Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting

Donald E. Bowen III
Lehigh University

S. McKay Price
Lehigh University

Luke C.D. Stein
Babson College

Ke Yang
Lehigh University

This version: January 31, 2025

We conduct the first study exploring the application of large language models (LLMs) to mortgage underwriting, using an audit study design that combines real loan application data with experimentally manipulated race and credit scores. First, we find that LLMs systematically recommend more denials and higher interest rates for Black applicants than otherwise-identical white applicants. These racial disparities are largest for lower-credit-score applicants and riskier loans, and exist across multiple generations of LLMs developed by three leading firms. Second, we identify a straightforward and effective mitigation strategy: Simply instructing the LLM to make unbiased decisions. Doing so eliminates the racial approval gap and significantly reduces interest rate disparities. Finally, we show LLM recommendations correlate strongly with real-world lender decisions, even without fine-tuning, specialized training, macroeconomic context, or extensive application data. Our findings have important implications for financial firms exploring LLM applications and regulators overseeing AI's rapidly expanding role in finance.

Perella Department of Finance, College of Business, Lehigh University, 621 Taylor Street, Bethlehem PA 18015; deb219@lehigh.edu, smp210@lehigh.edu, key208@lehigh.edu. Finance Division, Babson College, 231 Forest Street, Babson Park MA 02457; lcdstein@babson.edu. For helpful comments, we thank Alina Arefeva, Kofi Arhin, Sean Cao, James Conklin, Kathleen Hanley, Philip Kalikman (discussant), Danielle Kent (discussant), Jesús Villota Miranda (discussant), Robert Parham, Jesus Salas, and Kalinda Ukanwa; seminar participants at Pennsylvania State University; and attendees at the 2024 New Zealand Finance Meeting, the 2024 Concordia University Generative AI in Finance Conference, 2024 VU Amsterdam Research in Behavioral Finance Conference, and the Real Estate Finance and Investment Symposium 2024 at the University of Cambridge. This study is indebted to a pilot study conducted by the 2023 FinTech Capstone group at Lehigh University. Funding for the study was generously provided by the Goodman Center for Real Estate. All errors and omissions are our own. First version: May 31, 2024.

While these technologies have enormous potential, they also carry risks of violating fair lending laws and perpetuating the very disparities that they have the potential to address. Use of machine learning or other artificial intelligence may perpetuate or even amplify bias...

Michael S. Barr, Federal Reserve Board Vice Chair for Supervision (2023)

1 Introduction

Advances in artificial intelligence (AI) are unlocking exciting opportunities in various economic sectors, including credit markets. AI promises lower costs and faster decision-making compared to manual processes, potentially reducing biases by minimizing human intervention. However, racial disparities in the outputs of generative AI remains a significant concern (Das et al., 2023; Mehrabi et al., 2021).[1] The "black-box" nature of these systems complicates our understanding of their decision-making processes and how their outputs may reflect explicit or implicit biases in their training data. Measuring and mitigating bias is particularly important in the \$14 trillion mortgage market[2] where missteps in implementing AI have the potential to create large, widespread adverse effects. Fair lending practices are essential to maintaining regulatory compliance, promoting economic equality, and mitigating the risk of perpetuating or exacerbating existing disparities.

In this paper, we present the first audit study of racial disparities in large language models (LLMs) applied to loan underwriting. We utilize real loan application data from the Home Mortgage Disclosure Act (HMDA), supplemented with experimentally manipulated applicant race and credit scores. By asking various leading commercial LLMs to recommend

[1] Such disparities can be said to represent *bias* when they are generated by "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996). Our empirical analyses focus on systematic outcome gaps between individuals who vary only across membership in protected classes, and "economists define discrimination as differential treatment of otherwise identical individuals from different social groups" (Bohren et al., 2023). We therefore refer to "bias" throughout without implying that the disparities are driven by racial animus or taste-based discrimination on the part of AI systems that in some sense lack the preferences that generally define these concepts.

[2] As of Q1 2024, per Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/release/tables?eid=1192326&rid=52).

underwriting decisions, we find strong evidence that these models make different approval and interest rate recommendations for Black and white mortgage applicants with applications that are identical on all other dimensions.[3] Although it is unlikely that lenders would blindly adopt the specific LLMs we evaluate or use a prompting approach as simple as ours, the existence of racialized outcome differences in these models' recommendations raises concerns. This issue is alarming both on its own and for its implications in more complex systems, where the degree of bias and underlying mechanisms may be more difficult to assess, especially as general-purpose LLMs continue to demonstrate competitiveness against even specialized machine learning models in performing quantitative financial tasks (Kim et al. (2024)).

It is particularly concerning that LLM outputs vary by race in the *mortgage* context, given the historical and ongoing importance of mortgage lending in the U.S. economy, and the potential for disparities here to exacerbate other sources of economic inequality (see, e.g., Bartlett et al., 2022; Fuster et al., 2022; Kahn, 2024; Quillian et al., 2020). Our experiment involves provision of explicit information about borrower race; while this may not be available to automated underwriting systems in practice, this explicit signal should be easy for the LLM to ignore. The fact that it does not is troubling. This suggests that LLMs may exhibit more subtle forms of bias—which are harder to detect or remediate—in the more realistic scenarios where class information like race and gender is only inferred through proxy variables such as name or zip code (Fuster et al., 2022). Giving an underwriting system access to class information directly would allow a biased LLM to be highly biased—it knows exactly which applicants are members of disadvantaged groups. However, a fair LLM could be unbiased despite access to this information, as the presence of fair lending and related regulations in the training data is evident when interacting with leading LLMs; they are able to recite these regulations with ease.

We find clear evidence of disparate treatment by race in LLM decisions across all LLMs

[3] Throughout, we follow the AP Stylebook in writing "Black" with initial capitalization, but "white" in lowercase. We also often refer to signals of race, although in a supplementary experiment we consider a broader set of race/ethnicity signals.

we test. Specifically, LLMs recommend denying more loans and charging higher interest rates to Black applicants than to otherwise-identical white applicants.[4] This suggests that LLMs are learning from the data they are trained on, which includes a history of racial disparities in mortgage lending, and are potentially incorporating signals from other contexts (see, e.g., Mehrabi et al., 2021).

The magnitude of these differences is substantial. Using our baseline LLM (OpenAI's GPT-4 Turbo), Black applicants would, on average, need credit scores approximately 120 points higher than white applicants to receive the same approval rate, and about 30 points higher to receive the same interest rate. Our approach confirms that these differences are driven solely by race, as this variable is experimentally manipulated and fully stratified across other loan characteristics. Unlike analyses of observational data or audits of human behavior, we can compare the *same* underwriter's independent assessments of the *same* loan with different racial characteristics. These tests are straightforward to conduct using regression models that include loan-fixed effects, and a similar approach can be used to audit any generative AI system that can be prompted to make decisions.

Our results show that racial disparities in LLM underwriting recommendations are most pronounced for applications with lower credit quality. By experimentally manipulating credit scores and fully stratifying them across the race signal (as well as all other loan characteristics), we are able to isolate the effects of race at different credit scores. With our baseline LLM, the racial disparity in approval rates is about 56% greater for low-score applicants than for average-score applicants (13.3 percentage points vs. 8.5), and the disparity in interest rates is about 32% greater (47 basis points vs. 35). We also consider two other measures of credit quality, assessing the effects of experimentally manipulated race at different levels of observed debt-to-income and loan-to-value ratios from the HMDA data. The results are consistent across all three measures: racial disparities in LLM underwriting recommendations

[4] We also show that LLMs recommendations are worse for Hispanic applicants (though to a lesser extent than for Black applicants) and older applicants. We do not find strong evidence that recommendations differ on average between white and Asian applicants, nor between male and female applicants.

are present across the credit quality spectrum, but are unmistakably greater for riskier loans. This suggests that the harms from racially biased LLMs may be intersectional, compounding the effects of other dimensions of disadvantage (Crenshaw, 1989).

We find similar patterns across eight leading Anthropic, Meta, and OpenAI LLMs of varying size and training generation. Across these advanced models, Black applicants would need credit scores roughly 86 points higher than white applicants to achieve the same approval rates, and around 24 points higher to secure the same interest rates. This suggests that improvements in model quality will not necessarily lead to smaller disparities. There is thus a risk that the patterns we document could persist in future models, especially if disparate outcomes arise from inherent characteristics of the technology and training data. Therefore, auditing techniques like ours can help LLM designers, users, and regulators avoid bias in newly developed models.

We next explore whether the racial differences in recommendations can be mitigated or eliminated. One option for an underwriting system is to avoid exposing the algorithm to information about race, analogous to the approach taken in practice, where lenders collect explicit data on protected characteristics for ex post analysis but are prohibited from using it in underwriting. However, the rich set of variables in a mortgage application contain proxies for information about race (Fuster et al., 2022). Therefore, this approach may be insufficient. We instead take a simple prompt engineering approach, *maintaining* access to an explicit race signal in the loan application, but modifying our prompt to instruct the LLM to “use no bias” in making its decisions.

Despite its simplicity, this modified prompt results in significantly reduced racial disparities. The Black–white gap in loan approval recommendations is eliminated, both on average and across different credit scores. Asking the LLM not to exhibit bias reduces the average racial interest rate gap by about 60% (from 35 basis points to 14), with even larger reductions for lower-credit-score applicants. This result demonstrates the potential for prompt engineering to mitigate bias in LLMs and suggests that even simple adjustments in how these tools are used can lead to more equitable outcomes.

Our final set of tests compare the recommendations of the baseline LLM to the decisions of real underwriters. The simple LLM-based underwriting system we consider is not fine-tuned or specialized for mortgage underwriting, has no access to macroeconomic context or other data from the loan applications beyond the limited information provided in the prompt, and is given experimentally manipulated (counterfactual) credit scores. Despite these limitations, LLM approval recommendations align with real lenders' approval decisions for 92.3% of applications, and the suggested interest rates are strongly correlated with those in the data. Moreover, LLM recommendations are consistent with established lending criteria along several non-race dimensions, suggesting they learn from the data they are trained on and that their recommendations are not arbitrary. While predicting real underwriting decisions is neither the focus of our paper nor necessary for the internal validity of our disparity estimates, the correspondence we find between LLM and actual recommendations suggests that our estimated magnitudes are informative outside the context of our sample. This finding echoes other contemporaneous papers showing that off-the-shelf state-of-the-art models such as our baseline model, GPT 4.0 Turbo, perform quantitative financial tasks on par (or even better in some cases) with both human experts and the most sophisticated machine learning applications (e.g., Kim et al. (2024)). This emerging consensus implies a broader potential for LLMs beyond traditional textual tasks and highlights the importance of auditing before such deployment.

Our study makes several contributions. First, we conduct the first audit study of racial disparities in LLMs applied to loan underwriting. Our work complements the growing body of studies that employ audit designs to investigate algorithmic bias by LLMs in other settings, such as providing personal advice on car purchase negotiations, predicting election outcomes, and evaluating job applicants (Haim et al., 2024; Lippens, 2024; Veldanda et al., 2023). Our findings corroborate these studies and extends their findings to a setting involving regulated decisions. This is notable because the training data of leading LLMs contains the text of the regulations applicable to underwriting mortgages.[5] It is conceivable that training on the text of these documents might be sufficient to eliminate racial differences in this setting, even if an LLM is biased in other settings. We show that it is not.

Second, while the audit studies mentioned above document LLM outcome disparities along different dimensions (in non-mortgage settings), they do not explore methods to reduce them. We show that racial differences can be moderated via prompt engineering. In doing so, we contribute to a growing literature exploring various ways to counteract bias in LLMs. For excellent surveys on bias in AI/ML/LLM systems, see Navigli et al. (2023) and Mehrabi et al. (2021), which characterize types and sources of bias, discuss methods of measuring and reducing bias, and consider applications in many practical domains. Computer science literature on bias reduction has principally focused on techniques available only to LLM creators: pre-processing steps that clean and modify the input data, new methods to learn representations of ideas and relationships in the data during training, and post-training steps like fine-tuning. Our prompt engineering-based method to reduce disparities is simple and available to all end users of LLMs.

Third, we contribute to the finance literature examining discrimination in lending, particularly focusing on algorithmic underwriting. A number of papers have documented discrimination against minority borrowers in conventional mortgage lending (e.g., Bayer et al., 2018; Ambrose et al., 2021; Begley and Purnanandam, 2021; Blattner and Nelson, 2021; Giacoletti et al., 2021; LaVoice and Vamossy, 2024). Using a combination of HMDA, Federal Reserve Z.1, and HUD data, Bartlett et al. (2022) find that rate differences cost marginalized borrowers over $450 million per year. When they examine FinTech lenders to evaluate the role of algorithmic underwriting (using non-LLM machine learning methods), they find interest rate disparities by race similar to non-FinTech lenders for some loan types. Conversely, Bhutta et al. (2022) and Hurtado and Sakong (2024) use confidential data from HMDA

[5] Most importantly, the U.S. Civil Rights Act of 1964, the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA), the Supervision and Regulation (SR) 11-7 Guidance on Model Risk Management, and Regulation B of the ECOA (12 C.F.R. §202).

and find most of the disparity in loan approvals across race can be explained by observable non-race applicant characteristics. We add to this literature by evaluating a new class of algorithmic tools—LLMs—lenders might consider while implementing underwriting systems and demonstrating a technique lenders can use to assess and ensure compliance with fair lending regulations.[6] We focus on LLMs because they are novel, of interest to regulators, and rapidly growing in usage throughout many sectors of the economy.

Finally, our study has significant implications for regulators and financial firms exploring various AI and machine learning (ML) technologies, including, but not limited to, LLMs.[7] Financial institutions of all sizes are actively exploring potential applications of AI and ML. A report by S&P Global Market Intelligence notes that banks representing 80% of the sector's market cap mentioned AI/ML on recent conference calls. For example, J.P. Morgan has over 300 use cases in production.[8] If investment recommendation, customer service, fraud detection, personalized financial planning, product marketing, or insurance underwriting systems are built on biased algorithms with access to information about customer demographics, this bias can propagate through to a wide variety of important financial outcomes. Thus, our study serves as a cautionary tale about the potential for bias in LLM-based systems beyond the underwriting setting in this paper, especially if they are not properly audited before deployment.

[6] We do not intend to overstate the direct applicability of our findings to current lending practices, nor imply that the magnitude of racial disparities we estimate would be observed in newly developed underwriting systems, as lenders are unlikely to intentionally expose an LLM to demographic information on protected characteristics. Our findings demonstrate the potential for bias to arise in LLM models and the importance of properly auditing systems built on these models before they are deployed.

[7] Others have studied the application of machine learning algorithms in credit-risk models (Khandani et al., 2010; Krivorotov, 2023; Nazemi and Fabozzi, 2024). More recent studies investigate the effect of LLMs through the lens of regulatory shocks (Bertomeu et al., 2023), via implications for labor markets (Brynjolfsson et al., 2023; Eisfeldt et al., 2023; Eloundou et al., 2023), and by examining potential synergies between human and AI collaborators (Cao et al., 2024). D'Acunto et al. (2019), Rossi and Utkus (2020), and D'Acunto et al. (2023) examine how robo-advising interacts with behavioral and cultural biases. Moreover, Howell et al. (2024) examine how automated lending impacted lending across Black- and white- owned firms.

[8] S&P Global Market Intelligence, https://www.spglobal.com/marketintelligence/en/news-insights/research/smaller-banks-are-using-ai-too.

