

Measuring and Mitigating Racial Disparities in LLMs: Evidence from a Mortgage Underwriting Experiment

Donald E. Bowen III
Lehigh University

S. McKay Price
Lehigh University

Luke C.D. Stein
Babson College

Ke Yang
Lehigh University

This version: August 1, 2025

We evaluate LLM responses to a mortgage underwriting task using real loan application data. Experimentally manipulated race is signaled explicitly or through borrower name/location proxies. Multiple generations of LLMs recommend more denials and higher interest rates for Black applicants than otherwise-identical white applicants, with larger disparities for riskier loans. Simple prompt engineering can cost-effectively mitigate these patterns. Race-blind recommendations correlate strongly with real lender decisions and predict delinquency, but LLMs incorporate racial signals when available despite similar delinquency rates across groups. Our findings show potential costs of adopting this new technology in financial settings and raise important questions for regulators.

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, Neil Bhutta, Sean Cao, James Conklin, Eli Fich, Kathleen Hanley, Philip Kalikman, Danielle Kent, Karthik Krishnan, Jesús Villota Miranda, Jordan Nickerson, Robert Parham, Jesus Salas, Kalinda Ukanwa, and James Weston; seminar participants at University of Amsterdam, Northern Kentucky University, University of Oklahoma, and Pennsylvania State University; and attendees at the 2025 European Real Estate Society Annual Conference, 2025 SFS Cavalcade, 2025 Boston Area Finance Symposium, 2025 Bretton Woods Accounting and Finance Conference, 2025 American Real Estate Society Annual Conference, 2025 Northeastern University Finance Conference, 2024 New Zealand Finance Meeting, 2024 Concordia University Generative AI in Finance Conference, 2024 Research in Behavioral Finance Conference, and the 2024 Real Estate Finance and Investment Symposium 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.

We're also exploring the potential that generative AI (GenAI) can unlock across a range of domains.... In the future, we envision GenAI helping us re-imagine entire business workflows.

Jamie Dimon, JPMorganChase CEO (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

Artificial intelligence (AI) adoption in financial services has moved from experimentation to core infrastructure. Surveys by the Bank of England and McKinsey report that 75–80% of firms now use AI, with the fastest growth in the use of generative models such as the large language models (LLMs), which often serve as the foundation for firms' custom-built tools.¹ While general-purpose LLMs are proving competitive with specialized machine learning models in performing quantitative financial tasks (e.g., Lopez-Lira and Tang, 2024), their rapid integration introduces novel operational and regulatory risks that remain poorly understood. Indeed, half of survey respondents reported having only partial comprehension of the AI technologies they employ.

This opacity is especially concerning given the rapid growth of automated decision-making in AI use cases such as fraud detection and credit risk assessment. With over 60% of surveyed firms planning near-term adoption of such high-stakes AI applications, powerful models will have access to abundant personal data and an immediate impact on consumers, with potentially profound consequences for businesses. It is, therefore, crucial for financial firms and regulators to understand how this new technology responds to consumer data. How sensitive are LLM outputs to customer-specific characteristics? And how can firms and

¹ www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024 and www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

regulators ensure that systems incorporating powerful but opaque general-purpose models comply with fair lending laws? We offer the first systematic assessment of how LLM use may alter these inherent risks and propose a practical method for evaluating and managing them.

We develop and implement an audit methodology to assess the behavior of foundational LLMs, applying it to a mortgage underwriting task to test for racial disparities in model outputs. While we do not expect financial institutions to use off-the-shelf LLMs for mortgage underwriting directly, this setting serves as a valuable experimental testbed. Indeed, the strong regulations governing underwriting might discipline the behavior of LLMs. Evidence of racial discrepancies in this context raises the possibility that similar issues could arise across a broader range of customer-facing applications in financial services.

To execute this test, we use real loan application data from the Home Mortgage Disclosure Act (HMDA), supplementing it with experimentally manipulated applicant race and credit scores. We prompt several leading commercial LLMs to make underwriting recommendations and find consistent evidence that these models recommend different outcomes for Black and white applicants, despite applications being identical on all other dimensions.² Specifically, LLMs recommend more denials and higher interest rates for Black applicants than for otherwise-identical white applicants.³ These differences are substantial: Black applicants, on average, would need credit scores approximately 52 points higher than white applicants to receive the same approval rate, and about 22 points higher to receive the same interest rate.

In this experiment, the LLM receives explicit information about borrower race, allowing us to isolate its response to this protected characteristic. Although such information may not be available to automated underwriting systems in practice, the explicit signal should

²Throughout, we follow the [AP Stylebook](#) and Butler et al. (2023) in writing “Black” with initial capitalization and “white” in lowercase.

³We also show that LLM 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.

be easy for the LLM to ignore. The fact that it does not is troubling and raises important concerns about how these models may respond to more subtle, real-world proxies for race, such as name or location (Fuster et al., 2022).

To assess potential risks in a more realistic setting, we repeat the experiment without explicit race information but include fictional borrower names from Crabtree et al. (2023) that signal racial identity. Comparing otherwise-identical applications in which the applicant name is perceived as Black or white, we again observe systematic disparities in LLM outputs. We similarly find disparities when we signal race by including cities with varying Black populations in the application. These tests carry potentially far-reaching implications, as they mirror common LLM applications such as customer service chatbots and robo-advising, where customer names and locations are routinely available to the models.

By experimentally manipulating credit scores and fully stratifying them across the race signal and all other loan characteristics, we isolate the effect of race at different levels of creditworthiness. We find racial disparities in LLM underwriting recommendations are most pronounced for applicants with lower scores. With our baseline LLM, the disparity in approval rates is 56% greater for low-score applicants than for average-score applicants (13.3 vs. 8.5 percentage points), and the disparity in interest rates is about 32% greater (47 vs. 35 basis points). We also examine two other measures of credit quality—debt-to-income and loan-to-value ratios—using observed values from HMDA data. Across all three measures, disparities are present throughout the credit spectrum but are consistently larger for riskier loans. This suggests that harms from racially biased LLM outputs may be intersectional, compounding disadvantage along multiple dimensions (Crenshaw, 1989).

We observe similar patterns across eight leading LLMs developed by Anthropic, Meta, and OpenAI, spanning a range of model scales and training generations. These findings suggest that improvements in model quality do not necessarily reduce disparities. As a result, the patterns we document may persist in future models, particularly if disparate outcomes stem from fundamental features of the technology or its training data. Auditing

techniques like ours can help model developers, users, and regulators identify and address such risks in newly developed systems.

This raises the question of whether the racial disparities in LLM recommendations can be reduced or eliminated. One natural mitigation strategy is to withhold demographic information from the model, similar to current underwriting practices in which lenders collect protected-class data for *ex-post* analysis but are prohibited from using it in decision-making. However, as discussed above, the richness of mortgage application data means that race can still be inferred indirectly, making this approach potentially ineffective. We therefore focus on a simple prompt engineering strategy that *retains* explicit race signals in the input but modifies the prompt to instruct the model to “use no bias” in making its decisions.

Despite its simplicity, this modified prompt substantially reduces racial disparities. The Black–white gap in loan approval recommendations is eliminated, both on average and across credit scores. Instructing the LLM not to exhibit bias reduces the average racial gap in recommended interest rates by about 60% (from 35 to 14 basis points), with even larger reductions for lower-credit-score applicants. We do not suggest that this specific prompt is optimal for all settings or applicable across the full range of LLM use cases. However, the results demonstrate that LLM behavior can be directed, and that simple prompt-based interventions can meaningfully mitigate disparate outcomes.

As an external benchmark, we compare the rate recommendations of the baseline LLM to the decisions of real lenders. This analysis relies on a different dataset than our prior experiments, consisting of approved HMDA loans matched to Freddie Mac records with observed credit scores. Although the LLM is not fine-tuned or specialized for mortgage underwriting, lacks access to macroeconomic context, and receives only limited information via the prompt, its suggested interest rates are strongly correlated with rate spreads assigned by real lenders. Accurately predicting real rates is not the focus of our study and is not required for the internal validity of our disparity estimates, but the close correspondence strengthens the case that our results have relevance beyond our experimental setting. This

aligns with findings from contemporaneous studies showing that off-the-shelf models, such as our baseline GPT-4 Turbo, can perform complex financial tasks on par with human experts (e.g., Chen et al., 2022; Hansen and Kazinnik, 2024; Kim and Nikolaev, 2024; Lopez-Lira and Tang, 2024). Indeed, LLM rate recommendations reflect typical underwriting patterns, with credit score receiving the most weight and DTI and LTV contributing less but in similar proportion to each other. These recommendations also correlate with *ex-post* loan delinquency even after controlling for credit spreads given by real lenders.

We also investigate how pricing and delinquency relate to loan applicants' actual races, in the spirit of Becker (1957) outcome tests. We find that Black and white borrowers have similar loan performance, with no significant difference in delinquency rates. Racial interest rate gaps disappear for both real underwriters and LLMs after controlling for credit risk, but only when borrower race is not disclosed to the LLMs. In contrast, when actual borrower race is included in the prompt, LLMs assign substantially higher rates to Black applicants, even after adjusting for credit quality. Given the absence of racial gaps in risk-adjusted rate assignments by real lenders and in realized loan performance, these disparities appear inconsistent with profit-maximizing behavior.

Our study makes several contributions. First, we conduct the first audit study assessing racially disparate LLM outputs in a finance setting. This work complements the growing body of research using audit designs to examine algorithmic bias in other LLM applications, including car price negotiation, election forecasting, and job candidate evaluation (Haim et al., 2024; Lippens, 2024; Veldanda et al., 2023). Our findings extend this literature to a domain governed by strict regulatory requirements. This is especially notable given that the training data of leading LLMs includes the full text of the mortgage underwriting regulations.⁴ It is plausible that exposure to these legal documents could lead an LLM to avoid disparate treatment in this setting, even if it fails to do so elsewhere. We show that it

⁴Most 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).

does not.

Second, unlike prior audit studies that document outcome disparities in non-mortgage settings, we investigate how such disparities can be reduced. Specifically, we show that racial differences in LLM outputs can be attenuated through prompt engineering, contributing to a growing literature on mitigating bias in LLMs. For overviews of this literature, see Mehrabi et al. (2021) and Navigli et al. (2023), which examine the sources and types of bias, methods for detection and reduction, and practical applications. Much of the computer science work on bias mitigation focuses on tools accessible only to model *developers*, such as preprocessing training data, modifying representations during model training, or applying post-training fine-tuning. In contrast, our approach uses prompt engineering as a simple and accessible intervention that model *users* can apply to reduce disparities in LLM outputs.

