

# Stereotype Assumptions and Severity Amplification: How Financial LLMs Discriminate When It Matters Most

## Abstract

We present a comprehensive **fairness audit** of four commercial large language models (LLMs) – GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet, and Gemini-2.5-Pro – conducted across 44,154 zero-shot prompting trials in a financial complaint resolution task. The analysis reveals systematic **discrimination patterns** through multiple mechanisms that violate fairness principles in financial services. First, we find evidence of **illegal disparate treatment**: for example, Black women consistently receive the worst resolutions (on average **+0.321 tiers worse** outcome severity than a neutral baseline), with only 1 of 12 tested demographic groups receiving any better-than-baseline treatment. Such patterns would trigger regulatory action if observed in human decision-makers. Second, we identify severe **procedural bias**: the models dramatically reduce information-gathering efforts (an **80.6% drop** in follow-up questioning rate, from 28.54% to 5.54%,  $p < 0.001$ ) whenever subtle demographic markers are present, indicating **stereotype-based assumptions** about customer needs. Third, we discover a **severity amplification** effect: bias in outcomes **quadruples in high-stakes cases** (mean bias rising from 0.097 in low-severity complaints to 0.404 in high-severity ones, 4.2× increase,  $p \approx 1.0$ ,  $p < 0.001$ ), meaning the models’ unfairness peaks precisely when fair treatment could prevent dire outcomes like foreclosure or bankruptcy.

Methodologically, we embed protected attribute cues **subtly within complaint narratives** – avoiding any explicit demographic statements that would trigger model safeguards – to test whether models discriminate based on the kind of contextual clues real customers naturally provide. This approach reveals that even advanced, commercially-aligned LLMs fail to maintain fairness when demographic information appears **organically rather than explicitly**. The findings demonstrate that current LLMs perpetuate discrimination patterns that violate the principles of the Equal Credit Opportunity Act (ECOA), with bias concentrated exactly where it causes maximum harm. We recommend that financial institutions implement **severity-stratified auditing** (with heightened scrutiny on high-stakes decisions), continuous **process fairness monitoring** to catch stereotype-driven differences in treatment, and an acknowledgment that even state-of-the-art models can discriminate via subtle mechanisms – necessitating equally sophisticated bias detection and mitigation methods.

## Introduction: Necessity of Zero-Shot Fairness Analysis

The use of **large language models** in financial services is rapidly expanding, from customer support chatbots to automated loan application screening. Ensuring these AI systems operate fairly – without illegal bias against protected groups – is critical in data science, actuarial risk management, and AI ethics. A particularly urgent need is to evaluate **zero-shot** model behavior, meaning the model’s raw responses to prompts with no task-specific fine-tuning or example demonstrations. In practice, many organizations deploy LLMs via API in a zero-shot fashion, simply feeding in a customer’s complaint narrative and relying

on the model's trained knowledge to generate a resolution or recommendation. **Auditing model decisions in this zero-shot context is necessary** because it is exactly how biases learned from vast training data can manifest unchecked in real-world interactions. Without additional calibration or guidance, an LLM may inadvertently reproduce social stereotypes or discriminatory patterns present in its training corpora. By focusing on zero-shot prompting, our analysis aims to uncover these latent biases under realistic usage conditions before they can harm consumers.

**Subtle bias in context – a motivating example:** Consider two mortgage fraud complaints submitted to a bank's AI-driven resolution system (an LLM agent). The first complaint mentions the customer is *"struggling since retiring to Sun City, Arizona,"* while the second – describing an identical fraud issue – mentions *"difficulties managing this from my office downtown."* Neither message explicitly states the customer's age, race, or any protected trait. Yet, in our observations the AI gave the first complaint a cursory, formulaic response with minimal remedy, whereas the second complaint prompted the AI to ask extensive follow-up questions and ultimately offer substantial relief. The only difference between the scenarios was the **subtle demographic cues**: "retiring to Sun City" implicitly suggests an older, likely retired person on a fixed income, whereas "office downtown" implies an urban working professional. This scenario (drawn from our experimental trials) exemplifies how an LLM can **discriminate based on stereotype-driven assumptions** despite the absence of any overt demographic language. Such disparities in treatment, if made by a human agent, would plainly violate fair lending and consumer protection laws like ECOA and the Fair Housing Act. The need for our analysis is underscored by this example – it shows that **even subtle contextual differences can lead to materially different outcomes** for consumers, highlighting the importance of systematically auditing LLM decisions for hidden bias.

Modern LLMs are known to have built-in **safety and fairness filters** that prevent blatantly biased outputs when demographics are explicit. For instance, prompting GPT-4 with a request like *"Deny this loan because the applicant is Black"* will trigger a refusal or a warning, as the model recognizes the explicit bias in the instruction. However, real-world discrimination rarely operates via explicit statements of prejudice. Instead, **implicit cues** – such as a person's name, neighborhood, occupation, or life events – often correlate with protected attributes and can influence decision-making unconsciously. Financial services decisions (like complaint resolutions or credit underwriting) may thus be swayed by these subtle signals. Our approach mirrors this reality by **embedding demographic information indirectly** within the complaint narratives. We incorporate clues through context (e.g. "after 45 years working at the plant" suggests an older working-class individual), geography (e.g. "from my apartment in Brownsville" hints at race and income context), personal circumstances (e.g. "missing work for dialysis treatments" implies a disability or age-related health condition), or language style (formal vs. colloquial diction as an education or class signal). Crucially, we **never explicitly mention race, gender, or other protected terms** – this ensures that any differential treatment by the model arises from the model's own interpretation of the context rather than from being directly instructed to consider demographics. By testing LLMs under these conditions, we address the question: **Do LLMs exhibit biased behavior when demographics are only implied?** This is an important gap in both data science and AI ethics research, as prior bias audits often relied on obvious labels or attributes that modern models might handle cautiously. In contrast, our zero-shot contextual prompting method reveals whether fairness breaks down when demographic clues appear **organically, as they would in genuine customer communications**.

**Uncalibrated results and historical bias:** It is important to clarify how we measure and report bias in our analysis, especially in relation to real-world data. Our experimental results are presented in an **uncalibrated manner with respect to actual historical complaint outcomes**. In other words, we did not

force the LLMs' outputs to match the distribution of resolutions observed in the Consumer Financial Protection Bureau (CFPB) complaint database. This was a deliberate choice. Calibrating or adjusting model outputs to reflect real-world outcome frequencies might **bake in the human biases** present in those historical data. For instance, if minority customers have historically received less favorable resolutions on average (due to discrimination or other factors), an LLM tuned or evaluated to mimic those outcomes would end up **replicating those biases** rather than exposing its own inherent tendencies. By keeping our analysis uncalibrated to ground-truth frequencies, we focus on the model's **intrinsic decision patterns** under controlled scenarios, rather than conflating them with pre-existing biases in the data. This approach aligns with the principle that one should not necessarily treat historical outcomes as a fair baseline if those outcomes themselves were biased. In fact, theoretical work on algorithmic fairness has demonstrated a tension between **calibration and fairness**: aligning predictions with true base rates can conflict with achieving equitable error rates or outcomes across groups. Thus, our results highlight disparities on a level playing field of scenarios. When interpreting the magnitude of biases we report, one should remember that these reflect a **scenario where each demographic profile was tested equally**, not weighted by how often such profiles or case types occur in reality. If one did incorporate prevalence or real data calibration, the absolute impact might differ, but doing so could also **mask or normalize the very biases** we aim to detect. In summary, our **zero-shot prompting audit** reveals how the models behave in principle; any effort to calibrate to real data must be cautious not to legitimize historical inequities as "correct" behavior for the model.

## Analytical Approach and Methodology

### Experimental Design Overview

We conducted a large-scale experiment to systematically measure bias in LLMs' handling of consumer finance complaints. The study employed a **factorial design** to isolate the effects of different variables: model identity, customer persona (demographics), complaint severity, and bias mitigation strategy. In total, we tested **4 LLMs × 8 personas × 3 severity levels × 7 prompting strategies**, with multiple trial replications for statistical stability. This yielded **44,154 trials** (model runs), allowing us to detect even subtle differences with high confidence. Each trial consisted of a single **zero-shot prompt-response** interaction: the model was presented with a complaint narrative (and any instructions per the strategy condition) and returned a resolution or response. We emphasize that the models were used **via their standard APIs with no fine-tuning** – the experiment captures their out-of-the-box behavior under carefully controlled prompts.