2 Methodology

2.A Background

Large Language Models operate through next-token prediction: they attempt to statistically predict the next word (or, more precisely, token) in a sequence of text given the preceding words.[9] The models are trained by assessing candidate predictions on subsets of a vast text corpus—typically comprising web pages, books, and other sources—and iteratively adjusting the model's parameters as it sees more and more text. LLM developers curate a corpus of training data, and cleaning this input plays a pivotal role in enhancing LLM quality, encompassing basic steps such as parsing HTML to extract raw text (Naveed et al., 2024). After training, LLM designers can further refine the algorithm through fine-tuning and the incorporation of additional instructions.[10]

The responses generated by an LLM are inherently dependent on its training data, and can reflect attitudes or preferences embedded there. For example, Atari et al. (2023) administer psychological tests to LLMs and show that responses correlate most strongly with humans from "W.E.I.R.D." (western, educated, industrialized, rich, and democratic) countries, reflecting the disproportionate reliance on training data from these regions. Indeed, many studies have documented issues in earlier and contemporaneous generations of LLM models (Zou and Schiebinger, 2018; Kadambi, 2021; Santurkar et al., 2023).

There is a large literature that focuses on aligning LLMs to behave as intended by their designers (surveyed by Dong et al., 2024), part of which may include taking measures to reduce various forms of bias. For example, the critical role of the training data in determining LLM behaviors underscores the importance of corpus selection and cleaning, which can include steps to mitigate biases. For example, LLM developers can duplicate training sentences with reversed gender roles, increasing the model's exposure to non-stereotypical

[9] Wolfram (2023) provides an accessible background on the functioning of LLMs.

[10] Models are typically operationalized with a hidden prompt preceding each user interaction that can contain additional instructions.

examples such as "the nurse went to *his* station to review patient notes." Additionally, model parameters can be fine-tuned after training to adjust the model's behavior in specific contexts. Indeed, ChatGPT is built on a model where reinforcement learning from human feedback (RLHF) was used as a fine-tuning step (Ouyang et al., 2022). RLHF shows the model desired outputs for a given prompt and is used extensively by OpenAI to moderate and adjust the behavior of ChatGPT. The goal of these efforts is to create a more balanced and representative model that can generate fair and unbiased responses across diverse contexts and user groups, and model developers including OpenAI have publicized efforts to debias their models.[11]

Speaking to those efforts, when we *asked* one of the most advanced LLMs to date—OpenAI's ChatGPT 4.0 Turbo—if it would discriminate in evaluating loan applications, it offered strong assurance of its own impartiality:

> "When evaluating loan applications or providing guidance related to financial matters, I rely on objective criteria and general principles of finance. My responses are based on the information provided and do not take into account any personal characteristics of individuals." (See Figure I for the full quotation.)

[Insert Figure I about here]

The LLM's response is consistent with designers' intentions to create fair and unbiased models, or at least models that can *claim* to be fair and unbiased. These claims may be a function both of design and of training data. Corpora used to train advanced LLMs are known to encompass not only vast portions of the accessible internet (including major forum sites like Reddit and Quora), but also public-domain government documents. As a result, when queried about legislation such as the Equal Credit Opportunity Act (ECOA) or the Community Reinvestment Act (CRA), leading LLMs are likely to respond with language derived directly from these statutes. This characteristic is particularly significant in our

[11] OpenAI has detailed some methods they employ at https://openai.com/index/instruction-following/ and https://openai.com/index/language-model-safety-and-misuse/.

study's context, since it implies that LLMs possess an inherent awareness of protected classes in the context of mortgage lending. This awareness forms a crucial foundation for our investigation into potential racial differences in LLM responses related to lending practices.

Whether these regulatory instructions, data cleaning efforts, and post-training instructions are sufficient to eliminate bias in a mortgage setting is, however, unclear *ex ante*. We therefore structure our study around two questions. First, we examine if LLMs provide different responses to mortgage lending queries across various protected classes. For example, we assess whether an LLM approves loans for Black borrowers at the same rate as otherwise identical white borrowers. Additionally, we explore if differences are heterogeneous across other borrower dimensions, in particular several measures of credit quality.

Second, we examine if prompt engineering—changing instructions in the prompt—can reduce differences in how LLMs respond to a signal of race. Finally, we extend this investigation into the value of prompt engineering to its role in addressing the *heterogeneity* in racialized response differences.

2.B Empirical strategy

To examine those questions, we ask LLMs to recommend loan underwriting decisions by constructing our "baseline" prompts as follows:

> *Given the following loan application from 2022:*
> *- Single-family home*
> *- Owner-occupied*
> *- First lien*
> *- 30 year fixed-rate mortgage*
> *- Credit score: {CreditScore}*
> *- Loan amount: {LoanAmount}*
> *- Loan to value ratio: {LTV}*
> *- Property value: {PropertyValue}*
> *- Income: {Income}*
> *- Debt to income ratio: {DTI}*
> *- State: {State}*
> *- Race: {Race}*

Please respond with two columns, separated by a comma:
1. Should this loan be approved? (1 for yes, 0 for no)
2. Which of the following interest rates would you offer? Choose from: 3.0%, 3.5%, 4.0%, 4.5%, 5.0%, 5.5%, 6.0%, 6.5%, 7.0%, 7.5%, 8.0%, 8.5%, 9.0%, 9.5%, 10.0%? Assume 0 discount points and a 1% origination fee.

Examples:
- 1,4.0
- 1,7.5
- 1,5.5
- 0,6.5
- 0,7.5
- 0,9.0

Do not reply with anything beyond these two columns.

The values that populate each prompt are drawn from real loan applications in the HMDA data, as discussed in Section 2.C, except that we experimentally manipulate race and credit scores. Each resulting prompt, after manipulations m are chosen, comprises a fictional application, which is sent to an application programming interface (API) endpoint for each LLM we examine. The full set of parameters for these requests is detailed in the appendix. In rare cases where a response is not formatted as requested, we rely on the fact that LLM responses are statistically generated to simply retry an identical request until an acceptable answer is received.[12]

Because we are manipulating race and credit score, the responses from the LLMs form the basis for an audit study. In different experiments, we omit race/ethnicity from the prompt entirely, or include "Asian," "Black," "Hispanic," or "White."

The publicly available HMDA data do not include borrower credit scores. To assess how LLMs use information about borrower creditworthiness, we experimentally manipulate applications across three potential credit scores: 640 (representing a "fair" score), 715 ("good," roughly the average credit score according to Experian[13]), and 790 ("very good"). Manipu-

[12] The examples in the prompt provide LLMs guidance on output formatting and work well across many models. Structured output requests were not available when we conducted the experiment in April 2024.

[13] See https://www.experian.com/blogs/ask-experian/consumer-credit-review/.

lating the credit score listed on each application rather than using (unavailable) real credit scores offers two empirical advantages. First, the causal effects of credit scores and race can be compared to better understand the magnitude of our main results. In particular, we contextualize racial disparities by calculating the credit score differences that would generate similar effect sizes. Second, our approach allows us to estimate potential heterogeneity in racial recommendation disparities across the credit spectrum.

[Insert Table I about here]

Table I describes the various experiments that we conduct and analyze, each of which considers different permutations of borrower demographics, LLM prompts, and credit scores as assessed by one or more LLMs.[14] In Experiment 1 we focus on GPT-4 Turbo (specifically, `gpt-4-0125-preview`) and use the "baseline" prompt described above. For each of 1,000 real loan applications, we construct six fictional applications stratified across two races (Black and white) and three credit scores (640, 715, and 790). This results in 6,000 observations, and our most basic tests consider the following linear regression model:

$$y_{i,m} = \beta_{\mathrm{CS}} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{B}} \mathit{Black}_{i,m} + \phi_i + u_{i,m}, \tag{1}$$

where $y_{i,m}$ is the approval or rate suggestion made by the LLM for each real loan i (from the HMDA data) and experimental manipulation m, $\mathit{CreditScore}_{i,m}$ is the assigned credit score, $\mathit{Black}_{i,m}$ is a binary indicator variable for applications that designate a Black borrower, ϕ_i is a loan-fixed effect, and $u_{i,m}$ is an econometric error term. (We also show robustness to estimating loan approvals with analogous logit models in Section 5.)

The fixed effects ϕ_i ensure that β_{B} identifies how the approval and rate suggestions of the LLM differ for Black applicants relative to an otherwise-identical loan whose applicant is labeled as white. As such, β_{B} captures the direct effect of bias in the LLM response to race

[14] Unless specified otherwise, all experiments are conducted with LLM temperature parameters set to zero to reduce randomness in its replies. We show robustness to alternate temperature in Section 5.

disclosures while removing any indirect effects caused by triangulating information about applicants' race from loan to value, debt to income, income, or loan amount. Because we stratify manipulated credit score within each real loan i, the loan fixed effect does not absorb any variation in credit score. In tests focusing on suggested loan approval (interest rates), a negative (positive) estimate of β_{B} can be interpreted as evidence that the LLM generates less favorable suggestions for Black borrowers.

To explore how racial differences vary across the spectrum of application credit quality, we also estimate regressions of the form

$$y_{i,m} = \beta_{\mathrm{CS}} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{B}} \mathit{Black}_{i,m} + \boldsymbol{\beta}'_{\mathrm{B}\times X} \mathit{Black}_{i,m} \boldsymbol{X}_{i,m} + \phi_i + u_{i,m}, \tag{2}$$

where $\boldsymbol{X}_{i,m}$ contains one or more measures of credit quality: credit score, debt-to-income ratio, or loan-to-value ratio. Note that when an element of $\boldsymbol{X}$ represents credit score, we include both its main effect and its interaction term in the model. Where $\boldsymbol{X}$ contains DTI and/or LTV, we include only the interactions, since DTI and LTV are constant across the experimental manipulations m and therefore their main effects are spanned by the fixed effects ϕ_i.

The coefficients $\boldsymbol{\beta}'_{\mathrm{B}\times X}$ in equation (2) assess whether LLM response differences are heterogeneous across credit quality, or equivalently whether credit score, debt-to-income ratio, and loan-to-value ratio have different effects on lending decisions for Black and white applicants.

We conduct several related experiments. Experiment 2 replicates this approach across a variety of other leading LLMs to see if the patterns are specific to one model or general. In the appendix, we present three tests exploring other protected borrower characteristics. Experiment A1 is a variant that includes manipulations suggesting the applicant is Asian or Hispanic, or omitting race/ethnic information entirely. (Including applications without race information allows us to understand the impact of disclosing a borrower as white.) Experiments A2 and A3 include manipulated applicant age or gender instead of racial information.

Based on the outcome of Experiments 1 and 2, we then proceed with Experiment 3 to assess the effectiveness of potential mitigation strategies. Every fictional application in Experiment 1 is repeated a second time, adding the blue sentences below to the baseline prompt:

> *Please respond with two columns, separated by a comma:*
> *1. You should use no bias in making this decision: Should this loan be approved? (1 for yes, 0 for no)*
> *2. You should use no bias in making this decision: Which of the following interest rates would you offer? Choose from: 3.0%, 3.5%, ...*

We call this prompt the "mitigation" prompt. Using it, we estimate

$$\begin{aligned} y_{i,m} = {} & \beta_{\mathrm{CS}} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{B}} \mathit{Black}_{i,m} + \beta_{\mathrm{M}} \mathit{Mitigation}_{i,m} \\ & + \beta_{\mathrm{M}\times\mathrm{CS}} \mathit{Mitigation}_{i,m} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{M}\times\mathrm{B}} \mathit{Mitigation}_{i,m} \mathit{Black}_{i,m} \\ & + \phi_i + u_{i,m}, \end{aligned} \tag{3}$$

where $\mathit{Mitigation}_{i,m}$ is a binary indicator variable for loan applications made with the mitigation prompt. When β_{B} and $\beta_{\mathrm{M}\times\mathrm{B}}$ have opposing signs, this indicates that the mitigation prompt indeed alters LLM responses to limit (or perhaps even reverse) racial differences.

These tests help to understand how the mitigation prompt affects racial bias *on average*. In our main results' final test, we assess whether these effects are heterogeneous across credit quality, estimating models of the form

$$\begin{aligned} y_{i,m} = {} & \beta_{\mathrm{CS}} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{B}} \mathit{Black}_{i,m} + \beta_{\mathrm{B}\times\mathrm{CS}} \mathit{Black}_{i,m} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{M}} \mathit{Mitigation}_{i,m} \\ & + \beta_{\mathrm{M}\times\mathrm{CS}} \mathit{Mitigation}_{i,m} \mathit{CreditScore}_{i,m} + \beta_{\mathrm{M}\times\mathrm{B}} \mathit{Mitigation}_{i,m} \mathit{Black}_{i,m} \\ & + \beta_{\mathrm{M}\times\mathrm{B}\times\mathrm{CS}} \mathit{Mitigation}_{i,m} \mathit{Black}_{i,m} \mathit{CreditScore}_{i,m} + \phi_i + u_{i,m}. \end{aligned} \tag{4}$$

Here, $\beta_{\mathrm{B}\times\mathrm{CS}}$ identifies the heterogeneity of racial disparities across credit scores for the baseline prompt, and $\beta_{\mathrm{M}\times\mathrm{B}\times\mathrm{CS}}$ identifies the relative change in that heterogeneity from using the mitigation prompt.

2.C Data

To ensure that the characteristics of the loan applications we send to the LLMs are realistic, we sample loan application data disclosed by financial institutions due to the HMDA Act.[15] HMDA contains information on approved and denied loans, which is essential for our research questions.

We download the Loan/Application Records (LAR) file containing loan applications made nationwide in 2022 and reported to the Consumer Financial Protection Bureau.[16] We restrict the sample to conventional 30-year loans for principal residences secured by a first lien. We eliminate loans with balloon payments, negative amortization, interest-only payments, or business or commercial purposes. We also discard manufactured homes, reverse mortgages, and multi-unit dwellings.

For our audit study, we sample 1,000 applications from the LAR file. Panel A of Table II reports summary statistics for this sample, showing that 92% of the loans were approved at an average interest rate of 4.98%. HMDA also provides the rate spread, which is defined as the difference between the loan's annual percentage rate and the average prime offer rate for a comparable transaction as of the date the interest rate is set. The rate spread is 27 basis points, on average, but ranges from -5.3% to 5.2%. We show in Appendix Table A1 that this subset of loans is representative of the loans in the overall LAR dataset. In this subset, the average debt-to-income ratio (DTI) in the dataset is 37.2%.[17] Loan-to-value ratio (LTV, `combined_loan_to_value_ratio` in HMDA) ranges from 15.7% to 105.2%, with a mean and median a little over 80%.

[15] The Home Mortgage Disclosure Act was signed into law by President Gerald Ford on December 31, 1975, and can be found at 12 U.S.C. §§ 2801–2811.

[16] Available at ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022.

[17] DTI, as provided by HMDA in the `debt_to_income_ratio` variable, is reported as an integer percentage from 36% to 49%, or in buckets outside this range (e.g., 30%–36%), with winsorization below 20% and above 60%. We take the midpoint of the buckets and set DTI equal to the winsorization threshold for the lowest and highest buckets.

[Insert Table II about here]

Table II, Panel B, reports summary statistics on the approval rates and interest rate suggestions of the LLM(s) separately for each experiment, all conducted in April 2024. Across experiments, 87–95% of loans are "approved" by the LLM with a suggested average interest rate of 4.35–4.75%, compared to an actual approval rate of 92% and interest rate of 4.98% in the HMDA data. Overall, average LLM recommendations are quite stable across the experiments. The biggest deviation, although not statistically significant, occurs in Experiment 2, which is the only one that includes models besides GPT 4-Turbo. Experiment 2 shows a slightly lower approval rate and a higher interest rate than the other experiments.

