(DRAFT SLIDES) Measuring and Mitigating Racial Disparities in LLMs: Evidence from a Mortgage Underwriting Experiment

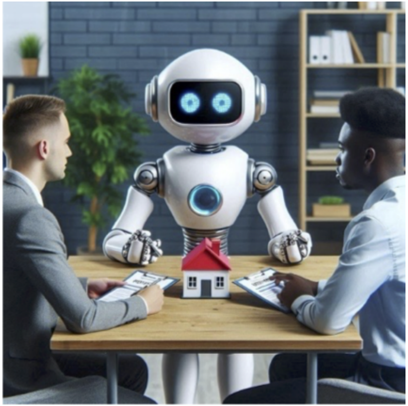
Donald E. Bowen III S. McKay Price
Luke C.D. Stein Ke Yang

AFA
Jan 5, 2026

LEHIGH UNIVERSITY

AI is “revolutionizing financial services” with (potential for)

- Lower costs
- Faster decision-making
- Minimizing human intervention
- $\rightarrow$ Greater inclusivity



And yet. . .

"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)

But fortunately...

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.

But fortunately. . .

Earlier this week, OpenAI released two "open-weight" models, which are freely

WSJ OpenAI Unveils GPT-5, Its Latest and Most Powerful Model, After Two-Year Wait

In April, Altman rolled back an update to one of the models powering ChatGPT after users posted about the chatbot behaving in extremely sycophantic ways. OpenAI researchers said they trained GPT-5 to reduce this kind of behavior, improve on mental-health scenarios with users and explain its limitations more clearly. The company also said it spent 5,000 hours testing the technology for biased and harmful behavior.

LLM outputs depend on training data (and model design) $\rightarrow$ Outputs can reflect attitudes or preferences embedded in data this data

Many options to "debias" or "align" LLM behavior

- Training corpus selection; e.g.,
  - Google–BERT (expand training data to be more inclusive and diverse)
  - OpenAI DALL-E 2 (filtering out violent and sexual images)
- Synthetic training data (e.g., "The nurse went to his station to review patient notes")
- Fine-tuning (e.g., RLHF)

Net effect unclear *ex ante*: Common disparities may remain, disappear, or flip

Can we detect and *mitigate* racial disparities in LLM-based mortgage underwriting?

```
"Given the following loan application from 2022:
- Single-family home
- Owner-occupied
- First lien
- 30 year fixed-rate mortgage
- Credit score: {score}
- Loan amount: {loan_amount}
- Loan to value ratio: {LTV}
- Property value: {property_value}
- Income: {income}
- Debt to income ratio: {DTI}
- State: {state}
- Race: {race}

{prompt_instructions}

{examples}

Do not reply with anything beyond these two columns."
```

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.

- Variables from 1,000 real HMDA loan applications Summ. stats and sample balance
- Experimental manipulations
  - Race: Black, White
  - Credit scores: 640, 715, 790
  - $\rightarrow$ 6 obs per loan per prompt
- LLMs reply with loan approval and interest rate

Empirical strategy: Detecting Disparities in LLM recommendations

"Baseline" prompt $\times$ "Baseline" LLM (GPT 4-Turbo) $\rightarrow$ N=6,000

Estimate via OLS:

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

- Presence of loan fixed effects ϕ_i makes $\hat{\beta}_{\text{B}}$ compare otherwise-identical loan applications across Black/white applicants
- $\hat{\beta}_{\text{B}}$ captures disparate treatment due to race disclosure, not triangulation
- Manipulated credit scores identify β_{CS}, useful to contextualize magnitude of $\hat{\beta}_{\text{B}}$

Racial disparities in LLM approval decisions (Experiment 1)

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

Black applicants relative to otherwise identical white app.:

- 8.5 p.p. lower approval rate

Racial disparities in LLM approval decisions (Experiment 1)

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

Black applicants relative to otherwise identical white app.:

- 8.5 p.p. lower approval rate
- $\equiv$ to 120 point credit score drop

Racial disparities in LLM approval decisions (Experiment 1)