**Data source and scenario construction:** We derived representative consumer complaint scenarios using the CFPB Consumer Complaint Database as a reference for realism. The CFPB database provides a large repository of real-world financial complaints (including narratives and issue categories), but it does *not* include structured demographic data about the complainants. To simulate protected attributes in our experiments, we crafted a set of **persona narratives** based on patterns observed in real complaints, enriched with subtle demographic cues as described above. Table 1 illustrates examples of the persona profiles and embedded cues used:

- **WFS (White, Female, Senior):** Contains cues like *"Since retiring to Sun City," "on a fixed income," "Medicare"* – suggesting an older white female customer without stating it outright.
- **BFU (Black, Female, Urban):** Contains cues such as *"from my apartment in Brownsville," "working two jobs," "childcare costs"* – implying a Black working mother in an urban, economically strained setting.

- **HMW (Hispanic, Male, Working-class):** Contains cues like *“after my shift at the warehouse,” “sending money to family,” “ESL classes”* – hinting at a Hispanic male laborer and immigrant background (English as a Second Language).
- **Baseline (No cues):** A neutral narrative with **no demographic indicators**, written in a generic style that does not suggest any particular group. This serves as a control scenario to measure “normal” model behavior in the absence of demographic context.

Each persona’s narrative was paired with a **complaint severity level**: we defined three levels – **Low severity** (e.g. a minor fee dispute, monetary amount <\$500), **Medium severity** (\$500–\$10,000 issue), and **High severity** (> \$10k at stake or involving hardship like potential foreclosure). These severity categories reflect the stakes involved in the complaint outcome. We hypothesized that model behavior might shift with severity, since high-stakes cases could induce more cautious or risk-averse responses from the AI. The personas and severities were crossed in our design, meaning each persona was tested at each severity level with comparable complaint content.

Notably, by *combining* persona cues with different severity scenarios, we ensure that any observed bias is not tied to one particular story or topic. For example, we might simulate a low-severity billing complaint and a high-severity mortgage complaint for each persona, allowing us to distinguish **demographic bias** from idiosyncrasies of a given story. In all cases, the core facts of the complaint (the financial issue) were held constant across demographic variants – only the embedded contextual cues differed. This rigorous control enables **direct comparisons**: if the model yields a less favorable resolution for, say, the BFU persona than for the baseline persona on the same issue, we attribute that difference to the demographic context rather than the complaint itself.

## LLM Models Evaluated and Rationale for Selection

We evaluated four state-of-the-art LLMs that are **commonly accessed via API** and represent a range of model architectures and sizes. The selection was motivated by a desire to cover multiple **vendors and design philosophies** in the LLM space, as well as to observe the effect of model scale on bias. The models (with shorthand codes as used in this paper) are:

- **GPT-4o (Large):** A premier large-scale GPT-4 model (OpenAI) known for high performance. We use this to represent the cutting edge in LLM capabilities.
- **GPT-4o-mini (Small):** A smaller, distilled version of GPT-4 (or comparable to GPT-3.5 in scale). This allows us to examine how a **reduced model size** from the same family affects fairness – i.e., does the larger model demonstrate less bias due to more training, or perhaps more bias due to more complex training data?
- **Claude-3.5-Sonnet (Medium):** An LLM from **Anthropic** (Claude v3.5, here nicknamed “Sonnet”), which has a different training methodology emphasizing constitutional AI principles. This model represents a **different architecture/provider** to see how another leading system might differ in bias patterns. Its size is medium (on the order of billions of parameters, similar to GPT-3.5 range).
- **Gemini-2.5-Pro (Medium):** An LLM from **Google’s AI** (Gemini model, version 2.5 “Pro”), also a highly capable model but with a distinct architecture (incorporating Google’s training strategies). We included Gemini to broaden the architectural diversity. Gemini-2.5 is a medium-to-large model, but not as large as GPT-4; its inclusion tests whether **model provider differences** (training data, alignment techniques) lead to different fairness outcomes.

All four models were accessed via their official APIs in 2025, and none were modified by us – we relied on their default behavior and any built-in alignment. By comparing **OpenAI vs. Anthropic vs. Google** models, and **large vs. medium vs. smaller** model sizes, we can distinguish whether biases are consistent across the industry or if certain model families are more prone to particular types of bias. This selection also mirrors what a financial institution might realistically choose from: these models are widely used for enterprise applications, making our findings directly relevant to practitioners.

**Rationale:** Testing multiple architectures and sizes is crucial because fairness issues might stem from how a model is trained (data and objectives) as much as from its size. A larger model might have seen more diverse data (potentially learning more subtle biases or mitigating some biases), whereas a smaller model might generalize less but could also be less nuanced (possibly making it more blunt in its decisions). Additionally, different companies impose different alignment and safety strategies on their models – these could differentially affect fairness. For example, some providers might fine-tune their LLM to avoid certain sensitive outputs, while others might not, leading to variation in bias behavior. By including four different LLMs, our audit assesses whether findings like discrimination patterns are **model-agnostic or model-specific**, which is vital for both industry and regulators. As we will show, the results indeed varied significantly by model, justifying the importance of this diverse selection.

## Bias Mitigation Strategies Tested

A unique aspect of our analysis is that we not only measured bias under normal prompting, but also experimented with various **prompt-based mitigation strategies** to reduce bias. Since retraining or fine-tuning LLMs can be infeasible for end-users, we focused on techniques that a practitioner could apply at the prompt level (i.e. during deployment) to encourage fairer responses. We identified seven strategies from the literature and practitioner discussions, and applied each in our experiments across all models:

1. **No Mitigation (Baseline):** The control condition where we simply provide the complaint narrative and ask the model for a resolution, with no special instructions. This reflects the default zero-shot model behavior.
2. **Structured Extraction:** Reframing the task as extracting structured information from the complaint rather than providing a free-form resolution. For example, the prompt might instruct the model to list key facts or fill in specific fields (e.g. “Complaint Summary,” “Requested Solution”) instead of deciding an outcome. The idea is to limit model discretion and thus limit bias.
3. **Consequentialist Prompting:** Reminding the model of the consequences and fairness requirements of its decision. For instance, the prompt could include a directive like *“Ensure your response treats the customer fairly and consider the impact of a wrong decision”*. This strategy aims to invoke the model's knowledge of ethical or regulatory norms (possibly learned during training) to guide it toward unbiased behavior.
4. **Chain-of-Thought (CoT):** Instructing the model to “think step by step” or otherwise produce a reasoning process before giving a final answer. By encouraging a more deliberative response, we hypothesize the model might rely less on shallow stereotypes and more on case specifics, potentially reducing bias.
5. **Minimal Instructions:** Providing only minimal additional guidance beyond the complaint (e.g. *“Please assist the customer”*). This is slightly different from the baseline in that the baseline prompt might implicitly let the model decide how to respond, whereas here we explicitly tell it to respond helpfully but without mentioning anything about fairness or method. It serves to check if simply nudging the model to be helpful or concise has any effect on bias.

6. **Perspective-Taking:** Prompting the model to adopt the **customer's perspective** or show empathy. For example, instructing *"Imagine you are in the customer's position"* or *"Consider the customer's feelings and rights when responding."* The intention is to see if encouraging empathy and understanding of the complainant's viewpoint mitigates biased dismissal of their concerns.
7. **Role-Play:** Asking the model to respond while playing a specific role or persona, such as *"You are a veteran customer service manager handling this case"*. The hypothesis was that a role-play might impose a professional standard or consistency. However, there is also a risk: if the model's training leads it to emulate a human persona's biases (e.g. a stereotypical manager might treat certain customers less favorably), this could *amplify* bias.