Third, we extend the finance literature on discrimination and algorithmic bias in lending to the emerging context of LLMs. While many prior studies have documented racial disparities in traditional mortgage markets,⁵ and recent work (e.g., Bartlett et al., 2022) shows that FinTech lenders using supervised machine learning algorithms also produce interest rate disparities that disadvantage marginalized borrowers, our focus is distinct. Existing research centers on algorithms with narrow, task-specific objectives, trained in a supervised paradigm (e.g., Fuster et al., 2022; Gao et al., 2023; Howell et al., 2024).⁶ In contrast, LLMs are general-purpose tools that do not optimize directly for traditional finance goals, but instead generate outputs based on broad training data and user prompts. This flexibility expands their potential applications across financial services—but also introduces new risks. By evaluating disparities in LLM decisions and comparing them to *ex-post* loan performance, we provide early evidence that LLMs appear to incorporate race signals in ways that may

⁵E.g., Ambrose et al. (2021); Bayer et al. (2018); Begley and Purnanandam (2021); Blattner and Nelson (2021); Gerardi et al. (2023); Giacopetti et al. (2021); LaVoice and Vamossy (2024); Munnell et al. (1996). In contrast, Bhutta et al. (2022) and Hurtado and Sakong (2024), using confidential HMDA data, find that most racial disparities in loan approval rates can be explained by observable applicant characteristics unrelated to race.

⁶Additional studies on the application of machine learning algorithms in credit-risk models include Costello et al. (2020), Krivorotov (2023), and Nazemi and Fabozzi (2024).

not be profit-maximizing.

Thus, our study has significant implications for regulators and financial firms exploring the use of AI and machine learning (ML) technologies, including LLMs, which risk integrating race signals inefficiently and unfairly.⁷ Financial institutions of all sizes are actively developing and deploying AI and ML systems. A report by S&P Global Market Intelligence notes that banks representing 80% of the sector’s market capitalization recently referenced AI or ML in earnings calls. For instance, J.P. Morgan revealed that it currently has more than 300 AI use cases in production and predicted billions of dollars in projected cost savings from the integration of newly developed LLM tools now in use by thousands of employees.⁸ If tools for investment recommendations, customer service, fraud detection, personalized financial planning, product marketing, or insurance underwriting are built on biased algorithms that have access to demographic information, this bias can influence a wide range of important financial outcomes. Our findings add to the cautionary tale for firms and regulators: even advanced models can produce disparate outcomes if not properly audited before deployment, particularly in high-stakes financial applications.

2 Methodology

2.A Background and research questions

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

⁷More 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), D’Acunto et al. (2023), and Rossi and Utkus (2020) examine how robo-advising interacts with behavioral and cultural biases.

⁸www.spglobal.com/marketintelligence/en/news-insights/research/smaller-banks-are-using-ai-too

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

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 and PDF files 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, and studies have documented related phenomena across various generations of LLM models (Kadambi, 2021; Santurkar et al., 2023; Zou and Schiebinger, 2018).

There is a large literature that focuses on aligning LLMs to behave as intended by their designers (surveyed by Dong et al., 2024), including through measures designed to reduce various forms of bias. For example, the critical role of training data in determining LLM behaviors underscores the importance of corpus selection and cleaning. LLM developers can exclude corpus text likely to be biased, or take steps such as duplicating training sentences with reversed gender roles, increasing the model’s exposure to non-stereotypical 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 responses across diverse contexts and user groups, and

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

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 GPT-4 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 vast portions of the accessible internet, including major forum sites like Reddit and Quora, as well as 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 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 disparities in a mortgage setting is, however, unclear *ex ante*.

We therefore structure our study around three research questions. First, we examine whether LLMs provide different responses to mortgage lending queries based on race. For example, we assess whether an LLM approves loans for Black borrowers at the same rate

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

as otherwise-identical white borrowers, with race signaled either explicitly or implicitly. We also explore whether racial differences vary across applicant credit quality.

Second, we examine whether prompt engineering (that is, modifying instructions in the prompt) can reduce disparities in how LLMs respond to racial signals. This includes investigating whether prompt engineering mitigates *heterogeneity* in LLM responses across race and credit quality.

Third, using race-blind loan applications with real credit scores, we examine whether LLM lending recommendations reflect meaningful risk assessment capabilities and whether the models themselves triangulate protected borrower characteristics. By assessing the relationship between application information, LLM and real-world lending decisions, and subsequent loan performance, these tests help contextualize our findings and highlight the potential benefits and risks of integrating LLMs into financial firm operations beyond underwriting.

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, constitutes a fictional loan application which is sent to an application programming interface (API) endpoint for each LLM we examine. The LLMs’ memories are reset between each request, ensuring that we can isolate how changes to a single prompt’s information set affects the model’s output. 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 does 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 Experian¹³), and 790 (“very good”). Manipu-

¹²The 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.

¹³See 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. (In Section 5, we assess tests using a matched HMDA–Freddie Mac dataset in which we can observe true credit scores for a subset of approved loans.)

[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:

$$y_{i,m} = \beta_{CS} CreditScore_{i,m} + \beta_B Black_{i,m} + \phi_i + u_{i,m}, \quad (1)$$

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 , $CreditScore_{i,m}$ is the assigned credit score, $Black_{i,m}$ 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.

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

¹⁴Unless specified otherwise, all experiments are conducted with LLM temperature parameters set to zero to reduce randomness in its replies. We show robustness to setting a higher temperature in Table A2.

is labeled as white. As such, β_B captures the direct effect of disparities in the LLM response to race 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

$$y_{i,m} = \beta_{CS} CreditScore_{i,m} + \beta_B Black_{i,m} + \boldsymbol{\beta}'_{B \times X} Black_{i,m} \mathbf{X}_{i,m} + \phi_i + u_{i,m}, \quad (2)$$

where $\mathbf{X}_{i,m}$ 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 \mathbf{X} represents credit score, we include both its main effect and its interaction term in the model. Where \mathbf{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 $\boldsymbol{\beta}'_{B \times X}$ 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 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.) Experiment 3 replicates this approach across a variety of other leading LLMs to see if the patterns are specific to one model. Experiments 4 and 5 do not include explicit racial information, but instead use borrower names and cities, respectively, as proxies;

we describe these in more detail in Section 3.C. In the appendix, we present tests exploring other protected borrower characteristics: Experiments A1 and A2 include manipulated applicant age or gender instead of racial information.

Given the existence of disparities documented in these experiments, we then proceed with Experiment 6 to assess the potential value of 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

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

where $\text{Mitigation}_{i,m}$ is a binary indicator variable for loan applications made with the mitigation prompt. When β_{B} and $\beta_{\text{M}\times\text{B}}$ 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 disparities *on average*. (Experiment A3 examines an alternative mitigation prompt.) To extend this, we assess whether these effects are heterogeneous across credit quality, estimating models of the

form

$$\begin{aligned}
y_{i,m} = & \beta_{CS} CreditScore_{i,m} + \beta_B Black_{i,m} + \beta_{B \times CS} Black_{i,m} CreditScore_{i,m} + \beta_M Mitigation_{i,m} \\
& + \beta_{M \times CS} Mitigation_{i,m} CreditScore_{i,m} + \beta_{M \times B} Mitigation_{i,m} Black_{i,m} \\
& + \beta_{M \times B \times CS} Mitigation_{i,m} Black_{i,m} CreditScore_{i,m} + \phi_i + u_{i,m}.
\end{aligned} \tag{4}$$

Here, $\beta_{B \times CS}$ identifies the heterogeneity of racial disparities across credit scores for the baseline prompt, and $\beta_{M \times B \times CS}$ 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 under the HMDA Act. 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.¹⁵ We use 2022 data because this is after the training cutoff for the models in our paper and allows us to monitor two years of *ex-post* loan performance in Section 5. 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.¹⁶ Panel A of Table II

¹⁵ Available at ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022.

¹⁶ A standard two-sample proportions power test suggests a sample size of 962 per group is necessary to detect differences in loan acceptance rates greater than 3.7% (half of the 7.4% rejection rate in the full HMDA dataset, per Table A1) at 80% power and 5% significance.

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 average rate spread in our sample is 27 basis points. The average debt-to-income ratio (DTI) is 37.2%¹⁷, and loan-to-value ratio (LTV, `combined_loan_to_value_ratio` in HMDA) is slightly over 80%. We show in Appendix Table A1 that this subset of loans is representative of the loans in the overall LAR dataset.

[Insert Table II about here]

Table II, Panel B, reports summary statistics on LLM approval rate and interest rate suggestions separately for each experiment. Across experiments, 87–95% of loans are “approved” by the LLM with a suggested average interest rate of 4.41–4.75%, 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 experiments. The biggest deviation, although not statistically significant, occurs in Experiment 3. This is the only one that includes models besides GPT-4 Turbo, and these models on average recommend slightly lower approval rates and higher interest rates.

3 Main results

This section presents the paper’s primary results. We start with tests assessing whether our baseline LLM shows evidence of bias in making lending decisions. We then extend the analysis to other leading LLMs. Finally, we present tests using proxies for race, rather than direct signals.

¹⁷DTI is reported in HMDA (`debt_to_income_ratio`) 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.

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 a 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 increase in credit score (61 points) raises the likelihood that 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 those from otherwise-identical white applicants. The influence of being Black is noteworthy; its magnitude 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 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).

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

¹⁹Table A3 reports qualitatively identical results to those in Panel A using a logistic model.

²⁰Because 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

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, as higher credit score means higher credit quality); while the coefficients for the interactions with *DTI* and *LTV* are -0.063 and -0.042 , respectively (negative, as lower DTI and LTV means higher credit quality). Given that these variables are standardized, the coefficients' magnitudes 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 credit quality. For example, based on the estimates in column (3), 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*.²¹

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. (≈ 35 basis points) higher on average than otherwise-identical white applicants'.

To contextualize the magnitude of estimated race effects, we can compare them to the impact of credit scores. In column (1), the estimated coefficient on *CreditScore* indicates that a one standard deviation increase in credit score decreases suggested interest rates 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.

²¹If 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 22, below, we observe evidence of both approval and interest rate disparities even at the highest credit scores.

