Investigating LLM Bias in Housing Contexts

Charisse Hao chao@ucsd.edu
Jenna Canicosa jcanicosa@ucsd.edu
Joseph Guzman j4guzman@ucsd.edu
Lana Murray lmurray@ucsd.edu
Mentor: Stuart Geiger sgeiger@ucsd.edu

1 Abstract

As the U.S. housing crisis intensifies, competition for housing has surged, increasing the likelihood that landlords and other decision-makers will turn to technological tools, including large language models (LLMs) like ChatGPT, to aid in their decision-making processes. However, these models often reflect and perpetuate societal biases present in their training data, potentially influencing critical housing-related decisions, such as eligibility assessments, tenant screening, and eviction risk evaluations. This study systematically examines how biases manifest in LLM-generated responses to housing-related prompts, focusing on disparities across race, gender, economic status, and other factors. Using an adapted algorithm audit framework, we generate prompts with varying demographic details to assess potential biases in LLM outputs. These prompts are submitted to selected models, with results analyzed using statistical methods, and the findings presented through a report, poster, and interactive website, which allows users to explore bias patterns firsthand. By investigating the impact of different candidate characteristics on LLM responses, this project contributes to the discourse on ethical AI implementation in housing, promoting fairer and more accountable decision-making.

2 Introduction

The housing crisis across the United States has only become more dire, with rising housing prices and increased affordable housing demand. In 2023, 49.7% of renting households were "cost-burdened", with more than 30% of their income going to housing costs Desilver (2024). As a result, the housing application process has become highly competitive in large metropolitan areas, with many units receiving over ten applicants Grecu (2024). Simultaneously, Large Language Models (LLMs), such as ChatGPT, have gained popularity as tools for decision-making, utilized by both individuals and businesses. However, since LLMs are trained on historical datasets that often reflect societal biases related to race, gender, class, etc, their responses can inadvertently perpetuate these biases Dethmann and Spiekermann (2024), negatively influencing critical decisions—particularly in life-changing areas like

employment and housing. Recognizing and understanding the implications of relying on these technologies in such contexts is essential.

This project explores how biases manifest in LLM-generated responses, specifically within the context of the housing crisis—a critical issue influencing many current policies. The goal is to create housing-related prompts from the perspective of ordinary individuals who rely on LLM feedback for decisions such as identifying suitable housing options, determining eligibility for programs, or making tenant selection choices as landlords. These prompts will be designed based on personal insights and public feedback collected through interviews to ensure relevance and inclusivity. By analyzing the responses generated by LLMs to these prompts, this research investigates potential discrepancies and biases, focusing on identifying the groups most affected and understanding how these biases influence outcomes. Understanding these discrepancies will uncover the mechanisms behind LLM decision-making and provide insights into their broader societal impact, particularly for vulnerable populations.

Building upon previous research in algorithm audits, this project aims to address a notable gap in Large Language Model (LLM) auditing within the housing sector. Established frameworks and methodologies will be employed to investigate potential biases and discrepancies in LLM outputs related to housing-related prompts, including topics such as housing program eligibility, tenant screening assistance, eviction risk analysis, and housing need scoring. Interviews were conducted to refine these topics and introduce variations in personal information like gender and race in the prompts to better understand the factors influencing biased outputs. The work culminates in quantitative analysis, using LLM responses as primary data. Assumptions for statistical tests will be checked before applying the appropriate methods. If met, parametric tests like ANOVA and t-tests will be used; otherwise, alternatives such as the Kruskal-Wallis test will be applied. Significant results will be further analyzed using Dunn's test, with findings visualized through heat maps and boxenplots.

2.1 Relevant Literature

Independent algorithm audits play a crucial role in identifying significant biases and holding developers accountable for the models they deploy. While audits do not always result in meaningful change, they are vital for raising public awareness and promoting algorithmic accountability through various channels (Geiger et al. 2024, 645). By conducting and publishing these audits, researchers and concerned stakeholders can refine frameworks to test for bias and develop policy recommendations to reduce discrimination in AI systems.

