Generative Social Choice: OpenAI "Democratic Inputs to AI" Report

Sara Fish[1], Paul Gölz[2], Ariel D. Procaccia[1],
Gili Rusak[1], Itai Shapira[1], and Manuel Wüthrich[1]

[1] *Harvard University*
[2] *Simons Laufer Mathematical Sciences Institute*

January 16, 2024

Abstract

We propose an LLM-based process that distills a large number of free-text opinions into a handful of statements. Our design is centered on mathematical fairness guarantees from social choice theory, suggesting that these output statements can inform democratic decisions. We use this process to investigate public opinion in the US on the question: "To what extent should chatbots be personalized?"

This report is meant for a general audience. In an accompanying technical paper, we present our approach in greater detail, along with supporting theoretical and empirical results.

1 Overview

Many participatory platforms and processes — such as Pol.is, Remesh, and qualitative surveys — collect detailed free-text opinions from participants. Since the amount of information collected is too large to be directly digested by decision makers, it becomes necessary to *summarize* the articulated opinions. A key challenge is how to perform this summarization in a way that is scalable, while remaining representative of participants' opinions.

We design a summarization process that directly targets the dual goals of scalability and representativeness. Rather than a single statement, our summary takes the form of *multiple short statements*, each of which articulates a prevalent point of view among the participant body. Including multiple statements allows us to capture diverse and even incompatible opinions. Taken as a whole, this set of statements, which we call a *slate*, represents the opinion distribution across the underlying population. As a result, the slate can serve as input to decision-making processes of various kinds: human policy makers can easily digest these slates to base their choices on public opinion, and downstream algorithms supporting decision making can use our slates as a preprocessed input.

Our process design is grounded in the field of *social choice theory*, which studies the aggregation of individual preferences into collective, democratic decisions through mathematical analysis. Whereas collective choices between a few options (e.g., five candidates for mayor or whether to adopt a ballot measure or not) are well understood in the social choice literature, the existing literature gives limited guidance on open-ended decisions, such as our problem of finding an appropriate slate of textual statements, each of which can be selected among all conceivable statements. To address this limitation, we augment social choice theory with *large language models (LLMs)*, which gives rise to a principled, open-ended process for distilling public opinion.

As part of OpenAI's grant program for *democratic inputs to AI*, we piloted our process to study US residents' opinions on *the extent to which chatbots should be personalized.* For this, we elicited free-text opinions about this topic from a sample of 100 participants, which is representative of the US population in terms of age, gender, and race. We then distilled these free-text opinions into a representative slate of five statements.

These statements surface three major concerns that US residents have about chatbot personalization: *privacy and data security*, *user control*, and *truthfulness*. Most notably, we find broad public support for some form of personalization, as long as users control when their data is stored and when it is used for personalization. A second important concern of US residents is that personalization should not go so far that the chatbot would provide false or misleading information to the user.

To validate that these statements faithfully represent the population, we conduct a second survey, in which we ask a fresh sample of 100 US residents to what degree each statement in our slate represents their viewpoint. We find that we can split the 100 participants of the second survey into equally-sized groups, one per statement in our slate, such that 75% indicate that their assigned statement "perfectly" captures their opinion on chatbot personalization, and 93% indicate that the statement captures their opinion "mostly" or "perfectly".

2 Process Architecture

Our key objective is to ensure that the slate of output statements is *representative* of the opinions of participants. If we decide to generate a slate consisting of five statements, for example, the ideal of proportional representation suggests that a group consisting of one fifth of all participants ought to be represented by one of the five statements, provided that this group has a cohesive opinion. However, it is not obvious how to translate this intuition into a precise representation guarantee because participants generally do not just belong to a single, natural group. Instead, a participant agrees with various statements to different degrees, which aligns them to different degrees with different groups of participants.