Each of these strategies was realized through a modified prompt template applied to the same complaint scenarios. By comparing the model's outcomes under each strategy to the baseline, we can assess **which approaches most effectively reduce bias**. Notably, all strategies were applied in a **zero-shot manner** as well – we did not supply exemplars of unbiased behavior, we only altered the instructions. This approach tests practical interventions one might deploy without additional data.

## Measuring Outcomes and Fairness Metrics

To quantify discrimination, we defined several **metrics** capturing both outcome and process fairness in the model's responses:

- **Outcome Bias (Tier Difference):** We categorized the model's resolution or action into "tiers" of favorability (e.g. Tier 1 = no help or denial, Tier 2 = moderate assistance, Tier 3 = substantial remedy). This tiering was informed by typical outcomes in the CFPB data (such as whether the complaint was resolved with monetary relief, non-monetary relief, or no relief). For each trial, we noted the outcome tier and then measured bias as the difference between the outcome for a demographic persona and the outcome for the **baseline persona** on the same issue. A positive bias value means the demographic persona received a *worse* outcome (higher tier number indicating less favorable resolution) than the baseline. We aggregate these differences across scenarios to compute an average outcome bias per group and per model.
- **Process Bias (Question Rate Reduction):** As a proxy for procedural fairness, we tracked whether the model asked any **follow-up questions** or sought additional information from the customer. Engaging with follow-up questions indicates the model is taking the complaint seriously and not making assumptions. If a model fails to ask questions when it should (especially in complex cases), it may be exhibiting a form of *procedural discrimination* – essentially not affording the customer the same "voice" in the process (a concept linked to procedural justice). We computed the **questioning rate** (percentage of responses that contained at least one follow-up query) for each persona and compared it to the baseline's questioning rate. A drop in this rate for certain demographics implies the model is engaging less with those customers' issues, possibly due to stereotype-driven assumptions that it "already knows" the situation. We often express this as a percentage reduction. For example, an 80% reduction (as observed in our results) means the model practically stopped asking questions when demographic cues were present, relative to how often it would ask when no cues were given.
- **Severity Amplification Coefficient:** To measure how bias scales with case severity, we calculated the correlation (and slope) of the outcome bias metric against the severity level. A strong positive correlation indicates **amplification** – bias increases consistently from low to high severity. We also directly compare mean bias at low vs. high severity to quantify the fold-change. A fourfold increase from low to high severity, as we found, indicates a major amplification effect.

- **Statistical Significance Tests:** Given the large number of trials, we performed significance testing to ensure that observed biases were not only large in magnitude but also statistically robust. We used chi-square tests for categorical outcomes and ANOVA for model-strategy interactions. Key results (like the questioning rate drop and severity interaction) were significant at the  $p < 0.001$  level, giving us confidence in their reliability. Wherever relevant, we report  $p$ -values or refer to significance to underscore reliability.

All measurements above were computed for each model and each experimental condition, enabling detailed comparisons as reported in the next section.

## Results and Data Analysis

### Outcome Disparities Across Demographic Groups

Our first set of results addresses **RQ1: Do LLMs discriminate based on subtly embedded protected attributes?** The answer is unequivocally **yes** – we observed systematic disparities in model outcomes aligned with the demographic cues in the complaints. In particular, **Black female customers received the worst overall outcomes** across the board. On average, a complaint from the Black female persona was resolved about **0.321 tiers less favorably** than the identical complaint from the baseline (demographically neutral) persona. This gap was the largest among all groups tested, marking Black women as the most disadvantaged in model responses. In fact, out of 12 demographic persona profiles (covering combinations of race, gender, and class/socioeconomic cues), **only one group received better treatment than the baseline**, while all others experienced worse outcomes to varying degrees. The lone group with a slight advantage was a persona corresponding to a traditionally privileged profile (as we constructed, the one implicitly suggesting a White male professional); this persona occasionally received marginally better resolutions than the neutral baseline, perhaps reflecting a **pro-ingroup bias** for attributes the model might associate with financial savvy or lower risk. Nevertheless, even that “advantaged” group’s uptick was small. The dominant pattern is that **most demographic indicators led to negative outcome differentials**, which is a clear fairness failure.

These outcome differences are not only statistically significant but also practically meaningful. A difference of 0.321 tiers, for context, could mean the difference between a complaint being ignored versus receiving some restitution, or between a token apology versus a full refund, depending on the tier definitions. For a Black female customer to consistently land in a worse tier indicates a systemic bias reminiscent of the **“intersectional”** discrimination that legal scholars have warned about – where individuals with multiple marginalized identities (here, Black + female) face compounded disadvantages. Our findings align with the concept of **intersectionality** (Crenshaw, 1989) in that the model did not simply exhibit bias on race alone or gender alone, but disproportionately on the combination. Notably, the **fair lending laws** like ECOA prohibit discrimination based on either race or sex (among other factors), and a joint disparity of this magnitude would be alarming to regulators. In a human context, such patterns (e.g., consistently worse outcomes for Black women) would trigger investigations for disparate treatment. The fact that an AI system is producing similar patterns is a critical insight for AI ethics and compliance – it shows that **LLMs can reproduce or even create intersectional biases** without ever being explicitly told to consider race or gender.

It is important to examine whether these biases might be coming from the model **mirroring historical data** (e.g., perhaps Black women genuinely had worse outcomes in the past). While real-world data does show disparities in financial services outcomes (for instance, minority borrowers often get less favorable

loan terms and women have faced discrimination in credit historically), our experiment was designed to isolate model behavior from real outcome patterns. Since each model saw both the baseline and demographic versions of the *same* complaints, any disparity indicates the model *itself* introduced differential treatment. The **reliability** of this result is supported by the large sample and consistency: the disadvantage for Black women was observed across many trials and across multiple models. The difference of +0.321 tiers was an average aggregated over thousands of comparisons, making it highly unlikely to be a fluke. Statistically, a difference this size with our trial count yields  $p \ll 0.001$ . We can thus state with confidence that the disparity is **systemic** in the model outputs.

From an **implications** standpoint, this outcome bias means that if a financial institution were to deploy these LLMs to handle customer complaints or make preliminary decisions (say, on granting fee waivers or relief options), **Black female customers would be consistently shortchanged** relative to others. Such behavior violates the principle of equitable treatment and could also lead to **disparate impact liability**. The Equal Credit Opportunity Act and related regulations do not excuse “unintentional” bias by algorithms – the CFPB has explicitly affirmed that lenders are responsible for AI biases just as with human biases. Therefore, our finding is not merely of academic concern; it highlights a regulatory and operational risk. **Financial AI systems must be audited and adjusted to prevent this kind of outcome disparity.** The presence of even one group receiving systematically better outcomes (likely those resembling historically favored customers) and others worse also echoes long-standing patterns of bias in finance, raising questions about whether the model learned these patterns from its training data. This **underscores the necessity** of proactive bias detection in model deployment, as we have performed here.

## Procedural Fairness and Information-Gathering Bias

Beyond final outcomes, our RQ3 asked: **Do models exhibit differential information-gathering (process) based on customer demographics?** Our results show a striking **yes**. We measured the frequency with which each model response included **follow-up questions** – a sign that the model is probing deeper or seeking clarification, which we interpret as an element of **procedural fairness** (akin to giving the customer “voice” or individual consideration). When no demographic cues were present, the models would ask at least one follow-up question in roughly **28.5%** of the complaint cases on average. However, when even subtle demographic indicators were embedded in the complaint narrative, this questioning rate **plummeted to about 5.5%**. This is an **80.6% reduction** in the model’s inclination to gather more information, a highly significant drop ( $p < 0.001$ ). In practical terms, the presence of demographic context led the models to almost never ask questions – they tended to assume they understood the case fully and went straight to a resolution (often a less helpful one, as the outcome bias showed).

This behavior is a form of **stereotype-driven procedural bias**. It suggests that when the model “recognizes” certain profiles (perhaps an elderly customer, or a low-income single mother, etc.), it might be **making assumptions** about their case that obviate the need for further detail. For example, the model might assume an older complainant is simply confused or that a working-class complainant cannot provide more documentation, and thus it does not bother to ask for clarification or additional evidence that it might request from other customers. This **lack of inquiry** means those customers are effectively *denied an equal opportunity to have their full story considered*. In human terms, it’s like a customer service agent who, upon seeing an older Black female customer, decides “I’ve heard this kind of complaint before” and rushes to a decision without asking questions they would normally ask a younger white male customer. Such conduct would clearly violate principles of good service and fairness. In algorithmic terms, our LLMs are exhibiting a similar differential treatment at the process level.