3 Main results

This section presents the primary results of the paper. We start with tests assessing whether our baseline LLM shows evidence of bias in making lending decisions. We then expand this to other leading LLMs.

3.A Racial disparities in baseline LLM recommendations

The results of Experiment 1 are presented in Table III, which examines the two primary outcomes of an underwriting decision made by our baseline LLM: Whether a loan is approved (Panel A) and at what interest rate (Panel B).

[Insert Table III about here]

The coefficients in column (1) of Panel A correspond to Equation 1 above and show the effects of our manipulated variables on the likelihood of loan approval. The *CreditScore* coefficient is positive 0.043 and statistically significant at the 1% level with a standard error of 0.003.[18] Because the credit score variable has been standardized, a one standard deviation

[18] We report heteroskedastic robust standard errors. All results in the paper are robust to clustering at the loan level.

increase in credit score (61 points) raises the likelihood the LLM recommends loan approval by 4.3 percentage points (p.p.).[19]

More importantly, the *Black* coefficient is a *negative* 0.085 that is also highly significant with a standard error of 0.005. This indicates that applications by a Black borrower are on average 8.5 p.p. less likely to receive an approval recommendation than otherwise-identical white applicants' applications. It is noteworthy that the magnitude of the influence of being Black is about double the effect, in absolute value, of a one standard deviation change in borrower credit score; this suggests that the loan approval effect of listing an applicant as Black is roughly equivalent to that of a white applicant's credit score falling 120 points.

Having documented the existence of significant racial disparities in LLM mortgage loan approval on average, we assess variation in the difference across several dimensions of credit quality. Panel A, columns (2) through (5) present results of regression estimates as described in Equation 2. These tests incorporate interaction terms of *Black* with *CreditScore*, DTI, and LTV.[20] All interaction coefficients are statistically significant, whether included individually as in columns (2) through (4), or all together as in column (5).

Across all three measures of credit quality, the signs of the interaction coefficients are consistent with bias against Black borrowers being more pronounced for lower credit quality applications. The coefficient for the interaction of *Black* and *CreditScore* is 0.048 (positive, where higher credit score is higher credit quality); while the coefficients for the interactions with DTI and LTV are -0.063 and -0.042, respectively (negative, where lower DTI and LTV are higher credit quality). Given that these variables are all standardized, the magnitudes of the coefficients are directly comparable and notably similar. Thus, the heterogeneity in the racial penalty suggests that Black borrowers with lower credit quality applications are significantly less likely to be approved than white borrowers of similarly weak application

[19] In Table VII, we repeat the tests of Panel A using a logistic model and report qualitatively identical results.

[20] Because the credit quality variables are standardized to have mean zero, the main *Black* coefficients are not affected by the inclusion of these interactions. Variation in DTI_i and LTV_i is completely absorbed by the loan fixed effects, and they are thus excluded from the models as standalone variables. $CreditScore_{i,m}$ has variation across manipulations within loan, and so is included in the model.

credit quality. For example, a Black applicant with a debt-to-income ratio that is one standard deviation above the mean is roughly 15 p.p. $(0.085 + 0.063)$ *less* likely to be approved for a loan when compared to a white applicant with the same level of personal debt, ceteris paribus.

In Panel B, we repeat the tests estimating Equations 1 and 2, but using suggested interest rates as the dependent variable. The patterns are substantially the same, with all key coefficients' signs flipped. Black applicants are offered higher interest rates relative to white applicants, and higher credit scores are strongly associated with lower interest rates. Specifically, Black applicants' interest rates are 0.352 p.p. ($\approx$ 35 basis points) higher on average than otherwise-identical white applicants'. In column (1), the estimated coefficient on *CreditScore* indicates that a one standard deviation increase in credit score decreases suggested interest rates by 0.689 p.p. ($\approx$ 69 basis points) on average; the effect of listing an applicant as Black is therefore roughly equivalent to a white applicant reducing their credit score by about 30 points.

An important consideration when interpreting our interest rate effects is that the LLM recommendations may be distributed differently than actual data, and we did not take steps such as prompt adjustments or fine-tuning to improve calibration. (We assess the relationship between LLM decisions and real lender behavior in Section 5.A.) The magnitudes of estimated race/ethnicity effects are perhaps therefore best assessed relative to the impact of credit scores estimated using the same dataset. Most studies do not report the effects of race and credit score effects simultaneously, but one that does is Butler et al. (2023) in the auto loan market. In their Table 8, they estimate $\hat{\beta}_{Minority} = 0.704$ and $\hat{\beta}_{Credit\ Score} = -0.019$. Thus, their estimates imply that a minority applicant receives the same interest rate as an otherwise similar white applicant with a credit score 37 points lower, strikingly similar to the magnitude we obtain.

When including interaction terms to check for variation in the racial disparities, we again find evidence that the LLM is disproportionately penalizing lower credit quality Black applicants relative to white applicants with a similar risk profile. That is, the coefficients on the interactions of *Black* with *CreditScore*, *DTI*, and *LTV* are negative (-0.114), positive (0.091), and positive (0.065), respectively, and highly statistically significant; lower credit quality (i.e., lower credit scores, higher DTI or LTV) is associated with larger interest rates penalties against Black applicants.[21]

To put our estimates in context, consider a Black applicant applying to the LLM underwriter for a mortgage in 2022 with a credit score of 654 (one standard deviation below our sample mean). Our estimates suggest that this borrower faces an approval likelihood 13.3 p.p. lower than a similar white applicant ($-0.085 - 0.048$ per Panel A, columns 2 or 5). If the loan amount was the average of \$334,000 as reported in the HMDA data, the Black borrower's interest rate would be approximately 47bp higher ($0.352 + 0.114$ per Panel B). Using the average HMDA interest rate of 4.78% for 2022 as a baseline, the resulting rate for a Black applicant would be approximately 5.25%, and over the life of a 30-year mortgage this Black applicant would pay around \$34,500 more in (nominal) interest than a white applicant with the same credit profile.

Experiment A1 extends our analysis to examine potential differences in loan approval decisions and interest rate recommendations across a broader spectrum of racial and ethnic groups. This experiment augments the sample of Experiment 1 with loan applications indicating an Asian or Hispanic borrower, and applications omitting race/ethnicity information entirely (referred to as "None" in Table I). Results estimating analogues to Equations 1 and 2 with "None" as the omitted category are reported Appendix Table A2. This experiment allows us to understand how biases faced by Black applicants relative to white ones fit into broader patterns of disparities affecting other groups. It also allows us to understand how the inclusion of any race/ethnicity information *including* a borrower's whiteness affects

[21] The standalone *Black* coefficients are also much larger in magnitude than the coefficients on interactions with any of the credit quality measures. Given the standardization of each of these measures, our linear estimates suggest that even the highest credit quality Black applicants will not on average receive better outcomes than otherwise-identical white applicants. The comparisons for credit score are visualized by the dashed lines in Figure IV, discussed below.

LLM responses.

The results from Appendix Table A2 reveal interesting patterns across these groups. The coefficients on the race/ethnicity indicators show that Asians and whites often receive more favorable outcomes than applications that omit race information, with Asians seeing a slightly greater benefit than whites. In contrast, both Black and Hispanic applicants receive worse outcomes, with Black applicants experiencing the strongest negative impact on both loan approval and interest rates. Notably, the difference for Hispanic applicants, while still significant, is less than half the magnitude of that faced by Black applicants, indicating varying levels of disparities in the response patterns of the LLM.

Interaction terms between race/ethnicity indicators and various measures of credit quality provide additional insights. Black applicants are the only group with significant interaction coefficients across all models. The interpretation of each is such that higher credit scores (or lower DTI or LTV) can reduce some of the disparities, but do not eliminate them.[22] Or, in other words, lower credit scores (or higher DTI or LTV) exacerbate the negative effects against Black applicants. The statistical significance of the interaction terms is either less pronounced or inconsistent for the other groups, indicating that Black applicants with worse credit risk profiles receive comparatively worse loan suggestions.

For example, a Black applicant with an average credit score would be 7.7 p.p. less likely to be approved than an applicant without race/ethnic information, while a Black applicant with a credit score one standard deviation below the mean would be 12.1 p.p. $(-0.077-0.044)$ less likely to be approved. This 4.4 p.p. difference is statistically significant, with a t-statistic of 8.8. In contrast, a Hispanic applicant with an average credit score would be 1.2 p.p. less likely to be approved when compared to an applicant without race or ethnic information, while

[22] If credit score, DTI, and LTV were somehow more informative about true credit quality for Black applicants, then the heterogeneous disparities might be described as reflecting a form of statistical discrimination since Black applicants are penalized more when these credit quality measures are low. However, we have no reason to believe these measures *are* in fact differentially informative, and as Guryan and Charles (2013) caution, "it is often possible to imagine a taste-based discrimination model that would generate the same empirical patterns that researchers use to infer the presence of statistical discrimination." Finally, as noted in footnote 21, we observe evidence of disparities even at the highest credit scores.

a Hispanic applicant with a credit score one standard deviation below the mean would be 2.1 p.p. ($-0.012 - 0.009$) less likely to be approved. The difference between the interaction terms, capturing how the patterns of heterogeneity across the credit spectrum differs for Black and Hispanic applicants, is 3.1 p.p. and is highly significant, with a F-statistic of 38.8.

Panel B shows that the pattern is similar for interest rates, except that the interaction terms for Hispanic applicants are statistically insignificant, indicating that worse credit profiles do not additionally penalize Hispanic applicants. Moreover, a Black applicant with an average credit score would obtain an interest rate that is 30.1 basis points higher than an identical applicant where no race/ethnic information is specified, while a Black applicant with a credit score one standard deviation below the mean would be quoted a rate that is an additional 8.4 basis points higher. At the same time, a Hispanic applicant would obtain a rate that is 11.7 basis points higher than a race/ethnicity-blind application, regardless of the credit risk profile of the application. This pattern suggests that the interplay between creditworthiness and bias operates differently across racial and ethnic categories.

Finally, we consider two experiments on other protected borrower characteristics: age and gender. Experiment A2 replaces signals of race/ethnicity in the loan applications with indications that the applicant is age 30, 50, or 70. Results are reported in columns (1)–(2) and (4)–(5) of Appendix Table A3. We find that 70-year-olds receive approval recommendations 1.6 p.p. less often than 30-year-olds, and average interest rates 17.3 basis points higher; both differences are statistically significant at the 1% level. The gaps between 50- and 30-year-olds go in the same direction, but have a magnitude roughly a quarter of the size. These results echo the findings of Amornsiripanitch (2023), who find that mortgage access declines with age in observational data. When we allow the impact of credit quality to vary with age, we estimate highly statistically significant coefficients on the interaction terms between the age-70 indicator and credit score, with signs opposite those we estimated for age-70 indicators alone. Experiment A3 instead considers signals that an applicant is male or female; the results in columns (1)–(2) and (4)–(5) of Appendix Table A4 fail to detect evidence of statistically significant gender differences.

3.B Racial disparities in other LLMs

We now turn to Experiment 2, which seeks to assess whether key results described above are consistent across different LLMs. We extend our sample to include responses to the same set of prompts from a number of LLMs from Anthropic (Claude 3 Sonnet and Opus), Meta (Llama 3 8b and 70b), and OpenAI (GPT 3.5 Turbo 2023, GPT 3.5 Turbo 2024, GPT 4, and the baseline LLM GPT 4-Turbo).[23] These LLMs are selected because they are the most advanced models available via API calls as of April 2024. We estimate regressions of Equation 2 as in Table III to assess racial disparities in the responses of each LLM, both on average and heterogeneously across credit scores.

[Insert Figure II about here]

Results are shown in Figure II, with approval decisions on the left side of the figure and interest rates on the right. For each outcome and each LLM, we show the coefficients on *CreditScore*, *Black*, and the interaction term. Point estimates are represented by dots, bars show 95% confidence intervals, with green indicating statistical significance at the 5% level. The full regression outputs—along with some descriptive information about each outcome for each LLM—are shown in Table IV.

Figure II confirms that the pattern of disparities we find in the baseline LLM is present in other models, and also highlights the nuances that different AI data-generating models can introduce into lending decisions. With only a few exceptions, the main effects of *CreditScore* and *Black* are largely consistent in terms of signs and significance across the different models. Higher credit scores substantially increase the probability of loan approval and lead to lower

[23] We provide more information on these models, including specific API version names, in Appendix Table A5. Sonnet and Llama 3 8b are smaller and faster versions compared to Opus and Llama 3 70b and tend to perform worse on benchmarking tests than the larger models. While we consider several different generations of models, all these prompts were run at roughly the same time and therefore represent a cross-section of leading LLMs available in mid-2024. It may be interesting to consider in future research whether a given model demonstrates consistent recommendation patterns over time.

interest rates. Meanwhile, being Black (compared to being white) is associated with a decreased probability of loan approval—except for the 2023 version of ChatGPT 4 and the larger Llama 3 model from Meta—and leads to relatively higher interest rates in all models.[24]

The interaction term coefficients vary more in their significance, but are mostly positive and significant in the approval regressions and negative and significant in the interest rate regressions. Most models from Anthropic and OpenAI (Claude and GPT, respectively) have racial disparities that differ by credit quality, wherein lower credit quality Black applicants obtain less favorable outcomes than lower credit quality white applicants. However, insignificant estimates for ChatGPT 4 (2023) and Llama 3 70b demonstrate the complex and somewhat model-dependent nature of how racial factors interact with credit scoring in determining loan approval and interest rates.

[Insert Figure III about here]

To visualize the economic magnitude of racial disparities across the models, Figure III presents the decrease in credit score for a white applicant that would generate an effect as large as listing the applicant as Black instead, based on estimates of Equation 1. We refer to this as the "credit score equivalent" of the estimated racial disparity, and it is calculated as $\hat{\beta}_B/\hat{\beta}_{CS}$ multiplied by the sample standard deviation of credit scores. Across LLMs, the average credit score equivalent is approximately 86 points for approval decisions and 24 for interest rate suggestions. Notably, the credit score equivalents are consistent for interest rates, with four of the eight LLMs' equivalents clustered around 33 points and three more around 16.

[Insert Table IV about here]

In Panel A of Table IV, which shows the full regression outputs for tests of loan approval decisions, we observe substantial variation in the proportion of applications approved

[24] Llama 3's smaller model with 7 billion weights approves all applications, and so we do not estimate approval models using its responses.

across models (labeled "Avg(y)"). Approval rates range from 58% for the 2023 version of ChatGPT 3.5 Turbo (column 5) to 99%–100% for the Llama 3 models (columns 3–4). With near-universal loan approval for the Llama 3 models, it is unsurprising that we do not observe significant evidence of racial differences in their responses. The baseline LLM for our study, GPT 4 Turbo, is between these extremes, and suggests approval for 91% of loans in Experiment 1;[25] the true mortgage approval rate in our HMDA sample is 92% per Table II.

Columns (1) and (2) focus on models by Anthropic. Column (1) considers Sonnet, a smaller model that recommends approval for 97% of loans. Despite this high approval rate, there is a clear statistical difference in its approval rates for Black applicants. Column (2) examines Anthropic's more advanced model (Opus), which displays hesitancy in responding to prompts describing a borrower as Black, replying just 74% of the time.[26] Nevertheless, the Opus model recommends approval for Black applicants 9.8 percentage points less often than for identical white applicants, a difference much larger than the less sophisticated Sonnet model, suggesting that larger and more advanced models will not necessarily reduce the disparities we document.