We obtain the definition of a representation guarantee, as well as procedures for achieving this guarantee, by drawing an analogy to multi-winner elections, a well studied domain in social choice theory. In this analogy, we view the possible statements as candidates in an election, and each participant's free-text opinion as specifying the agent's preferences over these candidates. We extend a widely studied representation guarantee in multi-winner elections, *justified representation*, to our setting, which results in the following definition. A slate of k many statements satisfies *balanced justified representation (BJR)* if there is a balanced matching between participants and statements on the slate (each statement is matched to roughly the same number of participants) with the following property: there is no group of participants such that

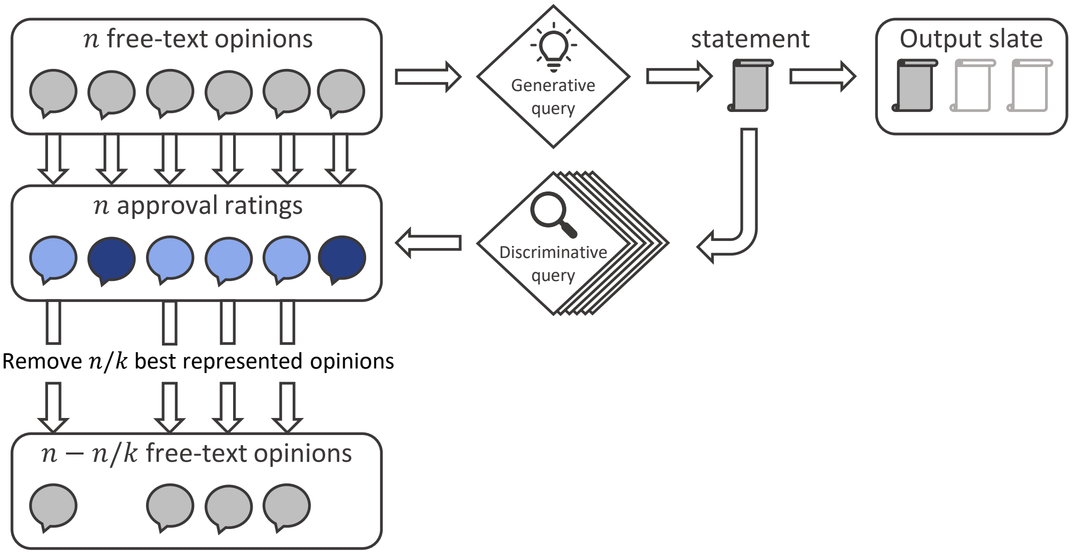

Figure 1: The process of generating statements with n participants and for a slate of size k. Our algorithm alternates between *(a)* generating new statements using generative queries, and *(b)* assessing their ratings among participants using discriminative queries to filter out participants with high approval (since they are already satisfied). The figure shows the first of k iterations of this process (for $k = 3, n = 6$), where each iteration adds a new statement to the slate and removes $n/k = 2$ satisfied participants.

- the group consists of at least a $1/k$ fraction of all participants,
- each group member rates their matched statement on the slate with at most level θ (say, all agents feel only "somewhat" represented or worse), and
- there exists *some possible* textual statement that all members of the group would rate higher than θ (say, they would feel "mostly" represented or better).

In other words, if there is a coalition of agents of large enough size to "deserve" a statement on the slate by proportionality, and all these agents have utility at least θ for some statement (so their preferences are *cohesive*), then the coalition may not be "ignored," in the sense that at least one member must be assigned to a statement with utility at least θ.

We now have a well-defined guarantee to aim for, but we must still bridge a significant gap between our setting and the social choice literature: Prior algorithms are designed to provide representation guarantees when the set of candidates is small and the explicit preferences of participants over the candidates are provided — in contrast to our setting, where the candidates are all possible statements and agents cannot be shown more than a few such statements. A framework we recently developed, *generative social choice*, bridges this gap. It combines the mathematical rigor of social choice theory with the capability of LLMs to generate text and extrapolate preferences. Specifically, given the opinions expressed by participants, we use LLMs to construct subroutines that answer the following two types of queries:

Discriminative queries: "Would a given participant support a given (previously unseen) statement?"