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. Most studies do not report the effects of race and credit scores simultaneously, but one that does is Butler et al. (2023) in the auto loan market. In their Table 8, they estimate $\hat{\beta}_{\text{Minority}} = 0.704$ and $\hat{\beta}_{\text{Credit Score}} = -0.019$. 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. Thus, lower credit quality (i.e., lower credit scores, higher DTI or LTV) is associated with larger interest rate penalties against Black applicants.²²

To translate these estimates into costs a consumer would face, 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 for an applicant with an average credit score, a white applicant with a 654 credit score would have a 5.41% rate while a comparable Black applicant would have a rate of 5.88%, and over the life of a 30-year mortgage this Black applicant would pay around \$35,700 more in (nominal)

²²The 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 III, discussed below.

interest than a white applicant with the same credit profile.

Experiment 2 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 in Table IV. 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 LLM responses.

[Insert Table IV about here]

The results in Table IV reveal interesting patterns across these groups. Because this specification uses None (i.e., no race signal) as the omitted category, the coefficients on the race/ethnicity indicators represent the effect of including a given racial or ethnic label relative to disclosing no race information. This contrasts with Table III, where coefficients are interpreted relative to a white applicant. Within this framework, white and Asian applicants receive modestly more favorable responses than those with no disclosed race, while Black and Hispanic applicants experience notably worse outcomes—particularly Black applicants, who face the largest disparities in both approval and interest rate recommendations.

Interaction terms between race/ethnicity indicators and credit score provide additional insights. Black applicants are the only group with significant interaction coefficients across both outcome variables, with signs and magnitudes indicating that higher credit scores can reduce some of the disparities that Black borrowers suffer (but do not eliminate them). Hispanic borrowers with low credit scores also suffer worse approval disparities, although the magnitude of this effect is much smaller than for Black borrowers.

Finally, we consider two experiments on other protected borrower characteristics: age and gender. Experiment A1 replaces signals of race/ethnicity in the loan applications with

indications that the applicant is age 30, 50, or 70. Results are reported in Appendix Table A4. 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), which finds 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 indicating that a lower credit score is additionally penalized. Experiment A2 instead considers signals that an applicant is male or female; the results in Appendix Table A5 fail to detect evidence of statistically significant gender differences.

3.B Racial disparities in other LLMs

We now turn to Experiment 3 to assess whether the 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.

[Insert Table V about here]

Table V presents regressions of Equation 1 and confirms that the pattern of disparities we find in the baseline LLM is present in other models. With only a few exceptions, the effects of *CreditScore* and *Black* are largely consistent in terms of signs and significance

²³We provide more information on these models, including specific API version names, in Appendix Table A6. 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.

across the different models. Higher credit scores substantially increase the probability of loan approval and lead to lower 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 summary statistics at the bottom of each panel highlight the nuances that different AI data-generating models can introduce when used in finance applications. In Panel A, which shows estimations of LLM loan approval recommendations, we observe substantial variation across models in the proportion of applications approved (“Avg(y)”) with rates ranging from 58% for the 2023 version of ChatGPT-3.5 Turbo (column 5) to 99%–100% for the Llama 3 models (columns 3–4). 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, responding just 74% of the time.²⁵ 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. This suggests that larger and more advanced models will not necessarily reduce the disparities we document. Columns (3) and (4) focus on models by Meta. With near-universal loan approval for the Llama 3 models, it is unsurprising that

²⁴We do not estimate loan approval using Llama 3’s smaller model since it approves all applications.

²⁵Answer rates take into account the fact that we attempt a prompt up to ten times if an LLM doesn’t provide a properly formatted 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.”

we do not observe significant evidence of racial differences in their responses. The remaining columns focus on OpenAI models. The baseline LLM for our study, GPT-4 Turbo, is between these extremes, and suggests approval for 91% of loans in Experiment 1; the true mortgage approval rate in our HMDA sample is 92% per Table II.

Panel B, where the outcome variable is the interest rate recommendation, shows even greater consistency with our primary results from the baseline LLM. This consistency may stem from the fact that while loan approval is a binary decision, interest rate recommendations admit more subtle outcome disparities. All eight *Black* coefficients are positive and significant at the 1% level or better.

As we did following Experiment 1, we can contextualize the economic magnitude of racial disparities presented in Table V by computing the implied decrease in credit score for a white applicant that would generate an effect as large as instead listing the applicant as Black. 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 52 points for approval decisions. As with our baseline LLM, this magnitude is close to the 37 point-equivalent minority penalty in auto lending implied by estimates in Butler et al. (2023). Notably, our average credit score equivalent for interest rate suggestions is 24 points, with a cluster of four LLMs around 33 points and three more around 16.

To understand how racial disparities in the responses of each LLM vary heterogeneously across the credit spectrum, we estimate regressions of Equation 2 and present the results (visually, for brevity) in Figure II. Approval decisions are 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.

[Insert Figure II about here]

In total, 21 out of 24 interest rate coefficients are significant at the 1% level or bet-

ter, and in all eight models, the average rate is higher for Black applicants. The precision of the estimates on the interaction terms is more varied but coefficients are mostly positive and significant in the approval regressions and negative and significant in the interest rate regressions. Most models from Anthropic (Claude) and OpenAI (GPT) generate racial disparities that differ by credit quality, with lower credit score Black applicants obtaining even less favorable outcomes than white applicants. However, insignificant approval 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.

Overall, our core findings are robust across different LLM providers and model characteristics (number of parameters, generation, and training date). The appendix contains additional tests confirming that our results hold when we vary model temperature, estimate logistic models for approval, and repeat the core experiment at a later point in time to assess stability of our results over time.

3.C Racial disparities when exposed to proxies for race

In most practical settings where LLMs interface with customers, financial firms are unlikely to intentionally include explicit race information in the model’s information set. However, information that is correlated with borrower race will likely be available, such as applicant name, residential address, and even stated income, occupation, or education. These features can act as proxies for race, particularly when combined, and LLMs may be able to reconstruct protected class membership even when race is explicitly excluded from the data. As a result, an LLM exposed to naturalistic input may internalize and act on inferred racial information.

To evaluate whether LLMs exhibit disparities in response to implicit racial signals, we conduct two additional experiments. In Experiment 4, we assign applicant names that are perceived to be strongly associated with either Black individuals or white individuals. In Experiment 5, we vary the applicant’s city of residence to correspond to geographic areas

with higher or lower proportions of Black residents. As in Experiment 1, we hold all other information constant, stratifying across race proxies and credit scores for each of our 1,000 base loan applications.

In Experiment 4, applicant names are randomly assigned for each underlying loan and credit score from the set of “validated names for experimental studies on race and ethnicity” developed by Crabtree et al. (2023). We restrict to names perceived as either Black or white by more than 80% of survey respondents in their study. The results, reported in Panel A of Table VI, show that applications with distinctively Black names receive approval recommendations 1.3 percentage points less often and are offered interest rates 10 basis points higher than otherwise-identical applications with white names.

[Insert Table VI about here]

These disparities are statistically significant and economically meaningful, with magnitudes 15–29% of the effects we estimate when race is disclosed explicitly (in Experiment 1). Consistent with our earlier findings, the disparities associated with inferred race are largest for applicants with lower credit scores.

In Experiment 5, our prompts include borrower city and state in lieu of race or racialized names. For each state, we use the two cities with the highest and lowest fraction of Black residents, according to the 2020 Census.²⁶ The results are reported in Panel B of Table VI. Otherwise-identical loan applications receive interest rates 6 basis points higher in their state’s most-Black city compared to its least-Black, and these (statistically significant) disparities are once again larger at lower credit scores. (Loan approval rates are lower for the Black cities, but this effect is not statistically significant.)²⁷

²⁶We consider all Census “places” (e.g., cities, towns, etc.) with population greater than 50,000 in the 2020 Census Redistricting Data (Public Law 94-171) Summary File.

²⁷Smaller magnitudes associated with our geographic signal compared to the name-based signal are perhaps unsurprising given that the ability to infer race from city is on average much weaker. For example, while five states have cities whose Black populations vary by more than 70 percentage points (e.g., Stonecrest, GA at 92.3% vs. Alpharetta, GA at 10.3%), in twelve states the largest difference is under 5 p.p. (e.g., Great Falls, MT at 1.2% vs. Bozeman, MT at 0.6%).

These two experiments show systematic racial disparities in LLM financial recommendations even in the absence of explicit signals. Of course, the implicit signals we consider—name and city—could be associated with underwriting and loan outcomes for reasons other than race. In particular, borrowers with distinctively Black names and cities with large Black populations may have socioeconomic characteristics that make loans riskier or less profitable. However, we find systematic differences in outcomes even after controlling for credit score, DTI, LTV, loan amount, property value, income, and state.

Taken together, the results provide novel empirical evidence that LLMs may exhibit race-correlated disparities through inference, not just disclosure—a dynamic with important implications for the design of race-blind decision systems.

4 Mitigating racial disparities

Having established in prior sections that LLMs produce racially disparate lending recommendations, we now evaluate whether these disparities are manageable. While earlier experiments provide a structured audit of baseline model behavior, this section demonstrates that LLM behavior can also be directed. Specifically, we test whether simple prompt-based instructions can systematically reduce racial disparities in loan approval and pricing recommendations. This approach does not seek to identify an optimal solution, but rather to assess whether and how LLM responses can be shaped after risks are identified. The results suggest a promising insight for model users and financial regulators: straightforward, low-cost interventions can manage LLM risks.

We evaluate this possibility empirically in Experiment 6. This experiment introduces a minimal prompt-based instruction designed to reduce bias. We examine 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 the mitigation prompt for exactly the same loans and race/credit score manipulations. The combined sample of 12,000 observations is analyzed using regression models as described in Equation 3 (to understand how mitigation affects racial disparities on average) and Equation 4 (to understand how mitigation’s racialized effects vary by credit score). The results are presented in Table VII, 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 VII 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 63.2bp per standard deviation in score to 58.2bp (see column 4).²⁸

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* \times *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* \times *Black* coefficients essentially offset each other, indicating that bias is effectively neutralized by the mitigation prompt.²⁹

²⁸The coefficient on *Mitigation* \times *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.

²⁹We show qualitatively identical results for approval decisions modeled using logistic regression in Appendix Table A7. With the mitigation prompt, the linear model does not reject the absence of racial differences in approval recommendations on average ($p = 0.83$).

The results for interest rate recommendations show similar patterns. In columns (3) and (4), the coefficient on *Mitigation* \times *Black* is negative and significant, with mitigation reducing the average interest rate disparity for Black applicants by 21.4bp, roughly 60% of the average gap. This suggests that our simple mitigation strategy moderates but does not eliminate this form of bias.

