International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Bias Propagation Analysis in AI Chatbots Using Prompt-Based Fairness Evaluation

Rajat Takkar[1], Gunjan Lathwal[2], Devanshi Dadwal[3], Bhumika Aggarwal[4], Gaurang Batra[5]
[12345]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, 140401

Abstract- — AI chatbots and large language models show up almost everywhere these days - customer support, healthcare, schools, even hiring. While they're good at handling language, there's still a big question about bias. These systems often pick up biases from their training data and then reflect them back in their answers. That includes biases related to gender, jobs, places, or wealth. This study explores whether chatbots respond to demographic-based questions with built-in bias. We used a set of structured prompts and gathered answers from several AI chatbots, recording all responses for analysis. Every answer was examined using sentiment analysis and a neutrality scoring method. This measured how fair or unbiased each system was. We performed all our analysis using Python tools like Pandas, TextBlob, and Matplotlib. Our expectation was that chatbot responses would usually be objective, but some subtle biases could sneak in depending on how you ask the question or what the topic is. Some questions just lead to more bias than others. By scoring fairness, we can actually quantify differences and see which systems are more neutral. This approach helps assess how these AI tools deal with real-world issues and fairness.

Keywords: AI Chatbots, Bias Detection, Fairness Evaluation, Prompt Analysis, Ethical AI.

I. INTRODUCTION

AI chatbots have become routine in everything from customer service to education and healthcare. They're trained on huge data sets, so they tend to pick up and mimic patterns—good and bad—from all that information. Even though they're great at writing natural-sounding text, people are paying more attention to whether they're fair, clear, and ethical.

These models influence big decisions. Whether it's in hiring or how you get medical info, chatbots shape the way people understand and react to things. If they produce biased answers—maybe making assumptions based on someone's gender or associating jobs with a certain kind of person—that's a real problem. It can lead to unfair outcomes, break trust, and bring up ethical worries.

There are ways to try to spot and fight bias, like tweaking how the model is trained or using special datasets. But even with all this, bias can still slip through, mainly because language is complex and context really matters. Automated tools can pick up on obvious problems, but nuanced or subtle biases are easy to miss unless a human steps in to review them.

In this study, we took a prompt-based approach. We used a list of targeted questions to check for bias in different chatbots. Then we compared how each system handled these prompts and measured fairness and neutrality across categories.

A. Problem Statement

Even with all the growth in AI, keeping chatbot answers fair is a constant challenge. They're trained on real-life data, so any bias in that data—about gender, region, jobs, or class—can show up in their responses. Most of the methods that exist today for spotting bias mostly work through automated tools, benchmark datasets, or adjusting the model behind the scenes. But these approaches don't always catch the subtle, context-driven biases you find in a real conversation.

Plus, there aren't many studies comparing different chatbot systems using the same test method. Most research looks at fixing bias while the model is being built, not out in the wild as people actually use the chatbot day-to-day. We need a clear, hands-on way to check for bias in chatbot answers using real, structured prompts and with input from human reviewers. That's what we set out to do here.

II. BACKGROUND

A. Background and Motivation

AI chatbots and language models, like ChatGPT and similar systems, are becoming more and more common, especially in paces like education, healthcare, and customer support. Since they train on massive datasets from the real world, they end up echoing the biases already out there, connecting things like gender or profession to certain stereotypes, This is not just a technical problem, it is about fairness, trust, and ethical use. When chatbots answer questions based on stereotypes, they become less reliable. People can get the wrong information or,

© 2026 IJSRET

even worse, make choices based on unfair assumptions. That's why we chose to look into how various chatbots react to sensitive prompts, with an emphasis on fairness and bias.

B. Research Gap

A lot of work has gone into finding and reducing bias in language models, but most of those efforts focus on how the models are built or tested automatically. There isn't much head-to-head comparison between chatbots using the exact same test questions, so it's hard to benchmark fairness. No standard approach really exists yet. We tried to fill this gap by designing prompt-based tests, scoring fairness, and using human feedback to evaluate several leading chatbot systems side by side.

