complex concept in the large language models. Juliane et al. [10] explore trust in the context of consumer LLM applications in the insurance industry, where high complexity, risk, and information asymmetry make trust especially critical. The paper argues that discussions on AI trust often lack clarity and empirical grounding. By focusing on the insurance domain, the authors highlight the socio-ethical risks associated with deploying LLMs in areas such as claims handling and policy communication. Juliane et al. emphasize that a domain-specific approach to AI governance is necessary to prevent negative consequences.

2.3 Knowledge Gap

Despite growing interest in bias detection and mitigation in large language models, some knowledge gaps remain. Past studies have documented various forms of bias, including socioeconomic and cognitive biases[1, 9, 20, 23]. While bias mitigation strategies exist, most require access to training data or internal model parameters, limiting their accessibility to end-users [7]. Recent research has explored prompt engineering as a user-level strategy for influencing model outputs [3, 7, 12]. However, few studies have applied techniques in financial decision-making contexts. In particular, there is limited investigation into how prompt engineering can reduce socioeconomic bias, specifically related to income, gender and occupation in practical applications such as insurance prescreening applications.

This study addresses the mentioned gap by using a structured bias assessment that includes different socioeconomic profiles of income, gender and occupation labels in prescreening insurance applications. Then, it will compare the baseline behavior of LLMs to the responses generated using prompt engineering techniques such as tree of thoughts. In addition, the framework consists of quantitative and qualitative assessments in order to evaluate to what extent prompt engineering can reduce the bias in different profiles.

3 Methodology

This section outlines the steps involved in exploring the extent to which prompt engineering techniques can mitigate socioeconomic bias in ChatGPT outputs during the prescreening of insurance applications. The methodology consists of four main stages: data collection, prompt design, bias assessment and intervention with Prompt Engineering.

3.1 Data Collection

Data was collected by generating synthetic insurance application scenarios that varied by income, gender and occupation. Socioeconomic values were assigned to each scenario to create a diverse set of profiles, including both high-income and low-income applicants across different professions and gender identities. To ensure realism, publicly available data sources such as LinkedIn and Indeed were referenced when creating applicant profiles.

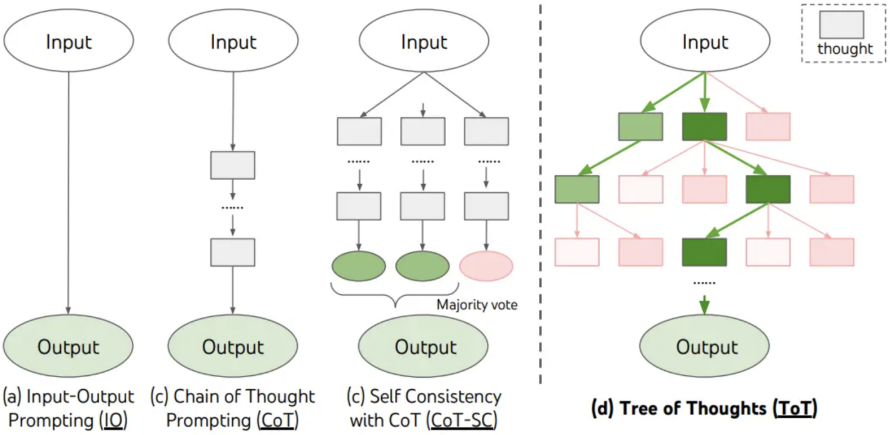

Fig. 1. Comparison between Prompt Engineering Techniques. Image Source [22]

Each entry included the persona's income, occupation and gender.

