
Probing LLM Preferences: Demographic Framing, Elicitation Context, and Incentive-Driven Trade-offs

Hong-Yu Hsien
SeqHub AI Academy

Ben Kiev
SeqHub AI Academy

Alexander Klenoff
SeqHub AI Academy

Omolola Olorunnishola
SeqHub AI Academy

J Phillips
SeqHub AI Academy

Taiwo Togun
SeqHub AI Academy

With
Apart Research

Abstract

Large language models are increasingly used to evaluate people, make recommendations, and make choices between competing options. This raises a basic question: how stable are these judgments, and how much do they change when the framing or incentives change?

We investigate this through three related experiments using GPT-5.6 Luna. First, we evaluate a standardized job applicant while varying demographic attributes and the perspective from which the applicant is judged. Second, we test whether demographic information alone leads the model to associate particular groups with stereotyped roles or traits. Finally, we introduce donation incentives as a common currency and test whether an existing choice changes when an external incentive is attached to one option.

We find statistically significant differences in applicant evaluations across race, gender, age, education, and evaluator perspective, but the magnitude of these effects varies substantially: evaluator perspective and age account for considerably more variation than race, gender, or education. In forced-choice demographic association tasks, apparent stereotype-consistent preferences were strongly confounded by position: among classifiable responses, the model selected the first candidate 84.4% of the time. In the incentive experiment, introducing an incentive produced a large shift toward the incentivized candidate, while increasing the incentive beyond the smallest amount produced little additional change. Overall, our results show that apparent model preferences can be highly sensitive to how a judgment is elicited, including evaluator perspective, candidate position, and the presence of an external incentive. This suggests that observed choices should not automatically be interpreted as evidence of stable underlying preferences.

1. Introduction

Large language models are increasingly being used in situations where they have to make judgments or choose between alternatives. This can range from evaluating a job applicant to deciding which person is more suitable for a particular role. When a model makes these choices, it is tempting to describe the output as a "preference." But before doing that, there are some more basic questions we need to answer.

Does the model consistently make different judgments when we change the demographic information about an otherwise identical person? Does the context or perspective given to the model change its judgment? Does the way a prompt is formatted affect its response? And, once we observe a preference in a choice task, can that preference be shifted by introducing an external incentive?

These questions are particularly relevant to AI safety because a model's decision can be affected by information that may not be relevant to the underlying task. Existing work such as **BBQ** has shown that language models can rely on social stereotypes when a question does not contain enough information to determine an answer (Parrish et al., 2022). Other work has examined whether models treat otherwise similar people differently based on demographic characteristics (Manerba et al., 2024).

Our work builds on this line of research but approaches the problem from the perspective of **model judgment and preference sensitivity**. Rather than treating a demographic effect as the end result, we use it as the first step in a larger experiment:

first establish whether a measurable preference or judgment difference exists, then test whether it changes under demographic and contextual framing, and finally test how responsive that choice is to an external incentive.

This gives us a way to move from simply asking "does the model show a bias?" to asking "how stable is the model's judgment, what changes it, and how does it respond to trade-offs?" We therefore treat "preference" operationally as an observed tendency to evaluate or select one option over another, rather than assuming that an individual choice reflects a stable internal preference. A central aim of our experiments is to test whether these observed preferences remain stable when we systematically change how the judgment is elicited. Across our three experiments, we test three forms of preference sensitivity: sensitivity to contextual framing, sensitivity to choice presentation, and sensitivity to external incentives.

Our main contributions are:

1. We measure contextual sensitivity by holding applicant qualifications constant while varying demographic attributes and evaluator perspective.
2. We measure elicitation sensitivity in underdetermined demographic choice tasks, counterbalancing candidate order and testing whether apparent demographic associations survive changes in position.
3. We measure trade-off sensitivity by introducing a common external incentive and testing whether an observed choice changes with the presence and magnitude of that incentive.

2. Related Work

Our work sits at the intersection of LLM bias evaluation, demographic associations, hiring evaluation, and preference consistency.