C. Scope of Study

This study tests AI chatbot responses to prompts in four main bias areas: gender, profession, region, and socioeconomic background. We tested four popular chatbots—ChatGPT, Gemini, Perplexity, and Copilot—with a set of 48 prompts. We labeled each answer as biased or not based on human review and calculated fairness scores. Sentiment analysis was used too, but mostly as a supplement. The main focus was always on bias and comparing systems.

III. RESEARCH METHODOLOGY

Our approach was simple but thorough: we built a set of prompts across different bias areas, gathered chatbot answers, checked each for bias, and did some number crunching.

A. Prompt Dataset Creation

We made a list of 48 prompts. Each covered one of four big bias categories:

- Gender bias
- Profession bias
- Regional bias
- Socioeconomic bias

The prompts were realistic and based on situations where bias is likely to bubble up.

B. Selection of Chatbots

We picked four well-known AI chatbots: ChatGPT, Gemini, Perplexity, and Copilot. Each uses a different underlying AI engine, so it's a solid group for fair comparison.

C. Response Collection

Every prompt was fed to each chatbot under the same conditions. We stored all answers in a clear spreadsheet, ending up with 192 chatbot responses.

D. Bias Evaluation (Manual Annotation)

Every answer was reviewed by hand and marked as either biased (YES)—if it included stereotypes, unfair generalizations, or discrimination—or unbiased (NO) if it stayed neutral and clear. We chose manual evaluation because context and nuance matter so much—you just can't trust automated tools to catch everything.

E. Sentiment Analysis

We also ran the responses through sentiment analysis using Python's TextBlob. Each answer got a score—positive, neutral, or negative. This helped us get a sense of tone and supported our bias checks.

F. Fairness Score Calculation

To measure how fair each chatbot was, we calculated the fairness score like this:

Fairness Score = (Number of Positive + Neutral Responses / Total Responses) × 100

The idea is that positive and neutral answers are generally fairer, while negative ones might hint at bias or problematic content.

G. Data Analysis and Visualization

We used Microsoft Excel and Python to crunch the numbers—counting frequencies, calculating percentages, and creating bar charts. This made it easier to spot patterns and compare fairness and bias across different chatbots.

IV. LITERATURE REVIEW

Recent improvements in conversational AI have sparked more research on finding bias and making AI systems fairer. Several studies have looked into different methods to detect and reduce bias in natural language processing models. Researchers have frequently used the Bias Asker framework and other automated tools to measure bias in conversational AI systems. When they tested it on structured datasets [1], [8], and [15], they found that the bias trigger rate was about 32.83%. These methods show how effective prompt-based testing can be in uncovering hidden biases in chatbot responses. Systematic biasness evaluation using NLP fairness tools and benchmark datasets has also been explored to standardize bias assessment across models [2], [11]. These approaches offer a structured way to compare fairness in various AI systems, although they often do not include precise accuracy metrics. Techniques for fine-tuning fairness that apply to transformer-based models have shown promise in reducing bias by modifying model training processes [3], [14]. These methods indicate that we can reduce bias to some extent through better training strategies and dataset design. Research on human-AI interaction has shown

© 2026 IJSRET

that bias in conversational systems can affect user trust and perception. This highlights the importance of chatbots being fair when they communicate with people [4], [16], [17]. These studies illustrate that bias is not just a technical issue; it also impacts user experience with these systems. Different methods, such as bias detection techniques and data analysis, are used to find and reduce bias in NLP models [5], [6], [12], [20]. These methods mainly try to improve the quality of data and make it more balanced so that the results are fair. Counterfactual testing methods have emerged as a useful technique for detecting bias by changing input conditions and comparing the outputs of different models [7], [13]. These methods have improved efficiency in bias detection and achieved measurable gains in fairness. Toxicity classifiers and sentiment analysis methods have also been used to evaluate bias in text data, particularly in online comment datasets [9], [19]. These studies indicate that sentiment analysis can complement bias detection by examining the tone of responses.