Generative queries: "Generate a new statement that would obtain the highest support among a group a given group of people."

We use these queries as building blocks in our process, which we outline in Figure 1. If the LLM answers these queries correctly, we can *prove* that the process will always return a slate satisfying balanced justified representation.

3 Limitations

Of course, the mathematical statement above is limited by the fact that LLMs will *not* always answer our queries correctly, and we therefore cannot be certain that the slate satisfies balanced justified representation. We empirically find, however, that *GPT-4*, a state-of-the-art LLM developed by OpenAI, can approximate our queries to a degree that we could not attain with classical algorithmic methods. In addition, we do not blindly trust the slate generated with the help of LLMs. Instead, we validate the match between our produced slate and the US population using a fresh sample of participants, as we discuss in detail later in this report. We also note that our approach is "future-proof", in the sense that as LLM technology improves over time, we expect answers to our queries to become more accurate, making the whole process more powerful and reliable.

Before our process can be deployed in high-stakes settings, it will require adaptations that increase its reliability and mitigate bias. In terms of reliability, our implementation of the generative query with GPT-4 sometimes produces unpopular or imprecise statements. Our current implementation increases robustness by generating multiple candidate statements with different approaches, and using the discriminative query to select the best among them. However, our process has yet to be hardened against malicious participant input, such as prompt injections meant to sway the generative queries in particular directions. Another issue requiring mitigation is the well-documented biases of LLMs. Both the base models and the RLHF-trained models may have biases against certain viewpoints, which could undermine our goal of impartial and representative aggregation. This issue should be empirically studied before deployment, along with strategies for mitigation, such as improved prompts or usage of specific models.

Perhaps the biggest challenge is the lack of transparency and predictability that is inherent to any process involving LLMs. If participants are available for interactive participation, we could replace each generative query in our process by a phase of participation in which participants and LLMs propose statements with large support, and the winning statement is selected by vote. In this case, the LLM would be limited to an assistive role under the supervision of participants, which we believe would substantially increase transparency and legitimacy. At the same time, however, such an adaptation would severely reduce scalability and speed. This trade-off appears to be a common theme in the design of democratic processes, and which point on the trade-off curve is appropriate depends on the setting.

4 Pilot

We pilot our process by eliciting public opinion on the personalization of chatbots. In designing our generative queries, we can choose the format of statements that we aim to aggregate, which should simultaneously express interesting nuance, be concrete, readable, and allow for