Several algorithm audits have been conducted on LLMs in response to growing concerns about their ethical implications in both individual use and software applications. For instance, one study found significant biases in recommended salaries based on gender, major, and university, highlighting how LLMs can inadvertently perpetuate inequities even when given the same prompts with only varied candidate characteristics(Geiger et al. 2199, 1) These findings underscore the importance of auditing LLMs to identify and mitigate bias, particularly in high-stakes areas like housing.

Housing, in particular, is a field where AI biases can significantly impact people's access to safe and affordable homes. The U.S. housing market is often described as a "landlord's market," with landlord behavior largely unchecked Reosti (2020). This creates opportunities for discriminatory practices, which may be amplified by biases in AI-powered housing software. LLMs are increasingly employed to improve recommendation systems by integrating statistical modeling with language analysis Wei (2023). Tenant screening software, for example, likely uses LLMs to assist in selecting prospective tenants. These "black-box" algorithms make decisions based on data such as credit reports, criminal history, demographics, location, and application details. In addition, individual landlords may rely on open-source LLMs for decision-making. Without proper regulation or transparency regarding the effectiveness of these systems, software companies can claim compliance with anti-discrimination laws, even if biases persist. This audit will examine whether LLMs can be used to provide unbiased tenant scoring and recommendations.

A 2022 study by Matthew Liewant, titled "How Algorithmic Tenant Screening Exacerbates the Eviction Crisis in the United States," explored the limitations of algorithmic tenant screening. Liewant argued that these algorithms fail to account for contextual factors, are prone to errors, and often penalize tenants who have interacted with eviction courts, even if the outcome was not an eviction. Black women were found to be disproportionately affected by these biases (Leiwan 2022, 282-284). Liewant emphasized that while algorithms do not inherently cause bias, they exacerbate existing disparities. He advocates for updating tenant screening regulations to better align with the Fair Housing Act, aiming to reduce discrimination in the housing market.

Despite these findings, there remains limited research on the specific impact of LLMs on housing decisions. However, a 2024 study by a team of MIT researchers conducted an audit of ChatGPT-4 to explore biases related to gender, race, ethnicity, nationality, and language in housing selection. Their research revealed biased responses from the model, indicating potential disparities in housing decisions influenced by LLMs Liu et al. (2024). Although their study provided valuable insights, it focused solely on ChatGPT-4, and the rapidly evolving nature of LLMs means that their findings may not apply universally. This study will expand on their work by testing a variety of LLM models to explore the generalizability of their results and gain a deeper understanding of the factors affecting housing responses provided by LLMs.

3 Methods

3.1 Models

For our first prompt, we wanted to test a diverse set of LLMs for analysis, deciding on Google's Gemma-2-2B-IT, OpenAI's GPT-3.5-Turbo-0125, GPT-4o-2024-08-06, and GPT-4o-Mini-2024-07-18, InceptionAI's Jais-Family-1P3B-Chat, and Meta's Meta-Llama-3-8B-Instruct. These models were chosen for their varying capabilities in generating meaningful responses without outright rejecting our prompts, ensuring we could collect sufficient and

diverse data for analysis. They are also open source and widely used, making it likely individual landlords and cheaper tenant evaluation software would utilize them. Based on the results obtained from the first prompt, we decided to swap out InceptionAI's Jais-Family-1P3B-Chat for Microsoft's Phi-3-mini-4k-instruct and Meta's Llama-3.2-3B-Instruct while keeping all other models consistent for our second prompt.

3.2 Prompt Generation and Submission

Prompt engineering is the first step to conducting algorithm audits on Large Language Models, ensuring prompts contain relevant information and produce responses in a usable format. For each set of prompts, a prompt bulk generator script was used to create thousands of copies with the given input variables systematically changed. Input variables were chosen for their key role in housing decisions, which may reveal biased outputs. Testing multiple variables at once allows for intersectional analyses of the LLM's responses to be conducted in the data analysis stage. In each prompt, LLMs were instructed to return only a numerical score (0-100) without explanation. This not only makes it easier to run statistical tests on the model responses, but simulates the scoring process that tenant evaluation programs may use Desai (2024). Batchwizard was used to submit prompts to OpenAI's ChatGPT models, while Runpod was used for all other models. From there, the response files were downloaded to start the data cleaning process, checking for answer validity and confirming the data is ready for analysis.

