The steerability of large language models toward data-driven personas

Junyi Li[1], **Charith Peris**[2]*, **Ninareh Mehrabi**[2]*, **Palash Goyal**[2], **Kai-Wei Chang**[2], **Aram Galstyan**[2], **Richard Zemel**[2] **and Rahul Gupta**[2]
[1]University of Maryland, College Park
[2]Amazon
junyili.ai@gmail.com
{perisc, ninarehm, palashg, kaiwec, argalsty, rzemel, gupra}@amazon.com

Abstract

Large language models (LLMs) are known to generate biased responses where the opinions of certain groups and populations are underrepresented. Here, we present a novel approach to achieve controllable generation of specific viewpoints using LLMs, that can be leveraged to produce multiple perspectives and to reflect the diverse opinions. Moving beyond the traditional reliance on demographics like age, gender, or party affiliation, we introduce a data-driven notion of *persona* grounded in collaborative filtering, which is defined as either a single individual or a cohort of individuals manifesting similar views across specific inquiries. As individuals in the same demographic group may have different personas, our data-driven persona definition allows for a more nuanced understanding of different (latent) social groups present in the population. In addition to this, we also explore an efficient method to steer LLMs toward the personas that we define. We show that our data-driven personas significantly enhance model steerability, with improvements of between $57\% - 77\%$ over our best performing baselines.

1 Introduction

In the recent past, Large Language Models (LLMs; Brown et al. 2020; Ouyang et al. 2022; OpenAI 2023; Touvron et al. 2023) have shown exceptional generation capabilities across a range of tasks (Zhao et al., 2023) that have led to their adoption in applications pertaining to multiple high-stakes fields such as healthcare (Singhal et al., 2023), education (Tan et al., 2023) and finance (Wu et al., 2023). Given this context, it is of utmost importance to avoid biases towards specific underrepresented populations, and leverage LLMs in a way that enables generation across a broad spectrum of viewpoints in a balanced way.

In practice, LLMs have been shown to generate responses that represent a wide range of opinions, with tendencies to over-represent the opinions of certain populations while under-representing those of others. For example, Santurkar et al. (2023) showed that LLMs under-represent the opinions of individuals aged 65 and over, Mormons, and the widowed, which constitute significant portions of the US population. The reason for this is that the typical fine-tuning of an LLM, done across datasets at hand, leads to a model gaining a randomized viewpoint based on the nature of the dataset.

Instead of fine-tuning towards such a randomized viewpoint, it is desirable to enable LLMs to have *controllable generation* that can be steered towards specific viewpoints. This can be leveraged to produce multiple perspectives and in turn encourage diversity through the curated inclusion of a broad spectrum of viewpoints. Such a diverse set of perspectives, enabled via controllable generation, can be extremely helpful in diminishing polarization and preventing the marginalization of the voices of minority groups.

Prior work has attempted to control LLM generation by aligning models toward demographic groups that are defined based on features such as age, gender, political party affiliation (Santurkar et al., 2023; Hwang et al., 2023; Simmons, 2022; Jiang et al., 2022; Feng et al., 2023; Salewski et al., 2023). However, we argue that a simple group definition based on demographic features might not be sufficient to represent the nuances of the underlying different social groups present in a given population.

In this paper, we present a new approach to achieve controllable generation of specific viewpoints using LLMs. We hypothesize that for a given dataset comprising of responses provided by a population of individuals to a set of questions, there is a space of differing characteristic opinions and beliefs. We map this space to an embedding

* These authors contributed equally to this work

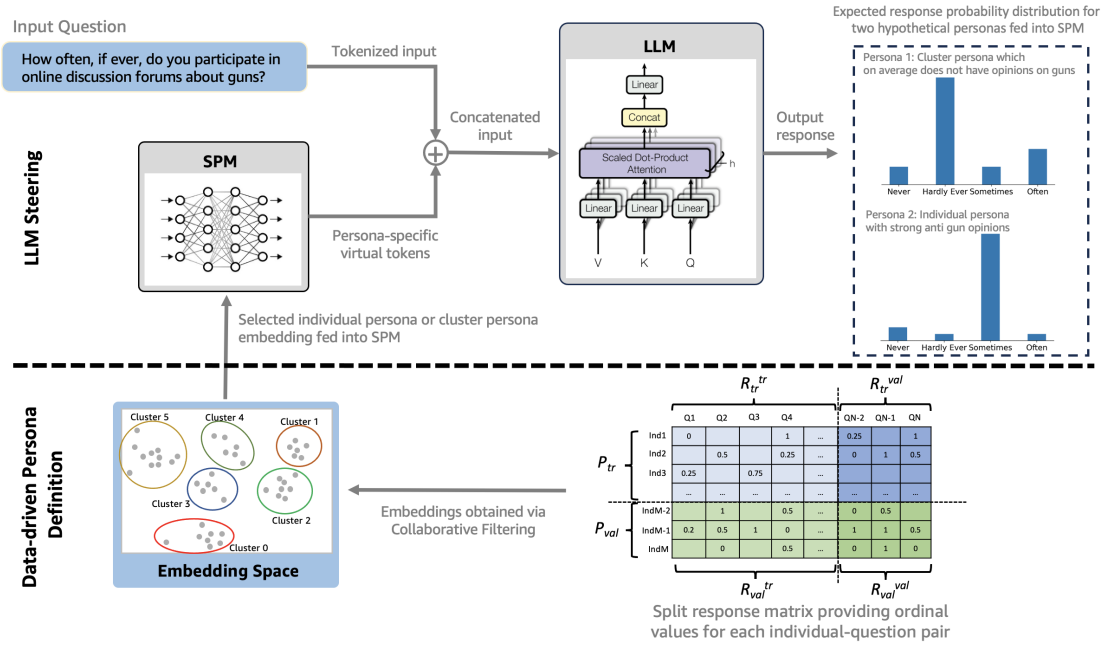

Figure 1: A schematic of our framework for steering LLMs toward data-driven personas. The bottom-half illustrates the formation of data-driven personas, and the top-half illustrates LLM steering. A persona is defined by generating individual embeddings via collaborative filtering. The persona can be a single individual embedding (grey dots) or the centroid of a group of embeddings, referred to as a cluster persona (denoted by the circled clusters). To steer the LLM we pass an embedding to a soft-prompting model (SPM), which maps the embedding to a set of persona-specific virtual tokens. Finally we prepend these virtual tokens to the tokenized input sequence and pass this into the LLM to obtain a persona-specific response.

space. We then propose the notion of *personas*, as examples from within this personality space. We use the term *persona* to refer to a portion of the embedding space which represents similar opinions and beliefs over a set of questions in our dataset of choice. This could range from a single individual embedding, referred to as an *individual persona*, to a group of individual embeddings represented by their centroid in embedding space, referred to as a *cluster persona*. To create this mapping we take a data-driven approach using collaborative-filtering to embed individuals into a continuous vector space which, to our knowledge, has not been done before. Compared to the use of traditional demographic traits, our definition of personas allows for nuanced understanding of different social groups in the population and makes the notion of steerability more meaningful.