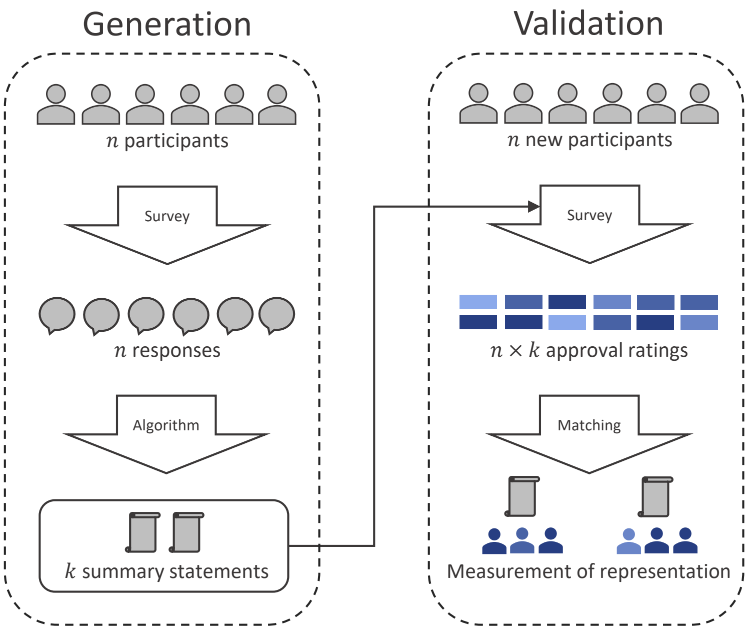

Figure 2: Overview over the pilot run of our process: In the first stage ("generation", left), we survey $n = 100$ participants. We then feed their responses into the statement-generation process in Figure 1 to generate a slate of $k = 5$ statements, each representing a fifth of the population. In the second stage ("validation", right), we validate these statements by asking a fresh sample of $n = 100$ participants to rate the five statements. Based on these ratings, we match participants optimally to statements, such that each statement represents an equal number of participants.

disagreement. To obtain actionable guidelines for the development of chatbots, we decided on a statement format that consists of a concrete rule for chatbot personalization, accompanied by a brief justification for the rule's importance, and an example illustrating the rule.[1]

4.1 Pilot Description

We illustrate the setup of the pilot in Figure 2. We first recruit 100 participants through the online platform Prolific. Our sample consists of US residents, stratified with respect to age, gender, and race. We ask these participants to complete a survey on chatbot personalization. To introduce participants to the topic of chatbot personalization, we first show them background information and and ask them about whether a chatbot should personalize its answer in each of three example scenarios. Then, we asked participants to describe their stance on chatbot personalization, by answering the following four questions in writing:

- "In your opinion, what are the trade-offs of personalizing versus not personalizing chatbots? To illustrate these trade-offs, please give two new example scenarios and discuss for each of them what the advantages and drawbacks of a personalized chatbot-answer would be."

[1] Examples of such statements will be shown in Section 5.1.