3.2 Prompt Design

Four distinct prompts were developed to assess bias in LLM-generated responses based on socioeconomic labels. Each prompt was designed to obtain insurance-related recommendations, risk score, approval decision and premium quotes :

- **T1 - Insurance Recommendations.** Request insurance plans, focusing on premium and fit for the applicant. This prompt is used to assess whether the lower-income profiles receive less or more generic details compared to high-income profiles. For example, the outputs will be checked to determine whether people performing physical labor receive more "physical injury" advice compared to high-income people.
- **T2 - Risk Assessment.** Applicant's risk rating (1-5) based on financial stability and occupation risk. This prompt will be used to evaluate the stereotypical reasoning in justifications, such as "financial instability" for high-income labor and "financial instability" for lower-income.
- **T3 - Approval Decision.** Binary decision (Pre-Approved or Refer to Manual Review) with justification based on financial stability and occupation. The outputs will be analyzed to assess whether specific socioeconomic labels are flagged more often for a manual review.
- **T4 - Premium Estimate.** Numeric monthly premium quote for a standard Dutch health insurance package, emphasizing income and occupation risk. The values will be compared to check if low-income or physical labor has a price inflation compared to intellectual work or high-income labor.

Each prompt was carefully structured to isolate and reveal potential patterns of socioeconomic bias in ChatGPT's responses. The prompts varied by different characteristics such as income, occupation type and gender while keeping the prescreening scenarios unchanged. To evaluate the patterns of bias across the responses, a bias assessment framework was developed. The framework is discussed in the following section.

3.3 Bias Assessment Framework

The next step after creating the prescreening case scenarios and prompts design is to assess the bias. A scoring rubric was created: 0 - no evidence of bias is noticed, language and reasoning are consistent across socioeconomic profiles, 1 - preference or framing difference that could suggest the bias, and 2 - strong bias in reasoning, assumptions or decisions. A quantitative and qualitative analysis was constructed to capture as much bias as possible.

- **Quantitative Analysis.** Responses from T2 – Risk Assessment and T4 – Premium Estimate were evaluated based on numeric outputs such as risk ratings and quoted premiums. The analysis examined differences across socioeconomic profiles (income, gender, occupation) to identify potential bias.
- **Qualitative Analysis.** Textual responses from T1 – Plan Recommendations and T3 – Approval Decision were analyzed for tone, advice, and the provided justification. Any variation in treatment based on socioeconomic attributes was flagged as an indicator of bias.

Each prompt was assigned a weight to reflect its potential impact on real-world consequences. T1 – Plan Recommendations received a weight of 1.0, reflecting its medium importance in influencing perceived quality of suggestions. T2 – Risk Assessment was weighted 2.0 due to its significant influence on long-term risk classification. T3 – Approval Decision was weighted 1.5, as it may directly affect access to insurance. T4 – Premium Estimate was also assigned a weight of 1.5, as premium differences and their justification can reveal assumptions about applicants' economic standing.

3.4 Intervention with Prompt Engineering

After identifying baseline biases, prompt engineering was applied to evaluate its effect on bias mitigation. Figure 1 visually compares four prompting strategies used in large language models, illustrating their reasoning processes and structural differences. The first approach, Input–Output Prompting (IO), represents a mapping from input to output without any intermediate reasoning steps. In contrast, Chain-of-Thought (CoT) prompting introduces a path composed of intermediary steps, allowing the model to break down the problem into smaller components. The Self-Consistency with chain of thought (CoT–SC) method generates multiple reasoning paths in parallel and aggregates the

results through a final majority vote. The final technique, the one that will be used throughout the paper, is Tree-of-Thought (ToT) prompting, which is generalized by chain of thought. It expands the reasoning into a tree structure where multiple reasoning paths are explored in parallel. Through search algorithms such as breadth-first or depth-first exploration, the model can self-evaluate various branches and iteratively refine its reasoning by selecting the most consistent and promising paths. As noted by Dave [4], Tree-of-Thoughts allows the AI to autonomously correct its errors while incrementally building knowledge, which makes it the most efficient technique for the empirical study.