Given these data-driven personas, we then propose an efficient algorithm to steer LLMs towards both an individual and a cluster persona. In particular, we learn a soft-prompting model (SPM) which maps the embedding of a persona to a set of virtual tokens. These virtual tokens are prepended before tokens mapping to the actual input text, to steer the responses of the LLMs.

We conduct a number of experiments using the OpinionQA dataset (Santurkar et al., 2023) to evaluate the efficacy of our persona definition and the LLM steering algorithm. Our experiments show that LLMs steered via personas can align to the opinions of individuals and groups better than baseline methods. In particular, the personas defined using our approach result in more accurate prediction of opinions when compared to the use of personas based on demographic-traits.

We summarize our **contributions** as follows:

- We propose a hitherto unexplored paradigm in steerable generation where, instead of predefined demographic features, we use a data-driven notion of a persona to modulate the generation process. In particular, we use collaborative-filtering to embed opinions of individuals within a dataset into a continuous vector space (individual personas), and then cluster groups of individuals with similar embeddings (cluster personas).

- We propose a simple model and an efficient algorithm to align LLMs with these personas

(both individuals and clusters).

- Finally, we benchmark our approach using a selected set of LLMs.

2 Related Work

Steering LLMs: In the recent past, several studies have investigated opinions expressed by LLMs (Santurkar et al., 2023; Durmus et al., 2023; Scherrer et al., 2023; Santy et al., 2023). For example, Santurkar et al. (2023) curated the OpinionQA dataset based on public opinion polls and evaluated the alignment of LLM opinions with 60 U.S. demographic groups. They found a substantial misalignment between the views reflected by current LLMs and those of U.S. demographic groups (we mention a cross section of these in Section 1). In another case, Scherrer et al. (2023) designed a survey comprising both high-ambiguity and low-ambiguity moral scenarios to study the moral beliefs encoded in different LLMs.

A variety of methods have also been proposed to steer the generation of LLMs toward specific opinions (Hwang et al., 2023; Simmons, 2022; Durmus et al., 2023; Argyle et al., 2023; Deshpande et al., 2023; Jiang et al., 2022; Feng et al., 2023). For instance, Santurkar et al. (2023) added demographic information (such as Democratic or Republican affiliations) to prompts in order to steer the opinions of LLMs toward those groups. Hwang et al. (2023) worked towards aligning LLMs with individuals by incorporating their past opinions. Simmons (2022) constructed four prompts based on combinations of *liberal* and *conservative* political identities, as well as moral and immoral stances, to steer the moral beliefs of LLMs. Jiang et al. (2022) fine-tuned LLMs using tweets authored by Democrats and Republicans to improve the model's alignment with the opinions of these groups.

In contrast to this prior work, which steer LLMs based on demographic traits, our approach using data-driven personas allows for a more expressive and nuanced understanding of the different social groups present in the overall population. This, in turn, enhances the applicability of model steerability.

Parameter-Efficient Fine Tuning: Given the rapid increase in size of LLMs, fine-tuning the full model has become costly, giving rise to various methods that only tune a portion of model parameters (Hu et al., 2021; Lester et al., 2021; Li and Liang, 2021; Liu et al., 2021, 2022; Zhang et al., 2023; Ozdayi et al., 2023). In one such approach, Hu et al. (2021) utlize low-rank decomposition to represent weight updates with two smaller matrices (called update matrices), and then only fine-tune the decomposition matrix. Several approaches have been proposed where virtual tokens can be tuned to align a model towards a specific task. Lester et al. (2021) trains a sequence of virtual tokens that have been prepended to the input whereas Li and Liang (2021) optimize a task-specific vector of virtual tokens called the prefix. Zhang et al. (2023) address the instability at the initial training phase using a learnable gate to control the effect of virtual tokens over generation.

3 Framework

In this section, we propose the notion of data-driven personas (Section 3.1) and an efficient algorithm for steering an LLM towards them (Section 3.2). An illustration of our process is shown in Figure 1.

3.1 Data-driven persona definition

Instead of relying on demographic information such as age, gender or party affiliation, we use collaborative filtering to embed all individuals into a continuous vector space based on their opinions. Then a *persona* is defined as a portion of the embedding space which represents similar opinions and beliefs. In particular, an *individual persona* is represented by a single individual embedding, while a *cluster persona* is represented by the centroid of a cluster of individuals.

Assume we have a set of questions $\mathcal{Q}$, consisting of multiple-choice questions that feature options with an ordinal structure. Furthermore, assume that we also have responses for these questions, given by a set of individuals $\mathcal{P}$. We represent these responses as a matrix R. If individual i responds to question j, the element $r_{i,j} \in [0, 1]$ represents the individual i's response, where the responses are mapped to the interval $[0, 1]$. If no response is given, $r_{i,j}$ is set to null. $\mathcal{R} = \{(i, j), \text{where } r_{i,j} \text{ is not null}, i \in \mathcal{P}, j \in \mathcal{Q}\}$ denotes the full set of responses. We utilize collaborative filtering (CF) to learn a continuous representation for each individual. More specifically, we denote the individual embeddings as $\{u_i \in \mathbb{R}^d, i \in \mathcal{P}\}$ and the question embeddings as $\{q_j \in \mathbb{R}^d, j \in \mathcal{Q}\}$ and optimize the following

objective:

$$\underset{\substack{\{u_i \in \mathbb{R}^d, i \in \mathcal{P}\} \\ \{q_j \in \mathbb{R}^d, j \in \mathcal{Q}\}}}{min} \sum_{(i,j)\in\mathcal{R}} \mathcal{L}(\langle u_i, q_j \rangle, r_{i,j}) \quad (1)$$

where $\mathcal{L}$ is a loss objective, *e.g.* the mean square error; $\langle\cdot\rangle$ denotes the inner product.