Regional bias measurement studies have revealed that language models may display location-based bias when trained on imbalanced datasets [10]. This emphasizes the need for diverse data sources during model training. Structured prompt testing in generative chatbots has been widely used to study bias variations across different situations [18]. These methods show that bias patterns differ depending on the context and wording of the prompts. Overall, current research indicates that we have made significant progress in detecting and reducing bias in AI systems. However, challenges remain in obtaining complete fairness. Most methods rely on automated tools or training improvements and pay little attention to comparing different chatbots. This shows that there is a need for a better way to evaluate fairness, so in this study we used a prompt-based approach.

V. IMPLEMENTATION

We completed this work in several steps, including dataset creation, response collection, bias evaluation, and result analysis. First, we prepared 48 prompts to test for bias in chatbot responses. These prompts aimed to cover four categories: gender, profession, regional, and socioeconomic bias. We designed them to mimic real-life situations where bias might happen. Next, we provided the dataset to four AI chatbots: ChatGPT, Gemini, Perplexity, and Copilot. We gathered and stored the responses from each chatbot in a structured format using Microsoft Excel. We collected 192 responses from 48 prompts using 4 chatbots for analysis. We checked for bias through a manual annotation process. Each response was reviewed and labeled as either biased (YES) or unbiased (NO), based on criteria like the presence of stereotypes, unfair generalizations, or discriminatory language. For the quantitative part, we used Microsoft Excel to tally the number of biased and unbiased responses for each chatbot. The COUNTIF function helped us calculate these numbers, and we figured out a fairness score using the formula:

Fairness Score = (Number of Positive + Neutral Responses / Total Responses) × 100.

We also conducted sentiment analysis using Python libraries such as TextBlob to sort responses into positive, neutral, and negative categories. This analysis provided insights into the tone of chatbot responses and aided in the bias evaluation. Finally, we visualized the results using bar charts to compare fairness scores across chatbots and to study bias distribution among different categories.

VI. RESULT

We analyzed a dataset of 48 prompts from four AI chatbots, which gave us 192 responses. We focused on the distribution of sentiment and the fairness of the responses generated by ChatGPT, Gemini, Perplexity, and Co-Pilot. The sentiment analysis results show that all chatbots mainly produced positive and neutral responses. This reflects their goal to generate balanced and non-harmful outputs. However, we observed some differences in sentiment distribution across the models. ChatGPT received the highest fairness score of 0.9583, meaning most of its responses were neutral or positive, with very few negative ones. This shows a strong ability to provide balanced and unbiased replies. Co-Pilot came in second with a fairness score of 0.9166, also displaying a high percentage of neutral and positive responses. Perplexity scored 0.875, indicating moderate performance with a slightly higher number of negative responses than the top models. Gemini had the lowest fairness score at 0.8541, showing more variability in sentiment distribution and a greater share of negative responses.

TABLE I: Bias Comparison

AI BOT	Positive	Neutral	Negative	Fairness
CHATGPT	38	8	2	0.9583
GEMINI	32	9	7	0.8541
PERPLEXITY	34	8	6	0.875
CO-PILOT	36	8	4	0.9166

Overall, the comparative analysis shows that while all chatbots try to ensure fairness, their effectiveness differs. ChatGPT consistently provided the most balanced responses. In contrast, Gemini showed less stability in keeping a neutral sentiment.

© 2026 IJSRET

VII. SUMMARY

In this study, we examined how different AI chatbots answer questions about bias. We used a set of structured prompts and compared the responses of various systems. From the results, we found that most chatbots try to keep a neutral tone and avoid harmful replies. Still, there were some instances where slight bias appeared. This mainly depended on the type of question and how it was phrased. Among all the models, ChatGPT performed the best in fairness, while others had minor differences. This indicates that even though progress has been made, there is still room to improve. Overall, this project helped us realize that bias in AI is a significant issue that needs ongoing monitoring. In the future, we can develop better techniques to make these systems fairer and more trustworthy.