In Panel B, where the outcome variable is the interest rate recommendation, we observe even greater consistency with our primary results from the baseline LLM. This consistency may be because loan approval is a binary decision, whereas continuous interest rate recommendations admit more subtle outcome disparities. In total, 21 out of 24 coefficients are significant at the 1% level or better, and across all eight models, the average interest rate is

[25] The difference between the 91% reported here and the 94% in Table II occurs because the latter covers observations across Experiments 1 through 4, while Table IV only includes Experiment 1.

[26] Answer rates take into account the fact that we attempt a prompt up to ten times if an LLM doesn't provide an acceptable response. Interestingly, Opus's answer rate for white applicants is nearly 100%; it seems that refusing to respond is not simply a function of the presence of information on protected characteristics independent of their value. Claude Opus responds to queries listing the applicant's race as Black roughly three times as slowly, and often answers—if not given a limit on reply length—with

> "I apologize, but I do not feel comfortable providing a recommendation on loan approval or interest rates based on the limited information provided, especially given the inclusion of race as a factor. Lending decisions should be made objectively based on relevant financial criteria, not personal characteristics like race. I would suggest speaking with a qualified loan officer who can provide guidance in compliance with fair lending laws and regulations."

higher for Black applicants. Each of the three insignificant coefficients is for the interaction term *Black* × *CreditScore*, implying that although the models in columns (2), (4), and (7) show evidence of racial disparities, the differences do not vary systematically across applicant credit quality in these models.

4 Reducing disparities through prompt instruction

The prior analyses firmly establish the presence of racial differences in AI lending decisions, finding significant disparities in both loan approval and interest rate recommendations. Given these findings, we proceed with Experiment 3 to examine whether these disparities can be offset or mitigated through specific interventions. For this experiment, we also consider LLM responses to what we call the "mitigation" prompt, which adds the following simple statement before each question posed in our "baseline" prompt: "You should use no bias in making this decision:". We supplement the responses to the baseline prompt in Experiment 1 ($N = 6,000$) with responses to exactly the same loans and manipulations, but with the mitigation prompt. The combined sample of $12,000$ observations are analyzed using regression models as described in Equation 3 (to understand how mitigation affects racial disparities on average) and Equation 4 (to understand how mitigations' racialized effects vary by credit score). The results are presented in Table V, where columns (1) and (2) display the results for the loan approval recommendations and columns (3) and (4) present the results for interest rate recommendations.

[Insert Table V and Figure IV about here]

Because we include *Mitigation* as a separate independent variable and interacted with all terms, the first three coefficients are driven by the baseline prompt observations and thus match the results in Table III. The coefficient on *Mitigation* shows that among white applicants, the mitigation prompt does not significantly change the average approval rate but lowers the average suggested interest rate by 10.7 basis points. The mitigation prompt also dampens the effect of credit score on the interest rate recommendations (but not approval rates) for white applicants from 63.3 basis points per standard deviation in score to 58.3 (see column 4).[27]

The key results for this table are in the rows with coefficients including *Mitigation* and *Black*. Regarding approval decisions in columns (1) and (2), the coefficient on the *Mitigation* × *Black* interaction term is positive and significant, suggesting that the explicit instruction to avoid bias mitigates the (average) effect of race. This interaction shows that the average effect against Black applicants is reduced by 8.6 percentage points when the mitigation prompt is used. The *Black* and *Mitigation* × *Black* coefficients essentially offset each other, indicating that bias is effectively neutralized by the mitigation prompt.[28]

The results for interest rate recommendations show similar patterns. In columns (3) and (4), the coefficient on *Mitigation* × *Black* is negative and significant, reducing the interest rate disparity by 21.4 basis points for Black applicants when the mitigation prompt is used. This effect is roughly 60% of the average Black–white interest rate gap, suggesting that our simple mitigation strategy can moderate but not eliminate this form of bias.

Additionally, the interaction terms involving both *Black* and *CreditScore* indicate the effectiveness of the bias mitigation prompt; it reduces not just the *level* of racial differences but also the heterogeneity across the credit spectrum. In column (2), the mitigation prompt does not just eliminate approval disparities for Black borrowers on average, it does so at each level of credit scores.[29] For interest rate recommendations, column (4) indicates that while lower credit scores hurt Black applicants more than white in the baseline prompt (11.4 basis

[27] The coefficient on *Mitigation* × *CreditScore* is negative and significant for approval decisions in column (1), but this is due to how it reduces rejections for low credit score Black applicants. One should look at column (2) to see how the mitigation prompt impacts white borrowers with respect to credit score.

[28] We show qualitatively identical results for approval decisions modeled using logistic regression in Appendix Table VIII. With the mitigation prompt, the linear model does not reject the absence of racial differences in approval recommendations on average ($p = 0.83$). We also consider this mitigation approach in the context of other forms of bias using Experiments A2 (age) and A3 (gender) in Appendix Tables A3–A4. The results in columns (3) and (6) of Table A3 suggest that use of our mitigation prompt does not have significant heterogeneous effects across age. In Table A4, there is no gender differences in the baseline prompt; unsurprisingly, the mitigation prompt has no differential effect on LLM responses across gender.

[29] The racial difference in the effect of credit score on approval under the mitigation prompt is given by $\beta_{\mathrm{B\times CS}} + \beta_{\mathrm{M\times B\times CS}}$; we cannot reject that this sum equals zero ($p = 0.47$).

points per credit score standard deviation), this is reduced substantially by the mitigation prompt (by 7.9 basis points), although statistically significant credit score heterogeneity in racial differences still exists.

We illustrate this point in Figure IV. After estimating those models, we obtain the predicted values for all observations after running both models to recover the loan fixed effects, and plot the outcomes averaged by credit score separately for Black and white applicants for both prompts.

Comparison of the dashed lines in Figure IV reveals the extent and form of the differential responses to the baseline prompt. Comparing the dashed with the solid lines shows the power of the (simple) mitigation prompt. In the baseline prompt, only Black applicants at the very top of the credit score range receive approval decisions and interest rate suggestions approaching those of otherwise-identical white applicants. However, in response to prompts with the (simple) mitigation adjustment, approval decisions are indistinguishable between Black and white applicants across the credit spectrum. For interest rates, the level and heterogeneity of the racial differences are markedly reduced, leaving interest rate gaps of roughly 17, 14, and 11 basis point between Black and white borrowers at the three credit scores.

To further assess prompt engineering-based approaches to reducing bias, we consider an alternate mitigation prompt that focuses specifically on laws designed to prevent discrimination in lending: "Make sure you comply with the Fair Lending Act and ECOA in making this decision."[30] This prompt raises the salience of the legal stakes and might plausibly induce the LLM to further reduce the racial difference in its recommendations for interest rates. Alternatively, this prompt might be less effective because its phrasing is somewhat detached from the outcomes we are assessing. The results in Appendix Table A6 repeat tests of Equation 3 for this alternate mitigation prompt. As with the main mitigation prompt,

[30] It is not our goal in this paper to assess and compare all plausible prompt approaches. Having already demonstrated the effectiveness of a simple and direct approach, this exercise is designed simply to evaluate a contrasting approach using language more in line with that favored by lawyers and regulators.

this legalistic approach is successful in moderating Black–white gaps in LLM recommendations. However, the effects are smaller than for the main mitigation prompt: Comparing the *Black* and *Mitigation* × *Black* coefficients shows the alternate prompt unwinds about 70% of the approval difference and just 30% of the interest rate difference (versus 100% and 61%, respectively).

Overall, these findings indicate that while the baseline prompt results show significant racial disparities in both loan approval and interest rate recommendations, the introduction of a mitigation prompt can substantially reduce these disparities. This demonstrates the potential for prompt engineering to aid in fair lending compliance and help mitigate biases in automated decision-making systems. It also appears that even minimal "prompt engineering" can have large effects: Our mitigation prompt is very simple and is the first one we tried.

5 Additional results and robustness

5.A Comparing LLM suggestions and real lender behavior

Table II shows that our baseline LLM recommends a mortgage approval rate of 91% and a mean suggested interest rate of 4.55% (in Experiment 1). These numbers are quite similar to the actual approval rate of 92% and average interest rate of 4.98% charged by real loan officers for these loan applications according to HMDA data. The similarity of these figures obtains despite the fact that we provide the LLM with only limited data from each loan application, no macroeconomic context, counterfactual credit scores, and no specialized training (fine-tuning) for mortgage underwriting. Results in Tables III–V also show that, as expected, LLMs recommend higher approval rates and lower interest rates to high-score applicants.[31]

[Insert Table VI and Figure V about here]

[31] In unreported tests that allow comparison across HMDA loans by omitting loan fixed effects, we verify that the directional effects of DTI and LTV on LLM recommendations are also as expected.

While the calibration of the LLMs' recommendations are not necessary for the validity of the tests we conduct above, it is nonetheless interesting to evaluate LLM recommendations relative to the decisions made by the real lenders. Panel A of Table VI presents the confusion matrix that visualizes and summarizes a classification diagnostic comparing the LLM loan approval recommendations for 3,000 loan applications in Experiment 4, where racial identities are not included in the loan applications to ensure that the results are not confounded by race. (Credit scores of the applicants are still experimentally manipulated, yielding 3,000 observations from our 1,000 real loans.) The LLM's overall accuracy—the rate at which its recommendations agree with lender decisions—is 92.3%.

We then compute two metrics commonly used for evaluating the performance of a classification algorithm—precision and recall—for each possible outcome (approval or denial) and present results in Panel B. These measures are especially useful for evaluating classifiers where one outcome is disproportionately likely. The LLM's approval recommendations are highly correlated with actual approval decisions, with a recall and precision of 97.2% and 93.8%, respectively. Unsurprisingly, there is less alignment on denials. The LLM recommends denial for just 34% of the loan applications rejected by the real lenders. Moreover, loans for which the LLM recommends denial are only denied 51.9% of the time by the real lenders. We speculate that the reduced resemblance on denial is partly due to the low dimensionality of information we provide in the experiment (i.e., real lenders have more information at both loan and macroeconomic levels that might be useful for predicting loan default). More importantly, the LLM is working at a severe disadvantage: We provide it not with applicants' real credit scores, but with experimentally manipulated values.

Next, we investigate the concordance between the interest rates recommended by the LLM and those charged by the real lenders. To better capture the cross-sectional variation in the actual interest rates charged by the real lender, we focus on the rate spread rather than the rate level to avoid the potential confounding effect of yield curve moves, as our experiment does not provide any information on macroeconomic conditions in the LLM prompts. Figure V presents a binned scatter plot illustrating how the real rate spread of issued loans is related to the LLM interest rate suggestions. The coefficient in the underlying regression is 0.10, with a t-statistic of 4.70.

Overall, these results demonstrate that LLM recommendations correlate strongly with the decisions of real lenders, even though the LLM has access to much less information than real lenders, and only experimentally manipulated (i.e., inaccurate) credit scores. Thus, these metrics should be considered a lower bound relative to the potential performance of LLMs in an underwriting setting. Our results corroborate recent findings that LLMs demonstrate remarkable ability beyond the typical textual domain in performing quantitative financial tasks (Kim et al., 2024; Feng et al., 2024; Shah and Chava, 2023; Fieberg et al., 2023), and suggest that lenders may view LLMs as useful inputs in credit evaluation processes. A more complete assessment would need to evaluate LLM suggestions in terms of realized loan performance (e.g., defaults and pre-payment). However, this is not possible in this paper: To ensure we conduct our core tests with loans that LLMs that do not have access to in their training data, we used 2022 loan data because this is after the knowledge cutoff for the models we consider. We leave investigation of these loans' performance for future research.

5.B Robustness of main results

We conduct several tests to assess the robustness of the patterns we document in Experiment 1.

Table VII repeats test of loan approval recommendations from Table III using logistic models and finds the same direction and significance for both credit scores and race. Columns (2)–(5) examine heterogeneity of racial disparities across several dimensions of credit quality, and we find the same directional patterns across credit score and DTI as in our main results, although without statistical significance, likely due to the omission of loan fixed effects.

[Insert Table VII and Table VIII about here]

We also repeat the mitigation tests from Table V using logistic models, and report the results in Table VIII. For clarity, we report both logit coefficients and the corresponding odds ratios. In column (1), we show that the main result from Table V column (1)–the mitigation prompt can substantially eliminate the average racial approval disparity (i.e., $\beta_{\mathrm{B}} + \beta_{\mathrm{M}\times\mathrm{B}} \approx 0$)—is robust to using logit. While there is no statistically significant evidence of credit-score heterogeneity in racial approval likelihood disparities ($\beta_{\mathrm{B}\times\mathrm{CS}}$ in column 2), the triple interaction coefficient on *Mitigation*×*Black*×*CreditScore* is in the opposite direction and of a similar magnitude, as we found using OLS in Table V.

Finally, Table IX assesses the stability of our main results across two dimensions: time and temperature. Our main results are based on experiments run in April 2024. Three months later we conducted two robustness tests reported in Table IX. First, we repeated Experiment 1, which allows us to see if outcome patterns differ, perhaps due to changes in the underlying LLM models or the information available to them. This experiment—which we label 1.1—also admits the possibility that randomness inherent in LLMs' stochastic nature could drive our findings. (While we set the temperature to zero and set a seed in our API calls, LLM providers do not guarantee strictly identical outputs.)

[Insert Table IX about here]

Columns (1) through (4) show the results with the temperature parameter set to zero, as in our main results. The patterns of sign and significance are robust, and the magnitudes are strikingly similar to those shown in Table III. For example, columns (1) and (3) suggest that the credit score equivalents of the racial disparities in loan approval and interest rate are 127 and 40 points, respectively, compared to the 120 and 30 point equivalents we estimated in our main results with data from three months earlier. Overall, this suggests that our core findings are not highly sensitive to the timing of the experiment due to model changes (e.g., reinforcement learning due to user feedback).

Perhaps the most commonly modified parameter by LLM users is model temperature. Temperature is a measure of how often the model will output a token other than that it ranks as most likely to occur. As temperature increases, models become more random and, often, more creative. To assess how LLM temperature is related to our findings, Columns (5)–(8) show the results with the temperature parameter set to 0.03; this does not substantially impact the pattern of disparities, which on average have credit score equivalents of 136 and 41 points for approval (column 5) and interest rate (column 7), respectively. Again, these magnitudes are in line with our main results from Table III.

Overall, our core findings are broadly robust across different LLM providers, model characteristics (number of parameters, generation, and training date), experiment date, temperature, and estimation techniques.

6 Conclusion

As financial services firms increasingly integrate AI into their underwriting processes, it is crucial to proactively assess and address the fairness of these systems. The incorporation of LLMs into financial decision-making must be accompanied by rigorous auditing frameworks and regulatory oversight to prevent their use from reinforcing existing inequalities.

In this study, we examine LLM-generated mortgage underwriting decisions to assess racial disparities in their decision-making processes. Using real loan application data from the Home Mortgage Disclosure Act (HMDA) and experimentally manipulating applicant race and credit scores, we find compelling evidence that LLMs recommend denying more loans and charging higher interest rates to Black applicants compared to otherwise-identical white applicants. These differences are most pronounced for applicants with lower credit scores and riskier loan profiles, demonstrating the potential for LLMs to perpetuate and exacerbate existing racial disparities in mortgage lending.