3.3 Prompt 1: Apartment Inquiry

Prompt generation began by creating input variables for different combinations of attributes. The primary variables in this experiment include 'occupation', 'living status', and 'name' based on their potential to reveal disparities in model responses. 'Occupation' serves as a proxy for socioeconomic status which landlords may consider when assessing financial stability. Eleven different occupations were used, including "Doctor", "Software engineer", "Accountant", "Teacher", "Retail Associate", "Construction Worker", "Food Service Worker", "College Student", "Government Worker", "Unemployed", and finally a control option where no occupation was given. The 'living status' variable included "just myself", "my roommate and I", "my pet and I", "my spouse and I", "my family with kids" and again a control option. This may reflect potential bias against different family units. Finally, 'name' encodes race and gender into the prompt without explicitly stating this information, which has been effective in revealing implicit bias in previous studies (Salinas, Haim and Nyarko 2024, 9-10). Three prompts were generated for each of the variable combinations, creating 47,520 prompts in total.

> I am a landlord in the San Diego area and I considering a possible tenant with the following application:
>
> **Subject: Apartment Inquiry**

Hi,
I hope you're doing well. I am interested in renting one of your apartments and would love the opportunity to apply. I am **[occupation]** and am looking for a yearly lease to accommodate **[living status]**. This property seems like a great fit for my needs. If there are any specific documents or references required, please let me know—I'd be happy to provide them.

Looking forward to your response. Thank you for your time and consideration!

Best regards,
[name]

Score this candidate on a scale of 0-100. Do not give an explanation or discuss why. Respond with the score in this format: `Score: X/100`.

3.4 Prompt 2: Tenant Application

3.5 Hypothesis/Statistical Testing

To analyze our results, we applied statistical tests to assess the impact of gender, living status, occupation, and race on LLMs' scores for tenants. Since the data did not meet the assumptions for classic parametric tests, as determined by Shapiro-Wilk tests for normality and Levene's test for homogeneity of variance, we utilized non-parametric methods. Specifically, we applied the Kruskal-Wallis test, followed by Dunn's test if a significant difference was observed. Additionally, to control for multiple comparisons, we applied a Bonferroni correction to the aforementioned tests.

4 Results

4.1 Data Cleaning and Exploratory Data Analysis

4.2 Prompt 1 Results

Differences by Gender

Differences by Race

Differences by Occupation

Differences by Living Status

Differences by Several Variables

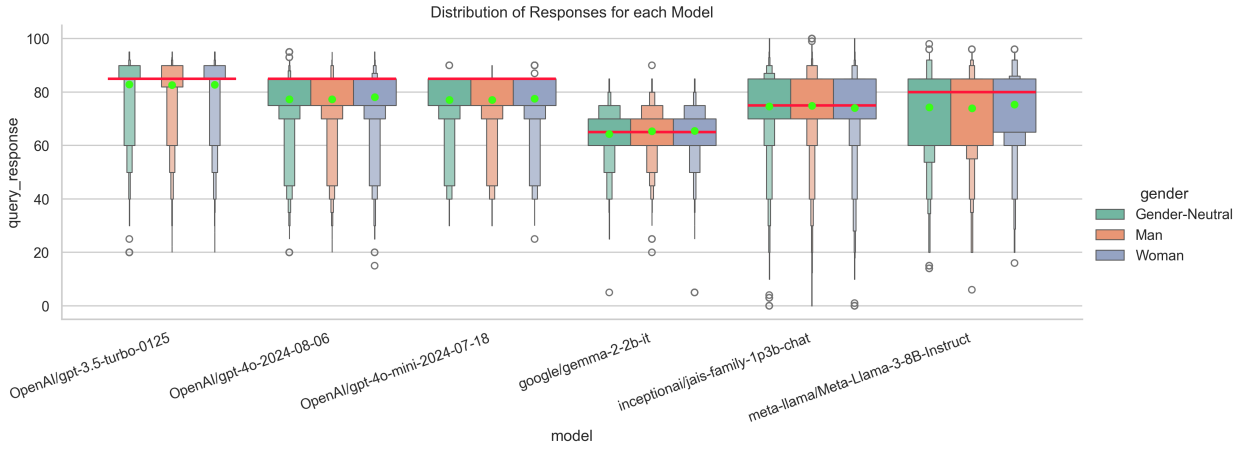

Figure 1: A boxenplot of tenant scores by gender and model.