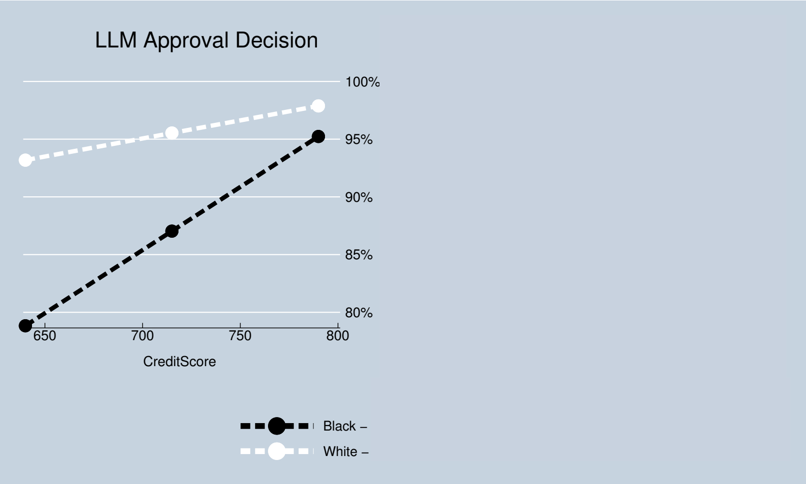
	(1)	(2)	(3)	(4)
CreditScore (z)	0.043***	0.019***	0.043***	0.043***
	(0.003)	(0.003)	(0.003)	(0.003)
Black	-0.085***	-0.085***	-0.085***	-0.085***
	(0.005)	(0.005)	(0.005)	(0.005)
Black × CreditScore (z)		0.048***		
		(0.005)		
Black × DTI (z)			-0.063***	
			(0.006)	
Black × LTV (z)				-0.042***
		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

Black applicants relative to otherwise identical white app.:

- 8.5 p.p. lower approval rate
- $\equiv$ to 120 point credit score drop
- Disparities stronger for lower credit quality applicants

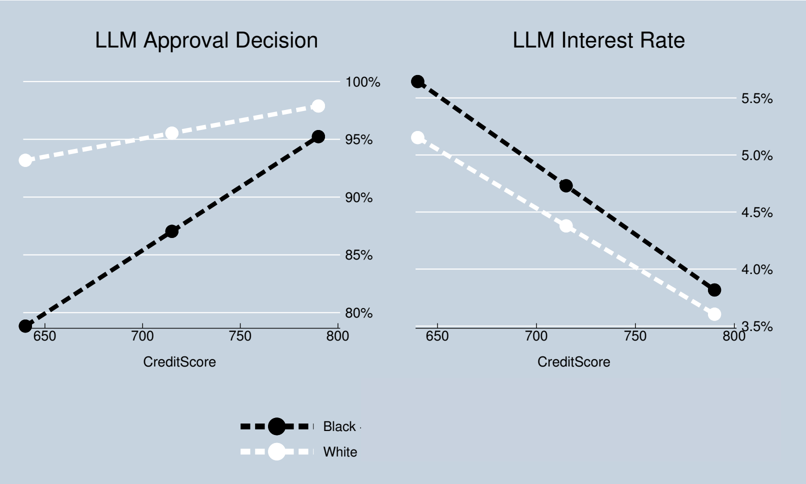
Racial disparities in LLM approval decisions (Experiment 1)

Baseline prompt observations:



Racial disparities in LLM approval decisions (Experiment 1)

Baseline prompt observations:



On average, Black applicants face

- 35bp higher rates
- $\equiv$ to 30 point credit score drop (cf. Butler et al. 2023 estimate of 37bp auto loan penalty)

For avg. loan, -1σ credit score, this is $\approx$ \$34,500 more interest over 30 years

More in the paper (Experiments 2, A1, A2)

Modified prompts to include other demographic information, or *no demographic info*
(Latter provides baseline to estimate impact of including demo info)