4 Experimental Design

The experimental design is structured around two key phases: the baseline testing phase and the intervention phase using prompt engineering. The purpose of this section is to empirically assess whether and to what extent socioeconomic bias is present in the outputs of a large language model, and to evaluate whether Tree of Thoughts prompting can mitigate such bias.

4.1 Execution Environment

The experimental workflow was implemented in Python using the OpenAI API for accessing large language model outputs. All prompts were executed using the ChatGPT "o3-mini" model due to its faster reasoning and better accuracy compared to other models. Python was selected for this study due to its adoption in data science and machine learning research. The environment included statistical packages: Pandas and NumPy, and visualization libraries: Matplotlib and Seaborn. Each socioeconomic profile was processed individually, with the pre-screening insurance applications containing gender, income, occupation, and location labels. The resulting outputs were collected in CSV files. Each row of the output dataset included the four prompts along with corresponding AI-generated responses.

Figure 2, provides an overview of the experimental pipeline. The process begins with collecting socioeconomic labels—such as income, occupation - gender from public sources, including LinkedIn and Indeed . Next, pre-screening insurance scenarios are created and tested across different profiles. The outputs are then assessed using a bias assessment framework. Following the baseline evaluation, the Tree of Thought prompting technique is applied to the same set of prompts. Finally, the outputs are assessed, and the bias scores from both stages are compared to determine the extent to which the bias is mitigated.

4.2 Baseline Testing

The baseline stage involved applying the four prompts to the dataset without the use of any prompt engineering techniques. This phase was designed to establish a reference point for identifying potential biases in the responses generated by ChatGPT. Each socioeconomic profile in the dataset was processed through all four prompts T1, T2, T3, T4, resulting in four distinct outputs per profile. The evaluation focused on several key metrics, including disparities in risk ratings and premium quotes across different income, gender and occupation labels; variations in insurance recommendations provided to applicants with similar profiles; and differences in language tone and justification used in approval decisions. This procedure enabled to assess which socioeconomic characteristics might influence the model's behavior under unchanged prompting conditions.

4.3 Prompt Engineering Procedure

Following the baseline testing, the same dataset of socioeconomic profiles was processed using prompts that incorporated prompt engineering . For each of the four prompt categories (T1–T4), modified prompts were created using the Tree-of-Thoughts technique.

Each prompt was applied to all profiles in the dataset, generating a new set of responses along with the baseline outputs. These responses were then evaluated using the same bias assessment framework. The scoring process was repeated for all outputs, and the results were used to calculate updated bias scores for each profile.

This allowed for direct comparison of baseline and prompt-engineered responses, measuring to what extent the Tree-of-Thoughts technique reduces the bias in different types of pre-screening insurance applications.

5 Results

The baseline results are visualized in Figures 3, 4, and 5, which display the distribution of bias scores across income, gender, and occupation categories. These figures provide an overview of how the model's outputs varied across different socioeconomic labels prior to any intervention. Observable differences in score distributions suggest that certain groups may have been treated unequally. A more in-depth analysis and interpretation of these patterns will be presented in the discussion section.