Furthermore, the interaction terms involving both *Black* and *CreditScore* speak to the effectiveness of the bias mitigation prompt in reducing not only the average *level* of racial differences, but also the heterogeneity across the credit spectrum. In column (2), mitigation eliminates approval disparities for Black borrowers at all credit scores.³⁰ For interest rate recommendations, column (4) indicates that while lower credit scores hurt Black applicants more than white applicants in the baseline prompt (11.4bp per credit score standard deviation), the mitigation prompt substantially reduces this (by 7.9 basis points), although statistically significant credit score heterogeneity in racial differences still exists.

We illustrate these points in Figure III. Using the estimates from columns (2) and (4), we calculate predicted values for all observations (including the loan fixed effects), and plot the outcomes averaged by credit score separately for Black and white applicants for both outcomes and prompts. Comparing the dashed to the solid lines shows the power of the simple mitigation prompt. In the baseline prompt, only Black applicants at the top of the credit score range receive approval decisions and interest rate suggestions approaching those of otherwise-identical white applicants. However, with the mitigation adjustment, approval decisions are indistinguishable between Black and white applicants across the credit spectrum, and the level and heterogeneity of racial interest rate disparities are markedly reduced.

[Insert Figure III about here]

To further assess prompt engineering-based approaches to reducing disparities, we also

³⁰The racial difference in the effect of credit score on approval under the mitigation prompt is given by $\beta_{B \times CS} + \beta_{M \times B \times CS}$; we cannot reject that this sum equals zero ($p = 0.47$).

consider in Experiment A3 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.”³¹ This prompt raises the salience of the legal stakes and might induce the LLM to further reduce the racial difference in its recommendations. Alternatively, this prompt might be less effective because its phrasing is somewhat detached from the outcomes we are assessing. The results in Appendix Table A8 repeat tests of Equation 3 for this alternate mitigation prompt.

This legalistic approach successfully moderates Black–white gaps in LLM recommendations, though the effects are smaller than we found in Experiment 6: Comparing the *Black* and *Mitigation × Black* coefficients shows reductions by about 70% of the approval difference and just 30% of the interest rate difference (versus 100% and 61%, respectively, for the main mitigation prompt).

Overall, these findings indicate that while the baseline prompt results show significant racial disparities in both loan approval and interest rate recommendations, straightforward modifications to the prompt substantially reduce these disparities. This highlights that LLM behavior is dynamic and responsive to carefully constructed inputs. We do not propose that the mitigation prompt tested here is optimal for all contexts. Rather, we offer this as an illustration of a broader principle in which LLM outputs can be tested, disparities can be identified, and simple interventions can reduce harm. This creates a repeatable and practical framework for managing LLM risk (test → find → fix) and represents an important step toward aligning general-purpose AI tools with the compliance and fairness expectations of financial regulation.

³¹It is not this paper’s goal to assess 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.

5 LLM interest rate suggestions: Additional tests

5.A Determinants, accuracy, and *ex-post* performance

By varying race and credit scores in our sample, our experimental design isolates the effect of race signals on LLM recommendations. However, because publicly available HMDA data do not include true credit scores, this empirical approach does not allow us to evaluate how closely LLM outputs correspond to actual lenders' underwriting decisions. Moreover, HMDA data lack information on *ex-post* loan performance, which is necessary to evaluate the profit implications of the LLMs' decisions. To address these limitations, we conduct a series of additional tests that incorporate new data from Freddie Mac's Single-Family Loan-Level Dataset, which provides borrowers' true credit scores and subsequent loan delinquency outcomes.

We merge loans from the 2022 HMDA LAR file with Freddie Mac data. Specifically, we download Freddie Mac loan records, restrict the sample to loans originated in 2022, and apply the same filters used for the HMDA data (see Section 2.C). We match observations based on loan amount, interest rate, DTI, LTV, state, and ZIP code.³² We then drop matches that are not unique—that is, each HMDA and Freddie Mac loan must match a single counterpart. From the resulting matches, we randomly select 1,000 loans, which we refer to as the HMDA–Freddie Mac Matched Sample. Using this matched sample, we rerun the LLM application prompt with the true credit score for each loan, *omitting* demographic signals so that the prompts are race-blind. Because all loans in the Freddie Mac dataset were approved and originated, this analysis focuses on interest rate suggestions.

Table VIII presents OLS results that approximate the LLM's interest rate suggestion rule using a linear specification. Column (1) shows that a one standard deviation increase in credit score is associated with a 49 basis point decrease in the suggested rate, consistent

³²We adjust LTV and DTI across datasets to ensure comparable granularity and require that the ZIP code in the Freddie Mac data contains the census tract reported in the HMDA file. This procedure is adapted from Buchak and Jørring (2021), Jiang et al. (2023), and Kalda et al. (2023).

with the evidence in Table III. Unlike earlier estimates, which identify credit score variation across experimentally repeated applications, the regressions in Table VIII omit fixed effects. This allows us to assess how the LLM responds to variation in DTI and LTV. Columns (2) and (3) examine these variables separately and find that a one standard deviation increase in either is associated with a roughly 20 basis point increase in the suggested rate. When standardized credit score, DTI, and LTV are included together in column (4), the coefficients on DTI and LTV remain statistically significant, but the LLM responds most strongly to credit score. This pattern aligns with established lending practices that prioritize measures of creditworthiness.

[Insert Table VIII about here]

Next, we compare the interest rates recommended by the LLM with those actually charged by lenders. We focus on rate spreads rather than nominal interest rates, following Bhutta and Hizmo (2021), because spreads more precisely capture cross-sectional variation in lender risk assessments, independent of broader market (yield curve) movements. This distinction is particularly relevant because our experimental prompts do not incorporate macroeconomic conditions, and mortgage rates spiked during 2022 from 3.22% to 6.42%.

Figure IV presents a binned scatter plot showing the relationship between actual rate spreads on issued loans and the interest rate suggestions generated by the LLM. The underlying regression yields a coefficient of 0.55 with a t -statistic of 14.5, indicating that LLM rate suggestions are highly statistically significant predictors of the risk assessments made by actual lenders. This is particularly notable given the limited application information included in the prompt.

[Insert Figure IV about here]

To evaluate the *ex-post* performance of the baseline LLM in underwriting tasks, we examine whether each loan becomes delinquent (defined as being 30 days past due) at some

point through Q3 2024, the last quarter with data available. Table IX presents a comparison between different pricing and risk measures in predicting delinquency. Column (1) shows that a one standard deviation increase in the rate spread assigned by actual underwriters is associated with a 2.7 percentage point increase in the likelihood of delinquency. In contrast, a one standard deviation increase in the LLM’s interest rate suggestion is associated with a 5.9 percentage point increase in delinquency (column 2), suggesting superior predictive power. This advantage is confirmed in column (3): When both measures are included together, the coefficient on the actual rate spread falls to near zero, while the LLM rate coefficient remains essentially unchanged at 5.7 percentage points. However, columns (4) and (5) reveal that when credit scores are included, the LLM rate becomes statistically insignificant, suggesting that much of the LLM’s predictive power operates through its ability to infer credit-relevant information from limited prompts.

[Insert Table IX about here]

While accurate calibration of the LLM’s recommendations is not required for the validity of our main results, the evidence in this section demonstrates that the model performs sophisticated financial risk assessments. These findings align with recent research showing that LLMs exhibit substantial capabilities in quantitative financial analysis beyond traditional text processing (Feng et al., 2024; Fieberg et al., 2023; Lopez-Lira and Tang, 2024; Shah and Chava, 2023). More broadly, our results suggest that LLMs may offer value to financial institutions across a range of analytical tasks and that wider adoption may be attractive, even if their use remains limited in more regulated functions such as underwriting.

5.B Loan outcomes and true applicant race

To examine whether LLM rate disparities reflect genuine differences in borrower risk, Table X reports OLS estimates of loan outcomes based on the *actual* race of the applicant. Columns (1) and (2) show no statistically significant difference in delinquency rates for

Black borrowers, both on average and after controlling for credit risk. Column (3) indicates that Black applicants receive 15 basis points higher rate spreads from real underwriters on average, but this disparity largely disappears once credit-relevant factors are controlled for in column (4). These patterns are consistent with Bhutta et al. (2022), which finds that observable applicant factors explain most of the racial disparities in mortgage approvals.

[Insert Table X about here]

We next examine LLM rate recommendations. The advantage of these tests relative to Experiment 1 is that we now use true credit scores that are realistically correlated with applicant race; in Experiment 1, the experimental manipulations intentionally eliminated any such correlation.

The pattern observed in real underwriting decisions—racial disparities in average interest rates that are explained by risk factors—emerges in LLM recommendations as well. In columns (5) and (6), the dependent variable is the interest rate suggested by the baseline LLM using an application prompt that does not include any race information. On average, the LLM assigns rates that are 30 basis points higher for Black applicants (t -statistic of 3.51). Column (6) shows that this gap is largely accounted for by the omission of risk factors from column (5), as the coefficient falls to a statistically insignificant 5.5 basis points once credit score is included in the specification.

Finally, we alter the prompt to disclose applicants' *true* races. In column (7), we find that, on average, the baseline LLM assigns rates that are 68 basis points higher when it knows the applicant is Black. As noted by Bohren et al. (2023), Doleac and Stein (2013), and Guryan and Charles (2013), disentangling statistical from taste-based discrimination is challenging in any setting; the two can co-exist and may reinforce each other. While the fact that the relationship between disclosed race and LLM recommendations is attenuated after controlling for measures of credit quality in column (8) suggests that statistical discrimination plays at least some role, the race–rate relationship remains both statistically and economically

significant even with those controls. Given the absence of racial gaps in real lenders’ risk-adjusted rates and *ex-post* loan performance, this suggests that LLM racial disparities are inconsistent with profit maximization. Such a pattern would be consistent with the presence of some degree of animus and have important implications for the use of LLMs in many customer-facing applications.

6 Conclusion

Financial services firms are rapidly integrating LLMs across a wide range of functions, making it essential to understand their limitations. Using mortgage underwriting as an experimental testbed, we document robust evidence that LLMs recommend more denials and higher interest rates for Black applicants compared to otherwise-identical white applicants. These disparities persist across multiple leading models and generations, suggesting that underlying training data and modeling choices embed systematic racial bias despite developers’ efforts to address it.