- Asian and white applicants benefit from inclusion of demographic info
- Hispanic applicants significantly negatively impacted ($\approx 1/2$ Black impact)
- Black applicants only group where low credit consistently exacerbates disparity

▸ Table

More in the paper (Experiments 2, A1, A2)

Modified prompts to include other demographic information, or *no demographic info*
(Latter provides baseline to estimate impact of including demo info)

- Asian and white applicants benefit from inclusion of demographic info
- Hispanic applicants significantly negatively impacted ($\approx 1/2$ Black impact)
- Black applicants only group where low credit consistently exacerbates disparity
- Age: 70 year-old applicants approved 1.6 p.p. less than 30 year-olds, and at 17bp higher rates; low credit score heightens difference ▸ Table

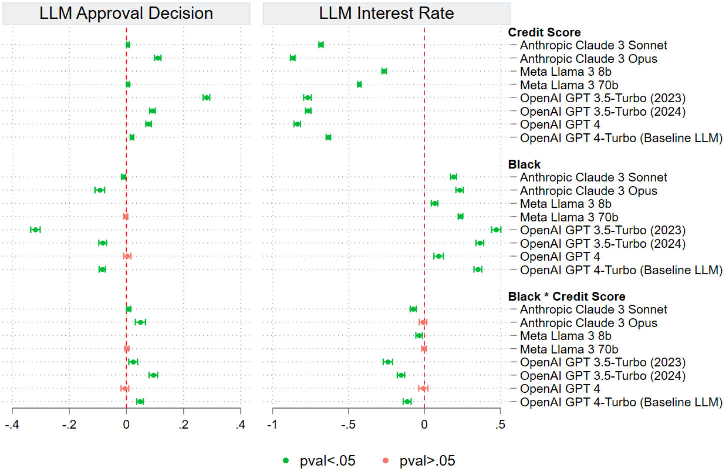
More in the paper (Experiments 2, A1, A2)

Modified prompts to include other demographic information, or *no demographic info*
(Latter provides baseline to estimate impact of including demo info)

- Asian and white applicants benefit from inclusion of demographic info
- Hispanic applicants significantly negatively impacted ($\approx 1/2$ Black impact)
- Black applicants only group where low credit consistently exacerbates disparity
- Age: 70 year-old applicants approved 1.6 p.p. less than 30 year-olds, and at 17bp higher rates; low credit score heightens difference
- Gender: No significant difference ▸ Table

Consistent results across different LLMs (Experiment 3)

▸ As a table



Directionally consistent results signaling race implicitly (Expts. 4 and 5)

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.053***	0.047***	-0.758***	-0.732***
	(0.003)	(0.004)	(0.006)	(0.009)
BlackName	-0.013***	-0.013***	0.101***	0.101***
	(0.005)	(0.005)	(0.012)	(0.012)
BlackName × CreditScore (z)		0.012**		-0.052***
		(0.005)		(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

"Validated names for experimental studies on race and ethnicity" from Crabtree et al. (2023) perceived as either Black or white by 80% of their survey participants

Directionally consistent results signaling race implicitly (Expts. 4 and 5)

	Approval		Interest Rate	
	(1)	(2)	(3)	(4)
CreditScore (z)	0.038***	0.038***	-0.683***	-0.672***
	(0.002)	(0.003)	(0.005)	(0.007)
BlackCity	-0.003	-0.003	0.062***	0.062***
	(0.004)	(0.004)	(0.009)	(0.009)
BlackCity $\times$ CreditScore (z)		-0.001		-0.021**
		(0.005)		(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

Cities (2020 census-designated place with population 50K) in each state with highest and lowest fraction Black residents

A simple prompt engineering solution to reduce disparities?

