

Are Large Language Models (LLMs) Good Social Predictors?

Anonymous ACL submission

Abstract

001 With the recent advancement of Large Lan- 047
002 guage Models (LLMs), efforts have been made 048
003 to leverage LLMs in crucial social science 049
004 study methods, including predicting human fea-
005 tures of social life such as presidential voting.
006 Existing works suggest that LLMs are capable
007 of generating human-like responses. Neverthe-
008 less, it is unclear how well LLMs work and
009 where the plausible predictions derive from.
010 This paper critically examines the performance
011 of LLMs as social predictors, pointing out the
012 source of correct predictions and limitations.
013 Based on the notion of *mutability* that classifies
014 social features, we design three realistic set-
015 tings and a novel social prediction task, where
016 the LLMs make predictions with input features
017 of the same mutability and accessibility with
018 the response feature. We find that the promis-
019 ing performance achieved by previous studies
020 is because of input shortcut features to the re-
021 sponse, which are hard to capture in reality;
022 the performance degrades dramatically to near-
023 random after removing the shortcuts. With the
024 comprehensive investigations on various LLMs,
025 we reveal that LLMs struggle to work as ex-
026 pected on social prediction when given ordinari-
027 ly available input features without shortcuts.
028 We further investigate possible reasons for this
029 phenomenon and suggest potential ways to en-
030 hance LLMs for social prediction.

031 1 Introduction

032 Social Prediction is one of the crucial elements in
033 social studies (Hofman et al., 2017), with a body of
034 literature (Liben-Nowell and Kleinberg, 2003; Bak-
035 shy et al., 2011; Cheng et al., 2014) devoted to esti-
036 mating inaccessible features, either unobserved or
037 missing, based on observed ones. Historically, so-
038 cial prediction is made by statistical models such as
039 linear regression (Uyanık and Güler, 2013). With
040 the development of machine learning, supervised
041 methods have been adopted, e.g. random forest
042 and neural networks (Chen et al., 2021b). How-
043 ever, the classic machine learning methods notably
044 rely on extensive labeled training data, which is
045 labor-intensive, especially in social studies. Addi-
046 tionally, the predictive power of machine learning

methods is limited (Mackenzie, 2015; Athey, 2018)
and can hardly model the complicated phenomenon
in social life.

050 With the rapid advancement in Large Language
051 Models (LLMs), undertaking text-related tasks is
052 empowered with a new paradigm (Zhuang et al.,
053 2023; Tan et al., 2023; Nijkamp et al., 2022; Chen
054 et al., 2021a; Zhou et al., 2022; Wei et al., 2022).
055 The extensive world knowledge (Zhao et al., 2023)
056 and inference abilities (Creswell et al., 2022) en-
057 able LLMs to mitigate the limitations of classic ma-
058 chine learning methods in social prediction. Recent
059 works leverage LLMs in predicting or simulating
060 human responses, such as voting decisions (Argyle
061 et al., 2022; von der Heyde et al., 2023) and polit-
062 ical attitude (Rosenbusch et al., 2023). They take
063 advantage of LLMs to augment existing datasets
064 with previously inaccessible features due to unob-
065 servability, data missing, sensitivity and privacy
066 issues. Promising performance is reported. How-
067 ever, the common methodology of these works is
068 worth being skeptical about: it first creates datasets
069 with a well-constructed survey; next, except for
070 the response feature, any other features are candi-
071 dates as input, even if they and the response feature
072 are almost semantically equivalent and thus nearly
073 (in)accessible.

074 This methodology introduces a question: *If the*
075 *observed features are nearly equivalent and thus*
076 *nearly accessible, why did the original survey avoid*
077 *directly collecting the key response feature, yet*
078 *bother to predict by other features?* This issue
079 hinders the exploration of LLMs’ authentic social
080 prediction ability and the underlying mechanisms,
081 as well as the realistic and practical implementa-
082 tion of proposed methods. Our paper responds to
083 it, critically checking and revising social prediction
084 in a group of settings considering the accessibility
085 of features.

086 To study it, our preliminary investigation revis-
087 its the famous case of voting prediction (Argyle
088 et al., 2022) with LLMs. We define **shortcut** as the
089 observed features approximately equivalent and
090 (in)accessible with the response feature, which
091 should be masked in input. The result indicates that

LLMs' performance is bolstered by the presence of shortcuts to the desired response features. Specifically, the presence of shortcuts directly associated with the feature to be predicted leads to exceptional performance, even replacing LLMs with machine learning models. Unfortunately, this effectiveness comes with a decline when eliminating shortcuts (detailed in Sec.2 and Sec.3). This performance gap leads us to question the true capability of LLMs in social predictions, challenging the prevailing perception of their prowess (Argyle et al., 2023). The research community is urged to be more cautious and skeptical when employing this method. Furthermore, we shed light on the potential causes and solutions to the under-performance, hinting at future works to comply with the realistic settings when delving into social prediction studies.