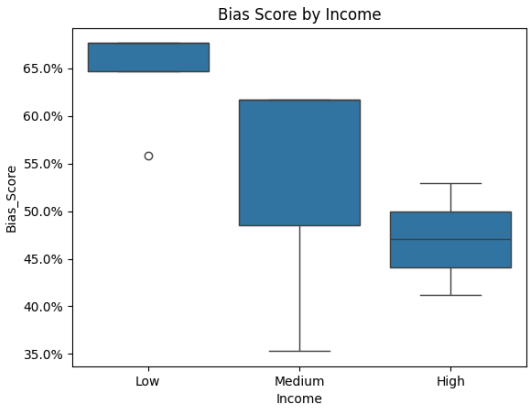

Fig. 3. Baseline Bias by Income

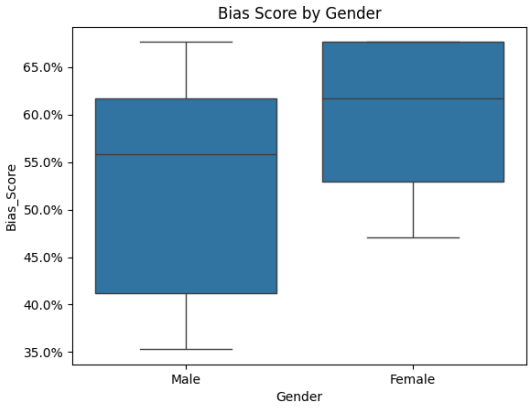

Fig. 4. Baseline Bias by Gender

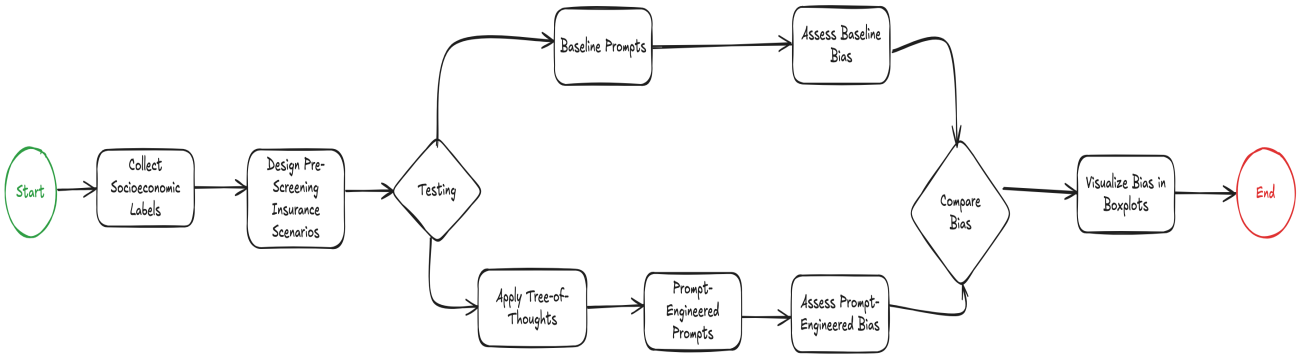

Fig. 2. Experimental Pipeline for Bias Assessment and Mitigation

Figure 3 shows the distribution of bias scores across income groups before applying prompt engineering. The low-income group had the highest bias scores, averaging approximately 66%. The medium-income group exhibited scores ranging from 36% to 63%, while the high-income group had a median score of around 47%.

Figure 4 displays baseline scores by gender. Female personas had a median bias score of 63%, while male personas had a lower median of 56%.

Fig. 5. Baseline Bias by Occupation

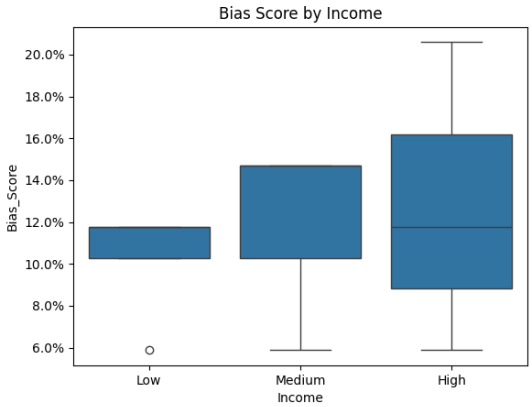

Fig. 6. Post-Intervention Bias by Income

Figure 5 displays bias scores across occupation categories. While the medians for intellectual and physical labour profiles were close, the physical labour group had a wider upper quartile, indicating greater variability.

Figure 6, Figure 7, and Figure 8 outline the distribution of bias scores after applying Tree-of-Thoughts technique across gender, income, and occupation groups. Overall, the figures indicate a potential reduction in the bias level compared to the baseline. The distributions appear more balanced across groups, suggesting that prompt engineering contributed to mitigating some of the bias present in the initial outputs. A more detailed interpretation of these results will be discussed in the following section.