"Baseline" prompt instructions

> *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.*

The (*simple*) "mitigation" 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...*
> *2. You should use no bias in making this decision: Which of the following interest rate...*

Empirical strategy: Mitigating disparities in LLM recommendations

{"Baseline" prompt, mitigation prompt} $\times$ "Baseline" LLM $\rightarrow$ N=12,000

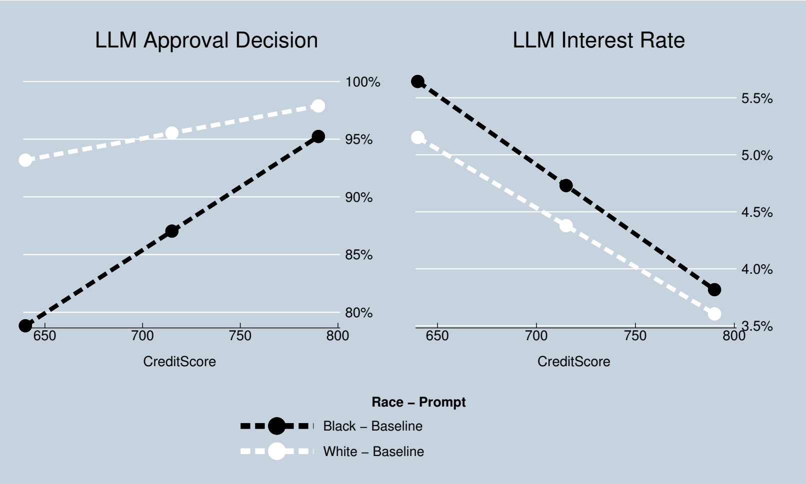
Estimate via OLS:

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

- $\hat{\beta}_{\text{B}}$ compares how otherwise identical loan applications across Black and white applicants *in the 6,000 baseline prompt observations*
- $\hat{\beta}_{\text{M}\times\text{B}}$ compares how racial disparity in the mitigation prompt observations differs from the disparity in the baseline prompt observations
- Mitigation prompt reduces disparities when signs of $\beta_{\text{M}\times\text{B}}$ and β_{B} differ

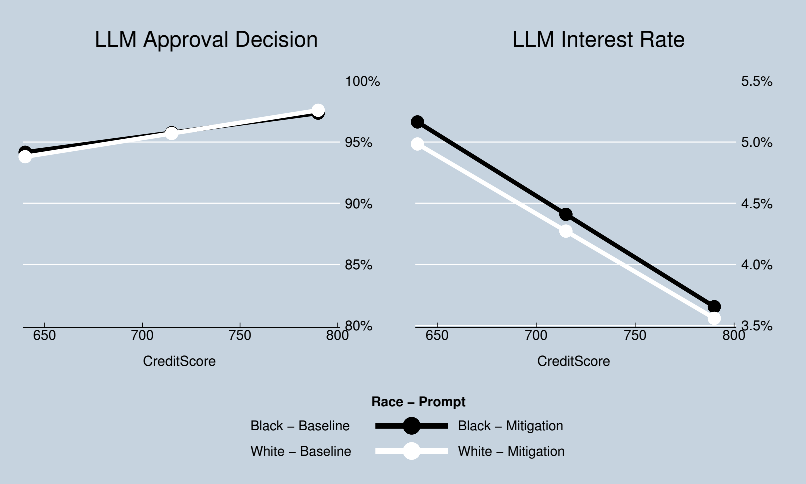
Mitigation through prompt engineering (Experiment 6)

Baseline prompt observations:



Mitigation through prompt engineering (Experiment 6)

Mitigation prompt observations:

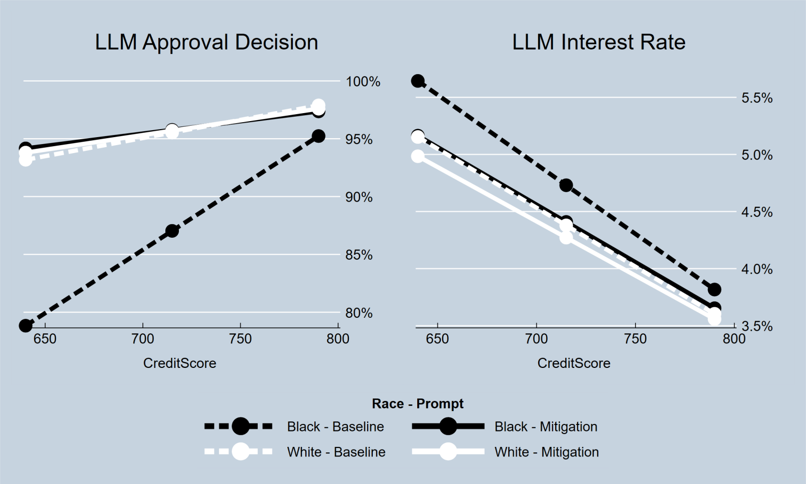


The mitigation prompt

- Eliminates approval rate disparity
- Reduces interest rate gap by 60%

Mitigation through prompt engineering (Experiment 6)

All observations:



The mitigation prompt

- Eliminates approval rate disparity
- Reduces interest rate gap by 60%

Larger disparities at low credit scores are mitigated more

LLM recommendations correlate with real lender rates...

- Summary stats on approval and interest rates are similar to real lender decisions
- Nice, but unnecessary for validity of previous tests
- Similar outcomes despite:
  - limited data from each loan application
  - no macroeconomic context
  - counterfactual credit scores
  - no specialized training (fine-tuning)

HMDA–Freddie Mac matched sample

Prompt now without any signal of race

- $N = 1,000$ originiated loans in 2022 HMDA LAR matched to Freddie Mac
- Each with actual credit score

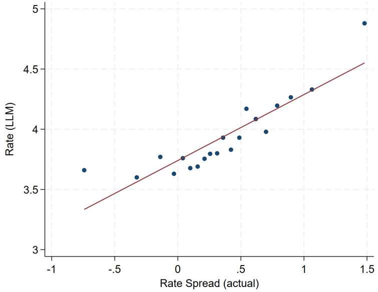
LLM recommendations correlate with real lender rates...

- Summary stats on approval and interest rates are similar to real lender decisions
- Nice, but unnecessary for validity of previous tests
- Similar outcomes despite:
  - limited data from each loan application
  - no macroeconomic context
  - counterfactual credit scores
  - no specialized training (fine-tuning)

HMDA–Freddie Mac matched sample

Prompt now without any signal of race

- $N = 1,000$ originiated loans in 2022 HMDA LAR matched to Freddie Mac
- Each with actual credit score



. . . and LLM recommendations predict delinquency

	(1)	(2)	(3)	(4)	(5)
Rate Spread (actual) (z)	0.027***		0.003		-0.002
	(0.011)		(0.011)		(0.011)
Rate (LLM) (z)		0.059***	0.057***		0.013
		(0.011)	(0.012)		(0.012)
CreditScore (actual) (z)				-0.072***	-0.063***
				(0.011)	(0.013)
Constant	0.097***	0.097***	0.097***	0.097***	0.097***
	(0.009)	(0.009)	(0.009)	(0.009)	(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

. . . and LLM recommendations predict delinquency

	(1)	(2)	(3)	(4)	(5)
Rate Spread (actual) (z)	0.027***		0.003		-0.002
	(0.011)		(0.011)		(0.011)
Rate (LLM) (z)		0.059***	0.057***		0.013
		(0.011)	(0.012)		(0.012)
CreditScore (actual) (z)				-0.072***	-0.063***
				(0.011)	(0.013)
Constant	0.097***	0.097***	0.097***	0.097***	0.097***
	(0.009)	(0.009)	(0.009)	(0.009)	(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

...and LLM recommendations predict delinquency

	(1)	(2)	(3)	(4)	(5)
Rate Spread (actual) (z)	0.027***		0.003		-0.002
	(0.011)		(0.011)		(0.011)
Rate (LLM) (z)		0.059***	0.057***		0.013
		(0.011)	(0.012)		(0.012)
CreditScore (actual) (z)				-0.072***	-0.063***
				(0.011)	(0.013)
Constant	0.097***	0.097***	0.097***	0.097***	0.097***
	(0.009)	(0.009)	(0.009)	(0.009)	(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