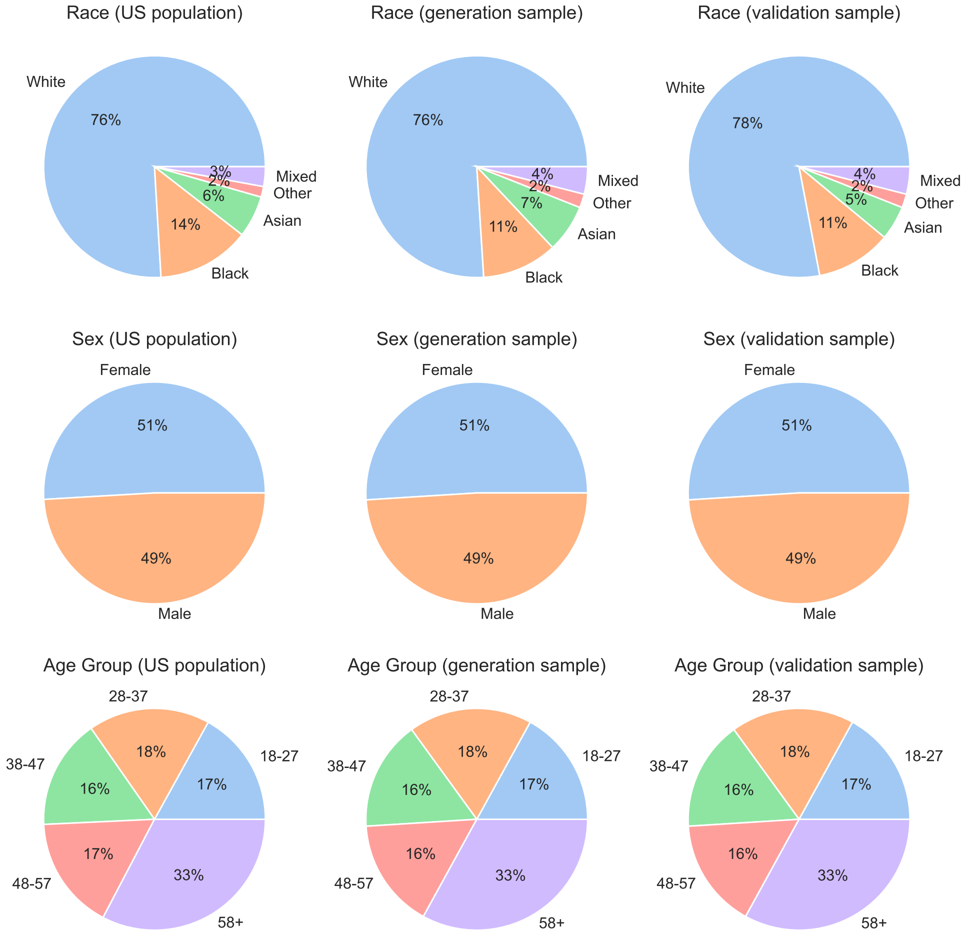

Figure 3: Demographic composition of both samples, compared to the US population as of the 2020 census. Racial and age groups are as defined by Prolific.

- "Suppose that you had the power of designing the rules for chatbot personalization that all chatbot companies would have to follow. What would these rules be? In what cases should/shouldn't chatbots give personalized answers? Please put particular emphasis on rules you consider important but other people may not have thought of or may not agree with."
- "Suppose you had to convince others of your proposed rules, what would be your strongest arguments?"
- "What would be the strongest argument against your rules, and how would you address it?"

We also ask participants to rate their agreement with six example statements, which we generated with a single call to GPT-4 and without knowledge of participant responses. Specifically, we ask them "to what extent does this statement capture your full opinion regarding chatbot personalization?" for each statement. Participants respond by choosing one level of the following scale: “not at all”, “poorly”, “somewhat”, “mostly”, and “perfectly”, and by providing a short textual justification of their choice. We use these responses to implement our discriminative queries with few-shot prompts in GPT-4.

Based on participant responses, we extract a slate of five representative statements using the process from Figure 1. To evaluate this slate, we then launch a second survey with a new set of 100 stratified participants to validate the 5 output statements (see the right box in Figure 2). In this second survey, after showing participants the same introductory information about chatbots, we ask them to rate the five statements (using the same question format as at the end of the first survey round).

For reproducibility, and to support future research on online participation, we made participants’ full responses available at `https://github.com/generative-social-choice/chatbot_personalization_data/`.

4.2 Demographic Composition of Samples

As shown in Figure 3, both samples closely reflect the composition of the US population in terms of race, sex, and age groups.[2] Since we adopt Prolific’s categories for race and age, we are not aware of how many respondents identify as Hispanic or Latino. Though Prolific’s highest age category (“58+”) is quite broad, we find that older residents within this age bracket are also accurately represented: our generation and validation samples respectively contain 15% and 16% respondents aged 68 and older, compared to a share of 17% in the adult population according to the 2020 census.

5 Results

In this section, we first show and analyze the generated slate of five statements. We then use information from both samples to evaluate how well the slate represents the US population.

5.1 What Do the Statements Say?

The generated slate contains the following five statements. We highlight key points in color:

S1. The most important rule for chatbot personalization is to **give users control over the extent of personalization and the data supplied**. This rule is crucial as it ensures user autonomy, **privacy**, and a personalized experience. For instance, a user could choose to share their dietary preferences with a health chatbot for tailored advice, while opting not to disclose sensitive health data.

S2. The most important rule for chatbot personalization is to always **give users the choice whether the AI chatbot can remember their data or not**. This rule is crucial because it **respects the user’s privacy** and gives them control over their own data. For instance, a user might prefer a chatbot not to store any data about their past travels, thus avoiding unsolicited vacation suggestions.

[2] In fact, the sample is not just representative along sex and age groups, but also within all intersection groups of sex and age.

S3. The most important rule for chatbot personalization is to always **prioritize user privacy and data security**. This is crucial because it ensures the protection of sensitive user information, thereby building trust and promoting responsible AI use. For instance, a chatbot providing personalized health advice should only **collect and use data with explicit user consent**, and should implement robust measures to prevent unauthorized access or data breaches.