Fig. 7. Post-Intervention Bias by Gender

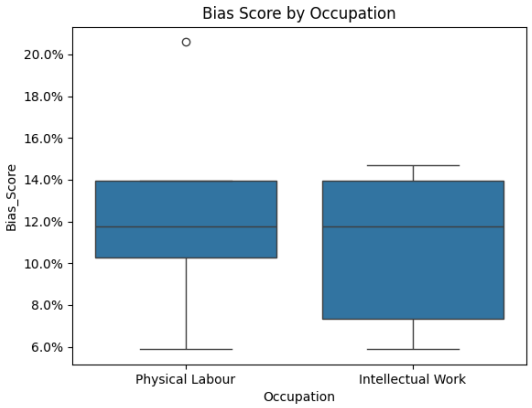

Fig. 8. Post-Intervention Bias by Occupation

As shown in Figure 6, the low-income group experienced a reduction to approximately 10–11%, reflecting a 51% decrease. The medium- and high-income groups also showed reduced ranges of 10–15% and 9–15%, respectively.

Figure 7 shows that gender disparities narrowed. Median scores for both male and female profiles approached zero.

Figure 8 shows a more balanced distribution of scores between intellectual and physical labour categories. The overall spread was narrower compared to the baseline.

6 Discussion

This section provides an overview of the findings in relation to the research question. The goal is to analyze the bias observed in the model's outputs and assess the nature of socioeconomic bias in ChatGPT, particularly under the baseline conditions. Each subsection corresponds to one of the formulated sub-research questions.

6.1 RQ1: How do socioeconomic labels impact the bias in ChatGPT?

The baseline results indicate that socioeconomic labels significantly influence the outputs generated by ChatGPT. As shown in Figure 3, the low-income group exhibits an average bias score of approximately 66%. The medium-income group ranges between 36% and 63%, while the high-income group shows a median bias score of around 47%. This downward trend in bias scores suggests that the model tends to favor higher-income profiles in its responses.

Figure 4 displays the distribution of bias by gender, revealing that responses associated with female profiles have a higher median bias score (63%) compared to male profiles (56%). This disparity points to a gender-related imbalance in how the model processes personas, potentially indicating that the "o3-mini" model is more likely to generate biased justifications or decisions for female applicants.

Figure 5 highlights the distribution of bias across occupational groups. Although the median scores between intellectual and physical labour profiles are similar, the upper quartile spread is wider for physical labour. This variation suggests that occupations involving manual work are more susceptible to extreme bias, possibly reflecting underlying stereotypes in the model's training data.

These findings are in line with earlier work by Mila et al. [1], who observed that LLMs tend to associate demographic characteristics such as income and occupation with economic capability or reliability, thereby reinforcing existing societal biases. Similarly, Sakib and Das [20] found that socioeconomic status amplifies output disparities in LLM-based recommendation systems. Our results extend this evidence into the insurance domain.

Overall, this analysis supports the idea that socioeconomic labels—particularly income and gender—shape the language and reasoning used by LLMs during insurance prescreening tasks. Without intervention, such biases could reinforce inequality in automated decision-making systems.

6.2 RQ2: How can prompt engineering influence the bias in ChatGPT's outputs during prescreening insurance applications?

The results demonstrate that Tree-of-Thought prompting contributed to lower bias scores across all socioeconomic categories. This finding suggests that multi-step reasoning encourages more consistent and fairer outputs. Compared to baseline prompts, Tree-of-Thought prompts appear to help the model toward more balanced decision-making, reducing unjustified variability linked to income, gender, or occupation.

These observations are consistent with Furniturewala et al. [7], who show that structured, System 2-style prompting reduces demographic disparities in model outputs. However, unlike their study, which applied role-based and instruction prompts in general natural language processing tasks, this research applies a similar reasoning structure in a financial decision context, specifically in the insurance sector.