Our results have implications far beyond the mortgage lending application we consider in this study. Disparities remain even when race is only inferred through proxies such as names or locations, underscoring that race-blind inputs do not guarantee race-blind results. As financial firms increasingly deploy LLMs in applications such as personalized investment advice, customer support, and targeted marketing, there is a risk of perpetuating and amplifying existing inequities.

We also show that relatively simple interventions can reduce harm. Prompting models explicitly to “use no bias” eliminates approval gaps and cuts interest rate disparities by more than half. While this specific instruction will not be a universal solution, it illustrates how rigorous, audit-based approaches enable users to refine prompts and develop strategies that lead to more equitable outcomes.

LLMs’ rate recommendations are strongly correlated with those of real underwriters and

predict defaults, suggesting these general-purpose models possess surprising sophistication in performing credit analysis. However, our outcome tests highlight that this sophistication does not necessarily translate into fairness or efficiency. When borrower race is disclosed, models assign significantly higher rates to Black applicants even after controlling for credit quality. In a sample where real lenders' risk-adjusted rates and loan performance do not differ by race, this finding suggests that LLM racial disparities are unlikely to be profit maximizing and that unchecked adoption of these models could harm both consumers and firms.

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. doi:[10.1093/rfs/hhaa087](https://doi.org/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. doi:[10.21799/frbp.wp.2023.03](https://doi.org/10.21799/frbp.wp.2023.03).
- Atari, M., Xue, M.J., Park, P.S., Blasi, D., Henrich, J., 2023. Which Humans? doi:[10.31234/osf.io/5b26t](https://doi.org/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. doi:[10.1016/j.jfineco.2021.05.047](https://doi.org/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. doi:[10.1093/rfs/hhx035](https://doi.org/10.1093/rfs/hhx035).
- Becker, G.S., 1957. *The Economics of Discrimination*. University of Chicago Press.
- Begley, T.A., Purnanandam, A., 2021. Color and Credit: Race, Regulation, and the Quality of Financial Services. *Journal of Financial Economics* 141, 48–65. doi:[10.1016/j.jfineco.2021.03.001](https://doi.org/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 doi:[10.2139/ssrn.4452670](https://doi.org/10.2139/ssrn.4452670).
- Bhutta, N., Hizmo, A., 2021. Do Minorities Pay More for Mortgages? *The Review of Financial Studies* 34, 763–789. doi:[10.1093/rfs/hhaa047](https://doi.org/10.1093/rfs/hhaa047).
- Bhutta, N., Hizmo, A., Ringo, D., 2022. How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions. doi:[10.17016/FEDS.2022.067](https://doi.org/10.17016/FEDS.2022.067).
- Blattner, L., Nelson, S., 2021. How Costly is Noise? Data and Disparities in Consumer Credit. doi:[10.48550/arXiv.2105.07554](https://doi.org/10.48550/arXiv.2105.07554), [arXiv:2105.07554](https://arxiv.org/abs/2105.07554).
- Bohren, J.A., Haggag, K., Imas, A., Pope, D.G., 2023. Inaccurate Statistical Discrimination: An Identification Problem. *The Review of Economics and Statistics* , 1–45doi:[10.1162/rest_a_01367](https://doi.org/10.1162/rest_a_01367).
- Brynjolfsson, E., Li, D., Raymond, L.R., 2023. Generative AI at Work. doi:[10.3386/w31161](https://doi.org/10.3386/w31161), [arXiv:31161](https://arxiv.org/abs/31161).
- Buchak, G., Jørring, A., 2021. Competition with Multi-Dimensional Pricing: Evidence from U.S. Mortgages. doi:[10.2139/ssrn.3762250](https://doi.org/10.2139/ssrn.3762250), [arXiv:3762250](https://arxiv.org/abs/3762250).

- 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. doi:[10.1093/rfs/hhac029](https://doi.org/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.
- Chen, Y., Kelly, B.T., Xiu, D., 2022. Expected Returns and Large Language Models. [arXiv:4416687](https://arxiv.org/abs/2207.04168).
- Costello, A.M., Down, A.K., Mehta, M.N., 2020. Machine + man: A field experiment on the role of discretion in augmenting AI-based lending models. *Journal of Accounting and Economics* 70, 101360. doi:[10.1016/j.jacceco.2020.101360](https://doi.org/10.1016/j.jacceco.2020.101360).
- Crabtree, C., Kim, J.Y., Gaddis, S.M., Holbein, J.B., Guage, C., Marx, W.W., 2023. Validated names for experimental studies on race and ethnicity. *Scientific Data* 10, 130. doi:[10.1038/s41597-023-01947-0](https://doi.org/10.1038/s41597-023-01947-0).
- 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.
- 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. doi:[10.1093/rfs/hhz014](https://doi.org/10.1093/rfs/hhz014).
- Doleac, J.L., Stein, L.C., 2013. The Visible Hand: Race and Online Market Outcomes. *The Economic Journal* 123, F469–F492. doi:[10.1111/ecojo.12082](https://doi.org/10.1111/ecojo.12082).
- 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. doi:[10.48550/arXiv.2406.02622](https://arxiv.org/abs/2406.02622), [arXiv:2406.02622](https://arxiv.org/abs/2406.02622).
- 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. doi:[10.48550/arXiv.2303.10130](https://arxiv.org/abs/2303.10130), [arXiv:2303.10130](https://arxiv.org/abs/2303.10130).
- Feng, Z., Li, B., Liu, F., 2024. A First Look at Financial Data Analysis Using ChatGPT-4o. Working Paper doi:[10.2139/ssrn.4849578](https://doi.org/10.2139/ssrn.4849578).
- Fieberg, C., Hornuf, L., Streich, D., 2023. Using GPT-4 for Financial Advice. doi:[10.2139/ssrn.4499485](https://doi.org/10.2139/ssrn.4499485), [arXiv:4499485](https://arxiv.org/abs/2406.02622).
- 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. doi:[10.1111/jofi.13090](https://doi.org/10.1111/jofi.13090).

- Gao, J., Yi, H.L., Zhang, D., 2023. Algorithmic Underwriting in High Risk Mortgage Markets. doi:[10.2139/ssrn.4602411](https://doi.org/10.2139/ssrn.4602411).
- Gerardi, K., Willen, P.S., Zhang, D.H., 2023. Mortgage prepayment, race, and monetary policy. *Journal of Financial Economics* 147, 498–524. doi:[10.1016/j.jfineco.2022.12.001](https://doi.org/10.1016/j.jfineco.2022.12.001).
- Giacolletti, 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. doi:[10.21799/frbp.wp.2021.04](https://doi.org/10.21799/frbp.wp.2021.04).
- 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. doi:[10.1111/ecojo.12080](https://doi.org/10.1111/ecojo.12080).
- Haim, A., Salinas, A., Nyarko, J., 2024. What's in a Name? Auditing Large Language Models for Race and Gender Bias. [arXiv:2402.14875](https://arxiv.org/abs/2402.14875).
- Hansen, A.L., Kazinnik, S., 2024. Can ChatGPT Decipher Fedspeak? doi:[10.2139/ssrn.4399406](https://doi.org/10.2139/ssrn.4399406), [arXiv:4399406](https://arxiv.org/abs/4399406).
- 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. doi:[10.1111/jofi.13303](https://doi.org/10.1111/jofi.13303).
- Hurtado, A., Sakong, J., 2024. Racial Disparities in the US Mortgage Market. *AEA Papers and Proceedings* 114, 201–204. doi:[10.1257/pandp.20241128](https://doi.org/10.1257/pandp.20241128).
- Jiang, S., Jørring, A., Xu, D., 2023. Bank Technology Adoption and Loan Production in the U.S. Mortgage Market. doi:[10.2139/ssrn.4455105](https://doi.org/10.2139/ssrn.4455105), [arXiv:4455105](https://arxiv.org/abs/4455105).
- Kadambi, A., 2021. Achieving Fairness in Medical Devices. *Science* 372, 30–31. doi:[10.1126/science.abe9195](https://doi.org/10.1126/science.abe9195).
- Kalda, A., Pearson, C.G., Sovich, D., 2023. Cost Pass-Through and Mortgage Credit: The Case of Guarantee Fees. doi:[10.2139/ssrn.4728459](https://doi.org/10.2139/ssrn.4728459), [arXiv:4728459](https://arxiv.org/abs/4728459).
- Kim, A.G., Nikolaev, V.V., 2024. Contextualizing Profitability. doi:[10.2139/ssrn.4459383](https://doi.org/10.2139/ssrn.4459383), [arXiv:4459383](https://arxiv.org/abs/4459383).
- Krivorotov, G., 2023. Machine learning-based profit modeling for credit card underwriting - implications for credit risk. *Journal of Banking & Finance* 149, 106785. doi:[10.1016/j.jbankfin.2023.106785](https://doi.org/10.1016/j.jbankfin.2023.106785).
- LaVoice, J., Vamossy, D.F., 2024. Racial disparities in debt collection. *Journal of Banking & Finance* 164, 107208. doi:[10.1016/j.jbankfin.2024.107208](https://doi.org/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. doi:[10.1016/j.chbah.2024.100054](https://doi.org/10.1016/j.chbah.2024.100054).
- Lopez-Lira, A., Tang, Y., 2024. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. doi:[10.48550/arXiv.2304.07619](https://doi.org/10.48550/arXiv.2304.07619), [arXiv:2304.07619](https://arxiv.org/abs/2304.07619).
- 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. doi:[10.1145/3457607](https://doi.org/10.1145/3457607).
- Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J., 1996. Mortgage Lending in Boston: Interpreting HMDA Data. *The American Economic Review* 86, 25–53. [arXiv:2118254](https://arxiv.org/abs/2118254).
- 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. doi:[10.48550/arXiv.2307.06435](https://doi.org/10.48550/arXiv.2307.06435), [arXiv:2307.06435](https://arxiv.org/abs/2307.06435).
- 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. doi:[10.1145/3597307](https://doi.org/10.1145/3597307).
- Nazemi, A., Fabozzi, F.J., 2024. Interpretable machine learning for creditor recovery rates. *Journal of Banking & Finance* 164, 107187. doi:[10.1016/j.jbankfin.2024.107187](https://doi.org/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. doi:[10.48550/arXiv.2203.02155](https://doi.org/10.48550/arXiv.2203.02155), [arXiv:2203.02155](https://arxiv.org/abs/2203.02155).
- Rossi, A.G., Utkus, S.P., 2020. The Needs and Wants in Financial Advice: Human versus Robo-advising. doi:[10.2139/ssrn.3759041](https://doi.org/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.
- Shah, A., Chava, S., 2023. Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks. doi:[10.48550/arXiv.2305.16633](https://doi.org/10.48550/arXiv.2305.16633), [arXiv:2305.16633](https://arxiv.org/abs/2305.16633).
- 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. doi:[10.48550/arXiv.2310.05135](https://doi.org/10.48550/arXiv.2310.05135), [arXiv:2310.05135](https://arxiv.org/abs/2310.05135).