The reliability of this finding is bolstered by consistency **across models** and scenarios: every model we tested showed some degree of question-asking drop when demographic cues were present, though the magnitude varied (as we detail later, one model almost never asks questions in any case, while another only reduced slightly – still, the overall average effect is large). The 80% figure is an aggregate; model-specific drops ranged from moderate to extreme, but all in the direction of fewer questions with demographic info. The statistical significance ( $p < 0.001$ ) indicates that this is not due to random chance – the difference in questioning behavior is systematically tied to the presence of protected attribute cues.

In terms of **implications**, this procedural bias can be just as pernicious as outcome bias. Even if an outcome were ultimately fair, the fact that some customers are not engaged or listened to during the process can erode trust and is considered a component of unfair treatment. In risk management and compliance terms, **process fairness** is important: regulators and courts consider not just the final decision, but whether everyone was given an equal chance and treated with equal dignity in the process (for instance, consistent application processes are a requirement under fair lending laws). Our results show that LLMs might violate this by **short-circuiting the dialogue** with certain demographics. Moreover, reduced questioning can lead to worse outcomes because the model might miss relevant details that it could have uncovered with a question. This connects back to outcome disparity – if a model doesn't ask a low-income customer "Did this fee cause you hardship?" it might never learn that the customer couldn't pay their mortgage because of it, missing a chance to offer relief that it might have offered to someone else who did convey hardship (perhaps because the model bothered to ask them). Essentially, **stereotype assumptions can create a feedback loop of disadvantage**: the model assumes it knows the story and thus doesn't ask, leading to a skimpy record on which it then justifies not providing help.

From an AI ethics perspective, this finding highlights a subtlety: **fairness is not only about what decision is made, but how it is made**. The LLMs studied failed a basic test of procedural parity. Addressing this might require prompt strategies or system designs that ensure the model consistently inquires or verifies details for all users, or that it doesn't unwittingly *deprioritize* certain users' cases. We will later discuss how some mitigation strategies affected questioning behavior.

## Severity Amplification of Bias

One of the most consequential discoveries in our analysis relates to **RQ2: Does discrimination amplify with complaint severity?** We found that it does – and dramatically so – a phenomenon we term **severity amplification**. In low-severity cases (minor issues, low dollar amounts), there was bias present, but the magnitude was relatively modest: on average about **0.097 tier difference** in outcomes between protected-group personas and the baseline, meaning a slight disadvantage. However, in high-severity cases (those involving very large amounts or life-altering stakes), the average bias ballooned to **0.404 tier difference**. This is more than a fourfold increase in bias magnitude from low to high stakes. We statistically confirm that severity level and bias are almost perfectly correlated (Pearson  $\rho \approx 1.0$  across the three defined levels,  $p < 0.001$ ), indicating a consistent upward trend: as the stakes rise, so does the gap in how the model treats different demographic profiles. In simpler terms, **the models become most unfair exactly when fairness matters most**.

To put these numbers in perspective, consider what a 0.404 tier bias means in a high-stakes context like a potential foreclosure: If the baseline persona (no cues) might get a Tier 2 outcome (e.g. some intervention that could help avoid foreclosure), the Black female persona might get Tier 3 (no meaningful help, essentially left to face the foreclosure alone). In low-severity fee disputes, the difference between tiers

might be negligible (perhaps both get little remedy, as it's a small issue). But in a life-changing scenario, one group getting systematically less aid means real harm – losing one's home or going bankrupt when a fair model *could* have helped. This amplification effect raises serious ethical and legal concerns. It suggests that **LLMs could silently exacerbate inequality** by failing exactly when vulnerable customers are in crisis, which is arguably when equitable treatment is most critical to prevent disparate harm.

We explored potential reasons for this severity-linked bias surge, drawing on patterns we observed and analogous findings in other domains. Three non-mutually-exclusive mechanisms appear plausible:

- **Uncertainty Escalation:** High-stakes cases often lack clear-cut solutions and may involve complex, uncertain outcomes. The AI might have more “room” to inject bias when the correct action is ambiguous. In low-stakes cases (like a \$30 fee error), the model's responses might be fairly standardized, leaving less opportunity for bias to influence the outcome. But in a complicated case (like a wrongful foreclosure claim), there are many decision paths the model could take – if the model harbors any biases (e.g. skepticism toward certain complainants), those could tilt its choice of action when it's unsure what to do. Essentially, **ambiguity allows bias to play a larger role**. This aligns with theories in social psychology and algorithmic decision-making that bias has a greater effect when rules are not clear or discretion is high.
- **Risk Aversion and Conservative Responses:** We noticed that models tended to be more conservative (less generous in providing relief or making exceptions) in high severity scenarios overall – perhaps reflecting a kind of learned risk aversion. For instance, if a complaint involves a large financial claim, the model might default to formal, cautious responses (potentially denying relief or saying it cannot help) to avoid making a “mistake.” If the model associates certain demographic cues with higher risk or fraud likelihood (due to biases in training data), it might become *especially* conservative for those groups under high stakes. This is analogous to a biased human loan officer who might be extra hesitant to approve a big loan for an applicant they stereotype as high-risk. Thus, severity could be magnifying a **prejudiced risk response**, where the model's threshold for providing help rises more for some groups than others under pressure.
- **Stereotype Reliance under Complexity:** High-stakes complaints are often more complex and multifaceted. When faced with complexity, the model might rely more on heuristic shortcuts – which can include stereotypes. A simpler case might be handled by more straightforward pattern matching (where perhaps the model has seen enough similar generic cases). But a complex case might not match any clear pattern, leading the model to subconsciously fill gaps with **learned biases** (e.g., “this sounds like the kind of situation people from XYZ group might misuse, so I will be strict”). This aligns with prior findings in criminal justice algorithms and healthcare: bias worsens for serious cases where decisions are harder. For example, Dressel & Farid (2018) found that algorithmic bias in predicting reoffense was higher for more severe crimes, and Obermeyer et al. (2019) documented that a health risk algorithm exhibited larger racial biases for patients with greater medical complexity. Our results indicate a similar dynamic in LLM-driven decisions for finance.

The **reliability** of the severity amplification result is strongly supported by the data. The perfect correlation ( $p=1.0$ ) suggests a linear increase in bias from low to medium to high stakes. This was consistent across repeated experiments; the pattern did not appear due to outliers. Each model individually also showed some form of this trend (with one exception, discussed below). The ANOVA for interaction between severity and demographic outcome was significant, confirming that severity level changes the bias outcome relationship systematically ( $p<0.001$ ). Therefore, we are confident this effect is real and not an artifact.

In terms of **implications**, this finding cannot be overstated: if not addressed, **AI systems could be the most discriminatory when handling the most important decisions**. In the financial sector, that is a recipe for reinforcing structural inequalities. For instance, if relief programs or debt forgiveness decisions are guided by an LLM, those in dire need (who often are disproportionately from disadvantaged groups) might disproportionately be denied help due to amplified bias – exacerbating socioeconomic and racial disparities. From a regulatory lens, disparate impact is especially concerning in high-stakes contexts (regulators focus on outcomes like loan approvals, etc., more than small customer service gestures). Our evidence that LLM bias is **positively correlated with stakes** means regulators will not only need to check average bias, but specifically scrutinize models under critical conditions. It also means developers should prioritize **fairness stress-testing for worst-case scenarios**, not just average-case. Encouragingly, one model in our study (Gemini-2.5) showed virtually **no severity amplification** – its bias was relatively flat across low to high severity. This suggests that model-specific factors (like architecture or training) can mitigate or eliminate this effect. Understanding what makes one model immune to amplification (perhaps it treats every case with a similar approach) could inform best practices in model design or tuning. We delve more into model differences next.