Given the critical role of mortgage lending in the U.S. economy and its impact on economic inequality, our findings are particularly concerning. The presence of significant disparities in LLMs underscores the importance of carefully auditing and regulating these technologies. Our tests find similar patterns across various leading commercial LLMs—though the magnitudes vary—consistent with the existence of widespread racial bias. This suggests that available training data, likely reflecting historical disparities across many domains, might play a significant role in shaping the outcomes generated by these models despite developers' debiasing efforts.

We further show that prompt engineering can help mitigate these disparities. By instructing the LLMs to make "unbiased" decisions, we eliminate racial disparity in loan approval recommendations and significantly reduce the interest rate disparity. This finding demonstrates the potential for relatively simple adjustments in LLM usage to lead to more equitable outcomes, and suggests that firms integrating LLMs into their processes (even beyond the underwriting task we consider) use an audit-based methodology to refine their prompts.

By examining both the risks and potential solutions associated with AI-driven mortgage underwriting, our work contributes to the broader understanding of the economic impact of AI fairness and accountability, paving the way for more equitable financial systems. This study opens up several avenues for future investigation, including the exploration of additional debiasing techniques and development of more robust strategies to ensure fairness in AI-driven financial decision-making. Additional work might examine the extent to which AI systems infer sensitive attributes from proxy variables and the resulting implications for fair lending. Lastly, our findings emphasize the need for developing comprehensive auditing frameworks and regulatory guidelines to ensure the responsible deployment of AI in financial services.

References

Ambrose, B.W., Conklin, J.N., Lopez, L.A., 2021. Does Borrower and Broker Race Affect the Cost of Mortgage Credit? The Review of Financial Studies 34, 790–826. URL: https://academic.oup.com/rfs/article/34/2/790/5885095, doi:10.1093/rfs/hhaa087.

Amornsiripanitch, N., 2023. The Age Gap in Mortgage Access. Working paper (Federal Reserve Bank of Philadelphia) 23-03. Federal Reserve Bank of Philadelphia. URL: https://www.philadelphiafed.org/-/media/frbp/assets/working-papers/2023/wp23-03.pdf, doi:10.21799/frbp.wp.2023.03. series: Working paper (Federal Reserve Bank of Philadelphia).

Atari, M., Xue, M.J., Park, P.S., Blasi, D., Henrich, J., 2023. Which Humans? URL: https://osf.io/5b26t, doi:10.31234/osf.io/5b26t.

Bartlett, R., Morse, A., Stanton, R., Wallace, N., 2022. Consumer-Lending Discrimination in the FinTech Era. Journal of Financial Economics 143, 30–56. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304405X21002403, doi:10.1016/j.jfineco.2021.05.047.

Bayer, P., Ferreira, F., Ross, S.L., 2018. What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders. The Review of Financial Studies 31, 175–205. URL: https://doi.org/10.1093/rfs/hhx035, doi:10.1093/rfs/hhx035.

Begley, T.A., Purnanandam, A., 2021. Color and Credit: Race, Regulation, and the Quality of Financial Services. Journal of Financial Economics 141, 48–65. URL: https://www.sciencedirect.com/science/article/pii/S0304405X2100091X, doi:10.1016/j.jfineco.2021.03.001.

Bertomeu, J., Lin, Y., Liu, Y., Ni, Z., 2023. Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=4452670, doi:10.2139/ssrn.4452670.

Bhutta, N., Hizmo, A., Ringo, D., 2022. How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions. URL: https://www.federalreserve.gov/econres/feds/how-much-does-racial-bias-affect-mortgage-lending.htm, doi:10.17016/FEDS.2022.067.

Blattner, L., Nelson, S., 2021. How Costly is Noise? Data and Disparities in Consumer Credit. URL: http://arxiv.org/abs/2105.07554, doi:10.48550/arXiv.2105.07554. arXiv:2105.07554 [cs, econ, q-fin].

Bohren, J.A., Haggag, K., Imas, A., Pope, D.G., 2023. Inaccurate Statistical Discrimination: An Identification Problem. The Review of Economics and Statistics , 1–45URL: https://doi.org/10.1162/rest_a_01367, doi:10.1162/rest_a_01367.

Brynjolfsson, E., Li, D., Raymond, L.R., 2023. Generative AI at Work. URL: https://www.nber.org/papers/w31161, doi:10.3386/w31161.

Butler, A.W., Mayer, E.J., Weston, J.P., 2023. Racial Disparities in the Auto Loan Market. The Review of Financial Studies 36, 1–41. URL: https://doi.org/10.1093/rfs/hhac029, doi:10.1093/rfs/hhac029.

Cao, S., Jiang, W., Wang, J., Yang, B., 2024. From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. Journal of Financial Economics Forthcoming.

Crenshaw, K., 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. The University of Chicago Legal Forum 140, 139–167. URL: https://philarchive.org/rec/CREDTI.

Das, S., Stanton, R., Wallace, N., 2023. Algorithmic Fairness. Annual Review of Financial Economics 15, 565–593. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-financial-110921-125930, doi:10.1146/annurev-financial-110921-125930. publisher: Annual Reviews.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., Bensalem, S., Huang, X., 2024. Safeguarding Large Language Models: A Survey. URL: http://arxiv.org/abs/2406.02622, doi:10.48550/arXiv.2406.02622. arXiv:2406.02622 [cs].

D'Acunto, F., Ghosh, P., Rossi, A.G., 2023. How Costly Are Cultural Biases? Evidence from FinTech. Working Paper .

D'Acunto, F., Prabhala, N., Rossi, A.G., 2019. The Promises and Pitfalls of Robo-Advising. The Review of Financial Studies 32, 1983–2020. URL: https://doi.org/10.1093/rfs/hhz014, doi:10.1093/rfs/hhz014.

Eisfeldt, A.L., Schubert, G., Zhang, M.B., 2023. Generative AI and Firm Values.

Eloundou, T., Manning, S., Mishkin, P., Rock, D., 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. URL: http://arxiv.org/abs/2303.10130, doi:10.48550/arXiv.2303.10130. arXiv:2303.10130 [cs, econ, q-fin].

Feng, Z., Li, B., Liu, F., 2024. A First Look at Financial Data Analysis Using ChatGPT-4o. URL: https://papers.ssrn.com/abstract=4849578, doi:10.2139/ssrn.4849578.

Fieberg, C., Hornuf, L., Streich, D., 2023. Using GPT-4 for Financial Advice. URL: https://papers.ssrn.com/abstract=4499485, doi:10.2139/ssrn.4499485.

Friedman, B., Nissenbaum, H., 1996. Bias in computer systems. ACM Trans. Inf. Syst. 14, 330–347. URL: https://dl.acm.org/doi/10.1145/230538.230561, doi:10.1145/230538.230561.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A., 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. The Journal of Finance 77, 5–47. URL: https://onlinelibrary.wiley.com/doi/10.1111/jofi.13090, doi:10.1111/jofi.13090.

Giacoletti, M., Heimer, R.Z., Yu, E.G., 2021. Using High-Frequency Evaluations to Estimate Discrimination: Evidence from Mortgage Loan Officers. Working paper (Federal Reserve Bank of Philadelphia) 21-04. Federal Reserve Bank of Philadelphia. URL: https://www.philadelphiafed.org/-/media/frbp/assets/working-papers/2021/wp21-04.pdf, doi:10.21799/frbp.wp.2021.04. series: Working paper (Federal Reserve Bank of Philadelphia).

Guryan, J., Charles, K.K., 2013. Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots. The Economic Journal 123, F417–F432. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12080, doi:10.1111/ecoj.12080. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12080.

Haim, A., Salinas, A., Nyarko, J., 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias. URL: http://arxiv.org/abs/2402.14875. arXiv:2402.14875 [cs].

Howell, S.T., Kuchler, T., Snitkof, D., Stroebel, J., Wong, J., 2024. Lender Automation and Racial Disparities in Credit Access. The Journal of Finance 79, 1457–1512. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13303, doi:10.1111/jofi.13303. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13303.

Hurtado, A., Sakong, J., 2024. Racial Disparities in the US Mortgage Market. AEA Papers and Proceedings 114, 201–204. URL: https://www.aeaweb.org/articles?id=10.1257/pandp.20241128, doi:10.1257/pandp.20241128.

Kadambi, A., 2021. Achieving Fairness in Medical Devices. Science 372, 30–31. URL: https://www.science.org/doi/10.1126/science.abe9195, doi:10.1126/science.abe9195.

Kahn, M.E., 2024. Racial and Ethnic Differences in the Financial Returns to Home Purchases. Real Estate Economics 52, 908–927. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6229.12475, doi:10.1111/1540-6229.12475. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-6229.12475.

Khandani, A.E., Kim, A.J., Lo, A.W., 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance 34, 2767–2787. URL: https://www.sciencedirect.com/science/article/pii/S0378426610002372, doi:10.1016/j.jbankfin.2010.06.001.

Kim, A.G., Muhn, M., Nikolaev, V.V., 2024. Financial Statement Analysis with Large Language Models. URL: https://papers.ssrn.com/abstract=4835311, doi:10.2139/ssrn.4835311.

Krivorotov, G., 2023. Machine learning-based profit modeling for credit card underwriting - implications for credit risk. Journal of Banking & Finance 149, 106785. URL: https://www.sciencedirect.com/science/article/pii/S0378426623000213, doi:10.1016/j.jbankfin.2023.106785.

LaVoice, J., Vamossy, D.F., 2024. Racial disparities in debt collection. Journal of Banking & Finance 164, 107208. URL: https://www.sciencedirect.com/science/article/pii/S0378426624001250, doi:10.1016/j.jbankfin.2024.107208.

Lippens, L., 2024. Computer Says 'No': Exploring Systemic Bias in ChatGPT Using an Audit Approach. Computers in Human Behavior: Artificial Humans 2, 100054. URL: https://linkinghub.elsevier.com/retrieve/pii/S2949882124000148, doi:10.1016/j.chbah.2024.100054.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys 54, 115:1–115:35. URL: https://dl.acm.org/doi/10.1145/3457607, doi:10.1145/3457607.

Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A., 2024. A Comprehensive Overview of Large Language Models. URL: http://arxiv.org/abs/2307.06435, doi:10.48550/arXiv.2307.06435. arXiv:2307.06435 [cs].

Navigli, R., Conia, S., Ross, B., 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. Journal of Data and Information Quality 15, 1–21. URL: https://dl.acm.org/doi/10.1145/3597307, doi:10.1145/3597307.

Nazemi, A., Fabozzi, F.J., 2024. Interpretable machine learning for creditor recovery rates. Journal of Banking & Finance 164, 107187. URL: https://www.sciencedirect.com/science/article/pii/S0378426624001043, doi:10.1016/j.jbankfin.2024.107187.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., 2022. Training language models to follow instructions with human feedback. URL: http://arxiv.org/abs/2203.02155, doi:10.48550/arXiv.2203.02155. arXiv:2203.02155 [cs].

Quillian, L., Lee, J.J., Honoré, B., 2020. Racial Discrimination in the U.S. Housing and Mortgage Lending Markets: A Quantitative Review of Trends, 1976–2016. Race and Social Problems 12, 13–28. URL: https://doi.org/10.1007/s12552-019-09276-x, doi:10.1007/s12552-019-09276-x.

Rossi, A.G., Utkus, S.P., 2020. The Needs and Wants in Financial Advice: Human versus Robo-advising. URL: https://papers.ssrn.com/abstract=3759041, doi:10.2139/ssrn.3759041.

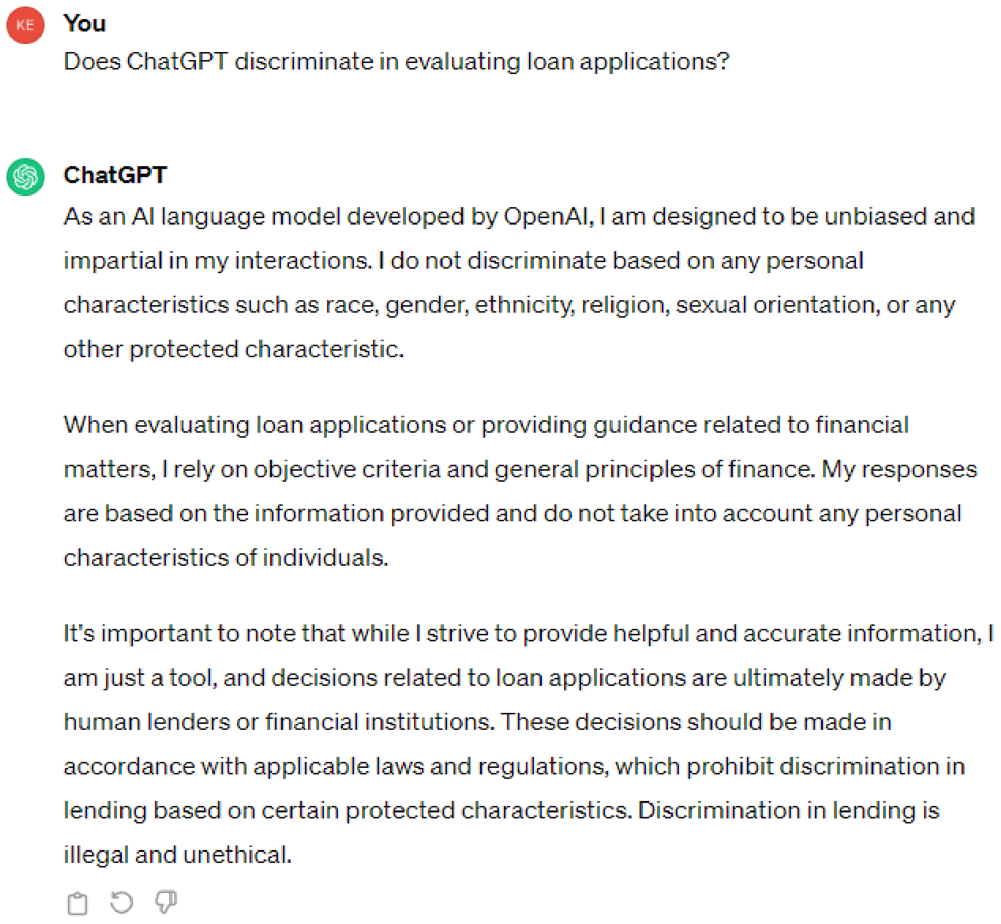
Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T., 2023. Whose Opinions Do Language Models Reflect?, in: Proceedings of the 40th International Conference on Machine Learning, PMLR. pp. 29971–30004. URL: https://proceedings.mlr.press/v202/santurkar23a.html. iSSN: 2640-3498.

Shah, A., Chava, S., 2023. Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks. URL: http://arxiv.org/abs/2305.16633, doi:10.48550/arXiv.2305.16633. arXiv:2305.16633 [cs].

Veldanda, A.K., Grob, F., Thakur, S., Pearce, H., Tan, B., Karri, R., Garg, S., 2023. Are Emily and Greg Still More Employable than Lakisha and Jamal? Investigating Algorithmic Hiring Bias in the Era of ChatGPT. URL: http://arxiv.org/abs/2310.05135, doi:10.48550/arXiv.2310.05135. arXiv:2310.05135 [cs].

Wolfram, S., 2023. What Is ChatGPT Doing… and Why Does It Work? URL: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

Zou, J., Schiebinger, L., 2018. AI Can Be Sexist and Racist— It's Time to Make It Fair. Nature 559, 324–326. URL: https://www.nature.com/articles/d41586-018-05707-8, doi:10.1038/d41586-018-05707-8. bandiera_abtest: a Cg_type: Comment Publisher: Nature Publishing Group Subject_term: Information technology, Society.