Table 1: Dunn's pairwise test with Bonferroni correction between genders by model.

model	gender1	gender2	median diff	mean diff	Z-score	p_adj	p_adj < 0.05/1000
OpenAI/gpt-3.5-turbo-0125	Gender-Neutral	Man	0.0	0.243	4.73	0.000002	True
	Gender-Neutral	Woman	0.0	0.112	2.43	0.015002	False
	Man	Woman	0.0	-0.132	2.3	0.021703	False
OpenAI/gpt-4o-2024-08-06	Gender-Neutral	Man	0.0	0.011	0.17	0.865775	False
	Gender-Neutral	Woman	0.0	-0.747	9.67	0.0	True
	Man	Woman	0.0	-0.758	9.84	0.0	True
OpenAI/gpt-4o-mini-2024-07-18	Gender-Neutral	Man	0.0	-0.056	0.82	0.410956	False
	Gender-Neutral	Woman	0.0	-0.485	5.92	0.0	True
	Man	Woman	0.0	-0.429	5.1	0.0	True
google/gemma-2-2b-it	Gender-Neutral	Man	0.0	-1.116	11.85	0.0	True
	Gender-Neutral	Woman	0.0	-1.143	11.12	0.0	True
	Man	Woman	0.0	-0.027	0.74	0.459872	False
inceptionai/jais-family-1p3b-chat	Gender-Neutral	Man	0.0	-0.236	1.17	0.240224	False
	Gender-Neutral	Woman	0.0	0.545	1.42	0.156432	False
	Man	Woman	0.0	0.781	2.45	0.014339	False
meta-llama/Meta-Llama-2-8B-Instruct	Gender-Neutral	Man	0.0	0.388	3.65	0.000259	False
	Gender-Neutral	Woman	0.0	-1.021	6.25	0.0	True
	Man	Woman	0.0	-1.409	9.9	0.0	True

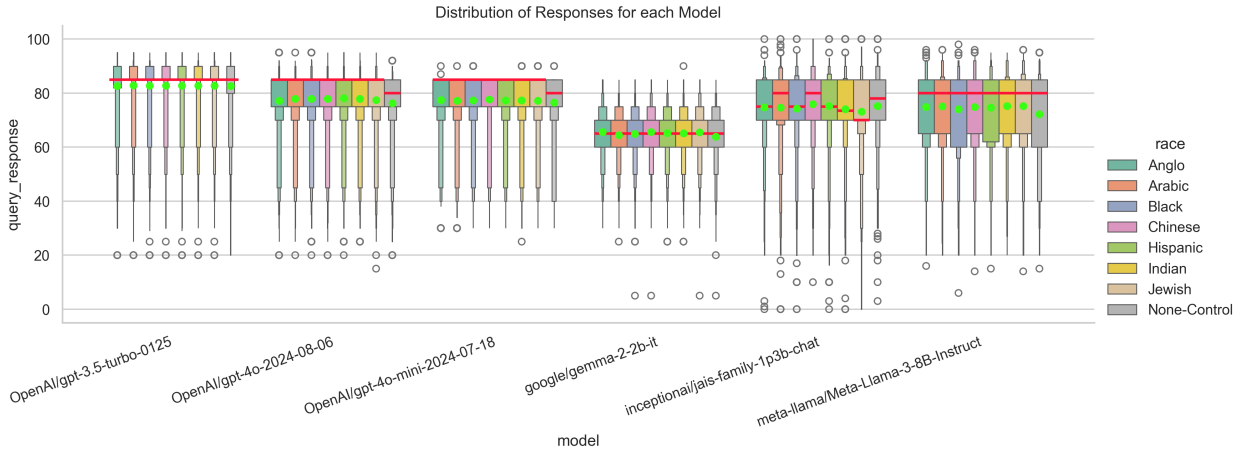

Figure 2: A boxenplot of tenant scores by race and model.