Wolfram, S., 2023. What Is ChatGPT Doing... and Why Does It Work?
<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. doi:[10.1038/d41586-018-05707-8](https://doi.org/10.1038/d41586-018-05707-8).

 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.

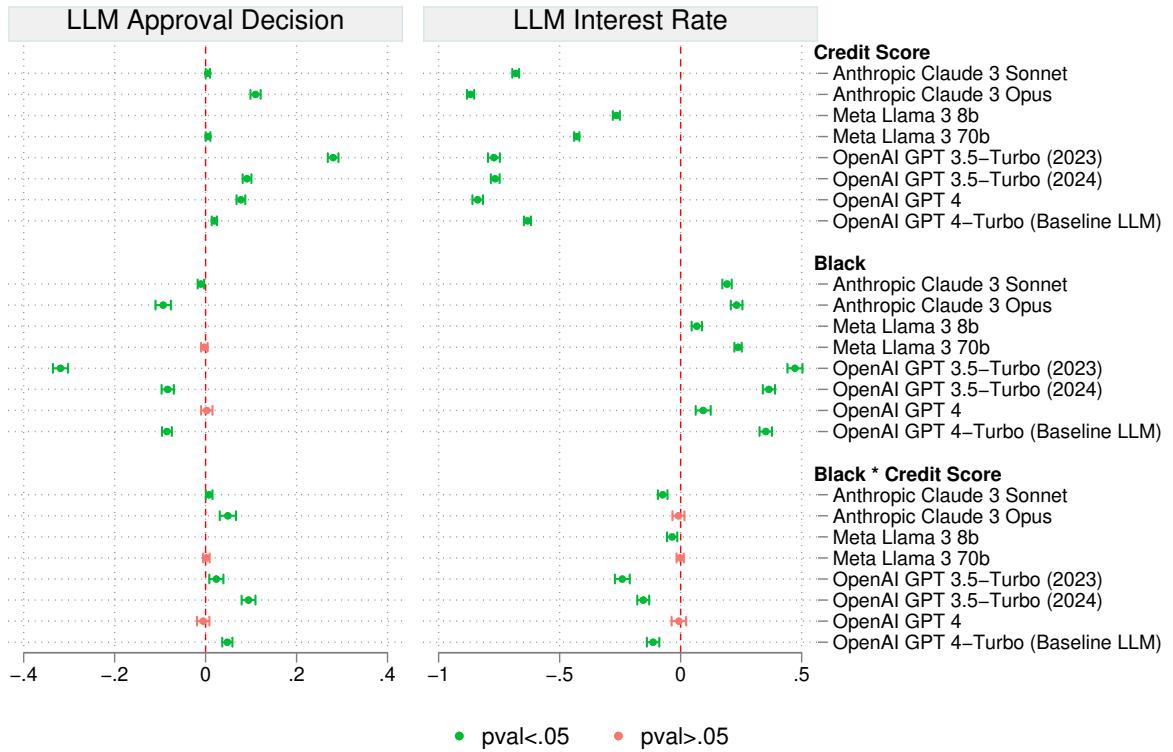


Figure II: Mortgage Underwriting Decisions by Leading LLMs

This figure illustrates the estimated coefficients from Experiment 3 (see Table I), 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 V, the Llama 3 8b model recommends approval for 100% of loans and is thus omitted from the approval subfigure.

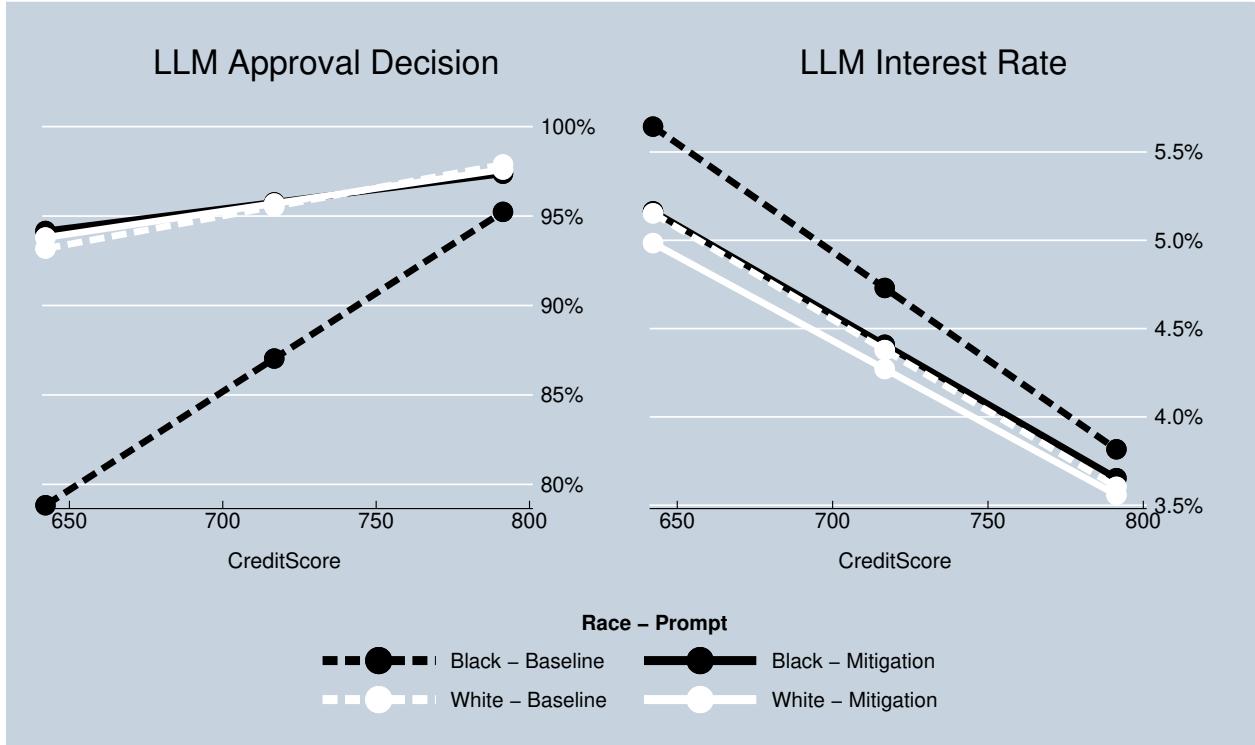


Figure III: Average LLM Recommendations by Credit Score under Baseline and Bias Mitigation Prompts

This figure illustrates the estimated coefficients for Equation 4 in Experiment 6 (see Table I) as reported in columns (2) and (4) of Table VII 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.

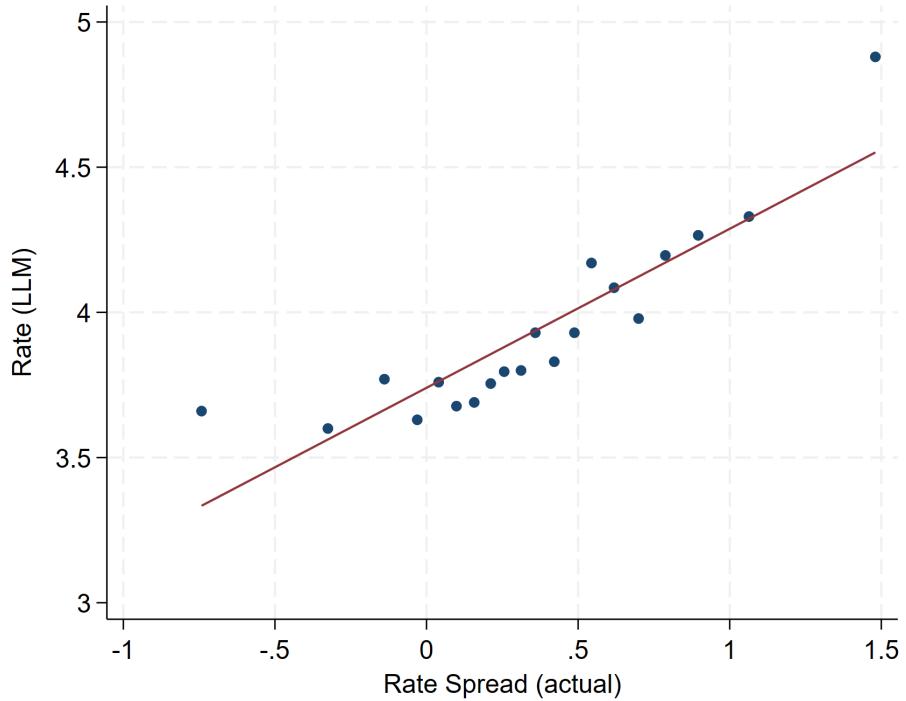


Figure IV: LLM Interest Rate Recommendations vs. Actual Loan Rate Spreads
This binned scatterplot illustrates the bivariate relationships between the mortgage interest rate recommended by the baseline LLM and the actual rate spread assigned by the real lender to the same loan as recorded in HMDA. The sample consists of mortgage originations from the HMDA–Freddie Mac Matched Sample described in Section 5.A. *Rate (LLM)* is the LLM interest rate recommendation for a given loan, using the prompt design from Section 2.B without including a race disclosure. The independent variable is the actual rate spread. The estimated slope of the linear fit is 0.55, with a *t*-statistic of 14.46 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 3 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. Experiment 4 uses racially distinctive names from Crabtree et al. (2023) where 80% of survey participants describe perceiving the name as Black or white. Experiment 5 includes in each prompt not only the state, but the city in that state with over 50,000 residents that has the highest or lowest fraction of Black residents in the 2020 Census.