By optimizing the objective in (1), we obtain the converged embeddings $\{u_i \in \mathbb{R}^d, i \in \mathcal{P}\}$ that encode important information about individual opinions. We explore steerability towards individual embeddings (*individual personas*), as well as towards clusters of individuals with similar opinions within the overall population (*cluster personas*).

3.2 Steering LLMs towards data-driven personas

For steering LLMs towards the personas defined in Section 3.1, we draw from the *prefix-tuning* technique (Li and Liang, 2021) which prepends and tunes a set of prefix-vectors to each model layer that yields improvements in model generation towards specific tasks (Lester et al., 2021; Ozdayi et al., 2023; Li and Liang, 2021; Liu et al., 2021, 2022). In our approach, we use a separate model that we refer to as the soft-prompting model (SPM), to map a given embedding (of a persona) from our continuous vector space, to a set of prefix-vectors. These are prepended to the tokenized input (in this case the tokenized question) and enables the steering of the generation process towards that specific persona. A schematic of this process if shown in the top-half of Figure 1.

Unlike vanilla prefix-tuning which trains a single set of virtual tokens for a given task, our method uses a single SPM to generate a sets of virtual tokens for all personas in our dataset. This is beneficial in that it is cost effective and also performant, since personas with proximate embeddings often reflect analogous opinions.

We train the SPM by optimizing the following objective:

$$\underset{\theta\in\Theta}{min} \sum_{(i,j)\in\mathcal{R}} \mathcal{L}(\text{LLM}(f(u_i;\theta), \text{Q}_j), \text{R}_{i,j}) \quad (2)$$

where $\mathcal{R}$ is the set of responses, Q_j is the tokenized representation of question j, $R_{i,j}$ is the tokenized representation of the response of individual i to question j, u_i is the embedding of individual i, $f(\cdot;\theta)$ denotes the SPM parameterized by θ, and $\mathcal{L}$ denotes the loss objective, which in this case is cross entropy loss. During SPM training, the LLM weights are frozen and we only use individual embeddings, then during inference, we fix the weights of the SPM and steer the LLM generation towards the opinions of a specific persona by feeding its embedding as the input to the SPM.

To test the robustness of our technique to other prompting methods, we also test if our method works when using prompt-tuning (Lester et al., 2021) (as opposed to prefix-tuning Li and Liang 2021). For details, please see Appendix A.5.

4 Experimental results

In this section we present our data, experimental setup and empirical results on data-driven persona definition and LLM steerability.

4.1 Dataset Details

We use the OpinionQA dataset (Santurkar et al., 2023) which includes opinions of a diverse set of individuals over a wide range of different topics, for our work. The OpinionQA dataset is curated from 15 American Trends Panel polls, and encompasses responses from 18,339 participants to 1,476 multiple-choice questions across 23 different topics. A full list of topics covered by OpinionQA is provided in Appendix A.1. The demographic information for each participant (i.e., Race, Ideology, Education etc.) are included in the dataset and we also include this information in Appendix A.1.

The response options for each question in the OpinionQA dataset is presented as ordinals, and we map each option to a numerical value. For example, for one set of response options *Worry a lot*, *Worry a little*, and *Not worry at all*, we assign values of 0, 0.5, and 1, respectively. We then use the numerical values of the responses as labels in equation (1) to learn the individual embeddings.

4.2 Analysis of cluster personas

Here, we present an analysis of the cluster personas obtained from our data-driven approach described in Section 3.1.

Cluster Definition: We employ the matrix factorization approach (Eq. (1)) to perform collaborative filtering. We embed individuals and questions into a 16-dimensional space and employ mean square error as our training objective. The learned individual embeddings are referred to as *individual personas*. As for *cluster personas*, we cluster individual embeddings using K-Means clustering.

We run K-Means with varying numbers of clusters, denoted as k, and for each value of k, we replace individual embeddings with the centroids of each cluster to evaluate Eq. (1). This measures how well the cluster centroids can represent individuals and we use it as the criteria to select k. Specifically, we utilize the *elbow* heuristic (Bishop, 2006) to choose $k = 6$, and we label these clusters as Cluster-X, where X ranges from 0 to 5, which represent the six cluster personas that we choose.

For each cluster persona, we present the demographic composition across 13 different traits. We also detail the characteristics of opinions within each cluster, specifying questions for which a cluster disagrees with the overall population and questions where the clusters disagree with each other.

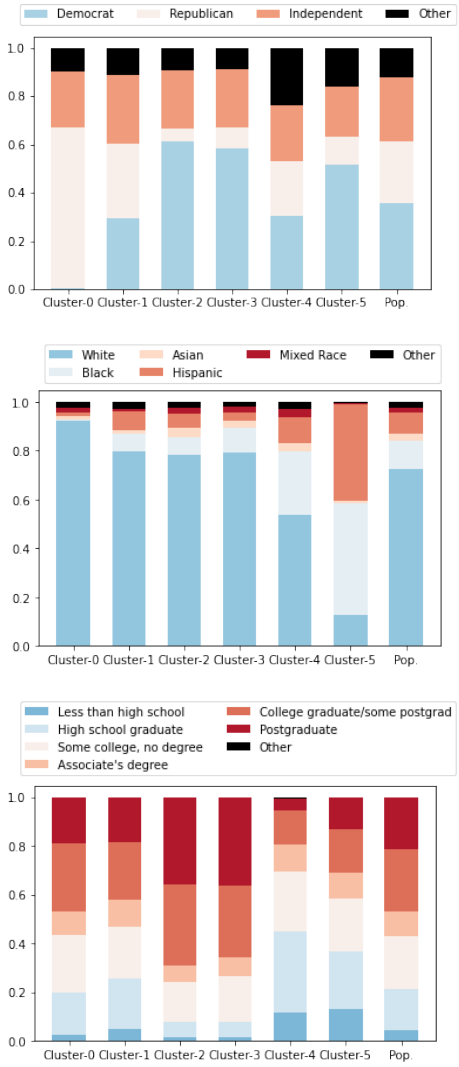

Figure 2: The Political Party, Race and Education composition (from top to bottom) of clusters and overall population.