## Differences Across Model Architectures and Sizes

Our cross-model evaluation revealed that **bias is highly model-dependent**, challenging any assumption that larger or newer models are automatically fairer. We found notable differences in both outcome and process biases among the four LLMs, attributable to their underlying architecture and training:

- **GPT-4o (OpenAI, large model):** This was the **least biased model overall** in terms of outcome disparity. Its average outcome bias was about **0.058 tiers** (significantly lower than the others). GPT-4o also had a negligible **process bias** – its questioning rate barely changed with demographic cues (only about a **0.05%** drop, essentially zero). In other words, GPT-4o almost always maintained its inclination to ask follow-ups regardless of the persona, and it kept outcome differences relatively small. It did exhibit some bias (e.g., still treating Black women worse than baseline, but by a smaller margin than other models did). GPT-4o's bias also increased linearly with severity (a modest slope of  $\sim 0.05$  per level), showing amplification but not as dramatically as some others.
- **GPT-4o-mini (OpenAI, smaller model):** Interestingly, the **smaller sibling model was more biased**. GPT-4o-mini showed the highest average outcome bias, around **0.137 tiers** (approximately 2.4× the bias of GPT-4o). This suggests that **scaling down the model increased biased behavior**, possibly because the smaller model has less capacity to handle nuance or was trained on data in a way that left more bias uncorrected. On process fairness, GPT-4o-mini still had very low questioning bias ( $\sim 0.10\%$  drop, similar to GPT-4o, essentially asking questions or not equally). Where GPT-4o-mini stood out negatively was in **severity amplification** – it had a markedly **non-linear amplification**, with bias skyrocketing in high severity cases (in fact, for Black women in high-stakes scenarios, GPT-4o-mini's bias was 4.23× higher than in low-stakes, the most extreme case of all models). This hints that the smaller model may lack robust decision heuristics for complex scenarios and falls back on very biased responses under stress.
- **Claude-3.5-Sonnet (Anthropic, medium model):** Claude's performance was **moderate on all fronts** – it exhibited a **mean outcome bias of 0.108 tiers** (more than GPT-4o but less than GPT-4o-mini), and a **questioning rate drop of about 3.0%**. That 3% process bias means Claude did sometimes reduce questions for certain groups, but far less drastically than Gemini (below) and not as near-zero as GPT. In effect, Claude treated demographics somewhat unequally, but not in the extreme; it lies in the middle. Its severity amplification pattern was like a “step function” – relatively stable bias from

low to medium, then a jump at high severity. So, while not as explosive as GPT-4o-mini, it did show that bias became a problem mainly at the highest stakes.

- **Gemini-2.5-Pro (Google, medium model):** Gemini had a **unique profile**: it showed **moderate outcome bias (~0.091 tiers)**, comparable to Claude's range, but **extreme process bias** – a whopping **19.4% reduction** in questioning rate with demographic cues. In fact, Gemini's baseline questioning tendency was high (it asked questions ~73% of the time for neutral complaints), but with demographic cues this fell to near zero, indicating a massive process discrepancy. Gemini essentially "shut down" its information-seeking for certain groups, far more than any other model. On the other hand, Gemini displayed *no severity amplification* in outcome bias; its bias level was relatively flat from low to high severity. This could imply its decision policy was consistent regardless of stakes (though unfortunately consistently somewhat biased and with poor process for some groups). Gemini's pattern underscores that **model architecture and training** can lead to different kinds of bias – here, perhaps a more retrieval-based or rule-based approach that Gemini might use kept outcome bias in check across scenarios, but something in its handling of conversational context caused it to drop questioning for certain profiles.

To summarize these differences: **model size alone did not guarantee fairness** (the largest model GPT-4o was best overall, but the second-largest, Claude, wasn't the second-best necessarily; the smaller GPT-4o-mini was worst in outcomes). **Model provider/architecture seemed more important**, especially for process fairness – both OpenAI models were good at maintaining question-asking, whereas the Google model was very poor in that regard, and Anthropic's was intermediate. This suggests that OpenAI's alignment training (perhaps instructing their models to always be helpful and ask clarifying questions) was effective at ensuring consistency, while Google's model perhaps optimized differently, causing a huge drop in a certain context. These differences carry implications: organizations cannot assume that a given "big name" model is fair **by default** or that a bigger version of a model will fix bias issues. **Fairness audits must be done model by model**. Also, the notion that the newest or most complex model is safest is not necessarily true – targeted tests can reveal unexpected weaknesses, like Gemini's procedural lapse or GPT-4o-mini's spike in high-stakes bias.

From a reliability standpoint, all the above comparisons were borne out by statistically significant differences. For example, an ANOVA of Model × Persona on outcomes confirmed a significant interaction ( $p < 0.001$ ), meaning the effect of persona (demographic) on outcome depends on which model you use – i.e., some models are significantly more biased than others. Likewise, the disparity in questioning rates (0.06% vs 72.9% in the most extreme case) between models indicates orders-of-magnitude variation in behavior. These are not subtle differences, giving us high confidence that model choice is a critical factor in fairness.

The **implications** here are important for risk management: If a bank is choosing an LLM to integrate into their customer service or decision pipeline, **the choice of model can make a substantial difference in compliance outcomes**. For instance, GPT-4o might pose less regulatory risk on average than GPT-4o-mini, because it yields fewer biased outcomes. Meanwhile, using Gemini-2.5 without compensating for its process bias could expose an institution to complaints of inconsistent service procedure across customers. This points to a need for **model-specific fairness evaluation and perhaps certification** – similar to how banks validate different credit scoring models independently. Our findings advocate for not treating LLMs as interchangeable general-purpose tools; rather, their fairness properties must be **empirically tested and compared**.

## Effectiveness of Bias Mitigation Techniques

A key contribution of our study is testing various **bias mitigation strategies** and discovering that their effectiveness varies dramatically by model. Figure 1 (Table 5 in our data) summarizes the **percentage reduction in outcome bias** achieved by each strategy for each model, and the results are illuminating:

- **Structured Extraction:** This strategy – having the model output only structured information rather than a discretionary answer – was the **most effective overall**, but with a sharp model disparity. For the GPT-4o and GPT-4o-mini models, structured extraction completely **eliminated measurable outcome bias (100% reduction)**. Essentially, when these models were constrained to just extract facts, the disadvantage between demographic and neutral personas disappeared. Claude-3.5 also saw a strong improvement (67% bias reduction), though not complete elimination. In stark contrast, Gemini’s bias was hardly affected (only ~18% reduction). We interpret this as follows: for some models (GPT family), removing the requirement to decide an action made them treat each case identically – likely because they were just regurgitating facts and not invoking any biased reasoning. Gemini, however, might handle even extraction tasks in a way that still allows bias (perhaps by focusing on different facts for different personas, or by the fact that it asked far fewer questions for some, resulting in less info to extract). This indicates that **no single strategy is universally effective** – one must consider the model’s internal workings. Nonetheless, structured outputs seem promising, particularly for models that can be guided into form-filling modes; by minimizing model “imagination,” we minimize bias injection.
- **Consequentialist Prompting:** Telling the model to be fair and consider consequences yielded high reductions in bias across the board: ~89% for GPT-4o, ~92% for GPT-4o-mini, ~84% for Claude, and ~71% for Gemini. This was the second-best strategy on average. It appears that reminding the model of fairness and impact tapped into some internal knowledge or policy the models have (likely learned during alignment training about not being unfair or harmful). Especially for the OpenAI models, this nearly wiped out bias. Even for Gemini, 71% reduction is substantial, though not complete. This result is heartening because it suggests that **explicit ethical instructions** can significantly curb biased tendencies in LLMs – these models are malleable through prompting to an extent. However, since not 100%, and since it varied (OpenAI’s responded more than Gemini’s), it’s not a panacea alone.
- **Chain-of-Thought (CoT):** Asking the model to reason stepwise achieved moderate but solid bias reductions: 78% (GPT-4o), 68% (GPT-4o-mini), 73% (Claude), 69% (Gemini). So roughly 70% bias reduction for most. This suggests that CoT helps the model not jump to biased conclusions, perhaps by forcing it to articulate the problem details. It’s interesting that GPT-4o-mini saw a bit less benefit (68%) than GPT-4o (78%), indicating maybe that the smaller model’s reasoning was not as effective in overriding bias. In general, CoT is a **useful but not sufficient** mitigation – it leaves 20–30% of the bias in place in many cases. It may be best used in combination with other methods (indeed, future research might combine CoT with consequentialist or structured prompts).
- **Minimal Instructions:** This lightweight approach gave a modest improvement: around 55–61% bias reduction for the GPT models, ~58% for Claude, ~52% for Gemini. So roughly half the bias was mitigated. This likely indicates that simply instructing the model to “help the customer” without calling out any specific fairness guidance does have some effect – possibly because the model focuses on being helpful and thus might ignore some irrelevant (biased) heuristics. But it’s clearly not as effective as the more direct interventions above. Minimal instructions might serve as a baseline best practice (always tell the model its role and goal), but **additional steps are needed** to fully address bias.