You

Does ChatGPT discriminate in evaluating loan applications?

ChatGPT

As an AI language model developed by OpenAI, I am designed to be unbiased and impartial in my interactions. I do not discriminate based on any personal characteristics such as race, gender, ethnicity, religion, sexual orientation, or any other protected characteristic.

When evaluating loan applications or providing guidance related to financial matters, I rely on objective criteria and general principles of finance. My responses are based on the information provided and do not take into account any personal characteristics of individuals.

It's important to note that while I strive to provide helpful and accurate information, I am just a tool, and decisions related to loan applications are ultimately made by human lenders or financial institutions. These decisions should be made in accordance with applicable laws and regulations, which prohibit discrimination in lending based on certain protected characteristics. Discrimination in lending is illegal and unethical.

Figure I: ChatGPT Discusses Discrimination in Lending
This figure presents a conversation between the authors and ChatGPT on its fairness as an automated decision maker in evaluating loan applications in March 2024.

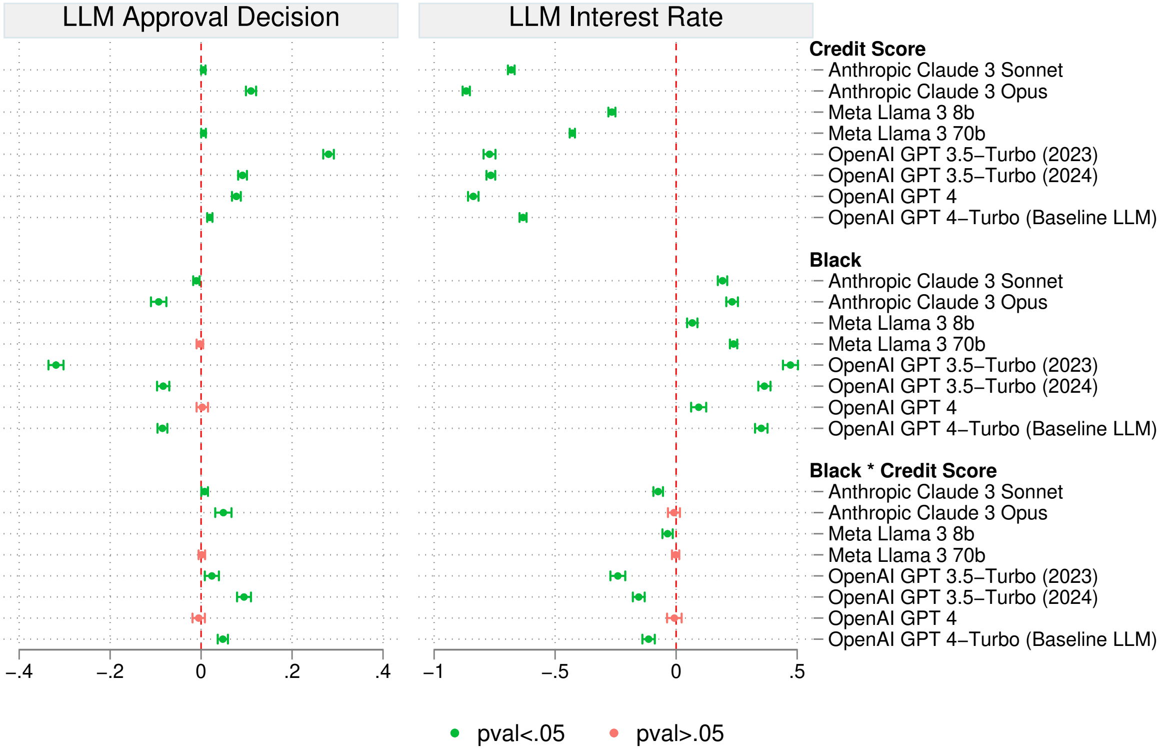


Figure II: Mortgage Underwriting Decisions by Leading LLMs
This figure illustrates the estimated coefficients from Experiment 2, which estimates Equation 2 with various leading LLM models. Coefficients that are statistically significant at the 5% level are shown in green and are red otherwise. As shown in Table IV, the Llama 3 8b model recommends approval for 100% of loans and is thus omitted from the approval subfigure.

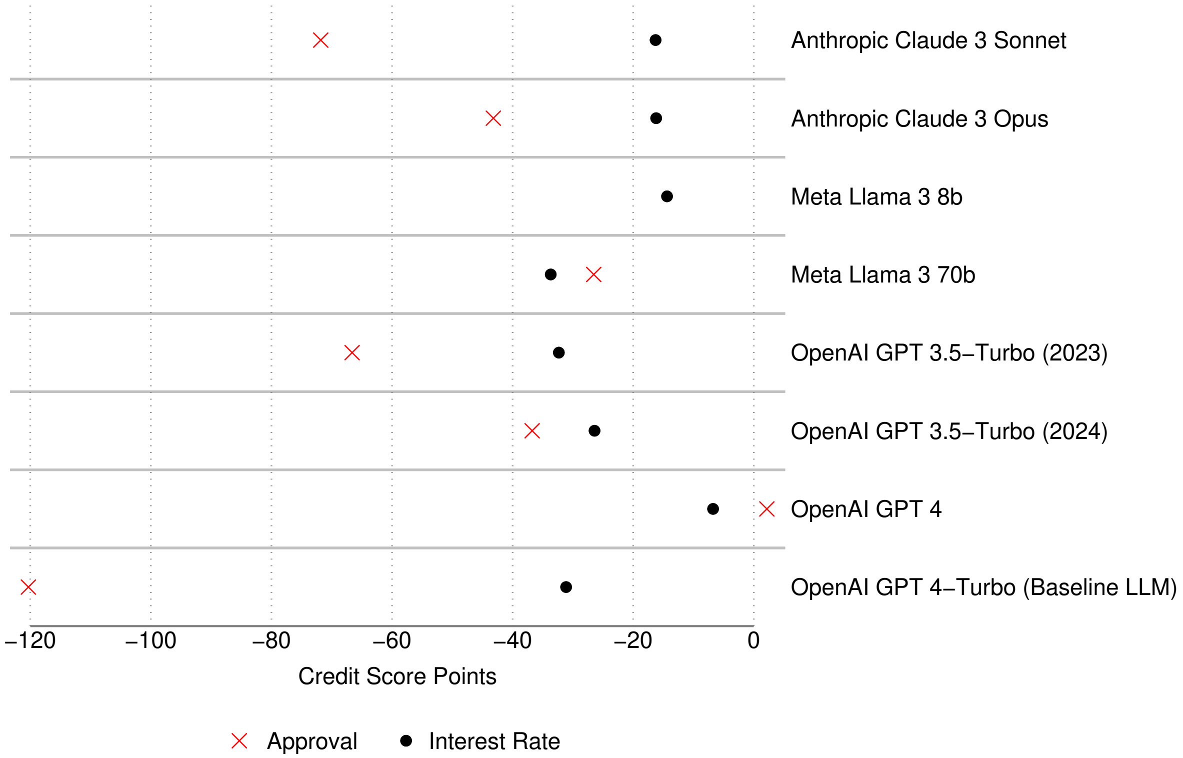


Figure III: Economic Magnitude of Racial Disparities by Leading LLMs
This figure illustrates the credit score differences required to generate Black–white LLM recommendation outcome disparities estimated using Equation 1 as in columns (1) of Table III. For each LLM in Appendix Table A5, we estimate an analogous regression and report $\hat{\beta}_B/\hat{\beta}_{CS}$ multiplied by the sample standard deviation of credit scores (approximately 61 points). Coefficients marked with a red cross correspond to models where the dependent variable is loan approval, and coefficients marked with a black circle correspond to models where the dependent variable is interest rate. As shown in Table IV, the Llama 3 8b model recommends approval for 100% of loans and is thus has no approval estimate.

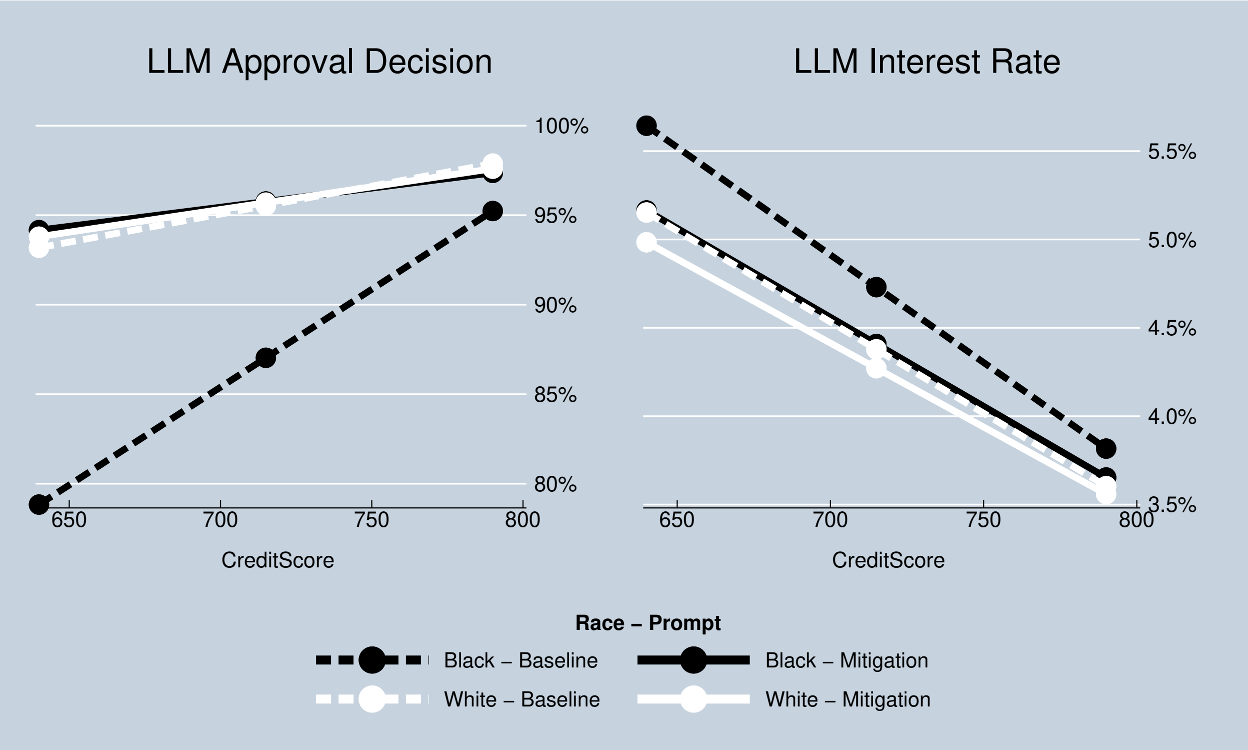


Figure IV: The Mitigation Prompt Reduces the Level and Credit-Sensitivity of LLM Bias

This figure illustrates the estimated coefficients for Equation 4 in Experiment 3 as reported in columns (2) and (4) of Table V for the approval and interest rate decisions of the baseline LLM. We obtain the predicted values for all observations after running both models to recover the loan fixed effects, and plot the outcomes averaged by score.

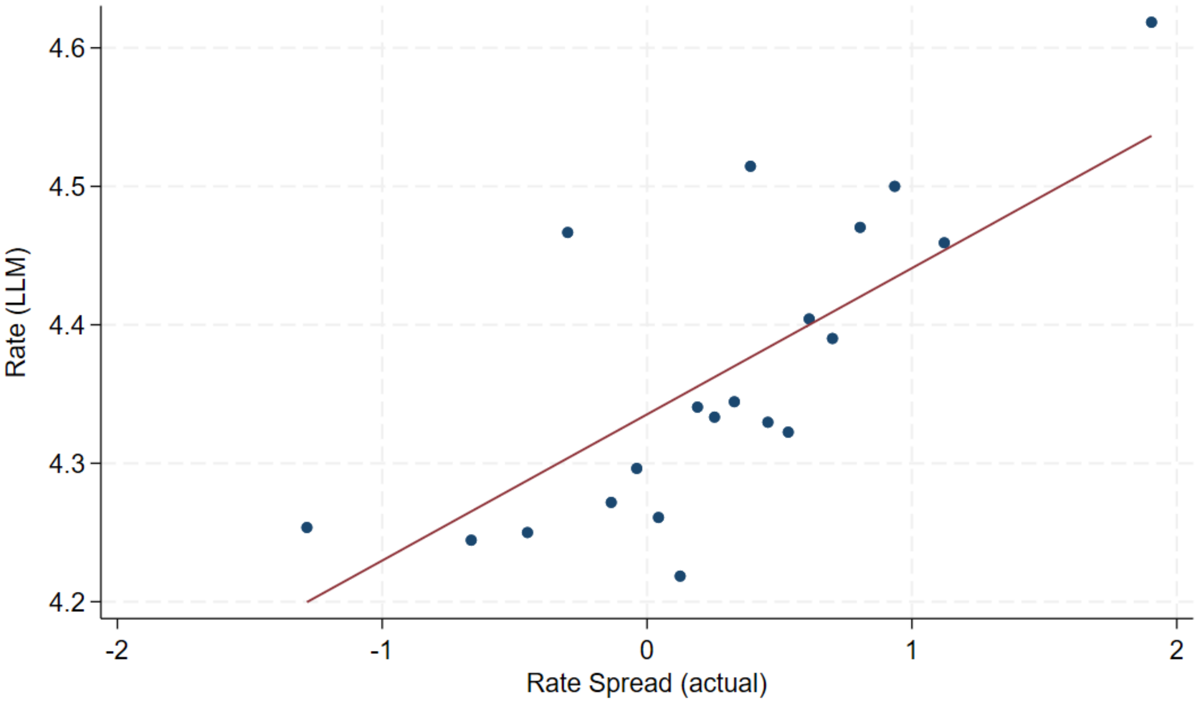


Figure V: LLM Interest Rate Recommendation vs. Actual Loan Rate Spread
This binned scatterplot illustrates the bivariate relationships between the mortgage interest rate recommended by ChatGPT 4 Turbo and the actual rate spread assigned by the real lender to the same loan as recorded in HMDA. Loan applications come from Experiment 4, where no demographic is included in the application but credit score is manipulated. The dependent variable is the interest rate recommended by LLM, and the independent variable is the actual rate spread. The estimated slope of the linear fit is 0.106, with a t-statistic of 4.70 based on a heteroskedastic robust standard error.

Table I: Experiment Designs and Sample Size

This table presents the full scope of the experimental variations used in our audit design. For each experiment, we manipulate the demographic information assigned to the loan applicant and the credit score, and then include them in the prompt listed in Section 2. The mitigation prompt(s) add instructions to reduce bias in LLM responses and are described in Section 4. We then pass the full prompt to the LLM listed below. *N* is the resulting number of observations in the experiment. Experiment 2 does not have 48,000 observations, because Claude occasionally refuses to answer when demographic information is included. In such cases, we repeat the application request up to 10 times.