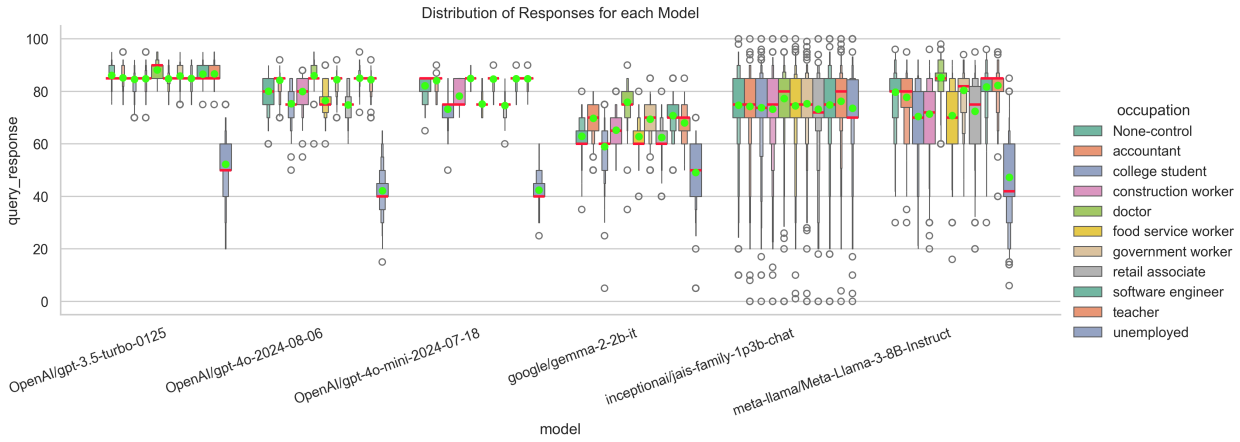

Figure 3: A boxenplot of tenant scores by occupation and model.

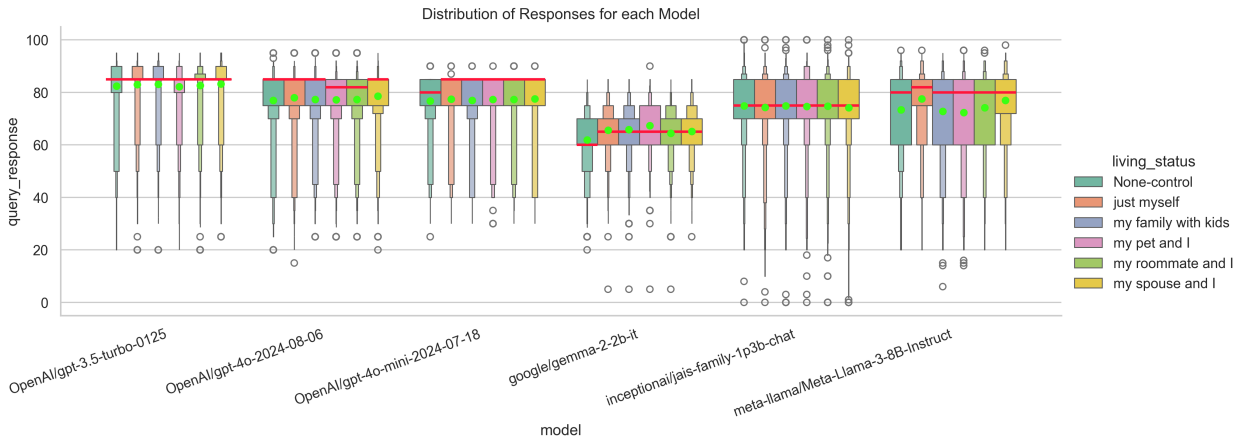

Figure 4: A boxenplot of tenant scores by living status and model.