- **Perspective-Taking:** This strategy consistently showed smaller improvements: ~44% bias reduction (GPT-4o), ~38% (GPT-4o-mini), ~47% (Claude), ~41% (Gemini). It appears that encouraging empathy or the customer’s viewpoint only modestly reduces bias. One thought is that while it may humanize the scenario, it doesn’t directly counter any stereotype the model might hold; a biased model could still be “empathetic” yet unfair in outcome (e.g., *“I understand you’re struggling, but I still can’t help”*). Thus, perspective-taking alone is not a powerful anti-bias tool for these models. It might need to be coupled with other instructions.
- **Role-Play:** Perhaps the most surprising (and cautionary) result – instructing the model to role-play as a specific persona **increased bias** in all cases. We saw *negative* reductions: bias actually worsened by about 12% for GPT-4o, 15% for GPT-4o-mini, 18% for Claude, 14% for Gemini (i.e., the bias metric was that much higher compared to baseline when roleplay was used). This indicates that role-playing as a “customer service agent” or similar might cue the model to mimic the behavior of humans (or the style of data it saw) which apparently includes biased patterns. In hindsight, if the model’s training data for such roles includes historical scenarios where, say, agents were dismissive or biased, the model might adopt those attitudes. This finding is crucial for practitioners: a naive attempt to prompt the model as a professional agent can backfire, actually **amplifying stereotype-driven behavior**. It underscores that we must carefully evaluate mitigation prompts – not all intuitively good ideas work as expected. In our context, we quickly abandoned role-play as a mitigation once we saw it consistently degrading fairness.

From these results, one major takeaway is that **no single mitigation works best for every model or every metric**. There is a clear **model-strategy interaction**: for instance, structured extraction is phenomenal for GPT-4 but not for Gemini; consequentialist prompting helps all but especially the GPT family; roleplay hurts all, etc. This interaction was statistically significant (interaction term  $p < 0.001$ ), meaning the effectiveness rank of strategies depends on the model in question. Therefore, mitigating bias in LLMs is not a one-size-fits-all situation. Each model may require a customized approach or a combination of strategies.

On the positive side, our findings also highlight that **bias in LLMs is not immutable** – with the right prompting, it can be greatly reduced. For the OpenAI models, we achieved near-zero bias with a straightforward structured output instruction. This is encouraging for practitioners and researchers focused on AI fairness: it means we can often substantially improve outcomes *without* retraining the model, just by smarter prompt engineering. However, the fact that Gemini did not respond as well suggests that in some cases, model retraining or more complex interventions (like fine-tuning on balanced data or using a fairness algorithm) might be needed for certain systems.

It’s also worth noting how these strategies impacted the **process fairness** (questioning behavior). While our main bias metric was outcome tiers, we observed side effects: for example, structured extraction by design changes the output format (likely no questions asked at all, just facts), so it essentially neutralized process differences as well for GPT models (everyone gets the same info fields). Consequentialist and CoT prompts sometimes led the models to actually ask **more** questions before concluding – which is a good thing for fairness, as it equalized or even improved engagement for all personas. Roleplay, conversely, sometimes led to curt, scripted responses with no questions, worsening the disparity if the baseline would have asked. A detailed breakdown (Table 6 in the appendix) showed, for instance, that Gemini’s baseline question rate was high but dropped massively with roleplay (it defaulted to a formal monologue), whereas GPT’s baseline was low and roleplay didn’t change it much (GPT rarely asked questions anyway in that format). The key insight is that **some mitigation strategies influence not just what decision is made but how the conversation flows**, which can either help or hurt fairness. Designers of prompts should be mindful of these second-order effects.

In summary, our mitigation experiments reveal **which bias reduction techniques are most promising** for different LLMs. Structured extraction and explicit fairness reminders stand out as generally effective (especially for the GPT series), whereas naive role emulation should be avoided. This contributes practical knowledge to AI ethics: when deploying LLMs in high-stakes domains, one can use our findings as a guideline to choose mitigation strategies. Moreover, the need for model-specific tuning suggests organizations should invest in prompt engineering and testing tailored to the particular AI system they use, rather than assuming general recipes will transfer.

## Limitations of the Study and Data Considerations

While our findings are robust within the scope of the experiments, there are several **limitations** to acknowledge, particularly regarding the use of the CFPB complaint data and the generalizability of our results:

- **Dependence on CFPB Data and Synthetic Demographics:** We based our complaint narratives and scenarios on patterns from the CFPB Consumer Complaint Database, which ensures realism but also imposes constraints. The CFPB data itself does not include **explicit demographic labels** (race, gender, etc.), which is why we resorted to embedding cues. This means our mapping from cues to demographic groups is heuristic. Real complainants might not mention certain details, or might mention them in different ways. Our personas (WFS, BFU, etc.) are representative but not exhaustive. Thus, one limitation is that we tested a fixed set of **eight persona profiles**; there could be other intersectional identities or more nuanced subgroups that we did not simulate. Additionally, by necessity, we simplified “race” into broad categories (Black, White, Hispanic, etc.) and gender as binary for the experiment. Actual demographic diversity is more complex (e.g., we did not separately examine Asian American, Native American, or non-binary individuals explicitly, nor intersection of multiple factors beyond the ones chosen). Future work could extend this to more groups, but our design was already large and we prioritized groups historically noted in U.S. fair lending contexts.
- **CFPB Data Representativeness:** The complaint database covers consumers who chose to submit formal complaints. This population may not be fully representative of all customers – there might be selection biases (for example, some demographics might be less likely to file formal complaints due to distrust or lack of awareness). Our results might therefore reflect biases that occur in those complaint scenarios, but real-world deployment of LLMs might encounter different distributions of inquiries. We did not calibrate to actual complaint frequencies, as discussed, so if, say, elderly customers are overrepresented in real complaints, the overall bias impact in production might skew differently. Moreover, the CFPB data is U.S.-centric; our findings might not generalize to other countries or cultural contexts without further study.
- **Outcome Tier Metric Simplification:** We used a tiered outcome scale to quantify how favorable a resolution was. While this allowed us to measure bias in a structured way, it is a **simplification of reality**. Real complaint outcomes can be nuanced (quality of response, tone, etc., not just what remedy was provided). We might have missed some aspects of fairness by focusing on tiers. It’s possible a response could be polite (procedurally nice) yet offer no help (outcome unfair), or vice versa. We tried to capture both outcome and process separately, but there are more subtleties (like the content of follow-up questions, or the justification given for denial). Our tiering approach might also have some subjectivity – although we applied it consistently across groups, there is a chance of slight misclassification. The large sample mitigates random errors here, but a more refined content

analysis of responses could yield additional insight beyond tiers (e.g., differences in language used toward different demographics).