Demographic Composition of Clusters: We calculate the demographic statistics for each cluster and summarize the results for Political Party, Race and Education composition in Figure 2 (compositions of other traits are summarized in Figure 3 in the Appendix A.5). Our analysis reveals that these clusters have distinct demographic compositions. For instance, in terms of political party, cluster-0 and cluster-1 lean toward Republican, while all other clusters lean toward Democrat. In terms of race, cluster-5 is dominated by Black and Hispanic races, while other clusters are dominated by the White race as in the overall population. In terms of education level, while cluster-0 and cluster-1 are dominated by members who received a college level education, cluster 2 and cluster-3 have a majority of members that have received post-graduate level education. In cluster-4 and cluster-5, the majority of members only received a high school level education. We also find that clusters are composed of a mixture of different demographic groups, which corroborates the fact that individuals with different demographic traits can hold similar opinions over many topics. This further verifies the necessity of defining groups based on actual opinion as opposed to depending solely on demographic characteristics.

Questions for which a cluster disagrees with the overall Population: To further investigate the characteristics of each cluster, we show questions where a cluster mostly disagrees with the opinion of the overall population. For each cluster, we calculate the response distribution to each question, where we denote the distribution of cluster c over question q as $D_{c,q}$ for $c \in [5], q \in \mathcal{Q}$. Similarly, we calculate the overall population response distribution over each question and denote it as D_q. For each question $q \in \mathcal{Q}$, we calculate the Total Variation (TV) between the cluster response distribution and the overall population response distribution $TV(D_{i,q}, D_q), i \in [5]$.

Note that a larger value of $TV(q)$ indicates a greater degree of disagreement between the cluster and the overall population for a given question. In Table 1, we show the three questions with the largest total variation between cluster response distribution and overall population response distribution for cluster-0. We show the corresponding questions for clusters 1 to 5 in Tables 12 to 16 respectively, in the Appendix. Taking Cluster-0 as an example, we find that its members feel that

Table 1: Comparison of responses between Cluster 0 and the overall population

Question	Cluster 0 Response	Overall Population Response
How much, if at all, do you think the following proposals would do to reduce economic inequality in the U.S.? Expanding government benefits for the poor	A great deal: 0.0% A fair amount: 2.27% **Not too much: 71.55%** Nothing at all: 26.16%	A great deal: 34.19% **A fair amount: 43.7%** Not too much: 19.83% Nothing at all: 2.26%
How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?	A great deal : 0.11% A fair amount: 22.86% **Not too much: 72.46%** Not at all: 4.55%	**A great deal : 46.86%** A fair amount: 38.02% Not too much: 14.51% Not at all: 0.6%
How well does the Democratic party represent the interests of people like you?	Very well: 0.0% Somewhat well: 0.0% Not too well: 11.94% **Not at all well: 88.05%**	Very well: 4.97% **Somewhat well: 44.87%** Not too well: 31.93% Not at all well: 18.21%

Table 2: Questions pertaining to immigration that exhibit the most disagreement among clusters

Question	Cluster	Response
If you were deciding what the federal government should do to improve the quality of life for future generations, what priority would you give to who come here legally? Allowing more immigrants into the U.S.	Cluster-0 Cluster-1 Cluster-2 Cluster-3 Cluster-4 Cluster-5	A lower priority (72.81%) Top priority (99.05%) Important, but not top priority (80.44%) Important, but not top priority (89.87%) Important, but not top priority (63.92%) Top priority (69.51%)
How much, if at all, do you think the following proposals would do to reduce economic inequality in the U.S.? Reducing illegal immigration	Cluster-0 Cluster-1 Cluster-2 Cluster-3 Cluster-4 Cluster-5	A great deal (60.63%) A great deal (95.58%) Not too much (68.95%) Not too much (68.67%) A fair amount (69.41%) A fair amount (69.67%)

expanding government benefits for the poor will not reduce economic inequality in the US, while we see over 75% individuals in the overall population believe helping the poor helps to reduce the economic inequality a great deal or a fair amount. They also believe that allowing people to easily obtain guns legally, does not contribute to gun violence in the country. In contrast, over 46% of the overall population believes that it does contribute to gun violence. Finally, when queried on how well the democratic party represented the interests of its members, 88.05% of individuals in cluster-0 responded with *Not at all well*. In contrast, only 18.21% of the overall population choose this option. Given that cluster-0 is mainly composed of Republicans, it is not surprising that the group has a clear mistrust of Democrats.

Questions that differentiate clusters: Finally, we show questions that elicit the most varied responses across the different clusters we have identified. More formally, we compute the average total variation between cluster pairs for each question as below:

$$TV_{Ave}(q) = \frac{1}{\binom{6}{2}} \sum_{0 \le i < j \le 5} TV(D_{i,q}, D_{j,q}) \quad (3)$$

Note that a larger value of $TV_{Ave}(q)$ indicates a greater degree of disagreement among clusters for that particular question. In Table 2, we present questions related to immigration topic that exhibit the greatest average total variation across clusters. The responses to these questions show interesting characteristics of each cluster. When considering the Immigration related topics, cluster-0, cluster-1 and cluster-5 have polarized attitudes, while the attitudes from other clusters are milder. In particular, Cluster-0 believes that a low priority should be assigned to allowing more legal immigrants, while placing importance on reducing the number of illegal immigrants. Cluster-1 and Cluster-5 believe that allowing more legal immigrants should be a top priority and that it is crucial to reduce illegal immigration. Cluster-2 and Cluster-3 assert that allowing more legal immigrants is not a top priority, and that reducing illegal immigration does not

significantly affect economic inequality. Cluster-4 believes that allowing more legal immigrants is not a top priority. However, it also believes that reducing illegal immigrants could reduce economic inequality. In addition to questions related to immigration, the clusters also show disagreement in other questions belonging to topics such as the Crime (Table 17) and Race (Table 18).

4.3 Steering LLMs towards personas

In this section, we benchmark the steerability of various LLMs toward personas defined in Sections 3.1. We follow the procedure described in Section 3.2 and use prediction accuracy for evaluating performance. Prediction accuracy, in this case, is defined as the macro average of individual prediction accuracy (i.e., the percentage of questions where the LLM correctly predicts a particular individual's responses, averaged over all individuals).

We compare the performance of our algorithm against the following baselines:

1. **Raw Q.** (Santurkar et al., 2023): LLMs are only prompted with questions.
2. **Demographics + Raw Q.** (Santurkar et al., 2023; Hwang et al., 2023): the demographic traits of each individual are provided as context information.
3. **Context + Raw Q.** (Hwang et al., 2023): a set of responses by the individual are provided as context information.