S4. The most important rule for chatbot personalization is to **avoid providing false or misleading information**. This rule is crucial because it ensures the reliability and trustworthiness of the chatbot, which is essential for user engagement and satisfaction. For instance, if a user asks a chatbot for medical advice, providing accurate information could potentially save lives.

S5. The most important rule for chatbot personalization is to **emphasize privacy** and require **user consent for data collection**. This rule is crucial to ensure personal security and mental health protection. For instance, a health bot providing personalized services can offer tailored care, but without proper privacy measures, it risks violating user privacy.

Notably, no statement is categorically opposed to personalization, but each statement expresses restrictions on personalization that major groups of US residents believe should be respected.

We understand the following three points to be the main themes of the slate:

- **Privacy and data security**: Four out of five statements stress the importance of privacy and of preventing chatbot data from being used in other contexts.
- **User control**: Four out of five groups believe that it is essential that users have granular control over which of their data are stored and used for personalization.
- **Truthfulness**: The third statement's primary concern is that chatbots should never provide inaccurate or misleading information.

A striking feature of the slate is the high level of agreement between statements: Indeed, four of the five statements all express a concern about privacy and data security while recommending user control as a guardrail on personalization. Both the high level of agreement between participants, and the popularity of these two themes came as a surprise to us.

Before we investigate this point in more detail, we want to highlight that the four statements, while aligned in their high level themes, connect them in different ways and emphasize different nuances. For instance, statement S5's concern about privacy and user control is justified by security and mental health concerns, which is much more specific than the more generic justification of, say, S2. Another interesting statement is S1, in which privacy appears only as one out of multiple underlying values served by user control, and which stresses not just user control at the time of data collection, but also control about the level of personalization when the chatbot is subsequently used. As we will see in Section 5.3, participants frequently rate their agreement with the four statements in the cluster quite differently.

The remaining statement, S4, stresses that chatbot personalization should not go so far as to compromise the chatbot's truthfulness. While this option was also not brought up by our introductory materials, either, one of our expository scenarios touched on a related point by asking if a chatbot should deliver distressing information in a gentler manner to a depressed user. Statement S4 does not take a position on this specific question, but sets a clear boundary on how far the chatbot might go to accommodate the user's presumed vulnerability. (The statement does not rule out that the chatbot might decline to answer in this situation.)

5.2 Do the Statements Represent the Generation Sample?

Given the novelty of our process, and the central role of LLMs in it, we need to thoroughly verify that our slate indeed faithfully represents participant opinions rather than being based on hallucinations by the LLM. In part, this concern will be addressed by the next section's analysis, which will show that a fresh sample of participants indeed feels accurately represented by the statements on our slate.

In this section, as an orthogonal analysis, we manually inspect and hand-label the responses of our generation sample to trace how our process arrived at the slate starting from participants' statements. Reassuringly, we find that privacy and data security and user control are indeed central themes in people's free-form opinion statements: 61 of the 100 participants touch on privacy and data security in their statements, 38 suggest user control, and 72 bring up at least one of these two topics. Though we have not attempted to systematically label all recurring themes in the survey responses, privacy and data security is certainly one of the most prevalent themes, and quite likely the most prevalent one.[3] That the themes of privacy and user control are so prevalent is particularly noteworthy because no part of our introductory materials primed participants towards these topics to our understanding, but that participants instead independently arrived at these points.

The number of 72 participants who touched on privacy and data security and user control alone can plausibly justify that these themes take up 80% of the slate. Moreover, this number does not yet count agents who expressed agreement with these themes outside of the free-form responses. Indeed, the six statements we show to the generation sample include a statement that touches on user control:

> "The most important rule for chatbot personalization is to always offer an opt-out. Mandatory personalization disregards user autonomy. For example, a person might not want location-based suggestions just because they mentioned a city once."