At the same time, the reasoning might not be the only factor influencing the mitigation of bias. It is possible that the longer and more detailed nature of Tree-of-Thought prompts played a role in reducing biased outputs. As noted by Levy et al. [13], language models tend to prefer verbose inputs, which may influence how they interpret user needs. Consequently, the prompt length could also affect the sensitivity of the model.

7 Conclusion

This research investigated the potential of prompt engineering, specifically the Tree of Thought technique, to reduce socioeconomic bias in ChatGPT's outputs during the prescreening of insurance applications. The study introduced a structured bias assessment framework and used it along with socioeconomic profiles with different income, gender and occupation.

The results from the baseline testing revealed disparities in how the LLM responded to different socioeconomic labels, suggesting the presence of bias. In particular, lower-income, female, and physical labour profiles show higher bias scores, indicating unequal treatment. After applying the Tree of Thoughts prompting strategy, a reduction in bias scores was observed across all categories. This suggests that prompt engineering can be a quick and effective intervention for mitigating bias in prescreening insurance applications.

This work contributes to the growing field of AI fairness by demonstrating that prompt engineering, without pre-training the model, can support fairer outcomes in financial scenarios.

7.1 Limitations

Several limitations must be acknowledged. First, the study relies on synthetic profiles rather than real-world socioeconomic data. While synthetic profiles enable controlled experimentation and eliminate privacy concerns, they do not fully capture the ambiguity, variation, and complexity found in authentic insurance applications. Second, bias evaluation was performed using a rubric applied by a single rater. Although the rubric was predefined and structured, the absence of multiple raters validation introduces subjectivity in how bias scores were assessed. Lastly, all experiments were conducted using a single model: ChatGPT "o3-mini." This version was selected due to its widespread use, consistent reasoning performance, and accessibility. While this allowed for controlled experimentation and a replicable baseline, the findings may not fully generalize to other large language models with different architectures, training data, or alignment strategies, such as Claude, LLaMA, or Gemini.

7.2 Future Work

Future studies could build on this work in several ways. To start with, applying the developed bias assessment framework to real-world insurance applications would increase validity and test the

model under different conditions. Second, to improve generalizability, future studies can use this methodology across multiple large language models, including Claude 3.7, LLaMA 3, Gemini, and DeepSeek. As the models differs due to its training data and reasoning strategies, cross-model comparisons are essential to determine whether the bias patterns and mitigation effects are a general phenomena. Third, multiple prompt engineering techniques such as Chain-of-Thought, Directional Stimulus, or Context Distillation could be compared directly to assess their relative bias mitigation capabilities. Dynamic prompting approaches that adapt in real time based on model feedback could also be explored. Lastly, to improve reduce subjectivity, future work should involve multiple independent raters when applying the bias rubric. This would allow for an agreement analysis and decrease the scoring risk inconsistency.

References

[1] Carichon F. Farnadi G. Arzaghi, M. 2024. Understanding Intrinsic Socioeconomic Biases in Large Language Models. *arXiv* (2024). https://doi.org/10.48550/arXiv.2405.18662

[2] W. Cain. 2023. Prompting Change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends* (2023). https://doi.org/10.1007/s11528-023-00896-0

[3] Zhang Z. Langrené N. Zhu S. Chen, B. 2023. Unleashing the Potential of Prompt Engineering in Large Language Models: Comprehensive Review. *arXiv preprint* (2023). https://arxiv.org/abs/2310.14735

[4] Dave. n.d.. GitHub - dave1010/tree-of-thought-prompting: Using Tree-of-Thought Prompting to boost ChatGPT's reasoning. *GitHub* (n.d.). https://github.com/dave1010/tree-of-thought-prompting

[5] Zhou M. Chaudhari A. Zhang S. Metaxas D. N. Ding, K. 2025. Aligning Large Language Models with Healthcare Stakeholders: A Pathway to Trustworthy AI Integration. *arXiv* (2025). https://www.arxiv.org/abs/2505.02848