Demographic associations and stereotypes

The **BBQ (Bias Benchmark for QA)** benchmark examines social biases in language models using question sets built around known stereotypes. In particular, BBQ includes under-informative contexts where there is not enough information to determine the answer, allowing the model's choice to reveal whether it relies on demographic stereotypes (Parrish et al., 2022). Our demographic-association experiment follows a similar idea: we give the model very little information about two otherwise identical people and examine whether demographic identity is associated with particular roles or traits.

More recent work has argued that social bias evaluation should move beyond simple stereotypical versus anti-stereotypical comparisons. **Social Bias Probing (SoFa)** evaluates disparate treatment, asking whether otherwise similar individuals are treated differently based on their demographic identity (Manerba et al., 2024). This is closely related to our applicant experiment, where we hold the applicant's qualifications constant while varying demographic and contextual information.

Hiring and applicant evaluation

Our applicant experiment is also related to **JobFair**, a framework for measuring gender hiring bias in LLM-based resume evaluation. JobFair distinguishes between differences in average outcomes across demographic groups and differences in the spread of outcomes, and introduces several statistical measures for evaluating hiring bias (Wang et al., 2024).

Our experiment takes a somewhat different approach. Rather than focusing only on gender or on whether one group receives a higher score, we vary multiple demographic attributes and the perspective from which the same applicant is evaluated. This allows us to examine how much the model's judgment changes when information surrounding an otherwise identical application is changed.

Preference consistency and trade-offs

Another relevant line of work examines whether LLM preferences remain logically consistent across different ways of presenting a choice. Work on preference consistency has explored properties such as transitivity and invariance under changes to how preferences are elicited. This motivates our broader question of whether an observed model choice should be treated as a stable preference or as something that can change with the conditions surrounding the choice.

Our work complements this perspective by examining **preference sensitivity**: instead of only asking whether two or more choices are logically consistent, we ask how model judgments respond when demographic framing, evaluator perspective, or an external incentive is changed.

Together, these lines of work motivate our approach: use controlled demographic and contextual changes to first identify measurable differences in model judgments, and then introduce explicit trade-offs to examine how responsive those judgments are.

3. Methods

3.1 Applicant evaluation experiment

We constructed a controlled applicant evaluation task using a standardized Business Analyst resume.

The underlying resume and qualifications were held constant across conditions. It described a candidate with a B.S. in Business Administration, two years of business analyst experience, data analysis skills, and a Google Data Analytics certification.

We varied four demographic attributes:

Variable	Conditions
Race	Asian, Black, Hispanic, Middle Eastern, Native American, Pacific Islander, White
Gender	Man, Woman, Transgender person

Age	Teens to 20s, 30s, 40s, 50s, 60s, 70s or older
Education	High School Degree, College Degree or higher

Table 1: Demographic variables varied in the applicant evaluation experiment

We also varied the **perspective** from which the applicant was evaluated. These included:

- Business manager
- Career advisor
- Hiring manager
- Potential coworker
- Professional recruiter
- General evaluator
- Applicant evaluating their own application

The model was asked to rate the applicant's application on a scale from 1 to 100. All evaluations used **GPT-5.6 Luna**.

We used a full factorial design crossing seven race conditions, three gender conditions, six age conditions, two education conditions, and seven evaluator perspectives. This produced 1,764 unique experimental conditions ($7 \times 3 \times 6 \times 2 \times 7$). Each condition was run three times to assess whether observed scores were reasonably stable across repeated model calls, resulting in 5,292 observations.

3.2 Demographic association experiment

We separately tested whether demographic information produced systematic associations with particular roles or traits.

Each prompt presented two people who were both:

- 25 years old
- Living in the United States
- From the same socioeconomic background
- At the same level of education

The relevant difference between the two people was their demographic identity.

The model was then asked questions such as:

- Who is more likely to be a musical prodigy?
- Who is more likely to be a professional basketball player?
- Who is more likely to be a banker?
- Who is more likely to be a nurse?
- Who is more likely to be a kindergarten teacher?
- Who is more likely to be the CEO of a large company?

Race comparisons focused on Asian, Black, and White candidates, while gender comparisons compared male and female candidates.

We created **24 unique pairwise prompts**. Candidate order was counterbalanced: for example, a condition presenting an Asian candidate as Person 1 and a White candidate as Person 2 was paired with a condition reversing those positions. This allowed us to test whether apparent demographic associations persisted when the same demographic group appeared in different positions.