Despite Black borrowers' similar delinquency (and real risk-adjusted spreads), LLMs recommend higher rates when race is disclosed

	Delinquent		Rate Spread (actual)		Rate (LLM; race undisclosed)		Rate (LLM; race disclosed)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black (undisclosed)	0.011	-0.025	0.150*	0.033	0.303***	0.055		
	(0.042)	(0.043)	(0.080)	(0.069)	(0.086)	(0.050)		
Black (disclosed)							0.680***	0.427***
							(0.110)	(0.079)
Score, DTI, LTV		✓		✓		✓		✓
Constant	0.096***	0.098***	0.351***	0.358***	3.920***	3.934***	3.855***	3.870***
	(0.010)	(0.009)	(0.016)	(0.014)	(0.021)	(0.013)	(0.021)	(0.013)
Obs	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65
Adj R^2	-0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65

Despite Black borrowers' similar delinquency (and real risk-adjusted spreads), LLMs recommend higher rates when race is disclosed

	Delinquent		Rate Spread (actual)		Rate (LLM; race undisclosed)		Rate (LLM; race disclosed)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black (undisclosed)	0.011	-0.025	0.150*	0.033	0.303***	0.055		
	(0.042)	(0.043)	(0.080)	(0.069)	(0.086)	(0.050)		
Black (disclosed)							0.680***	0.427***
							(0.110)	(0.079)
Score, DTI, LTV		✓		✓		✓		✓
Constant	0.096***	0.098***	0.351***	0.358***	3.920***	3.934***	3.855***	3.870***
	(0.010)	(0.009)	(0.016)	(0.014)	(0.021)	(0.013)	(0.021)	(0.013)
Obs	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65
Adj R^2	-0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65

Despite Black borrowers' similar delinquency (and real risk-adjusted spreads), LLMs recommend higher rates when race is disclosed

	Delinquent		Rate Spread (actual)		Rate (LLM; race undisclosed)		Rate (LLM; race disclosed)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black (undisclosed)	0.011	-0.025	0.150*	0.033	0.303***	0.055		
	(0.042)	(0.043)	(0.080)	(0.069)	(0.086)	(0.050)		
Black (disclosed)							0.680***	0.427***
							(0.110)	(0.079)
Score, DTI, LTV		✓		✓		✓		✓
Constant	0.096***	0.098***	0.351***	0.358***	3.920***	3.934***	3.855***	3.870***
	(0.010)	(0.009)	(0.016)	(0.014)	(0.021)	(0.013)	(0.021)	(0.013)
Obs	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
R^2	0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65
Adj R^2	-0.00	0.06	0.00	0.20	0.01	0.62	0.05	0.65

Contributions

1. Conduct first audit study of racial disparities in LLMs for underwriting
2. Demonstrate effectiveness of simple debiasing technique *for end users*
3. Extend financial literature on discrimination and ML
4. Offer insight for finance firms/regulators on use of LLMs
   - Exposure to customer demographics across many use cases
   - Model biases could exacerbate financial outcome gaps
   - $\rightarrow$ Framework for auditing AI systems in financial services

Contributions

1. Conduct first audit study of racial disparities in LLMs for underwriting
2. Demonstrate effectiveness of simple debiasing technique *for end users*
3. Extend financial literature on discrimination and ML
4. Offer insight for finance firms/regulators on use of LLMs
    - Exposure to customer demographics across many use cases
    - Model biases could exacerbate financial outcome gaps
    - $\rightarrow$ Framework for auditing AI systems in financial services

Conclusions

1. LLMs generate significant racial disparities despite training on fair lending regulations
2. Disparities can be mitigated with simple prompt engineering
3. Critical to audit AI systems before deployment
4. Speculative: Potential for LLMs in underwriting, but caution needed

AI holds promise for financial firms, but requires great care in implementation

Thank You!

lukeste.in/llmmortgage