We randomly split the individuals in the OpinionQA dataset into train ($\mathcal{P}_{tr}$) and evaluation ($\mathcal{P}_{val}$) partitions (represented in blue and green respectively, in Figure 1). Accordingly, the responses in each partition can be referred to as $\mathcal{R}_{tr} = \{(i,j), \text{where } r_{i,j} \text{ is not null}, i \in \mathcal{P}_{tr}, j \in \mathcal{Q}\}$ and $\mathcal{R}_{val} = \{(i,j), \text{where } r_{i,j} \text{ is not null}, i \in \mathcal{P}_{val}, j \in \mathcal{Q}\}$. Next, we randomly split the response sets in each partition above, into their own train and validation sets, which leaves us with the four sets that can be denoted as $\mathcal{R}^{tr}_{tr}$, $\mathcal{R}^{val}_{tr}$, $\mathcal{R}^{tr}_{val}$ and $\mathcal{R}^{val}_{val}$ (represented in light-blue, dark-blue, light-green and dark-green respectively, in Figure 1).

For our method, we first use the responses in $\mathcal{R}^{tr}_{tr}$ to optimize Eq. (1) to get $U^*_{tr} = \{u_i \in \mathbb{R}^d, i \in \mathcal{P}_{tr}\}$ and $Q^* = \{q_j \in \mathbb{R}^d, j \in \mathcal{Q}\}$. Next, we use $\{u_i \in \mathbb{R}^d, i \in \mathcal{P}_{tr}\}$ and $\mathcal{R}^{tr}_{tr}$ to train the SPM (Eq. (2)). We then report the average prediction accuracy of our steered LLMs over $\mathcal{R}^{val}_{tr}$ in Table 3. For the **Context + Raw Q.** baseline, we use $\mathcal{R}^{tr}_{tr}$ to provide context questions. In particular, for a given question, we identify the K most closely related questions within the training set, that the individual has responded to, to serve as the context. The similarity of these questions is measured using cosine distance bewteen the embeddings of the question (we employ the `text-embedding-ada-002` (OpenAI, 2023) model created by OpenAI to obtain the embeddings). The **Raw Q.** and **Demographics + Raw Q.** baselines do not use the train split. As in the case of our steered LLMs, we report the average prediction accuracy over $\mathcal{R}^{val}_{tr}$ for all baselines.

4.3.1 Individual opinion prediction

In Table 3, we present the average prediction accuracy of our method compared against the baselines. Note that for the **Demographics + Raw Q.** baseline, we include 13 different demographic traits in the prompt, and for the **Context + Raw Q.** baseline, we set $K = 5$ (an ablation study of different K values is presented in Table 8). Our method outperforms all baselines significantly, with improvements of $57\% - 77\%$ over the best performing baseline for each model. We observe that when LLMs are provided with only the questions, their responses have low alignment with the individuals' responses. Including demographics traits in addition to responses to related questions tends to improve alignment. However, the LLMs still lack information about the individual which causes lower prediction accuracy. In contrast, our method, which steers the LLM utilizing the embeddings learned via CF, that encode knowledge about the overall opinions of a given individual, show higher prediction accuracy.

4.3.2 Effectiveness of the SPM

We verify the effectiveness of training the SPM by comparing the prediction accuracy with and without SPM training. As shown in Table 4, the prediction accuracy clearly increases when the SPM is trained. This justifies the use of the SPM to map individual embeddings to the LLM embedding space.

4.3.3 Performance of cluster personas

We explore how well cluster persona embeddings can predict an individual's opinion. For this, we replace an individual's embedding with their corresponding cluster embedding (**CE**). For an individual belonging to cluster i, their corresponding

Table 3: Prediction accuracy across baselines and our experimental method.

Model	Raw Q.	Demographic + Raw Q.	Context + Raw Q.	Individual Embeddings (Ours)
GPT-Neo-1.3B	32.54%	33.40%	33.82%	**59.99%**
GPT-Neo-2.7B	31.09%	34.32%	30.43%	**60.59%**
GPT-j-6B	26.50%	31.86%	39.34%	**61.84%**
Falcon-7B-Instruct	36.10%	38.40%	37.96%	**60.78%**

Table 4: Prediction accuracy with and without training the SPM.

Model	Random SPM Weights	Trained SPM Weights
GPT-Neo-1.3B	7.92%	**59.99%**
GPT-Neo-2.7B	8.31%	**60.59%**
GPT-j-6B	23.26%	**61.84%**
Falcon-7B-Instruct	38.20%	**60.78%**

cluster embedding is the centroid of the embeddings of *all* individuals belonging to the cluster i. The prediction accuracy is shown in the last two columns of Table 5. We see **CE**s get close performance as individual embeddings (around $8\%-12\%$ difference). Furthermore, we also compare **CE**s with the demographic group embedding (**DE**) in Table 5. **DE** is the centroid of individual embeddings belonging to the same demographic group. For instance, for an individual who belongs to the Democrat party, we use a demographic embedding that is the average of the embeddings of all the individuals who belong to the Democrat party. We see that the use of **CE**s provide improvements of between $0.23\% - 2.59\%$ over **DE**s. Note that for **DE**s, we consider the political party trait that consists of six parties and shows the highest prediction accuracy, compared to other demographic traits (shown separately in Table 9 in the Appendix). For **CE**s, we present results using six clusters in Table 5, which enables us to obtain better prediction accuracy than any demographic trait. As shown in Table 11 in the Appendix, increasing the number of clusters can yield much higher prediction accuracies.

4.3.4 Generalization to unseen individuals

We also run an ablation test to see if the steered LLM can generalize to **unseen individuals**. Here, we use responses from the set $\mathcal{R}^{tr}_{val}$ to get embeddings U^*_{val} for individuals in $\mathcal{P}_{val}$. $\mathcal{R}^{tr}_{val}$ contains a small number of responses for each individual and the question embeddings are fixed as Q^* (learned based on $\mathcal{P}_{tr}$ as defined in Section 4.3). Note that the SPM, trained on $\mathcal{P}_{tr}$, has not seen any responses from $\mathcal{P}_{val}$. We explore the use of different numbers of responses (K) from $\mathcal{R}^{tr}_{val}$ to generate the embeddings for the unseen individuals. In Table 6, we report the prediction accuracies for our models over $\mathcal{R}^{val}_{val}$ for these experiments and for our baselines. We see, even in the case of using a single response per individual, our method generalizes better to unseen individuals compared to the baselines. Furthermore, the prediction accuracies increase as we use more responses (K) to get more accurate embeddings. See Appendix A.5 for further results.