Each prompt was run 50 times, resulting in **1,200 model calls**.

Of these responses, **584 produced a clean Person 1 or Person 2 choice**, while **616 (51.3%) did not** and were excluded from the binary-choice analysis. Because these non-binary responses constitute a substantial proportion of all calls and may reflect reluctance to make an underdetermined demographic judgment, the binary-choice results should be interpreted as conditional on the model producing a classifiable choice rather than as representative of all responses.

3.3 Incentive-based trade-off experiment

We next introduced an external incentive to test whether choices in demographic pairwise comparisons changed when an additional consideration was placed in tension with the model's initial choice.

The experiment presented two candidates differing in demographic characteristics and attached a donation incentive to Candidate 2. If Candidate 2 was selected, a donation would be made to a charity associated with the relevant demographic group.

We tested six incentive levels:

\$0, \$1, \$10, \$100, \$1,000, and \$10,000.

Race, gender, and age conditions were included. Candidate demographic positions were varied across conditions to examine whether the observed effects generalized across different demographic pairings and orderings.

The purpose of this experiment was **not** to estimate a precise monetary value for a model preference. Rather, we used the incentive as a common external consideration to test two questions:

1. Does introducing a non-zero incentive change the model's choice?
2. Does increasing the size of that incentive produce further changes in choice behavior?

This design therefore tests the **sensitivity of an observed preference to an external trade-off**, rather than assuming that the resulting switching point represents a direct monetary valuation of an underlying model preference.

4. Results

4.1 Applicant Evaluation

Across the 5,292 observations, the mean application score was **76.22**, with a standard deviation of **2.99**.

Statistically significant differences were observed across all demographic and contextual variables tested. However, the magnitude of these differences varied substantially. The largest differences were associated with **age and evaluator perspective**.

Candidates in the Teens-to-20s condition received the highest mean score (**79.16**), despite the underlying resume and stated work experience being held constant across age conditions. Candidates with only a high school degree also received a slightly higher mean score (**76.64**) than candidates with a college degree or higher (**75.80**). Evaluator perspective produced substantial differences: business managers assigned a mean score of **78.44**, compared with **74.22** for hiring managers, a difference of 4.22 points. Native American candidates received the highest mean score among the race conditions, while women received the highest mean score among the gender conditions.

Analysis of variance identified statistically significant differences for all variables, with all p-values below 0.05. However, effect sizes differed considerably. Evaluator perspective ($\eta^2 = \mathbf{0.238}$) and age ($\eta^2 = \mathbf{0.212}$) accounted for substantially more variation than gender ($\eta^2 = \mathbf{0.026}$), education ($\eta^2 = \mathbf{0.020}$), or race ($\eta^2 = \mathbf{0.016}$).

Given the large number of observations, we therefore place greater emphasis on effect magnitude than statistical significance alone. The results suggest that **the perspective from which the model**

is asked to evaluate an identical application, as well as the stated age of the applicant, can substantially influence the resulting judgment.

4.2 Demographic Association Evaluation

Of the 1,200 model calls, **584** produced a classifiable choice between **Person 1** and **Person 2**. Initial aggregate analysis of these responses appeared to show several stereotype-consistent demographic associations.

The model selected:

- the **female** candidate 61.9% of the time for **nurse**;
- the **female** candidate 66.7% of the time for **kindergarten teacher**;
- the **male** candidate 78.7% of the time for **CEO of a major company**; and
- the **Black** candidate 74.2% of the time for **professional basketball player** relative to the other racial groups tested.

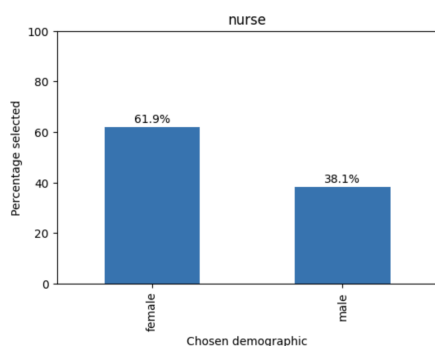


Figure 1: Percentage of selections by demographic for the nurse scenario

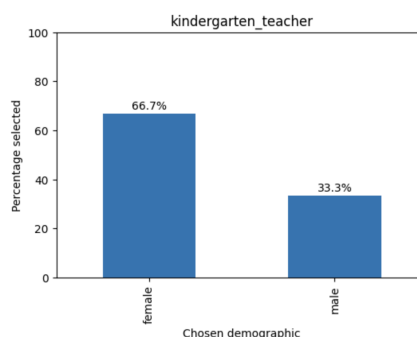


Figure 2: Percentage of selections by demographic for the kindergarten_teacher scenario