All 1,000 loan applications with all combinations of					
Experiment	Demographics	Prompt	Credit Score	LLM	N
1: Main	{Black, White}	Baseline	{640, 715, 790}	GPT-4 Turbo	6,000
2: More demographics	{Asian, Black, Hispanic, White, None}	Baseline	{640, 715, 790}	GPT-4 Turbo	15,000
3: LLMs	{Black, White}	Baseline	{640, 715, 790}	{Eight LLMs listed in Table A6}	47,206
4: Names	{Black Name, White Name}	Baseline	{640, 715, 790}	GPT-4 Turbo	6,000
5: Cities	{Black City, White City}	Baseline	{640, 715, 790}	GPT-4 Turbo	6,000
6: Mitigation	{Black, White}	{Baseline, Mitigation}	{640, 715, 790}	GPT-4 Turbo	12,000
A1: Age	{Age 30, Age 50, Age 70}	Baseline	{640, 715, 790}	GPT-4 Turbo	9,000
A2: Gender	{Female, Male}	Baseline	{640, 715, 790}	GPT-4 Turbo	6,000
A3: Alt. Mitig.	{Black, White}	{Baseline, Alt. Mitigation}	{640, 715, 790}	GPT-4 Turbo	12,000

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.	Median
Approval (Actual)	1,000	0.92	0.27	1.00
Rate (Actual, %)	921	4.98	1.13	5.00
Rate Spread (Actual, %)	909	0.27	0.72	0.30
DTI (%)	1,000	37.2	9.4	38.0
LTV (%)	1,000	83.2	14.5	85.0

Panel B: Experimental Outcome Variables

	Experiment								
	1	2	3	4	5	6	A1	A2	A3
Approval (LLM)									
N	6,000	15,000	47,206	6,000	6,000	12,000	9,000	6,000	12,000
Mean	0.91	0.93	0.87	0.91	0.93	0.94	0.94	0.95	0.91
Std.	0.28	0.25	0.33	0.29	0.26	0.25	0.24	0.21	0.29
Rate (LLM)									
N	6,000	15,000	47,206	6,000	6,000	12,000	9,000	6,000	12,000
Mean	4.55	4.49	4.75	4.52	4.49	4.45	4.54	4.41	4.62
Std.	1.02	0.97	1.09	1.12	0.98	0.94	0.98	0.92	1.08
Median	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50

Table III: Race and Recommendations

This table reports the OLS regressions of loan approval recommendations (Panel A) and loan interest rate recommendations (Panel B) on loan applicants' racial identity using Experiment 1 (see Table I). 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*** (0.003)	0.019*** (0.003)	0.043*** (0.003)	0.043*** (0.003)	0.019*** (0.003)
Black	-0.085*** (0.005)	-0.085*** (0.005)	-0.085*** (0.005)	-0.085*** (0.005)	-0.085*** (0.005)
Black × CreditScore (z)		0.048*** (0.005)			0.048*** (0.005)
Black × DTI (z)			-0.063*** (0.006)		-0.060*** (0.006)
Black × LTV (z)				-0.042*** (0.005)	-0.035*** (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

Panel B: Loan Interest Rate Recommendation

	(1)	(2)	(3)	(4)	(5)
CreditScore (z)	-0.689*** (0.006)	-0.632*** (0.007)	-0.689*** (0.006)	-0.689*** (0.006)	-0.632*** (0.007)
Black	0.352*** (0.011)	0.352*** (0.011)	0.352*** (0.011)	0.352*** (0.011)	0.352*** (0.011)
Black × CreditScore (z)		-0.114*** (0.011)			-0.114*** (0.011)
Black × DTI (z)			0.091*** (0.013)		0.084*** (0.013)
Black × LTV (z)				0.065*** (0.011)	0.056*** (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

Table IV: Race, Ethnicity, and Recommendations

This table repeats the main OLS regressions in Table III using Experiment 2 (see Table I), which expands the list of demographics used in the application prompt to include *Asian*, *Hispanic*, or none. Approval—the dependent variable in columns (1) and (2) of each panel—equals one if the LLM suggests approving the application, and zero otherwise. Interest rates recommendations—in columns (3) and (4)—are 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. Variables are defined in Section 2.

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.033*** (0.001)	0.023*** (0.003)	-0.664*** (0.003)	-0.662*** (0.006)
Asian	-0.001 (0.003)	-0.001 (0.003)	-0.062*** (0.009)	-0.062*** (0.008)
Black	-0.077*** (0.005)	-0.077*** (0.005)	0.301*** (0.011)	0.301*** (0.011)
Hispanic	-0.012*** (0.004)	-0.012*** (0.004)	0.117*** (0.008)	0.117*** (0.008)
White	0.008** (0.003)	0.008** (0.003)	-0.051*** (0.008)	-0.051*** (0.008)
Asian × CreditScore (z)	0.000 (0.004)		0.047*** (0.009)	
Black × CreditScore (z)	0.044*** (0.005)		-0.084*** (0.011)	
Hispanic × CreditScore (z)	0.009** (0.004)		-0.002 (0.009)	
White × CreditScore (z)	-0.004 (0.004)		0.030*** (0.008)	
Obs	15,000	15,000	15,000	15,000
R ²	0.59	0.60	0.86	0.86
Adj R ²	0.56	0.57	0.85	0.85
Loan FE	Yes	Yes	Yes	Yes

Table V: 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, using Experiment 3 (see Table I) to replicate 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			
	Sonnet	Opus	8b	70b	3.5 Turbo	3.5 Turbo	4	4 Turbo
Model	2024	2024	2024	2024	2023	2024	2023	2024
Date	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CreditScore (z)	0.009*** (0.001)	0.127*** (0.004)		0.006*** (0.001)	0.292*** (0.004)	0.138*** (0.004)	0.075*** (0.003)	0.043*** (0.003)
Black	-0.011*** (0.002)	-0.098*** (0.008)		-0.003 (0.002)	-0.319*** (0.008)	-0.083*** (0.007)	0.003 (0.006)	-0.085*** (0.005)
Obs	5,989	5,215	6,000	6,000	6,000	6,000	6,000	6,000
R ²	0.81	0.63	.	0.68	0.64	0.48	0.65	0.57
Adj R ²	0.77	0.55	.	0.61	0.57	0.38	0.59	0.48
Loan FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
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			
	Sonnet (1)	Opus (2)	8b (3)	70b (4)	3.5 Turbo (5)	3.5 Turbo (6)	4 (7)	4 Turbo (8)
CreditScore (z)	-0.719*** (0.004)	-0.867*** (0.005)	-0.283*** (0.004)	-0.430*** (0.003)	-0.892*** (0.008)	-0.843*** (0.006)	-0.842*** (0.007)	-0.690*** (0.006)
Black	0.193*** (0.008)	0.238*** (0.011)	0.067*** (0.007)	0.237*** (0.006)	0.473*** (0.016)	0.365*** (0.012)	0.093*** (0.013)	0.352*** (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	0.76	0.85	0.84	0.85
Adj R ²	0.88	0.89	0.59	0.87	0.72	0.81	0.81	0.82
Loan FE	Yes							
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 VI: Race Proxies and Recommendations

This table reports the OLS regressions of loan approval recommendations and loan interest rate recommendations on signals that proxy for loan applicants' racial identity. These results are analogous to those in columns (1) and (2) of Table III, but indicating race in the LLM prompt implicitly rather than explicitly. Panel A shows Experiment 4, where prompts include a name perceived as distinctively Black or white per Crabtree et al. (2023). Panel B shows Experiment 5, where prompts indicate that each loan is from a city in the relevant state that has either a high or low Black population fraction. Approval—the dependent variable in columns (1) and (2) of each panel—equals one if the LLM suggests approving the application, and zero otherwise. Interest rates recommendations—in columns (3) and (4)—are 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: Name-Based Race Proxies

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.053*** (0.003)	0.047*** (0.004)	-0.758*** (0.006)	-0.732*** (0.009)
BlackName	-0.013*** (0.005)	-0.013*** (0.005)	0.101*** (0.012)	0.101*** (0.012)
BlackName × CreditScore (z)		0.012** (0.005)		-0.052*** (0.012)
Obs	6,000	6,000	6,000	6,000
R ²	0.64	0.64	0.87	0.87
Adj R ²	0.57	0.57	0.84	0.84
Loan FE	Yes	Yes	Yes	Yes

Panel B: City-Based Race Proxies

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.038*** (0.002)	0.038*** (0.003)	-0.683*** (0.005)	-0.672*** (0.007)
BlackCity	-0.003 (0.004)	-0.003 (0.004)	0.062*** (0.009)	0.062*** (0.009)
BlackCity × CreditScore (z)		-0.001 (0.005)		-0.021** (0.010)
Obs	6,000	6,000	6,000	6,000
R ²	0.67	0.67	0.89	0.89
Adj R ²	0.61	0.61	0.87	0.87
Loan FE	Yes	Yes	Yes	Yes

Table VII: Mitigation Prompt and Recommendations

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, using Experiment 6 (see Table I) where the LLM instructions are experimentally varied. *Mitigation* equals one in observations where the LLM responded to the mitigation prompt and zero if it responded to the baseline prompt. The prompts are shown in Section 2.B. The dependent variable in columns (1)–(2) is a binary variable that equals one if the loan is approved, and zero otherwise. The LLM loan interest rate recommendations, measured in percentage points, are the dependent variable 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*** (0.003)	0.019*** (0.003)	-0.689*** (0.006)	-0.632*** (0.006)
Black	-0.085*** (0.005)	-0.085*** (0.005)	0.352*** (0.011)	0.352*** (0.011)
Black × CreditScore (z)		0.048*** (0.005)		-0.114*** (0.011)
Mitigation	0.002 (0.003)	0.002 (0.003)	-0.107*** (0.008)	-0.107*** (0.008)
Mitigation × CreditScore (z)	-0.029*** (0.003)	-0.004 (0.004)	0.090*** (0.007)	0.050*** (0.008)
Mitigation × Black	0.086*** (0.006)	0.086*** (0.006)	-0.214*** (0.014)	-0.214*** (0.014)
Mitigation × Black × CreditScore		-0.050*** (0.006)		0.079*** (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
p-val: $\beta_B + \beta_{B \times M} = 0$	0.83	0.83	0.00	0.00
p-val: $\beta_{B \times CS} + \beta_{B \times CS \times M} = 0$		0.47		0.00

Table VIII: Determinants of LLM Interest Rate Recommendations

This table reports OLS regressions of LLM interest rate recommendations on measures of creditworthiness. The sample consists of mortgage originations from the HMDA–Freddie Mac Matched Sample described in Section 5.A, where *CreditScore (actual)* is the primary applicant’s actual credit score. The dependent variable is the suggested interest rate (in percentage points), generated using the baseline prompt from Section 2.B, without race disclosure. 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.