5 Conclusion

In this work, we present an approach for steerable generation using LLMs, where we utilize a data-driven notion of a persona to modulate the generation process. We proposed an simple model and efficient algorithm to align LLMs with these personas (both individuals and clusters). We validate the efficacy of our algorithm using the OpinionQA dataset. For a select set of LLMs we show that our method out-performs traditional steering mechanisms supporting our hypothesis that LLMs align with individuals' opinions better when leveraging our data-driven personas.

Limitations

There are some limitations to the work we present here. Firstly, we rely on the QA format to perform collaborative filtering and embed individuals into an embedding space. Secondly, we only test our approach over one dataset due to time and resource constraints. However, we note that our method should scale well to other similar datasets. Thirdly, in this work we only test prefx-tuning (Lester et al., 2021) and prompt-tuning (Lester et al., 2021) for steering LLMs. Other parameter efficient fine tuning methods such as LoRA (Hu et al., 2021), $(IA)^3$ (Liu et al., 2022) *etc.* can also be incorporated into our approach.

Table 5: Prediction accuracy with different types of personas. For Demographic Embedding results we present only the political party trait (PARTY) which consists of 6 groups (see other demographic traits in Table 9).

Model	Demographic Embeddings [PARTY (6)]	Cluster Embeddings [6 Clusters]	Individual Embeddings (Ours)
GPT-Neo-1.3B	55.16%	55.57%	**59.99**%
GPT-Neo-2.7B	55.66%	55.79%	**60.59**%
GPT-j-6B	55.64%	55.82%	**61.84**%
Falcon-7B-Instruct	52.87%	54.24%	**60.78**%

Table 6: Prediction accuracy computed on unseen individuals for baselines and our method (K is the number of responses we use to generate embeddings for unseen individuals).

Model	Raw Q.	Demographic + Raw Q.	Context + Raw Q.	K = 1	K = 5	K = 10	K = 20	K = 50	K = 100
GPT-Neo-1.3B	32.22%	33.57%	33.97%	40.02%	46.83%	49.97%	53.27%	56.18%	57.74%
GPT-Neo-2.7B	31.57%	33.94%	30.54%	40.40%	46.58%	48.75%	50.90%	52.66%	53.29%
GPT-j-6B	26.58%	32.10%	39.32%	41.41%	49.29%	52.19%	55.08%	57.73%	58.75%
Falcon-7B-Instruct	36.15%	38.22%	38.12%	39.11%	48.12%	51.16%	53.87%	56.76%	57.83%

Ethical Considerations

We fine tune the LLMs over a dataset encoding the opinions of individuals and we acknowledge that the LLMs could reflect and even intensify the biases held by certain individuals in the training set. A careful audit to the training data is necessary before training in practice.

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

OpenAI. 2023. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model. Accessed: [insert date of access].

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Mustafa Safa Ozdayi, Charith S. Peris, Jack G. M. FitzGerald, Christophe Dupuy, Jimit Majmudar, Haidar Khan, Rahil Parikh, and Rahul Gupta. 2023. Controlling the extraction of memorized data from large language models via prompt-tuning. In *Annual Meeting of the Association for Computational Linguistics*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *arXiv preprint arXiv:2305.14930*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *arXiv preprint arXiv:2306.01943*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M Blei. 2023. Evaluating the moral beliefs encoded in llms. *arXiv preprint arXiv:2307.14324*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*.

Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Kehui Tan, Tianqi Pang, and Chenyou Fan. 2023. Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects. *arXiv preprint arXiv:2305.03433*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Additional details on the Dataset and the Experiments

A.1 Additional information on the OpinionQA dataset

We consider the coarse topics of OpinionQA which include: healthcare system, future, relationships and family, leadership, status in life, political issues, personal health, race, personal finance, crime or security, corporations, banks, technology and automation, self-perception and values, science, education, religion, discrimination, community health, immigration, global attitudes and foreign policy, economy and inequality, news, social media, data, privacy, job/career, gender and sexuality. The list of demographic traits available for each participant are: Education, Citizen, Marital, Income, Political Ideology, Region, Political Party, Sex, Age, Religion Attendance, Race and Religion.

A.2 Details of optimizing Eq. (1)

To optimize Eq. (1), we perform mini-batch gradient descent with the Adam optimizer (learning rate 0.001) and a batch-size of 2048. Furthermore, both individuals and questions are represented by 16-dimensional vector. For unseen users, we fix the question embedding and only optimize the individual embedding for new users.

A.3 Architecture of the SPM

The SPM that we use is a two-layer multilayer perceptron (MLP) with 32 hidden units. The output of the SPM is a vector of shape $[T, L\times 2\times D]$ where T is the number of virtual tokens, L is the number of layers and D is the token dimension. We follow the setting of prefix-tuning (Li and Liang, 2021) to insert virtual tokens before each transformer layer. Since the Falcon model uses multi-query attention (Shazeer, 2019), the output of the SPM becomes a vector of shape $[T, L\times 2\times D/H]$ where H is the number of heads. We set the number of virtual tokens to be 1 in our experiments. For training, we use the AdamW (Loshchilov and Hutter, 2017) optimizer with learning rate 0.001 and weight decay 0.001, and we train the SPM for 10 epochs with early stopping.

A.4 Prompt templates of baseline methods

In Table 3, we includes three types of baselines **Raw Q.**, **Demographics + Raw Q.** and **Context + Raw Q.**. For the **Raw Q.** baseline, only the question text is included in the prompt. For the **Demographics + Raw Q.** baseline, we prepend the demographics of an individual before the question text. An example prompt is shown on the left panel of Figure 4. For the **Context + Raw Q.** baseline, we prepend a set of responses provided by individual to other questions. An example prompt is shown on the right panel of Figure 4.

A.5 Additional experimental results

We provide further detail on our experimental results in this section.

In Figure 3, we show the demographic composition of clusters for other demographic traits not included in Figure 2.

In Table 9, we show the prediction accuracy when we use a single demographic-trait as the representation of an individual. We note that our method with the use of individual or cluster embeddings out-performs all demographic embeddings.

In Table 10, we present results obtained by using combinations of demographic groups to represent individuals. We note that individual or cluster embeddings out-performs the prediction accuracy of these combinations as well.

In Table 8 we show the prediction accuracy for the **Context + Raw Q.** baseline under different K values, *i.e.* the number of context samples.

As mentioned in Section 3.2, we draw from *prefix-tuning* tuning technique to implement our SPM. In order test the robustness of our method to other prompting methods, we also implement the SPM with the use of prompt-tuning (Lester et al., 2021). We show our results in Table 7.