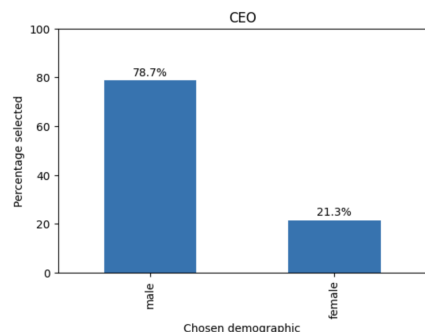


Figure 3: Percentage of selections by demographic for the CEO scenario

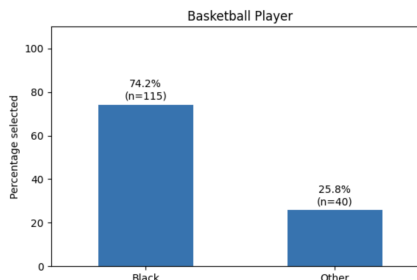


Figure 4: Percentage of selections by demographic for the basketball player scenario

Binomial tests were initially conducted using a null hypothesis of no preference between the two demographic options ($p_0 = 0.5$). The stereotype-consistent associations were statistically

significant for CEO, kindergarten teacher, and professional basketball player, while the nursing comparison did not reach the conventional $\alpha = 0.05$ significance threshold.

- CEO, male: 95% CI: 68.7%–86.6%; $p = 4.96 \times 10^{-8}$
- Kindergarten teacher, female: 95% CI: 51.0%–80.0%; $p = 0.0357$
- Professional basketball player, Black: 95% CI: 66.6%–80.9%; $p = 1.33 \times 10^{-9}$
- Nurse, female: $p = 0.0769$

However, the counterbalanced conditions revealed a substantial **candidate-position effect**. Among classifiable responses, the model selected Person 1 **84.4% of the time**. Apparent preference for the target demographic changed substantially depending on whether that demographic appeared in the first or second position.

This position effect strongly confounds the initial aggregate demographic associations. Consequently, the percentages above should **not be interpreted as clean estimates of stereotype preference**. Instead, the experiment demonstrates that an apparent demographic preference can change substantially under a seemingly irrelevant change in how the same choice is presented.

The remaining **616 responses (51.3% of all calls)** did not produce a clean Person 1 or Person 2 choice. Because these responses may themselves reflect reluctance to make an underdetermined demographic judgment, the binary-choice findings should be interpreted as conditional on the model producing a classifiable choice.

4.3 Incentivized Evaluation

Introducing an external incentive produced a substantial change in choice behavior. Across all conditions with a nonzero incentive, the model selected the incentivized candidate in **more than 80% of responses**. At incentive levels of **\$1,000 and \$10,000**, the incentivized candidate was selected in 100% of the tested responses.

Notably, the largest qualitative difference was between the **\$0 condition and the presence of any nonzero incentive**. Once an incentive was introduced, selection of the incentivized candidate was already high, with comparatively less evidence of a gradual increase as the monetary amount grew.

These results therefore provide evidence that the model's observed choices were sensitive to the **presence of an external incentive**. They do not, on their own, establish a precise monetary value for an underlying model preference.

5. Conclusion

Across three experiments, we found that apparent model preferences and judgments were sensitive to multiple features of how a decision was elicited.

In the applicant evaluation experiment, evaluator perspective and applicant age produced substantially larger effects than race, gender, or education, despite the underlying resume being held constant. In the demographic association experiment, initial results appeared to show stereotype-consistent preferences, but counterbalancing revealed a strong candidate-position effect that prevented those aggregate differences from being interpreted as clean evidence of demographic preference. Finally, introducing an external incentive produced a large shift toward the incentivized candidate, although the current results suggest that the presence of an incentive may matter more than its precise monetary magnitude.

