

# **Am I (AI) Biased**

Jason Tran, Michelle Duong, AaJanae Henry

Computer Science Department

Pacific University

9/23/2025

## **Abstract**

Our research project examines how large language models (LLMs) produce and reflect bias in their outputs. Because these systems are increasingly used in chatbots, classrooms, healthcare, and hiring platforms, the text they generate can shape decisions and affect people directly. Yet, models trained on internet data often reproduce stereotypes, leave gaps in representation, or present misleading claims. Our research project will test AI outputs across different contexts to uncover where bias shows up and why. These findings will inform more responsible development practices, offering both a clearer picture of the limits of current LLMs and practical steps for safer deployment. By documenting these patterns, the research project aims to help developers, researchers, and everyday users engage with AI in ways that are more ethical and trustworthy.

## **Problem Statement**

The problem this research project addresses is bias in LLMs. These models are trained on large amounts of internet data, which means they can pick up and reproduce the same patterns of bias found in that data [1]. As a result, they often generate outputs that unintentionally reinforce stereotypes, leave out underrepresented perspectives, or provide misleading information. This is not just a technical issue, but one that affects people in real ways. For example, biased outputs in hiring tools can disadvantage certain applicants, AI used in healthcare can overlook specific groups of patients, and educational tools can present narrow or one-sided views.

Anyone who interacts with AI is impacted. Casual users of chatbots may get answers that reflect bias, students and educators may rely on content that is incomplete, and organizations in areas like healthcare or hiring may make decisions based on skewed information. This matters because biased outputs can influence important choices and reduce trust in AI systems.

The goal of this research project is to systematically identify and analyze these biases in AI-generated text. By looking closely at when and how bias appears, the research project will provide insights that developers, researchers, and users can apply to make AI systems more fair and reliable. In the end, the research project seeks to support the responsible use of AI by offering clear strategies for reducing bias and improving the quality of AI-assisted decision-making.

## **Proposed Solution**

The focus of our research paper is to identify at least three measurable biases in text generated by LLMs. We will design a research framework that uses tools and datasets to evaluate bias. One tool we may use is the Holistic Evaluation of Language Models (HELM) [2], which shows how biases appear in real-world settings.





Our framework has three parts:

- Review existing studies on AI bias to see what has been done and where gaps remain.
- Collect data using resources such as StereoSet on HuggingFace [3], which labels bias type, context, and sentence choice. And add other datasets or methods as needed.
- Analyze the data alongside the literature review to answer our research question.





Our research question will be finalized after our background research but will focus on identifying and categorizing measurable biases in LLM outputs. A measurable bias will be defined as a bias type that can be shown through dataset metrics or repeated examples in generated text. By the end of the research project, we will present at least three distinct types of bias. For each bias, we will provide supporting data, describe the context, and explain how it appears in AI outputs.

## Methodology





### Task tracker

 Assignee	 Title	Description	Expected Weeks	 Date to Complete By	 Status	Deliverables
Everyone	Research	<ul style="list-style-type: none"> <li>- Understand more about LLMs, what their purposes are, and how we can explain this to someone who knows nothing about it.</li> <li>- Comprehend the language that needs to be made clear in the paper, how we are defining biases, how we will be able to determine if we are not using offensive language, what type of biases we are studying, and the categories of biases</li> <li>- Look into previous research done on biases in LLM, what they found, what's something similar we will have, and what's something different</li> <li>- Ethical background, make sure we ourselves are not inputting any biases, and we use neutral terms when indicating our findings</li> </ul>	4	Oct 27, 2...	Not st... ▾	Be able to describe our research paper in full, on what we need to complete, and we can comprehend what we have learned, like what gaps exist between previous studies and what we want to study
Everyone	Tools and Coding	<ul style="list-style-type: none"> <li>- Look into sources of training data, HELM, and SteroSet on HuggingFace</li> </ul>	2	Nov 10, ...	Not st... ▾	Be able to have a list and plan for

## Task tracker

 Assignee	 Title	Description	Expected Weeks	 Date to Complete By	 Status	Deliverables
		<ul style="list-style-type: none"> <li>- How to utilize the data sets</li> <li>- Look into coding background needs for the project's use of Python</li> <li>- Tools used for organizing and collecting data that we will need to display</li> <li>- How will we display data and findings, and what tools are needed?</li> </ul>				all of the tools and coding background we will need for this research project, and start to become familiar with them
Everyone	Data Collection	<ul style="list-style-type: none"> <li>- Generate sample outputs from LLM and apply SteroSet and HELM to collect data on bias</li> <li>- Document the outputs thoroughly, and have a system for note-taking</li> <li>- Ensure documentation is organized and well thought out</li> </ul>	4	Dec 8, 2023	Not started	Collected datasets and organized tables of documented outputs and notes on findings
Everyone	Data Analysis	<ul style="list-style-type: none"> <li>- Analyze the results of our findings and identify our three concluding biases</li> <li>- Validate findings using the tools and information we collected</li> <li>- Create data visualizations, bar charts, graphs, etc.</li> </ul>	4 (Excluding Breaks)	Feb 2, 2024	Not started	Be able to report our findings and create data visualizations