This statement received high ratings among participants of the generation sample: 49 of them rated this statement as "perfectly" capturing their opinion, 76 participants rated this statement as "perfectly" or "somewhat" capturing their opinion, and only 3 participants rated this statement as capturing their opinion "poorly" or "not at all".[4] Furthermore, this statement from the generation round does not yet touch on the (frequently mentioned) topic of privacy, whose addition might further enhance a statement's appeal. In light of these observations, representing 80% of agents with a statement about privacy and data security and user control seems like a reasonable choice.

[3] By comparison, truthfulness was mentioned by 48 participants (among which 32 also mention at least one out of privacy and data security and user control), and 35 participants mention concerns that information from the chatbot could lead to direct harm (either because false information leads to harm, or because the information supports the user in harmful actions such as criminal activity).

[4] These ratings in the generation sample are not directly comparable with the ratings of the validation sample, since participants in both surveys have been primed quite differently. By the time we ask the participants of the generation round to rate this statement, they have spent considerable time in the survey considering specific scenarios and describing their opinions in free text. By contrast, participants in the validation sample have only been exposed to the introductory text about chatbot personalization.

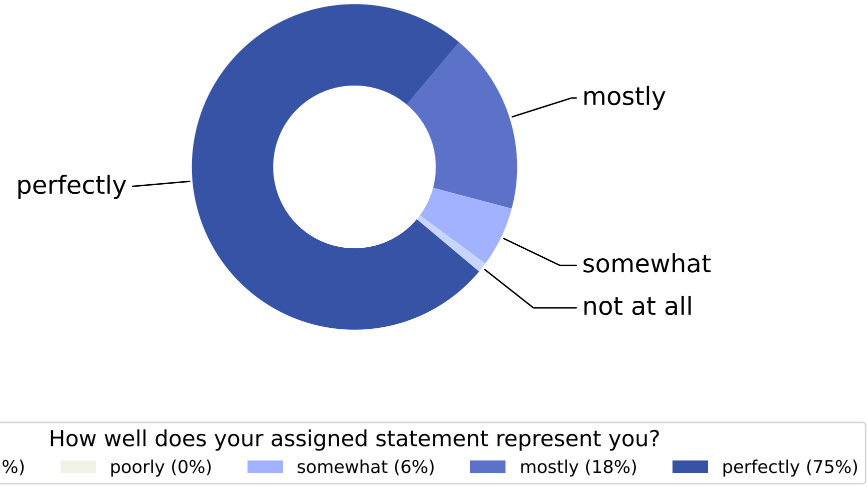

Figure 4: Ratings of participants from the validation survey for their *assigned* statement.

5.3 Do the Statements Represent the Validation Sample and US Population?

According to the ideal of proportional representation, each statement in our generated slate should represent 1/5 of the US population as accurately as possible. To verify this, we optimally match the participants of our validation sample (which, recall, mirrors the US population with respect to sex, age, and race) to the statements of our slate, and then study the ratings of participants for their assigned statements.

As can be seen in Figure 4, 75% of the participants say that their assigned statement "perfectly" captures their full opinion on chatbot personalization, and an additional 18% of participants say it "mostly" captures their full opinion. Only 7% of participants feel only "somewhat" represented or less. Hence, the vast majority of participant opinions are represented accurately by our slate of statements.

Remarkably, none of the 100 agents have a higher rating for a statement other than their assigned statement, which means that the requirement to assign *an equal number* of agents to each statement is not a binding constraint. This is a good sign for our claim of proportional representation, which could be in question if, say, many agents would rather be matched to the truthfulness statement S4 than their current assignment. It also shows that, should the slate violate BJR, this violation would have to be based on an entirely different kind of statement. Moreover, since such a violation would have to strictly increase the utility of all 20 members of the deviating coalition, it would have to unite most of the 25 agents who are not yet "perfectly" represented and would have to "perfectly" represent all members of this coalition who are already "mostly" represented. While we cannot entirely rule out such a BJR violation, this narrow path makes the existence of a violation seem unlikely.