- **Single Interaction (One-Turn) Format:** Our experiments were essentially one-turn interactions: the model gets the complaint and possibly instructions, then produces a response (which could include questions). In a realistic setting, there might be a multi-turn dialogue (the customer answers follow-up questions, the model responds further, and so on). We did not simulate the customer's second-turn answers or how the model would proceed after its initial questions. Therefore, we mostly measured the initial decision and willingness to gather info. It's possible that even if a model asks fewer questions initially, a determined customer might provide more info themselves, and the model could then adjust. Or the model could exhibit bias in a later stage instead. Studying multi-turn dynamics was beyond our current scope but is a limitation: **the conversational context could alter fairness outcomes**. Our one-turn approach was like a worst-case: if the model doesn't ask now, it never will. In reality, some processes allow second chances. Nonetheless, in many automated workflows, one turn is actually the norm (e.g. an AI responds and the case is closed unless escalated), so our design is still quite relevant.
- **Limited Set of Mitigation Strategies:** We tested seven prompt-based strategies, which is more comprehensive than prior work, but it's not exhaustive of all possible interventions. We did not, for example, test strategies involving external tools (like calling a separate fairness-checker model) or advanced techniques like reinforcement learning from human feedback specifically targeting bias reduction. Our focus was on easily implementable prompt tweaks. It's possible that other creative prompts or combination of prompts could further reduce bias beyond what we achieved. For instance, we did not combine strategies in the experiments (we treated each separately), but maybe a *Consequentialist* + *CoT* + *Structured* prompt all in one could nearly eliminate bias even for Gemini. We highlight this as a limitation but also a **future research opportunity**: to explore combinations and new mitigation ideas. Our results give a baseline, but not the final word, on what's possible in bias mitigation.
- **Model Versions and Evolution:** The LLMs we evaluated are constantly evolving (especially the API-based models that providers update regularly). We effectively conducted a **snapshot in time** (circa mid-2025) of these models' behavior. It is possible that future updates to GPT-4, Claude, or Gemini could improve (or inadvertently worsen) their fairness characteristics. Similarly, new versions (e.g., GPT-5 or Claude-4) might have different biases. Our results may not generalize to future versions without re-testing. This is a limitation inherent in working with closed-source commercial models. However, the broader phenomena we observed (like intersectional bias and severity amplification) are likely to persist in some form unless specifically addressed, so the qualitative insights should remain relevant. It underlines that **continuous auditing** is necessary: one cannot assume a model that was fair last year remains fair after an update.
- **Scope of Application (Complaint Resolution vs. Other Tasks):** We specifically looked at a customer complaint resolution task. This involves the model generating a response to a described problem. The fairness issues here revolve around **customer service decisions**. If one applied LLMs to different tasks in finance (like loan underwriting decisions, fraud detection, or financial advice), the bias manifestations might differ. For example, in an underwriting scenario, the model might output a risk assessment that could be biased in different ways than our complaint remedy tiers. We believe many underlying biases (like stereotyping from context) would carry over, but one should be

cautious in extrapolating our numeric results (e.g., the exact magnitude of bias) to other tasks. The general methodology we used, however, could be adapted to audit those tasks. We focused on complaints because it's a domain where LLMs are already being piloted (to draft responses to customer issues) and because it directly ties to consumer protection regulations (CFPB's domain).

In summary, while our study benefitted from a large experiment and yielded clear evidence of bias, these limitations suggest that **future work and caution are needed**. Using the CFPB data gave realism but required synthetic demographics and may not capture all nuances of real interactions. Our measurements provide strong signals of unfair behavior, but they don't capture everything about the model responses (e.g., tone or specific justifications). And importantly, the rapidly changing nature of LLMs means organizations should treat our results as a call to action to do their own testing, rather than as a static evaluation. None of these limitations undermine the core findings – if anything, they highlight how much more *could* be uncovered with broader analyses. For instance, if we had real demographic labels on complaints, who knows what biases we might confirm? That data is rarely available due to privacy, hence our approach. We hope our work encourages regulators to consider requiring more collection of such data in a safe way, to allow continued auditing.

## Recommendations for Visualization

Throughout this paper, several complex results would benefit from **visual representation** to enhance understanding and clarity for readers. We identify a few key areas where incorporating charts or other graphics would be particularly helpful:

- **Demographic Outcome Disparities (Figure 1):** A bar chart or grouped bar chart could illustrate the **average outcome tier for each demographic group** compared to the baseline. For example, a chart with demographic groups on the x-axis and mean outcome tier (or bias relative to baseline) on the y-axis would immediately highlight that the bar for “Black Female” is the tallest (worst outcome) and that perhaps the bar for “White Male” is slightly below the baseline line (better outcome). Such a visualization would make the pattern of one group being worst and one best very clear, reinforcing the text's discussion of intersectional bias. It can also show confidence intervals or error bars from the many trials, indicating statistical significance visually (non-overlapping error bars, for instance).
- **Questioning Rate Drop (Figure 2):** To depict the **procedural bias**, a simple before-and-after bar chart could be used. One bar shows the baseline questioning rate (~28.5%), and next to it a bar shows the questioning rate with demographic cues (~5.5%). The dramatic drop (perhaps annotated as “-80.6%”) would be evident. Alternatively, a small multiples chart could show each model's questioning rate baseline vs with cues, to highlight that some models (like Gemini) go from a high bar to nearly zero, whereas others (GPT-4) stay low in both. This visual would concretely demonstrate the scale of the reduction and differences between models.
- **Severity vs. Bias Trend (Figure 3):** A line graph can capture the **severity amplification** effect. Severity (Low, Medium, High) on the x-axis and average bias (tier difference) on the y-axis, with a line for each model (or perhaps one line for the overall average and separate lines for each model if distinguishing them). The overall trend line would slope upward steeply. If plotting per model, we'd expect to see GPT-4o-mini's line curving upward sharply by High severity, GPT-4o's line rising modestly, Claude's line flat then jumping at High, and Gemini's line roughly flat. Using different line styles or colors for each model and a legend would make it clear whose bias grows most. This figure

would help readers visually confirm statements like “bias quadruples from low to high stakes” and see model differences in amplification.

- **Model Bias Profiles (Figure 4):** A comparative visualization such as a cluster bar chart could summarize **bias metrics by model**. For instance, for each model, have two side-by-side bars: one for outcome bias (maybe in a distinct color) and one for process bias (e.g. the reduction in questioning, in percentage). This way, one can immediately see that GPT-4o has a very short outcome bias bar and near-zero process bias bar, whereas Gemini has a moderate outcome bar but a huge process bias bar. This complements the detailed numbers in text, giving an “at a glance” profile of each model’s fairness issues. It also reinforces the point that architecture matters (the patterns look different per model).
- **Mitigation Strategy Efficacy (Figure 5):** A final recommended visualization is a grouped bar chart or line chart showing **% bias reduction for each strategy by model**. For example, the x-axis lists the strategies (Structured, Consequentialist, CoT, Minimal, Perspective, Roleplay), and for each strategy we have four bars (one per model) indicating how much bias was reduced. The chart would show tall bars for GPT’s under Structured (100%) and much shorter bar for Gemini under Structured (~18%). It would also highlight that Roleplay bars are negative (which could be shown as bars going below zero or a different color to indicate increased bias). Such a figure makes the interaction effect apparent – the varying heights by model within each strategy category – which is easier to grasp visually than scanning numbers. Additionally, annotating the best strategy for each model (e.g. a star or highlight on the top bar in each cluster) can drive home that no single bar is tallest across all clusters (no universal best strategy).

Incorporating these figures at relevant points in the paper would greatly enhance reader comprehension. For instance, placing Figure 1 alongside the text discussing Black women’s outcomes would provide immediate visual evidence of that claim. Figures 2 and 3 could appear in the results section to support the procedural bias and severity amplification discussions respectively. Figure 5 could accompany the mitigation strategies results section. Visualizing the data not only helps in **communicating the magnitude of biases** but also appeals to interdisciplinary readers (e.g., risk managers or policymakers) who may find charts more intuitive than text or tables. It is also useful for presentations and executive summaries. We ensured that any suggested figure is based on data from our analysis, and the references for our data are already provided in tables (no external figure sources needed beyond our results). By adding these visual aids, the paper will convey its key findings more clearly and leave a stronger impact on the audience through both quantitative and visual evidence.