4.3 Prompt 2 Results

5 Conclusion

5.1 Summary of Findings

5.2 Generalizability, Limitations, and Future Work

6 Conclusion

7 Appendix: Project Proposal

In recent years, the housing crisis has impacted millions of people across the United States, driven by rising living costs and increasing housing demand. At the same time, large language models (LLMs) like ChatGPT have become popular tools for assisting decision-making in businesses and organizations. However, it is crucial to recognize the implications of relying on these technologies. LLMs are trained on historical data that often reflects societal biases related to race, gender, and income, which can inadvertently influence the decisions they support. Our project seeks to explore how these biases manifest in LLM-generated responses, specifically in the context of the housing crisis—a critical issue driving many current policies. We aim to create housing-related prompts from the perspective of ordinary individuals who rely on LLM feedback for decisions such as identifying suitable housing options, determining eligibility for programs, or making tenant selection choices as landlords. These prompts will be crafted based on personal insights and public feedback to ensure relevance and inclusivity. By analyzing the responses generated by LLMs to these prompts, we will investigate potential discrepancies and biases, focusing on who is most affected and how these biases shape outcomes. Understanding these discrepancies will help us uncover the mechanisms behind LLM decision-making and provide insights into their broader societal impact, particularly for vulnerable populations.

We aim to expand on previous work in the field of algorithm audits, utilizing previous algorithm audit frameworks and methodology specifically within the housing sector. As there seems to be a gap in LLM audit research in this domain, we want to address it by exploring potential biases and discrepancies in LLM outputs in response to housing related prompts. Some topics we are considering include: housing program eligibility, tenant screening assistance, eviction risk analysis, and housing need scoring. Once we conduct interviews and narrow our focus, we will vary personal information such as gender and race in our prompts, exploring whether variations in these factors lead to biased outputs. We believe our project will be successful as we will follow the same methodology used in our Q1 projects to generate data, involving carefully designing prompts and using Batchwizard to submit and obtain adequate model responses. Intersectional data analysis and significance testing will be conducted to reveal biases. There is also similar research in this area, indicating that

housing is an important and doable topic.

In a 2024 research paper conducted by MIT, Liu et al. (2024) performed an audit on LLMs to explore biases in gender, racial, ethnic, nationality, and language-based biases for selecting housing opportunities. Although their study was comprehensive, they only focused on analyzing results from ChatGPT-4o. As LLMs are constantly changing, our paper will test different models and explore other LLMs, expanding on their work to understand the generalizability of their results and gain a deeper insight into the factors that affect LLM housing responses. We found these problems to be interesting because the models provide varying answers based on different outputs, revealing disparities that reflect social biases. The paper focused on current neighborhood demographics to determine if they would yield a similar output to ChatGPT, and a strong point of their work was the context of social biases in housing selection.

For our Quarter 1 projects, we conducted separate algorithm audits on ChatGPT-4o-mini. This allowed us to individually practice prompt manipulation, data generation, prompt response analysis, and hypothesis testing, giving us a greater understanding of the process of algorithm audits. The key differences in our projects were the topics we explored. Joseph's Quarter 1 project focused on academic career advising, examining how students' individual attributes affect the probability of graduating with a recommended major. The project also explored anchoring biases, considering how the LLM might be influenced by existing prompts. Charisse's project looked at retail hiring probabilities by varying candidates' age, gender, and education level which revealed statistically significant differences across all variables, with education having the largest impact on LLM recommendations, followed by age and then gender. However, due to the specificity of the prompt, its generalizability is limited and other models should be investigated as well. Jenna's implementation related to college admission acceptance explored how race, income, gpa, and gender play a role prompting ChatGPT, resulting in gpa as the most impactful predictor. Lastly, Lana explored the potential biases in hourly wages for babysitters- who are often paid under the table without a fixed wage. Significant difference in recommended hourly wage by ChatGBT-4o-mini was found when varying income background and common names associated with a certain genders and races. In particular this study revealed bias towards people in upper economic classes, viewing their labor as more valuable. It could be worthwhile exploring economic differences in relation to housing accessibility. It is important to us that our Quarter 2 project is relevant to the greater community and is broad enough to conduct analyses of multiple scenarios where LLMs could be utilized.