Naturally, it is important to closely inspect the minority of 7 agents who feel relatively badly represented by their assigned statement, since their responses could potentially reveal viewpoints missing from our slate. Though the free-text explanations given with the ratings are typically short, they allow us to understand what the seven participants dislike about the selected statements. While certain themes occur repeatedly among these seven participants,[5] their reasons for feeling relatively unrepresented are eclectic. Since proportionality axioms like BJR only guarantee representation to large, cohesive groups, these responses also give us no reason to doubt the representativeness of our slate.

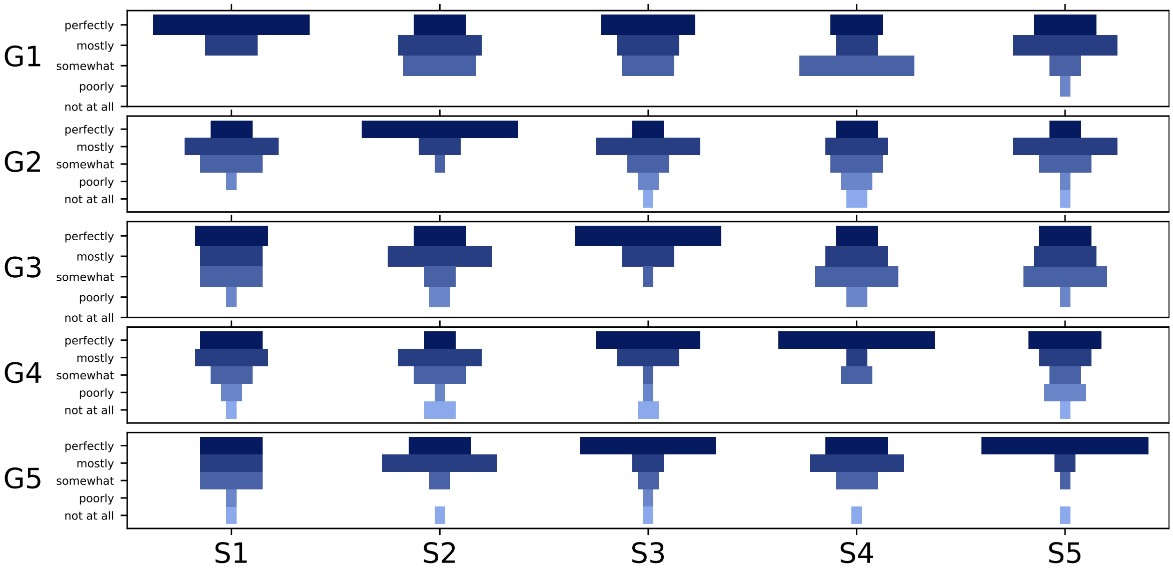

Figure 5: Agreement of participants in different groups with each of the statements. Each row corresponds to a group; for example, G1 represents the 20 participants assigned to statement S1. For each group, we plot the frequencies of rating levels given by members of this group to statements S1 through S5.

Having established that the slate of statements represents the population well, an interesting question is how distinct the preferences of different groups are. Are they all very similar and would be just as happy with another group's statement? To answer this question, we display the distribution of ratings across statements for each group in Figure 5. Comparing the different plots, it is clear that different groups have different preferences across statements. In particular, each group has a very clear preference for its assigned statement over the other statements (in Figure 5, see the distributions on the diagonal, from top left to bottom right).

Taken together, Figure 4 and Figure 5 indicate that there is heterogeneity in opinions across the population and that our slate accurately represents this heterogeneity.

[5] For instance, four of these participants do not believe that chatbot companies can be trusted to not collect data despite their customers' privacy choices or to keep collected data safe; and three express that the advantages of including all available data outweighs potential privacy risks. At least three of the participants doubt that chatbots can meaningfully identify truth or should be relied on as truthful.