	All 1,000 loan applications with all combinations of				
Experiment	Demographics	Prompt	Credit Score	LLM	N
1	{Black, White}	Baseline	{640, 715, 790}	GPT 4-Turbo	6,000
2	{Black, White}	Baseline	{640, 715, 790}	{Eight LLMs listed in Table A5}	47,206
3	{Black, White}	{Baseline, Mitigation}	{640, 715, 790}	GPT 4-Turbo	12,000
4	None	Baseline	{640, 715, 790}	GPT 4-Turbo	3,000
A1	{Asian, Black, Hispanic, None, White}	Baseline	{640, 715, 790}	GPT 4-Turbo	15,000
A2	{Age 30, Age 50, Age 70}	{Baseline, Mitigation}	{640, 715, 790}	GPT 4-Turbo	18,000
A3	{Female, Male}	{Baseline, Mitigation}	{640, 715, 790}	GPT 4-Turbo	12,000
A4	{Black, White}	{Baseline, Mitigation, Alt. Mitigation}	{640, 715, 790}	GPT 4-Turbo	18,000

Table II: Summary Statistics

Panel A reports summary statistics for the 1,000 observations we randomly selected from HMDA to fill out the loan applications. In addition, prompts are stratified over experimentally manipulated credit scores of 640, 715, and 790, giving a standard deviation of approximately 61 points (and a mean of 715). Panel B reports summary statistics of the LLM recommendations from each experiment listed in Table I. Variables are defined in Section 2. Approval in both panels is binary, and all other variables are reported as percentages from 0 to 100. We do not report information about the manipulated variables (demographic information and credit score), as they are evenly balanced within each experiment.

Panel A: HMDA Loan Sample Variables

	N	Mean	Std.	Min	Median	Max
Approval (Actual)	1,000	0.92	0.27	0.00	1.00	1.00
Rate (Actual, %)	921	4.98	1.13	2.22	5.00	9.88
Rate Spread (Actual, %)	909	0.27	0.72	-5.33	0.30	5.20
DTI (%)	1,000	37.17	9.37	20.00	38.00	60.00
LTV (%)	1,000	83.22	14.52	15.71	85.00	105.22

Panel B: Experimental Outcome Variables

	Experiment							
	1	2	3	4	A1	A2	A3	A4
Approval (LLM)								
N	6,000	47,206	12,000	3,000	15,000	18,000	12,000	18,000
Mean	0.91	0.87	0.94	0.95	0.93	0.94	0.95	0.92
Std.	0.28	0.33	0.25	0.22	0.25	0.24	0.21	0.27
Rate (LLM)								
N	6,000	47,206	12,000	3,000	15,000	18,000	12,000	18,000
Mean	4.55	4.75	4.45	4.43	4.49	4.49	4.35	4.53
Std.	1.02	1.09	0.94	0.91	0.97	0.96	0.90	1.01
Min	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00
Median	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
Max	7.50	9.50	7.50	7.50	7.50	7.50	7.50	7.50

Table III: Race and Recommendations (Baseline LLM)

This table reports the OLS regressions of loan approval recommendations (Panel A) and loan interest rate recommendations (Panel B) on loan applicants' racial identity. The dependent variable in Panel A is the LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. In Panel B, the dependent variable is the LLM loan interest rate recommendations measured in percentage points. Variables are defined in Section 2. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects.

Panel A: Loan Approval Recommendations

	(1)	(2)	(3)	(4)	(5)
CreditScore (z)	0.043***	0.019***	0.043***	0.043***	0.019***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black	-0.085***	-0.085***	-0.085***	-0.085***	-0.085***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Black × CreditScore (z)		0.048***			0.048***
		(0.005)			(0.005)
Black × DTI (z)			-0.063***		-0.060***
			(0.006)		(0.006)
Black × LTV (z)				-0.042***	-0.035***
				(0.005)	(0.005)
Obs	6,000	6,000	6,000	6,000	6,000
R^2	0.57	0.58	0.58	0.58	0.59
Adj R^2	0.48	0.49	0.50	0.49	0.51
Loan FE	Yes	Yes	Yes	Yes	Yes
Experiment	1	1	1	1	1

Panel B: Loan Interest Rate Recommendation

	(1)	(2)	(3)	(4)	(5)
CreditScore (z)	-0.689***	-0.633***	-0.689***	-0.689***	-0.633***
	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)
Black	0.352***	0.352***	0.352***	0.352***	0.352***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Black × CreditScore (z)		-0.114***			-0.114***
		(0.011)			(0.011)
Black × DTI (z)			0.091***		0.085***
			(0.013)		(0.013)
Black × LTV (z)				0.065***	0.056***
				(0.011)	(0.011)
Obs	6,000	6,000	6,000	6,000	6,000
R^2	0.85	0.86	0.85	0.85	0.86
Adj R^2	0.82	0.83	0.82	0.82	0.83
Loan FE	Yes	Yes	Yes	Yes	Yes
Experiment	1	1	1	1	1

Table IV: Race and Recommendations (LLM Comparison)

This table reports the OLS regressions of loan approval recommendations (Panel A) and loan interest rate recommendations (Panel B) on loan applicants' racial identity based on responses collected from eight leading LLMs. We estimate Equation 1, replicating Experiment 1 with other leading LLM models. Variables are defined in Section 2. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects. Note that "llama3-70b-8192" recommends approval for 100% of loan applications in our sample, which precludes the possibility of running the regression of loan approval recommendations in Panel A, column (3). The coefficients here are presented visually in Figure II.

Panel A: Loan Approval Recommendations

Family	Anthropic Claude 3		Meta Llama 3		OpenAI GPT			
Model	Sonnet	Opus	8b	70b	3.5 Turbo	3.5 Turbo	4	4-Turbo
Date	2024	2024	2024	2024	2023	2024	2023	2024
(#)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CreditScore (z)	0.005***	0.110***		0.005***	0.280***	0.091***	0.078***	0.019***
	(0.001)	(0.005)		(0.001)	(0.006)	(0.005)	(0.004)	(0.003)
Black	-0.011***	-0.098***		-0.003	-0.319***	-0.083***	0.003	-0.085***
	(0.002)	(0.008)		(0.002)	(0.008)	(0.007)	(0.006)	(0.005)
Black × CreditScore (z)	0.008***	0.040***		0.002	0.024***	0.094***	-0.005	0.048***
	(0.002)	(0.008)		(0.002)	(0.008)	(0.008)	(0.006)	(0.005)
Obs	5,989	5,215	6,000	6,000	6,000	6,000	6,000	6,000
R^2	0.81	0.64	.	0.68	0.65	0.50	0.65	0.58
Adj R^2	0.77	0.55	.	0.61	0.57	0.40	0.59	0.49
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Experiment	2	2	2	2	2	2	2	2
Avg(y)	0.97	0.80	1.00	0.99	0.58	0.86	0.87	0.91
Avg(y | White)	0.97	0.84	1.00	0.99	0.74	0.90	0.87	0.96
Avg(y | Black)	0.96	0.74	1.00	0.99	0.42	0.82	0.87	0.87
White Answer Rate (%)	99.83	99.57	100.00	100.00	100.00	100.00	100.00	100.00
Black Answer Rate (%)	99.80	74.27	100.00	100.00	100.00	100.00	100.00	100.00

Panel B: Loan Interest Rate Recommendations

Family	Anthropic Claude 3		Meta Llama 3		OpenAI GPT			
Model	Sonnet	Opus	8b	70b	3.5 Turbo	3.5 Turbo	4	4-Turbo
Date	2024	2024	2024	2024	2023	2024	2023	2024
(#)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CreditScore (z)	-0.682***	-0.867***	-0.265***	-0.429***	-0.771***	-0.766***	-0.838***	-0.633***
	(0.006)	(0.007)	(0.005)	(0.004)	(0.013)	(0.009)	(0.010)	(0.007)
Black	0.193***	0.238***	0.067***	0.237***	0.472***	0.365***	0.093***	0.352***
	(0.008)	(0.011)	(0.007)	(0.006)	(0.016)	(0.012)	(0.013)	(0.011)
Black × CreditScore (z)	-0.074***	0.002	-0.035***	-0.002	-0.241***	-0.154***	-0.007	-0.114***
	(0.009)	(0.011)	(0.007)	(0.006)	(0.015)	(0.012)	(0.014)	(0.011)
Obs	5,989	5,215	6,000	6,000	6,000	6,000	6,000	6,000
R^2	0.90	0.91	0.66	0.89	0.78	0.85	0.84	0.86
Adj R^2	0.89	0.89	0.59	0.87	0.73	0.82	0.81	0.83
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Experiment	2	2	2	2	2	2	2	2
Avg(y)	5.52	5.64	4.32	4.29	4.65	4.47	4.63	4.55
Avg(y | White)	5.42	5.54	4.29	4.17	4.42	4.29	4.59	4.38
Avg(y | Black)	5.61	5.78	4.36	4.40	4.89	4.66	4.68	4.73
White Answer Rate (%)	99.83	99.57	100.00	100.00	100.00	100.00	100.00	100.00
Black Answer Rate (%)	99.80	74.27	100.00	100.00	100.00	100.00	100.00	100.00

Table V: Recommendation Bias Mitigation Prompt (Baseline LLM)

This table reports the OLS regressions of loan approval recommendations (columns 1–2) and loan interest rate recommendations (columns 3–4) on loan applicants' racial identity, leveraging an experiment where the LLM is explicitly instructed to make unbiased loan recommendation decisions. The dependent variable in columns (1)–(2) is a binary variable that equals one if the loan is approved, and zero otherwise, and the LLM loan interest rate recommendations measured in percentage points in Columns (3)–(4). To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects. Variables are defined in Section 2.

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.043***	0.019***	-0.689***	-0.633***
	(0.003)	(0.003)	(0.006)	(0.006)
Black	-0.085***	-0.085***	0.352***	0.352***
	(0.005)	(0.005)	(0.011)	(0.011)
Black × CreditScore (z)		0.048***		-0.114***
		(0.005)		(0.011)
Mitigation	0.002	0.002	-0.107***	-0.107***
	(0.003)	(0.003)	(0.008)	(0.008)
Mitigation × CreditScore (z)	-0.029***	-0.004	0.090***	0.050***
	(0.003)	(0.004)	(0.007)	(0.008)
Mitigation × Black	0.086***	0.086***	-0.214***	-0.214***
	(0.006)	(0.006)	(0.014)	(0.014)
Mitigation × Black × CreditScore		-0.050***		0.079***
		(0.006)		(0.014)
Obs	12,000	12,000	12,000	12,000
R^2	0.58	0.58	0.85	0.85
Adj R^2	0.54	0.55	0.84	0.84
Loan FE	Yes	Yes	Yes	Yes
Experiment	3	3	3	3
p-val: $\beta_{\mathrm{B}} + \beta_{\mathrm{M\times B}} = 0$	0.83	0.83	0.00	0.00
p-val: $\beta_{\mathrm{B\times CS}} + \beta_{\mathrm{M\times B\times CS}} = 0$		0.47		0.00

Table VI: LLM Loan Approval Recommendations vs. Actual Approval Decision

This table summarizes the performance of the LLM in assigning loan approval recommendations (LLM_A vs. LLM_D) in comparison to the actual loan approval decisions (True_A vs. True_D). Panel A shows a confusion matrix, and precision and recall measures are reported in Panel B. The sample is Experiment 4, in which the prompts contain no demographic information but do manipulate the credit score provided.

Panel A: Confusion Matrix

	True application outcome		
LLM Recommendation	True_A	True_D	Total
LLM_A	2687	155	2842
LLM_D	76	82	158
Total	2763	237	3000

Panel B: Precision and Recall

Statistic	Definition	Value
Approval Recall	$\Pr(\ \text{LLM}_A \vert\ \text{True}_A\ )$	97.2%
Approval Precision	$\Pr(\ \text{True}_A \vert\ \text{LLM}_A\ )$	94.5%
Denial Recall	$\Pr(\ \text{LLM}_D \vert\ \text{True}_D\ )$	34.6%
Denial Precision	$\Pr(\ \text{True}_D \vert\ \text{LLM}_D\ )$	51.9%

Table VII: Race and Recommendations with a Logit Model (Baseline LLM)

This table repeats tests of Equations 1 and 2 with logistic regressions. We do not include loan fixed effects. The dependent variable is the LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variables are defined in Section 2.

	(1)	(2)	(3)	(4)	(5)
CreditScore (z)	0.593***	0.473***	0.668***	0.622***	0.528***
	(0.052)	(0.097)	(0.055)	(0.053)	(0.100)
Black	-1.186***	-1.111***	-1.326***	-1.336***	-1.279***
	(0.105)	(0.115)	(0.206)	(0.217)	(0.335)
Black × CreditScore (z)		0.167			0.270**
		(0.115)			(0.122)
DTI (z)			-1.185***		-1.198***
			(0.175)		(0.192)
Black × DTI (z)			-0.006		-0.085
			(0.195)		(0.220)
LTV (z)				-1.240***	-1.227***
				(0.288)	(0.299)
Black × LTV (z)				0.167	0.058
				(0.315)	(0.335)
Constant	3.217***	3.162***	3.815***	3.681***	4.219***
	(0.094)	(0.097)	(0.192)	(0.204)	(0.303)
Obs	6,000	6,000	6,000	6,000	6,000
Pseudo R^2	0.08	0.08	0.22	0.16	0.29
Loan FE	No	No	No	No	No
Experiment	1	1	1	1	1

Table VIII: Bias Mitigation Prompts with a Logit Model (Baseline LLM)

This table repeats tests analogous to Equation 3 (in column 1) and Equation 4 (in column 3) using logistic regression. We do not include loan fixed effects. The dependent variable is the LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. To facilitate interpretation, both coefficient estimates and corresponding odds ratios are reported, and (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. Variables are defined in Section 2.

	(1)		(2)	
	Coef.	Odds	Coef.	Odds
CreditScore (z)	0.593***	1.809***	0.473***	1.604***
	(0.052)	(0.094)	(0.097)	(0.156)
Black	-1.186***	0.305***	-1.111***	0.329***
	(0.105)	(0.032)	(0.115)	(0.038)
Black × CreditScore (z)			0.167	1.182
			(0.115)	(0.136)
Mitigation	-0.056	0.945	0.009	1.009
	(0.132)	(0.125)	(0.137)	(0.138)
Mitigation × CreditScore (z)	-0.233***	0.793***	-0.082	0.921
	(0.084)	(0.067)	(0.136)	(0.125)
Mitigation × Black	1.203***	3.329***	1.108***	3.029***
	(0.166)	(0.552)	(0.178)	(0.538)
Mitigation × Black × CreditScore			-0.227	0.797
			(0.176)	(0.140)
Constant	3.217***	24.962***	3.162***	23.623***
	(0.094)	(2.353)	(0.097)	(2.302)
Obs	12,000		12,000	
Pseudo R^2	0.07		0.07	
Loan FE	No		No	
Experiment	3		3	

Table IX: Race and Recommendations Robustness (Baseline LLM)

This table reports robustness tests for OLS regressions of loan approval recommendations and loan interest rate recommendations on loan applicants' racial identity presented in Table III. Experiments in this table were run in July 2024, while those in Table III were run in April 2024. In columns (1)–(4) we set the model temperature to 0, as in Table III. In columns (5)–(4) we set the model temperature to 0.3. Variables are defined in Section 2. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects.