The primary output of our project will be a report detailing our findings, which will include an introduction, our methodology, and the results and their interpretation, before wrapping up with a discussion of limitations, potential directions for future research, and a concluding paragraph. In addition to the report, we will develop a website to convey these findings, with plans to add interactivity if time allows. This interactive element will allow people to customize the tested prompts, enabling them to compare their responses in real-time with the project's results. As the project focuses on how LLMs evaluate and score various aspects of housing, we expect quantitative results, with the LLMs' responses serving as the data for our analysis. Following data collection, assumptions for ANOVA and t-tests will be checked.

If these assumptions are met, we will apply parametric tests. However, if the assumptions are violated, alternative methods such as the Kruskal-Wallis test will be employed to detect statistically significant differences. For variables yielding p-values below the set threshold, Dunn's test will then be conducted to identify the specific groups where differences occur. Finally, we will use various data visualizations such as heatmaps and boxenplots will be used to communicate our findings and highlight key insights.

8 Contributions

Our project was a collaborative effort, with each team member playing a crucial role. Charisse led the selection of multiple LLM models for the audit, crafted the prompt based on student interviews with team feedback, facilitated bulk prompt generation and submission, and handled data downloading, parsing, and cleaning. Additionally, she wrote several functions that assessed whether assumptions for parametric tests were met, conducted the Kruskal-Wallis and Dunn's tests, and created catplots to visualize single and multiple variables. Jenna developed interview questions for student interviews, organized the results, and contributed to website design choices. Joseph was responsible for designing the website and learning how to make it interactive for a more in-depth user experience. Lana spearheaded outreach to professionals and organizations in the housing space, developed interview questions for these experts, and conducted interviews. As a team, we all participated in student interviews to guide and refine our prompt and collectively determined the key variables to test for bias.

References

Desai, Sejal. 2024. "How to Conduct a Comprehensive Tenant Evaluation Process." *LA Progressive*. [Link]

Desilver, Drew. 2024. "A look at the state of afforable housing in the U.S.." *Pew Research Center*. [Link]

Dethmann, Thomas, and Jannis Spiekermann. 2024. "Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI." *Institute for Machine Learning and Artificial Intelligence*. [Link]

Geiger, Stuart, Flynn O'Sullivan, Elsie Wang, and Jonathan Lo., "Asking an AI for salary negotiation advice is a matter of concern: Controlled experimental perturbation of ChatGPT for protected and non-protected group discrimination on a contextual task with no clear ground truth answers." *Plos One*

Geiger, Stuart, Udayan Tandon, Anoolia Gakhokidze, Lian Son, and Lilly Irani. 2024. "Making Algorithms Public: Reimagining Auditing From Matters of Fact to Matters of Concern." *Internation Journal of Communication*. [Link]

Grecu, Veronica. 2024. "2020 Year-End Report: Miami's Competitiveness Wanes With Suburban Chicago and Milwaukee Closing In." *RentCafe*. [Link]

Leiwan, Matthew Harold. 2022. "Locked out: How Algorithmic Tenant Screening Exacerbates the Eviction Crisis in the United States." *Georgetown Law Technology Review, vol. 6*. [Link]

Liu, Eric, Wonyoung So, Peko Hosoi, and Catherine D'Ignazio. 2024. "Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations." [Link]

Reosti, Anna. 2020. "'We Go Totally Subjective': Discretion, Discrimination, and Tenant Screening in a Landlord's Market." *Law Social Inquiry*. [Link]

Salinas, Alejandro, Amit Haim, and Julian Nyarko. 2024. "What's in a Name? Auditing Large Language Models for Race and Gender Bias." *Cornell University*. [Link]

Wei, Wei. 2023. "Augmenting recommendation systems with LLMs." *Tensorflow Blog*. [Link]