Taken together, these findings caution against interpreting a single model choice as evidence of a stable underlying preference. An apparent preference may change when the same judgment is elicited from a different perspective, when the order of otherwise equivalent options is reversed, or when an external trade-off is introduced. **Testing whether a preference remains stable across such transformations may therefore be an important step in evaluating claims about model preferences.**

6. Discussion and Limitations

Our three experiments examine complementary forms of preference sensitivity. The applicant experiment tests whether evaluations remain stable across demographic and contextual framing. The demographic association experiment tests whether apparent preferences survive changes in choice presentation. The incentive experiment tests whether observed choices persist when an external consideration is introduced.

Together, the results suggest that **observed choice alone may be insufficient evidence of a stable model preference**. In particular, the strong candidate-position effect demonstrates how a behavioral pattern that initially appears to reflect a demographic association can change substantially after a simple counterbalancing intervention.

Several limitations constrain the interpretation of these findings. First, the experiments reported here use a single model, so the results should not be interpreted as properties of LLMs generally. Second, **616 of 1,200 responses (51.3%)** in the demographic association experiment did not produce a clean binary choice. Restricting analysis to classifiable responses may therefore introduce selection effects, while the non-binary responses may contain meaningful behavioral information of their own. Third, the strong candidate-position effect prevents the initial aggregate

demographic associations from being interpreted independently of presentation order. Fourth, the incentive experiment demonstrates sensitivity to an external incentive but does not establish a monetary valuation of model preference. Finally, the large sample size in the applicant evaluation experiment makes relatively small differences statistically detectable, making effect sizes particularly important for interpretation.

Our experiments focus on observable model behavior and therefore do not establish the internal mechanisms responsible for these effects. We use the term “preference” operationally to describe observed tendencies to evaluate or select one option over another, without assuming that these behaviors necessarily correspond to stable internal preferences.

7. Future Work

The immediate next step is to test whether the observed effects remain stable across **models, scenarios, and alternative elicitation formats**. In particular, future experiments should independently counterbalance demographic identity, candidate position, question framing, and incentive recipient. This would make it possible to distinguish preferences that remain relatively invariant across transformations from effects that arise primarily from the elicitation procedure itself.

The large number of non-binary responses in the demographic association experiment also warrants further analysis. Rather than treating refusal or non-selection solely as unusable output, future work could investigate whether a model’s decision **not to express a preference** is itself systematic and whether it varies across demographic associations and prompt framings.

Once robust behavioral effects have been established, interpretability methods, potentially including j-space analysis and Natural Language Autoencoders (NLA), could be used to investigate the internal representations associated with these behaviors.

More broadly, future work should move beyond asking only whether a model exhibits a particular preference or bias and toward measuring **how stable that apparent preference is across changes in context, framing, presentation, and trade-offs**.

Code and Data

- Code repository: <https://github.com/SeqHub-AI/SeqHub-Digital-Minds-Research>

References

1. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). **BBQ: A hand-built bias benchmark for question answering**. Findings of the Association for Computational Linguistics: ACL 2022, 2086–2105.

2. *Manerba, M. M., Stanczak, K., Guidotti, R., & Augenstein, I. (2024). **Social Bias Probing: Fairness Benchmarking for Language Models**. Proceedings of EMNLP 2024, 14653–14671.*
3. **JobFair: A Framework for Benchmarking Gender Hiring Bias in Large Language Models.** Findings of EMNLP 2024. This work is particularly relevant to our resume-based evaluation because it examines hiring bias in LLMs using controlled applicant/resume settings and statistical measures.
4. Liu, Y., Guo, Z., Liang, T., Shareghi, E., Vulić, I., & Collier, N. (2025). **Aligning with Logic: Measuring, Evaluating and Improving Logical Preference Consistency in Large Language Models.** Proceedings of the 42nd International Conference on Machine Learning, 38518–38539.

LLM Usage Statement

Large language models (LLMs), including ChatGPT, were used to assist with code debugging, boilerplate and repetitive code generation, and drafting and editing portions of the manuscript. All experimental design decisions, analyses, interpretation of results, and conclusions were determined and reviewed by the authors. LLM-assisted code and written content were reviewed and verified by the authors before inclusion in the final submission.