Table 12-Table 16 show the top-3 questions that a cluster mostly disagrees with, when considering the overall population. Finally, Table 17 and Table 18 show the questions that have the largest disagreement among clusters for the Crime or Security and Immigration topics.

Table 7: Prediction accuracy across baselines and our experimental method showcasing the use of both *prefix-tuning* and *prompt-tuning* techniques.

Model	Raw Q.	Demographic + Raw Q.	Context + Raw Q.	Prefix-Tuning (Ours)	Prompt-Tuning (Ours)
GPT-Neo-1.3B	32.54%	33.40%	33.82%	**59.99%**	**59.65%**
GPT-Neo-2.7B	31.09%	34.32%	30.43%	**60.59%**	**58.96%**
GPT-j-6B	26.50%	31.86%	39.34%	**61.84%**	**59.41%**
Falcon-7B-Instruct	36.10%	38.40%	37.96%	**60.78%**	**57.98%**

Table 8: Prediction accuracy for the Context + Raw Q baseline using different numbers of context samples K.

Model	K=3	K=5	K=8	K=10
GPT-Neo-1.3B	33.05%	33.82%	31.43%	31.41%
GPT-Neo-2.7B	32.99%	30.43%	29.76%	29.89%
GPT-j-6B	37.88%	39.34%	38.32%	38.49%
Falcon-7B-Instruct	36.78%	37.96%	36.59%	36.98%

Table 9: Prediction accuracy across different types of demographic traits (Single). We provide the number of groups in each demographic trait within parenthesis

Model	EDUCATION (7)	INCOME (7)	RACE (7)	RELIGION (16)	PARTY (6)	Cluster Embeddings	Individual Embeddings
GPT-Neo-1.3B	53.24%	53.09%	53.02%	53.85%	55.16%	55.57%	59.99%
GPT-Neo-2.7B	53.69%	53.43%	53.35%	54.31%	55.66%	55.79%	60.59%
GPT-j-6B	53.73%	53.25%	53.25%	54.32%	55.64%	55.82%	61.84%
Falcon-7B-Instruct	50.87%	51.65%	51.14%	51.78%	52.54%	54.24%	60.78%

Table 10: Prediction accuracy for different combinations of demographic traits.

Model Model	EDUCATION X PARTY (42)	INCOME X PARTY (42)	RELIGION X PARTY (96)	Cluster Embeddings	Individual Embeddings
GPT-Neo-1.3B	55.59%	55.78%	55.88%	55.57%	59.99%
GPT-Neo-2.7B	56.03%	56.06%	56.35%	55.79%	60.59%
GPT-j-6B	56.16%	56.14%	56.47%	55.82%	61.84%
Falcon-7B-Instruct	54.96%	54.78%	55.34%	54.24%	60.78%

Table 11: Prediction accuracy of cluster personas across different choices for number of clusters.

Model	6 Clusters	10 Clusters	20 Clusters	30 Clusters	50 Clusters	Individual Embeddings
GPT-Neo-1.3B	55.57%	56.65%	58.11%	58.41%	58.74%	59.99%
GPT-Neo-2.7B	55.79%	56.93%	58.61%	58.93%	59.42%	60.59%
GPT-j-6B	55.82%	57.13%	58.97%	59.52%	59.86%	61.84%
Falcon-7B-Instruct	54.24%	56.72%	58.99%	59.18%	59.39%	60.78%

Figure 3: The demographic composition of Clusters-0 to 5 and the Overall Population. From left to right and top to bottom, we show demographic composition for Ideology, Region, Age, Citizenship, Marital status, Religion, Sex, Religion attendance and Income.

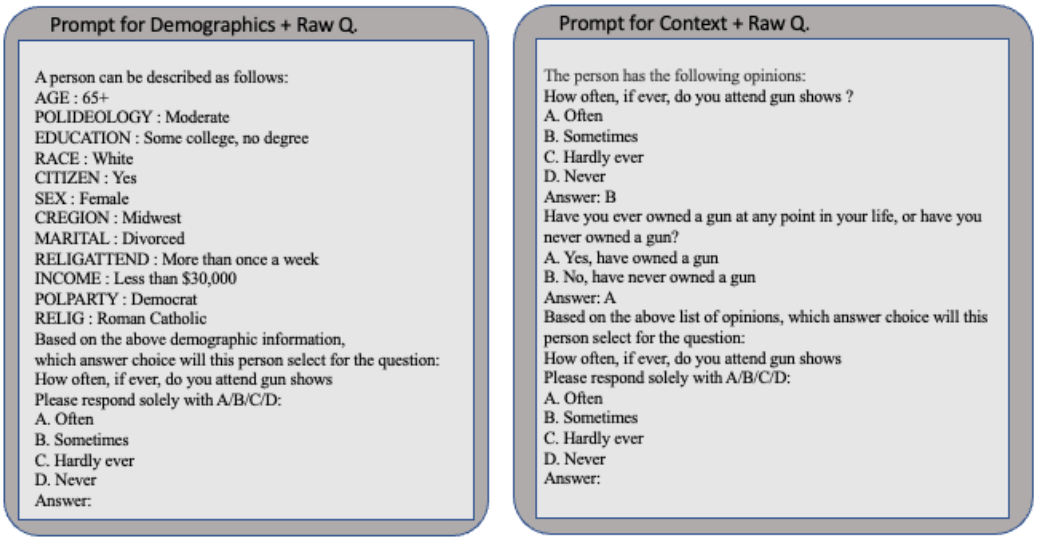

Figure 4: Example prompts of baseline methods. The left panel shows an example prompt used for the **Demographics + Raw Q.** baseline while the right panel shows an example prompt used for the **Context + Raw Q.** baseline.

Table 12: Comparison of responses between Cluster 1 and the Overall Population.

Question	Cluster 1 Response	Overall Population Response
How much control, if any, do you think you have over who can access the search terms you use online?	**A lot of control: 99.47%** Offline: does not have internet or A little control: 0.52% No control: 0.0%	A lot of control: 25.3% **Offline: does not have internet or A little control: 60.88%** No control: 13.8%
For each, please indicate if you, personally, think it is acceptable. A white person using the n-word	**Always acceptable: 87.38%** Sometimes acceptable: 12.52% Rarely acceptable: 0.09% Never acceptable: 0.0%	Always acceptable: 20.09% Sometimes acceptable: 2.95% Rarely acceptable: 25.03% **Never acceptable: 51.91%**
Overall, how does being Jewish affect people's ability to get ahead in our country these days?	**Helps a lot: 99.59%** Helps a little: 0.4% Neither helps nor hurts: 0.0% Hurts a little: 0.0% Hurts a lot: 0.0%	Helps a lot: 22.9% Helps a little: 21.6% **Neither helps nor hurts: 52.89%** Hurts a little: 2.59% Hurts a lot: 0.0%

Table 13: Comparison of responses between Cluster 2 and the Overall Population.