## Task tracker

 Assignee	 Title	Description	Expected Weeks	 Date to Complete By	 Status	Deliverables
Everyone	Synthesis and Writing	<ul style="list-style-type: none"> <li>- Synthesize all of our findings from background research, data collection, data analysis, and tools resources</li> <li>- Bring everything together into a research paper</li> <li>- Have detailed explanations of every step completed</li> </ul>	3	Feb 23, 2...	Not st... ▾	Have a draft of our research paper with visuals
Everyone	Review and Finalize	<ul style="list-style-type: none"> <li>- Revise and edit the research paper draft into a final paper</li> <li>- Ensure there are reviews from outside people for full well-roundedness</li> </ul>	2	Mar 9, 2...	Not st... ▾	Final Research Paper

## Research Impact

Our research addresses the real-world consequences of AI bias. This is a critical issue in fields like healthcare. Biased medical AI can lead to poor clinical decisions [4]. It can also worsen existing healthcare disparities. Addressing and identifying these biases is crucial for ensuring medical AI treats all patients equitably. Similarly, crime-prediction algorithms can be biased [5]. They are more likely to incorrectly label Black defendants as high-risk. Conversely, algorithms sometimes incorrectly label white defendants as low-risk, even if

they later re-offend. The consequences of these errors are dire. These errors could wrongly imprison a person or lead to the death penalty. Meanwhile, the system might set a more dangerous individual free.

Two solutions help address these biases: Stanford University's HELM and StereoSet. HELM provides a benchmark for evaluating language models using various metrics. The StereoSet dataset measures stereotypical (hence the name) bias in models. These two solutions are useful in identifying and measuring the extent of bias within an AI model, but they are primarily diagnostic tools. They do not help developers actively correct or mitigate the biases they discover.

Our solution will create a replicable and consistent framework for producing and testing these biases in any LLM. We can then ask crucial questions about why these biases occur, how they reflect societal prejudices, and what factors in an AI's development lead to harmful outcomes. By answering these questions, our research will contribute to a deeper understanding of how LLMs function and where they fall short. These findings can help developers and researchers make more informed decisions to build responsible and equitable AI.



## Resources Needed

- **Python:** The standard programming language for AI/ML work. We will use Python for data analysis, interacting with AI models, and building our testing framework. We will use Python with an IDE such as VS Code, PyCharm, or Google Colab.
- **OpenAI API:** The most popular AI LLM currently available. For OpenAI's standard GPT-5 model, the API will cost \$1.25 for input and \$10.00 for output per 1 million tokens [6]. More affordable versions (GPT-5 Mini and GPT-5 Nano) are also available, if needed.
- **Gemini API:** Another popular AI LLM that is comparable to OpenAI in terms of popularity and intelligence. Free for use [7].
- **Llama API:** If OpenAI's API and Gemini's API are not suitable for this project, Llama API can be a viable backup. Free for use, although you need to join a waitlist for access [8].
- **HELM:** An open source Python framework for holistic evaluation of LLMs [2]. We can use this tool to benchmark various LLMs on fairness, bias, and accuracy. These metrics provide a baseline for the project's own findings.
- **StereoSet:** This dataset measures stereotypical bias in LLMs [3]. Evaluates biases related to gender, race, religion, and profession. This dataset will help this project's goal of identifying and analyzing bias.

## Project Related Resources

- [1] Philip Resnik. 2025. Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics* 51, 3 (Sep. 2025), 885–906.  
<https://direct.mit.edu/coli/article/51/3/885/128621/Large-Language-Models-Are-Biased-Because-They-Are>
- [2] Percy Liang, Rishi Bommasani, Tony Lee, et al. 2023. Holistic Evaluation of Language Models. Retrieved September 18, 2025 from  
<https://openreview.net/forum?id=iO4LZibEqW>.
- [3] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. Retrieved September 18, 2025 from  
<https://doi.org/10.48550/arXiv.2004.09456>.
- [4] James L Cross, Michael A Choma, and John A Onofrey. 2024. Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health* 3, 11, Article e0000651 (Nov. 2024). <https://doi.org/10.1371/journal.pdig.0000651>
- [5] Julia Angwin, Jeff Larson, Suyra Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica*. Retrieved September 18, 2025 from  
<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [6] OpenAI. 2025. OpenAI API. Retrieved September 18, 2025 from  
<https://openai.com/api/>.
- [7] Google. 2025. Google AI for Developers. Retrieved September 18, 2025 from  
<https://ai.google.dev/gemini-api/docs/api-key>.

[8] Meta. 2025. Llama API. Retrieved September 18, 2025 from <https://www.llama.com/products/llama-api/>.