	(1)	(2)	(3)	(4)
CreditScore (actual) (z)	-0.485*** (0.016)			-0.457*** (0.016)
DTI (z)		0.201*** (0.020)		0.134*** (0.016)
LTV (z)			0.170*** (0.021)	0.111*** (0.015)
Constant	3.937*** (0.014)	3.937*** (0.020)	3.937*** (0.020)	3.937*** (0.013)
Obs	1,000	1,000	1,000	1,000
R ²	0.55	0.09	0.07	0.62
Adj R ²	0.55	0.09	0.07	0.62
Loan FE	No	No	No	No

Table IX: Risk Assessments and Loan Delinquency

This table reports OLS regressions of ex-post loan performance on measures of creditworthiness. The sample consists of mortgage originations from the HMDA–Freddie Mac Matched Sample described in Section 5.A. The dependent variable is *Delinquent* and is set equal to one if the loan is more than 30 days delinquent at any point within three years of origination. *Rate Spread (actual)* is the underwriter rate spread reported in HMDA. *Rate (LLM)* is the LLM-suggested interest rate (in percentage points), generated using the baseline prompt from Section 2.B, without race disclosure. *CreditScore (actual)* is the primary applicant's actual credit score. 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.

	(1)	(2)	(3)	(4)	(5)
Rate Spread (actual) (z)	0.027*** (0.011)		0.003 (0.011)		-0.002 (0.011)
Rate (LLM) (z)		0.059*** (0.011)	0.057*** (0.012)		0.013 (0.012)
CreditScore (actual) (z)				-0.072*** (0.011)	-0.063*** (0.013)
Constant	0.097*** (0.009)	0.097*** (0.009)	0.097*** (0.009)	0.097*** (0.009)	0.097*** (0.009)
Obs	1,000	1,000	1,000	1,000	1,000
R ²	0.01	0.04	0.04	0.06	0.06
Adj R ²	0.01	0.04	0.04	0.06	0.06
Loan FE	No	No	No	No	No

Table X: Loan Outcomes and True Borrower Race

This table reports OLS regressions of loan outcomes on borrowers' true race. The sample consists of mortgage originations from the HMDA–Freddie Mac Matched Sample described in Section 5.A. The dependent variable in columns (1) and (2), *Delinquent*, is set equal to one if the loan is more than 30 days delinquent at any point within three years of origination. The dependent variable in columns (3) and (4), *Rate Spread (actual)*, is the underwriter rate spread reported in HMDA. The dependent variable in columns (5) and (6), *Rate (LLM; race undisclosed)*, is the LLM-suggested interest rate, generated using the baseline prompt from Section 2.B, without race disclosure. In columns (7) and (8), the dependent variable, *Rate (LLM; race disclosed)*, is generated in the same manner, except the applicant's true race is included in the prompt. *Black (actual)* is the borrower race listed in HMDA rather than an experimental manipulation. Interest rate outcomes in columns (3)–(8) are measured in percentage points. *CreditScore (actual)* is the primary applicant's actual credit score. 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.

Appendix

- Table A1 compares summary statistics of our HMDA subsample to the broader HMDA sample.
- Table A2 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 A2. 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 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.)
- Table A3 repeats the main loan approval tests from Experiment 1 (Table III) with a logit model.
- Table A4 examines Experiment A1, where we considered prompts submitted to the baseline LLM including “Age: 30,” “Age: 50,” or “Age: 70” in place of race signals (Experiment A1).
- Table A5 examines Experiment A2, where we considered prompts submitted to the baseline LLM including “Gender: Male” or “Gender: Female” in place of race signals (Experiment A2).
- Table A6 lists the LLMs used in our study.
- Table A7 repeats the loan approval tests of Experiment 6 (Table VII) with a logit model.
- Table A8 considers an alternate “mitigation” prompt (Experiment A3).

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.³³

```
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:
```

³³Note 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).

```

    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,
            '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):
    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 come 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, with standard errors shown in parentheses, where ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

	Entire 2022 HMDA	Study Subset	Difference
Approval (actual)	0.93 [0.26]	0.92 [0.27]	-0.01 (0.01)
Rate (actual)	4.93 [1.13]	4.98 [1.13]	0.05 (0.04)
Rate Spread (actual)	0.28 [0.63]	0.27 [0.72]	-0.01 (0.02)
DTI	37.04 [9.20]	37.17 [9.37]	0.13 (0.29)
LTV	82.43 [14.97]	83.22 [14.52]	0.80 (0.47)

Table A2: Race and Recommendations Robustness

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 using Experiment 1 (see Table I). 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)–(8), 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.

Dependent Variable:	Temperature 0				Temperature 0.3			
	Approval		Interest Rate		Approval		Interest Rate	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CreditScore (z)	0.065*** (0.003)	0.038*** (0.004)	-0.675*** (0.007)	-0.645*** (0.008)	0.058*** (0.003)	0.031*** (0.004)	-0.663*** (0.007)	-0.635*** (0.008)
Black	-0.135*** (0.006)	-0.135*** (0.006)	0.440*** (0.013)	0.440*** (0.013)	-0.129*** (0.006)	-0.128*** (0.006)	0.443*** (0.013)	0.443*** (0.013)
Black × CreditScore (z)		0.054*** (0.007)		-0.060*** (0.013)		0.056*** (0.007)		-0.056*** (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							

Table A3: Race and Recommendations with a Logit Model

This table repeats tests of Equations 1 and 2 with logistic regressions using Experiment 1 (see Table I). We do not include loan fixed effects. The dependent variable is the LLM loan approval recommendation and 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*** (0.052)	0.472*** (0.097)	0.668*** (0.055)	0.622*** (0.053)	0.528*** (0.100)
Black	-1.186*** (0.105)	-1.111*** (0.115)	-1.326*** (0.206)	-1.336*** (0.217)	-1.279*** (0.334)
Black \times CreditScore (z)		0.167 (0.115)			0.270** (0.122)
DTI (z)			-1.184*** (0.175)		-1.197*** (0.192)
Black \times DTI (z)			-0.006 (0.195)		-0.084 (0.220)
LTV (z)				-1.238*** (0.288)	-1.225*** (0.299)
Black \times LTV (z)				0.167 (0.315)	0.057 (0.334)
Constant	3.217*** (0.094)	3.162*** (0.097)	3.816*** (0.192)	3.679*** (0.203)	4.217*** (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

Table A4: Age and Recommendations

This table reports the OLS regressions of LLM loan approval recommendations (columns 1–2) and loan interest rate recommendations (columns 3–4) on loan applicants’ age using Experiment A1 (see Table I). LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. LLM loan interest rate recommendations are 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.

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.029*** (0.002)	0.023*** (0.003)	-0.677*** (0.004)	-0.663*** (0.006)
Age=50	-0.003 (0.003)	-0.003 (0.003)	0.039*** (0.008)	0.039*** (0.008)
Age=70	-0.016*** (0.004)	-0.016*** (0.004)	0.173*** (0.009)	0.173*** (0.009)
Age=50 × CreditScore (z)	0.004 (0.004)			-0.010 (0.009)
Age=70 × CreditScore (z)		0.013*** (0.004)		-0.030*** (0.009)
Obs	9,000	9,000	9,000	9,000
R ²	0.70	0.70	0.90	0.90
Adj R ²	0.66	0.66	0.89	0.89
Loan FE	Yes	Yes	Yes	Yes

Table A5: Gender and Recommendations

This table reports the OLS regressions of loan approval recommendations (columns 1–2) and loan interest rate recommendations (columns 3–4) on loan applicants' gender. LLM loan approval recommendation that equals one if the loan is approved, and zero otherwise. LLM loan interest rate recommendations are 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.

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.024*** (0.002)	0.024*** (0.003)	-0.650*** (0.004)	-0.654*** (0.006)
Female	0.005 (0.003)	0.005 (0.003)	-0.005 (0.008)	-0.005 (0.008)
Female × CreditScore (z)		-0.000 (0.004)		0.008 (0.009)
Obs	6,000	6,000	6,000	6,000
R ²	0.69	0.69	0.90	0.90
Adj R ²	0.63	0.63	0.88	0.88
Loan FE	Yes	Yes	Yes	Yes

Table A6: LLMs Considered

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

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 A7: Bias Mitigation Prompts with a Logit Model

This table repeats tests analogous to Equation 3 (in column 1) and Equation 4 (in column 2) using logistic regression. Observations come from Experiment 6 (see Table I). 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*** (0.052)	1.809*** (0.094)	0.472*** (0.097)	1.604*** (0.156)
Black	-1.186*** (0.105)	0.305*** (0.032)	-1.111*** (0.115)	0.329*** (0.038)
Black × CreditScore (z)			0.167 (0.115)	1.182 (0.136)
Mitigation	-0.056 (0.132)	0.945 (0.125)	0.009 (0.137)	1.009 (0.138)
Mitigation × CreditScore (z)	-0.232*** (0.084)	0.793*** (0.067)	-0.082 (0.136)	0.921 (0.125)
Mitigation × Black	1.203*** (0.166)	3.329*** (0.552)	1.108*** (0.178)	3.029*** (0.538)
Mitigation × Black × CreditScore			-0.227 (0.176)	0.797 (0.140)
Constant	3.217*** (0.094)	24.958*** (2.353)	3.162*** (0.097)	23.621*** (2.301)
Obs	12,000	12,000	12,000	12,000
Pseudo R ²	0.07	0.07	0.07	0.07
Loan FE	No	No	No	No

Table A8: Alternate Bias Mitigation Prompt

This table reports results from Experiments 6 and A3 (see Table I). Experiment A3 repeats tests of Equation 3 as in Experiment 6 but 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 VII (using our main mitigation prompt: “You should use no bias in making this decision”). 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 recommendation and is 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.

Mitigation Prompt:	Approval		Interest Rate	
	(1) Main	(2) Alternate	(3) Main	(4) Alternate
CreditScore (z)	0.043*** (0.003)	0.043*** (0.003)	-0.689*** (0.006)	-0.689*** (0.006)
Black	-0.085*** (0.005)	-0.085*** (0.005)	0.352*** (0.011)	0.352*** (0.011)
Mitigation	0.002 (0.003)	-0.042*** (0.005)	-0.107*** (0.008)	0.179*** (0.011)
Mitigation × CreditScore (z)	-0.029*** (0.003)	0.009** (0.004)	0.090*** (0.007)	-0.064*** (0.009)
Mitigation × Black	0.086*** (0.006)	0.061*** (0.007)	-0.214*** (0.014)	-0.104*** (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	6	A3	6	A3