Question	Cluster 2 Response	Overall Population Response
How confident are you, if at all, that the actions taken by the international community will significantly reduce the effects of global climate change?	Very confident: 1.18% **Somewhat confident: 91.61%** Not too confident: 7.19% Not at all confident: 0.0%	Very confident: 25.32% **Somewhat confident: 41.1%** Not too confident: 31.73% Not at all confident: 1.83%
In general, how often, if ever, would you say you have parties or get-togethers with any of your neighbors?	Every day: 1.65% Several times a week: 16.71% **About once a week: 49.86%** About once a month: 30.25% Less than once a month: 1.51% Never: 0.0%	Every day: 12.33% Several times a week: 11.1% About once a week: 17.87% About once a month: 17.94% **Less than once a month: 28.73%** Never: 12.0%
How enthusiastic are you, if at all, about the possibility of using computer programs to make hiring decisions for society as a whole?	Very enthusiastic: 6.73% **Somewhat enthusiastic: 81.57%** Not too enthusiastic: 11.69% Not at all enthusiastic: 0.0%	Very enthusiastic: 27.55% **Somewhat enthusiastic: 33.07%** Not too enthusiastic: 38.58% Not at all enthusiastic: 0.78%

Table 14: Comparison of Responses between Cluster 3 and the Overall Population.

Question	Cluster 3 Response	Overall Population Response
How often, if ever, do you attend gun shows?	**Never: 90.7%** Hardly ever: 9.29% Sometimes: 0.0% Often: 0.0%	Never: 20.6% **Hardly ever: 39.74%** Sometimes: 20.36% Often: 19.29%
How important, if at all, is being a gun owner to your overall identity?	Not at all important: 5.14% **Not too important: 89.26%** Somewhat important: 5.58% Very important: 0.0%	Not at all important: 0.57% Not too important: 24.21% **Somewhat important: 42.63%** Very important: 32.57%
Thinking about the country today, would you say there are?	**Too few women in top executive business positions: 88.1%** About the right number of women in top executive business positions: 11.89% Too many women in top executive business positions: 0.0%	Too few women in top executive business positions: 18.73% **About the right number of women in top executive business positions: 41.96%** Too many women in top executive business positions: 39.3%

Table 15: Comparison of Responses between Cluster 4 and the Overall Population.

Question	Cluster 4 Response	Overall Population Response
Thinking about medical doctors, how often would you say they do a good job providing diagnoses and treatment recommendations?	All or most of the time: 20.0% **Some of the time: 77.84%** Only a little of the time: 2.15% None of the time: 0.0%	**All or most of the time: 70.3%** Some of the time: 29.61% Only a little of the time: 0.07% None of the time: 0.0%
If robots and computers do much of the work currently done by humans, do you think this would be	A very good thing for the country: 0.0% A somewhat good thing for the country: 15.88% **A somewhat bad thing for the country : 76.66%** A very bad thing for the country: 7.45%	A very good thing for the country: 23.29% A somewhat good thing for the country: 28.35% **A somewhat bad thing for the country : 47.21%** A very bad thing for the country: 1.13%
In the future, what kind of an impact do you think science and technology will have in solving the biggest problems facing the country?	A very positive impact: 15.49% **A somewhat positive impact: 81.37%** A somewhat negative impact: 3.13% A very negative impact: 0.0%	**A very positive impact: 63.59%** A somewhat positive impact: 36.1% A somewhat negative impact: 0.29% A very negative impact: 0.0%

Table 16: Comparison of Responses between Cluster 5 and the Overall Population.

Question	Cluster 5 Response	Overall Population Response
How much pressure, if any, did you feel from family members to marry your partner after you moved in together?	A lot of pressure: 2.13% **Some pressure: 87.26%** Not too much pressure: 10.59% No pressure at all: 0.0%	A lot of pressure: 21.97% Some pressure: 10.68% **Not too much pressure: 62.09%** No pressure at all: 5.24%
How much control do you think you have over the data the government collects about you?	A great deal of control: 2.87% **Some control: 87.09%** Very little control: 10.02% No control: 0.0%	A great deal of control: 21.91% Some control: 11.47% **Very little control: 60.06%** No control: 6.53%
When you are asked to agree to a company's privacy policy, how often do you read it before agreeing to it?	Always: 4.19% **Often: 87.42%** Sometimes: 8.38% Never: 0.0%	Always: 22.24% Often: 12.63% **Sometimes: 62.87%** Never: 2.24%

Table 17: Questions pertaining to crime that exhibit the most disagreement among clusters.

Question	Cluster	Response
How often, if ever, do you participate in online discussion forums about guns	Cluster-0	Hardly ever (72.24%)
	Cluster-1	Often (77.92%)
	Cluster-2	Hardly ever (77.21%)
	Cluster-3	Never (97.01%)
	Cluster-4	Hardly ever (72.74%)
	Cluster-5	Sometimes (84.22%)
How important, if at all, is being a gun owner to your overall identity?	Cluster-0	Somewhat important (84.41%)
	Cluster-1	Very important (95.35%)
	Cluster-2	Somewhat important (72.45%)
	Cluster-3	Not too important (89.26%)
	Cluster-4	Somewhat important (76.27%)
	Cluster-5	Very important (82.0%)

Table 18: Questions pertaining to race that exhibit the most disagreement among clusters.

Question	Cluster	Response
Are the country's current economic conditions helping or hurting people who are Hispanic?	Cluster-0	Helping a little (70.42%)
	Cluster-1	Helping a lot (94.54%)
	Cluster-2	Hurting a little (59.44%)
	Cluster-3	Hurting a little (65.74%)
	Cluster-4	Hurting a little (42.94%)
	Cluster-5	Helping a little (52.25%)
By 2050, a majority of the population will be made up of blacks, Asians, Hispanics, and other racial minorities. In terms of its impact on the country, do you think this will be	Cluster-0	A somewhat bad thing (52.33%)
	Cluster-1	A very good thing (97.6%)
	Cluster-2	A somewhat good thing (66.44%)
	Cluster-3	A somewhat good thing (75.7%)
	Cluster-4	Neither a good nor bad thing (48.82%)
	Cluster-5	A very good thing (55.46%)