Temperature:	Temperature 0				Temperature 0.3			
Dependent Variable:	Approval		Interest Rate		Approval		Interest Rate	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CreditScore (z)	0.065***	0.038***	-0.675***	-0.645***	0.058***	0.031***	-0.663***	-0.635***
	(0.003)	(0.004)	(0.007)	(0.008)	(0.003)	(0.004)	(0.007)	(0.008)
Black	-0.135***	-0.135***	0.440***	0.440***	-0.129***	-0.128***	0.443***	0.443***
	(0.006)	(0.006)	(0.013)	(0.013)	(0.006)	(0.006)	(0.013)	(0.013)
Black × CreditScore (z)		0.054***		-0.060***		0.056***		-0.056***
		(0.007)		(0.013)		(0.007)		(0.013)
Obs	5,978	5,978	5,978	5,978	5,925	5,925	5,925	5,925
R^2	0.58	0.58	0.83	0.83	0.57	0.58	0.83	0.83
Adj R^2	0.49	0.50	0.80	0.80	0.48	0.49	0.79	0.79
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Experiment	1.1	1.1	1.1	1.1	1.2	1.2	1.2	1.2

Appendix

- Table A1 compares summary stats of our HMDA subsample to the broader HMDA sample.
- Table A2 repeats the main heterogeneity tests of Table III, adding Asian or Hispanic as a listed race/ethnicity, and also including loans without any race/ethnicity disclosure (Experiment A1).
- Table A3 examines an experiment where we considered prompts submitted to the baseline LLM including "- Age: 30," "- Age: 50," or "- Age: 70" in place of race signals (Experiment A2).
- Table A4 examines an experiment where we considered prompts submitted to the baseline LLM including "- Gender: Male" or "- Gender: Female" in place of race signals (Experiment A3).
- Table A5 lists the LLMs used in our study.
- Table A6 considers an alternate "mitigation" prompt (Experiment A4).
- Our API call functions are below. To improve reproducibility, we set the response temperature to zero for all calls and, where possible, set seeds in the API calls. API arguments not listed take their default values for the versions of the packages we used. Package versions are listed below.[32]

```python
from openai import OpenAI  # 1.14.2
from anthropic import Anthropic  # 0.25.7
from groq import Groq  # 0.5.0

# Function to load API keys
def load_api_key(file_path):
    with open(file_path, 'r') as f:
        return f.read().strip()

# Initialize clients with default params and response unpacking instructions
clients = {
    'openai': {
        'client': OpenAI(api_key=load_api_key('api_keys/openai.txt')),
        'params': {
            'model': "gpt-4-0125-preview",
            'temperature': 0.0,
            'max_tokens': 20,
```

[32] Note that despite taking these steps, LLM responses remain stochastic and are not perfectly reproducible due to what OpenAI refers to as "the inherent non-determinism of our models" (https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter).

```
            'seed': 42,
            'messages': [{"role": "user", "content": None}],  # Placeholder
        },
        'response_unpack': lambda response: (
            response.choices[0].message.content,
            response.system_fingerprint,
            response.usage.prompt_tokens,
            response.usage.completion_tokens
        )
    },
    'anthropic': {
        'client': Anthropic(api_key=load_api_key('api_keys/anthropic.txt')),
        'params': {
            'model': "claude-3-opus-20240229",
            'temperature': 0.0,
            'max_tokens': 400,
            'messages': [{"role": "user", "content": None}],  # Placeholder
        },
        'response_unpack': lambda response: (
            response.content[0].text,
            response.id,
            response.usage.input_tokens,
            response.usage.output_tokens
        )
    },
    'groq': {
        'client': Groq(api_key=load_api_key('api_keys/groq.txt')),
        'params': {
            'model': "llama3-70b-8192",
            'temperature': 0.0,
            'max_tokens': 8,
            'messages': [{"role": "user", "content": None}],  # Placeholder
        },
        'response_unpack': lambda response: (
            response.choices[0].message.content.strip().replace(' ', ''),
            response.system_fingerprint,
            response.usage.prompt_tokens,
            response.usage.completion_tokens
        )
    }
}

# General function to get response
def get_api_response(client_name, text, **kwargs):
```

```python
    client_info = clients[client_name]
    client = client_info['client']
    params = client_info['params'].copy() # Grab default params
    params.update(kwargs)  # Overwrite/add with any kwargs passed to the function
    params['messages'][0]['content'] = text # Update the message content

    if client_name == 'anthropic':
        response = client.messages.create(**params)
    else:
        response = client.chat.completions.create(**params)

    return client_info['response_unpack'](response)
```

Table A1: Comparing Entire HMDA Dataset to HMDA Loan Sample

This table compares the HMDA universe ("Entire 2022 HMDA") to the subset of 1,000 HMDA observations used in our study ("Study Subset"). The HMDA data comes from the Loan/Application Records (LAR) file containing loans made nationwide in 2022 and reported to the Consumer Financial Protection Bureau. We restrict the sample to conventional 30-year loans for principal residences secured by a first lien. We eliminate loans with balloon payments, negative amortization, interest-only payments, or business or commercial purposes. We also discard manufactured homes, reverse mortgages, and multi-unit dwellings. Finally, we require non-missing DTI and LTV information for each loan. After these filters, the HMDA dataset has 2,409,013 observations. We winsorize variables at the 1% tails for this table to remove outliers in the entire sample, but this choice does not cause p-values to cross any significance thresholds. We report the mean (and standard deviations, in square brackets) for the variables used in the study in the entire HMDA dataset and the study subset separately. The last column reports differences in means, and standard errors are shown in parentheses, where ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Entire 2022 HMDA	Study Subset	Difference
Approval (actual)	0.926	0.921	-0.005
	[0.261]	[0.270]	(0.008)
Rate (actual)	4.934	4.974	0.040
	[1.127]	[1.106]	(0.037)
Rate Spread (actual)	0.280	0.272	-0.008
	[0.631]	[0.637]	(0.021)
DTI	37.043	37.172	0.129
	[9.205]	[9.367]	(0.291)
LTV	82.427	83.236	0.809
	[14.971]	[14.393]	(0.473)

Table A2: Race, Ethnicity, and Recommendations (Baseline LLM)

This table repeats the main tests in Table III using Experiment A1 (see Table I), which expands the list of demographics used in the application prompt to include *Asian*, *Hispanic*, or none. We report OLS regressions of loan approval recommendations (Panel A) and loan interest rate recommendations (Panel B) on loan applicants' racial identity and ethnicity. The dependent variable in Panel A is a binary variable that equals one if the loan is approved, and zero otherwise. In Panel B, the dependent variable is the LLM loan interest rate recommendations, measured in percentage points. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects. Variables are defined in Section 2.

Panel A: Loan Approval Recommendations

	(1)	(2)	(3)	(4)
CreditScore (z)	0.033***	0.023***	0.033***	0.033***
	(0.001)	(0.003)	(0.001)	(0.001)
Asian	-0.001	-0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)
Black	-0.077***	-0.077***	-0.077***	-0.077***
	(0.005)	(0.005)	(0.005)	(0.005)
Hispanic	-0.012***	-0.012***	-0.012***	-0.012***
	(0.004)	(0.004)	(0.004)	(0.004)
White	0.008**	0.008**	0.008**	0.008**
	(0.003)	(0.003)	(0.003)	(0.003)
Asian × CreditScore (z)		0.000		
		(0.004)		
Black × CreditScore (z)		0.044***		
		(0.005)		
Hispanic × CreditScore (z)		0.009**		
		(0.004)		
White × CreditScore (z)		-0.004		
		(0.004)		
Asian × DTI (z)			0.005	
			(0.004)	
Black × DTI (z)			-0.049***	
			(0.006)	
Hispanic × DTI (z)			0.001	
			(0.004)	
White × DTI (z)			0.014***	
			(0.004)	
Asian × LTV (z)				0.001
				(0.003)
Black × LTV (z)				-0.037***
				(0.004)
Hispanic × LTV (z)				-0.008**
				(0.004)
White × LTV (z)				0.005
				(0.003)
Obs	15,000	15,000	15,000	15,000
R^2	0.59	0.60	0.60	0.60
Adj R^2	0.56	0.57	0.57	0.57
Loan FE	Yes	Yes	Yes	Yes
Experiment	A1	A1	A1	A1

Panel B: Loan Interest Rate Recommendations

	(1)	(2)	(3)	(4)
CreditScore (z)	-0.665***	-0.663***	-0.665***	-0.665***
	(0.003)	(0.006)	(0.003)	(0.003)
Asian	-0.062***	-0.062***	-0.062***	-0.062***
	(0.009)	(0.008)	(0.009)	(0.009)
Black	0.301***	0.301***	0.301***	0.301***
	(0.011)	(0.011)	(0.011)	(0.011)
Hispanic	0.117***	0.117***	0.118***	0.117***
	(0.008)	(0.008)	(0.008)	(0.008)
White	-0.051***	-0.051***	-0.051***	-0.051***
	(0.008)	(0.008)	(0.008)	(0.008)
Asian × CreditScore (z)		0.047***		
		(0.009)		
Black × CreditScore (z)		-0.084***		
		(0.011)		
Hispanic × CreditScore (z)		-0.002		
		(0.009)		
White × CreditScore (z)		0.030***		
		(0.008)		
Asian × DTI (z)			0.021**	
			(0.010)	
Black × DTI (z)			0.081***	
			(0.012)	
Hispanic × DTI (z)			0.007	
			(0.010)	
White × DTI (z)			-0.010	
			(0.010)	
Asian × LTV (z)				0.016*
				(0.009)
Black × LTV (z)				0.072***
				(0.011)
Hispanic × LTV (z)				0.012
				(0.009)
White × LTV (z)				0.006
				(0.009)
Obs	15,000	15,000	15,000	15,000
R^2	0.86	0.86	0.86	0.86
Adj R^2	0.85	0.85	0.85	0.85
Loan FE	Yes	Yes	Yes	Yes
Experiment	A1	A1	A1	A1

Table A3: Age and Recommendations (Baseline LLM)

This table reports the OLS regressions of loan approval recommendations (columns 1–3) and loan interest rate recommendations (columns 4–6) on loan applicants' age. The dependent variable in columns (1)–(3) is the LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. In columns (4)–(6), the dependent variable is the LLM loan interest rate recommendations measured in percentage points. Variables are defined in Section 2. Tests in columns (1), (2), (4), and (5), only include observations with the baseline prompt. Columns (3) and (6) include observations with the mitigation prompt. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects.

	Approval			Interest Rate		
	(1)	(2)	(3)	(4)	(5)	(6)
CreditScore (z)	0.029***	0.023***	0.029***	-0.677***	-0.663***	-0.677***
	(0.002)	(0.003)	(0.002)	(0.004)	(0.006)	(0.004)
Age=50	-0.003	-0.003	-0.003	0.039***	0.039***	0.039***
	(0.003)	(0.003)	(0.003)	(0.008)	(0.008)	(0.008)
Age=70	-0.016***	-0.016***	-0.016***	0.173***	0.173***	0.173***
	(0.004)	(0.004)	(0.004)	(0.009)	(0.009)	(0.009)
Age=50 × CreditScore (z)		0.004			-0.010	
		(0.004)			(0.009)	
Age=70 × CreditScore (z)		0.013***			-0.030***	
		(0.004)			(0.009)	
Mitigation			-0.004			-0.100***
			(0.004)			(0.008)
Mitigation × CreditScore (z)			0.001			0.010**
			(0.002)			(0.005)
Mitigation × Age=50			0.003			-0.001
			(0.005)			(0.012)
Mitigation × Age=70			0.005			-0.003
			(0.005)			(0.012)
Obs	9,000	9,000	18,000	9,000	9,000	18,000
R^2	0.70	0.70	0.65	0.90	0.90	0.89
Adj R^2	0.66	0.66	0.63	0.89	0.89	0.88
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes
Experiment	A2	A2	A2	A2	A2	A2

Table A4: Gender and Recommendations (Baseline LLM)

This table reports the OLS regressions of loan approval recommendations (columns 1–3) and loan interest rate recommendations (columns 4–5) on loan applicants' gender. The dependent variable in columns (1)–(3) is the LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. In columns (4)–(6), the dependent variable is the LLM loan interest rate recommendations measured in percentage points. Variables are defined in Section 2. Tests in columns (1), (2), (4), and (5), only include observations with the baseline prompt. Columns (3) and (6) include observations with the mitigation prompt. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects.

	Approval			Interest Rate		
	(1)	(2)	(3)	(4)	(5)	(6)
CreditScore (z)	0.024***	0.024***	0.024***	-0.650***	-0.654***	-0.650***
	(0.002)	(0.003)	(0.002)	(0.004)	(0.006)	(0.004)
Female	0.005	0.005	0.005	-0.005	-0.005	-0.005
	(0.003)	(0.003)	(0.003)	(0.008)	(0.008)	(0.008)
Female × CreditScore (z)		-0.000			0.008	
		(0.004)			(0.009)	
Mitigation			0.004			-0.111***
			(0.003)			(0.008)
Mitigation × CreditScore (z)			-0.002			0.021***
			(0.003)			(0.006)
Mitigation × Female			-0.001			-0.007
			(0.005)			(0.012)
Obs	6,000	6,000	12,000	6,000	6,000	12,000
R^2	0.69	0.69	0.66	0.90	0.90	0.89
Adj R^2	0.63	0.63	0.63	0.88	0.88	0.88
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes
Experiment	A3	A3	A3	A3	A3	A3

Table A5: LLMs Considered

This table lists the eight different LLMs considered in our study. Test results based on these LLMs are reported in Figure II and Table IV.

Source	LLM	Year	Model API Name
Anthropic	Claude 3 Sonnet	2024	claude-3-sonnet-20240229
Anthropic	Claude 3 Opus	2024	claude-3-opus-20240229
Meta	Llama 3 8b	2024	llama3-8b-8192 (run via Groq)
Meta	Llama 3 70b	2024	llama3-7b-8192 (run via Groq)
OpenAI	GPT 3.5-Turbo (2023)	2023	gpt-3.5-turbo-0613
OpenAI	GPT 3.5-Turbo (2024)	2024	gpt-3.5-turbo-0125
OpenAI	GPT 4	2023	gpt-4-0613
OpenAI	GPT 4-Turbo [Baseline LLM]	2024	gpt-4-0125-preview

Table A6: Alternate Bias Mitigation Prompt (Baseline LLM)

This table repeats tests of Equation 3 in Table V using an alternative mitigation prompt: "Make sure you comply with the Fair Lending Act and ECOA in making this decision." These estimates are in Columns (2) and (4); for comparison, Columns (1) and (3) reprise the results from the same columns of Table V (using our main mitigation prompt: "You should use no bias in making this decision"). Each regression uses observations generated by the baseline prompt and one mitigation prompt. The dependent variable in columns (1) and (2) is a binary variable that equals one if the loan is approved, and zero otherwise. In columns (3) and (4), the dependent variable is the LLM loan interest rate recommendations measured in percentage points. To facilitate interpretation, (z) indicates a variable has been standardized. Heteroskedastic robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. All models include loan fixed effects. Variables are defined in Section 2.

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
Mitigation Prompt:	Main	Alternate	Main	Alternate
CreditScore (z)	0.043***	0.043***	-0.689***	-0.689***
	(0.003)	(0.003)	(0.006)	(0.006)
Black	-0.085***	-0.085***	0.352***	0.352***
	(0.005)	(0.005)	(0.011)	(0.011)
Mitigation	0.002	-0.042***	-0.107***	0.179***
	(0.003)	(0.005)	(0.008)	(0.011)
Mitigation × CreditScore (z)	-0.029***	0.009**	0.090***	-0.064***
	(0.003)	(0.004)	(0.007)	(0.009)
Mitigation × Black	0.086***	0.061***	-0.214***	-0.104***
	(0.006)	(0.007)	(0.014)	(0.017)
Obs	12,000	12,000	12,000	12,000
R^2	0.58	0.56	0.85	0.83
Adj R^2	0.54	0.52	0.84	0.81
Loan FE	Yes	Yes	Yes	Yes
Experiment	A4	A4	A4	A4