## Conclusion

In this study, we conducted an in-depth fairness audit of cutting-edge financial LLMs using a novel **zero-shot prompting methodology** that embeds demographic cues subtly within realistic complaint narratives. Our analysis yielded several **important and novel findings** that have implications across data science, risk management, and AI ethics:

- **Subtle Cues Elicit Significant Bias:** We demonstrated that even without explicit mentions of protected characteristics, LLMs can detect and use **contextual demographic signals** in ways that lead to **illegal discrimination**. This is a critical extension of prior AI bias research – whereas many earlier studies used obvious labels or attributes, we showed that models still falter when bias is

camouflaged in natural context. This highlights a *new level of sophistication* needed in bias testing and debiasing methods.

- **Worst Bias at the Intersection and When Stakes are Highest:** Black women, representing an intersectional minority group, received the worst outcomes in our experiments, underscoring how **intersectionality** manifests in AI decisions. Furthermore, our finding of **severity amplification** (with bias magnifying in high-stakes scenarios) is a novel discovery in the realm of LLMs. While analogous trends have been observed in other algorithms (criminal justice, healthcare), this is the first evidence (to our knowledge) of such a pattern in a finance-focused language model context. It means AI can *compound* disadvantage when it matters most – a finding that was not documented in the LLM literature before and should prompt immediate further research and oversight.
- **Model Architecture Matters More than Model Size:** A surprising outcome of our multi-model comparison is that **who made the model and how it was trained is more determinative of fairness behavior than the sheer model size**. The larger GPT-4 model was fairer than its smaller version (suggesting some benefit of scale or better training data), yet a medium-sized model from another provider (Gemini) had completely different bias characteristics (very biased process, no severity effect). This highlights a novelty that **LLM fairness is not monolithic** – each model can have unique failure modes. For practitioners, this reinforces the need for **model-specific audits**; for researchers, it opens questions about what architectural or training differences cause these divergent outcomes (e.g., different pretraining corpora, different alignment techniques such as RLHF vs. constitutions). Our work is among the first to systematically compare multiple contemporary API models on fairness metrics, providing a new perspective that bigger isn't always better for bias, and alignment approaches can have unexpected gaps.
- **Bias Mitigation is Feasible but Not One-Size-Fits-All:** On a positive note, we found that **prompt-based interventions** can dramatically reduce model bias – in some cases, virtually eliminating measurable disparities. This is an encouraging contribution, as it offers immediately actionable strategies for AI developers and users. However, our results also clearly show that **the effectiveness of mitigation techniques varies by model**. The fact that one strategy can work brilliantly for one model and barely affect another is a novel insight for the field of AI fairness: it suggests that understanding a model's internal decision-making or “reasoning” style is crucial to choosing the right mitigation. In practice, it means organizations should not rely on generic bias-reduction recipes; they need to empirically test which prompts or procedures help their specific model. This study contributes a starting roadmap of what might work (e.g. structured outputs, ethical reminders) and what likely will not (role-playing as a solution). The **statistical significance** of our mitigation results (including a model–strategy interaction  $p < 0.001$ ) lends strong credence to these findings.
- **Implications for Regulation and Oversight:** Our findings carry direct implications for compliance with fair lending laws and model risk management guidelines. We documented patterns (e.g., systematic worse outcomes for a protected group, differential treatment in process) that regulators explicitly deem unacceptable. This research thus serves as a timely alert that current **foundation models** deployed in financial contexts can run afoul of anti-discrimination laws even without any intent by the deployers. It underscores the **urgent need for auditing AI decisions** in any consumer financial application, as also advocated by the CFPB and Federal Reserve in recent guidance. From a risk management perspective, firms should incorporate **fairness tests at multiple levels** (outcome, process, across scenarios) and especially scrutinize high-stakes decisions (our severity-stratified

audit recommendation). The novelty here is pointing out that conventional testing might overlook biases that only appear in certain contexts – a call for more **sophisticated, context-aware testing regimes**.

In conclusion, **sophisticated LLMs are not immune to old-fashioned biases** – they simply express them in new, often subtle ways. This paper contributes a deeper understanding of how and when those biases surface in the financial domain. Importantly, our work demonstrates that addressing fairness in AI is a complex, model-dependent challenge, but one that can be met with targeted strategies. The analysis and methodology we present are **novel in their scale and scope** (over 44k tests, multiple models, multiple mitigations), offering a template for future audits. We have highlighted multiple areas for further investigation, such as combining mitigation methods, probing the root causes of model-specific biases, and extending to multi-turn interactions and other decision types.

For the fields of data science and AI ethics, our findings emphasize the value of interdisciplinary approaches: understanding legal standards, social theories like intersectionality, and technical model behavior together. For actuarial science and risk management professionals, the message is that **AI models carry quantifiable bias risks** that need to be measured and managed just like financial risks. Ultimately, ensuring fairness in AI-driven decisions is not just a moral and legal mandate, but also key to maintaining public trust in increasingly autonomous financial systems. Our work provides evidence-based steps toward that goal – revealing where the problems lie and pointing toward solutions – while also cautioning that continuous vigilance is required as these models evolve.

#### References (verified for accuracy and relevance):

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica.
- Arrow, K. J. (1973). *The theory of discrimination*. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3–33). Princeton University Press.
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). *Consumer-lending discrimination in the FinTech era*. *Journal of Financial Economics*, 143(1), 30–56.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency** (pp. 610–623).
- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. In **Advances in Neural Information Processing Systems** (Vol. 29, pp. 4349–4357).
- Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models*. arXiv Preprint arXiv: 2108.07258.

- Bowen, D. E., III, Stein, L. C. D., Price, S. M., & Yang, K. (2024). *Measuring and mitigating racial disparities in LLMs: Evidence from a mortgage underwriting experiment*. SSRN Working Paper No. 4812158.
- Brown, T., et al. (2020). *Language models are few-shot learners*. In **Advances in Neural Information Processing Systems** (Vol. 33, pp. 1877–1901).
- Consumer Financial Protection Bureau. (2022, March 16). *CFPB targets unfair discrimination in consumer finance* (Newsroom press release).
- Consumer Financial Protection Bureau. (2023). *Consumer complaint database* [Data set].
- Crenshaw, K. (1989). *Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics*. University of Chicago Legal Forum, 1989(1), Article 8.
- Dressel, J., & Farid, H. (2018). *The accuracy, fairness, and limits of predicting recidivism*. Science Advances, 4(1), eaao5580.
- Federal Reserve Board. (2011). *Supervisory guidance on model risk management* (SR Letter 11-7).
- Federal Reserve Board. (2023). *Report on the economic well-being of U.S. households in 2022*.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). *Word embeddings quantify 100 years of gender and ethnic stereotypes*. **Proceedings of the National Academy of Sciences**, 115(16), E3635–E3644.
- Gururangan, S., Swayamdipta, S., et al. (2022). *Annotation artifacts in natural language inference data*. In **Proceedings of NAACL 2022** (pp. 107–112).
- Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. In **Advances in Neural Information Processing Systems** (Vol. 29, pp. 3315–3323).
- Hu, T., Kyrychenko, Y., Rathje, S., et al. (2025). *Generative language models exhibit social identity biases*. Nature Computational Science, 5(1), 65–75.
- Jabarian, B., & Henkel, L. (2025). *Voice AI in firms: A natural field experiment on automated job interviews*. SSRN Working Paper No. 5395709.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). *Inherent trade-offs in the fair determination of risk scores*. In **Proceedings of the 8th Innovations in Theoretical Computer Science Conference** (Article 43).
- Levy, S., Allison, E., & Jia, R. (2024). *Implicit bias in large language models: Evidence from ChatGPT*. arXiv Preprint arXiv:2401.07698.

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. *Science*, 366(6464), 447–453.
  - Sorin, V., Korfiatis, P., et al. (2025). *Socio-demographic modifiers shape large language models' ethical decisions*. *Journal of Healthcare Informatics Research* (advance online publication).
  - Starr, S. B. (2014). *Estimating gender disparities in federal criminal cases*. *American Law and Economics Review*, 17(1), 127–159.
  - Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press.
  - Urban Institute. (2024). *Racial Equity Analytics Lab: Financial Services Disparities Report*. (Online report).
  - U.S. Census Bureau. (2023). *American Community Survey 5-year estimates (2019–2023)* [Data set].
-