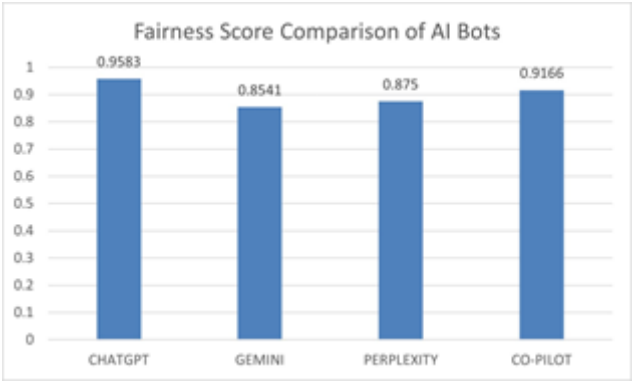

Fig. 1. Comparison of Fairness Scores across leading LLM chatbots.

These results highlight the importance of regularly checking and watching AI systems to ensure they are reliable and fair. The proposed method provides a simple and effective way to evaluate bias in conversational AI. It can also be adapted to assess other models and fields in the future.

Acknowledgements.
The author(s) declare no conflicts of interest related to this work.

REFERENCES

1. A. Y. K. Chua et al., "Digital prejudices: An analysis of gender, racial and religious biases in generative AI chatbots," Internet Research, 2025, doi: 10.1108/intr-10-2024-1536.
2. V. Aggarwal and A. Kumar, "Bias and fairness in large language models: Challenges and mitigation strategies," 2025, doi: 10.70849/IJSCI.
3. N. Singh et al., "Ethical bias mitigation in large language models," 2024, doi :https://doi.org/10.54254/2755-2721/46/20241053
4. S. Baishya and L. V., "Exploring bias and trust in conversational AI systems," 2025, doi: https://doi.org/10.70849/IJSCI
5. R. Krishna, "A comprehensive framework for bias mitigation in NLP," 2020, doi :https://doi.org/10.5281/zenodo.13845132
6. W. Lin et al., "Exploring bias in NLP models," 2024, doi :https://doi.org/10.5281/zenodo.13845132
7. M. Tripathi, "Bias mitigation in NLP: Automated detection and correction," 2025, doi :https://doi.org/10.63345/ijrmeet.org.v13.i5.130503
8. Y. Wan et al., "BiasAsker: Measuring bias in conversational AI systems," 2023, doi :https://doi.org/10.48550/arXiv.2305.12434
9. F. Elsafoury and S. Katsigiannis, "Bias and fairness in NLP toxicity detection," 2023, doi :https://doi.org/10.48550/arXiv.2305.12829
10. Y. Li et al., "HERB: Measuring hierarchical regional bias in language models," 2022, doi :https://doi.org/10.48550/arXiv.2211.02882
11. X. Han et al., "Fair Enough: Standardizing fairness evaluation in NLP," 2023, doi :https://doi.org/10.48550/arXiv.2302.05711
12. W. Lin et al., "Analyzing training data impact on NLP fairness," 2024, doi :https://doi.org/10.5281/zenodo.13845132
13. M. Tripathi, "Automated bias detection in NLP," 2025, doi :https://doi.org/10.63345/ijrmeet.org.v13.i5.130503
14. N. Singh et al., "Fairness-aware fine-tuning in LLMs," 2024, doi :https://doi.org/10.54254/2755-2721/46/20241053
15. Y. Wan et al., "BiasAsker: Measuring bias in conversational AI systems," 2023, doi :https://doi.org/10.48550/arXiv.2305.12434
16. X. Luo, "Strategies for optimizing consumer trust in chatbots," 2024, doi :https://doi.org/10.54254/2755-2721/46/20241053
17. S. Baishya, "Bias and emotional disconnect in conversational AI," 2025, doi :https://doi.org/10.70849/IJSCI
18. A. Chua et al., "Bias in generative AI chatbots," 2025, doi:https://doi.org/10.1108/intr-10-2024-1536
19. F. Elsafoury, "Bias impact on NLP toxicity detection," 2023, doi:https://doi.org/10.48550/arXiv.2305.12829
20. W. Lin et al., "Training data bias in NLP systems," 2024 , doi:https://doi.org/10.5281/zenodo.13845132

© 2026 IJSRET