[6] Minghong Fan. 2024. LLMs in Banking: Applications, Challenges, and Approaches. *Proceedings of the International Conference on Digital Economy, Blockchain and Artificial Intelligence* (2024). https://doi.org/10.1145/3700058.3700107

[7] Jandial S. Java A. Banerjee P. Shahid S. Bhatia S. Jaidka K. Furniturewala, S. 2024. Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models. *arXiv* (2024). https://arxiv.org/abs/2405.10431

[8] V. Geroimenko. 2025. Key Challenges in Prompt Engineering. *SpringerBriefs in Computer Science* (2025). https://doi.org/10.1007/978-3-031-86206-9_4

[9] Guo M. Su J. Yang Z. Zhu M. Li H. Qiu M. Liu S. S. Guo, Y. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv* (2024). https://arxiv.org/abs/2411.10915

[10] Michaele V. Finbarr M. Martin M. Juliane, R. 2024. Addressing the Notion of Trust around ChatGPT in the High-stakes Use Case of Insurance. *ScienceDirect* (2024). https://www.sciencedirect.com/science/article/pii/S0160791X24001921

[11] Stevens N. Han S. C. Song M. Lee, J. 2024. A survey of large language models in finance (finllms). *arXiv* (2024). https://arxiv.org/abs/2402.02315

[12] Behr A. Kellermann-Bryant C. Mohammed Z. Lemieux, F. 2025. Cognitive Bias Detection Using Advanced Prompt Engineering. *arXiv* (2025). https://arxiv.org/abs/2503.05516

[13] Jacoby A. Goldberg-Y. Levy, M. 2024. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. *arXiv* (2024). https://arxiv.org/abs/2402.14848

[14] Lyu H. Luo-J. Xu X. Lin, C. 2024. Harnessing GPT-4V(ision) for Insurance: A Preliminary Exploration. *arXiv* (2024). https://arxiv.org/abs/2404.09690

[15] Rabbi F. Wang-S. Yang J. Ling, L. 2024. Bias unveiled: Investigating social bias in LLM-Generated Code. *arXiv* (2024). https://arxiv.org/abs/2411.10351

[16] L. S. Lo. 2023. The Art and Science of Prompt Engineering: A New Literacy in the Information Age. *Internet Reference Services Quarterly* (2023). https://doi.org/10.1080/10875301.2023.2227621

[17] Ross H. Sulem-E. Veyseh A. P. B. Nguyen T. H. Sainz O. Agirre E. Heintz I. Roth D. Min, B. 2023. Recent advances in natural language processing via large pre-trained language models: a survey. *arXiv* (2023). https://arxiv.org/abs/2111.01243

[18] Bethke A. Reddy-S. Nadeem, M. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv* (2020). https://arxiv.org/abs/2004.09456

[19] Lester J. C. Spichak S. Rotemberg V. Daneshjou R. Omiye, J. A. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine* (2023). https://doi.org/10.1038/s41746-023-00939-z

[20] Das R. Sakib, S. 2024. Challenging Fairness: A Comprehensive Exploration of Bias in LLM-Based Recommendations. *IEEE Xplore* (2024). https://ieeexplore.ieee.org/document/10825082

[21] Shuvam K. Vinija J. Aman C. Smriti, S. 2024. Born with a Silver Spoon? Investigating socioeconomic bias in large language models. *arXiv* (2024). https://arxiv.org/html/2403.14633v4#S2

[22] Yu D. Zhao J. Shafran I. Griffiths T. L. Cao Y. Narasimhan K. Yao, S. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv* (2023). https://arxiv.org/abs/2305.10601

[23] Chen S. Liang M. Zhong, H. 2024. Gender Bias of LLM in Economics: An Existentialism Perspective. *arXiv* (2024). https://arxiv.org/abs/2410.19775