Our contributions are listed below:

- We introduce a novel social prediction task Soc-PRF Prediction (stands for Social Profile Prediction). Informed by theoretical social studies (Bailey, 1998), we categorize social features into two groups, and the degrees of feature accessibility comply with the principle of "*intra-group homogeneity, inter-group heterogeneity*". Prediction across the groups avoids shortcuts and delves deeper into LLMs' abilities.
- We conduct comprehensive experiments of social prediction with various LLMs, including closed-sourced models GPT 3.5 (OpenAI, 2022), GPT 4 (Achiam et al., 2023), and Gemini Pro (Anil et al., 2023), as well as lighter open-sourced models like Llama-7B, Llama-7B-chat (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023). The results reveal the incapability of LLMs in rigorous yet realistic settings.
- Our studies suggest that LLMs are reluctant to work on social prediction with ordinary input features without shortcuts. We further explore the potential reasons and future directions to enhance LLMs for social prediction.

2 Revisit Voting Prediction with LLMs

Large Language Models (LLMs) have demonstrated impressive performance in predicting voting decisions in the United States (Argyle et al., 2022;

Veselovsky et al., 2023). In this section, we revisit this voting prediction study with LLMs (Argyle et al., 2022) and take a further step beyond it.

2.1 Reflecting on Voting Prediction

The work of (Argyle et al., 2022) adopts the American National Election Studies (ANES) to construct the dataset. ANES is a survey conducted in every presidential election year, with features about American public views and political decisions. To elicit LLMs' prediction of the response feature (aka individual voting decision), this study selects 10 input features: racial/ethnic self-identification, gender, age, ideological self-identification, party identification, political interest, church attendance, if discussing politics with family/friends, patriotism feelings, state of residence. With these 10 input features and a question to elicit predictions, they build the prompts with an example as below:

Racially, I am white. I am male. Ideologically, I describe myself as conservative. Politically, I am a strong Republican ...
In 2016, I voted for: _____.

However, intuitively two of the input features are near-equivalent to voting decision, i.e. ideological self-placement and party identification. It is evident from political science studies (Miller, 1991; Dalton, 2016) that given the partisan nature of American politics, voting decision are closely related to these two features; besides, they share similar degrees of difficulty to capture due to privacy and costs. To validate this assertion, we first calculate their Cramer's V^1 with vote decision. The Cramer's V scores are 0.86 and 0.76 respectively, indicating these two features are highly correlated with vote decision.

Worse still, these features are rarely found and nearly (in)accessible with the response features. Referring to a survey on political social data mining (dos Santos et al., 2021), only **1.89%** of the studies conducted have access to election-related input features. Consequently, including features closely related to the election as inputs is also impractical.

¹Cramer's V is a measurement of association between features; the score 0 indicates no association and 1 indicates a perfect association.

We term features in this context as **shortcuts**, which are nearly semantic-equivalent and nearly accessible with the response feature, and thus should not be used as input features.

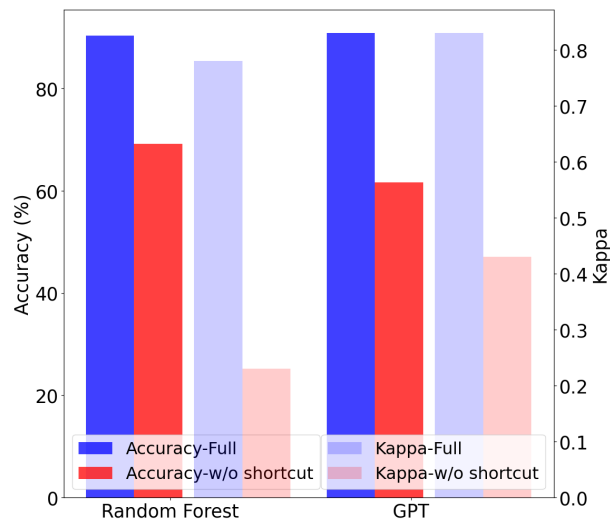


Figure 1: Performance of voting prediction. For each model: ■ = **Full** + Accuracy; ■ = **w/o shortcut** + Accuracy; ■ = **Full** + Kappa; ■ = **w/o shortcut** + Kappa.

2.2 Further Experiments of Voting Prediction

Next, we conduct further experiments on the impacts of shortcuts on social prediction. We choose both GPT-based approaches and classic supervised machine learning models. For GPT-based approaches, the backbone GPT 3.5 model and prompts are the same with (Argyle et al., 2022); for classic supervised machine learning, we choose the Random Forest Classifier². There are two settings on features: (1) **Full**, taking all the 10 input features; (2) **w/o shortcut**, taking input features except the 2 shortcuts. Given the balanced distribution of the voting decision (51.9% vs. 48.1%), the metric to use is Accuracy; in addition, Cohen’s **Kappa**³ κ is adopted to evaluate the agreement between the predicted and true voting decision.

As shown in Fig. 1, the GPT-based approach with all input features achieved the accuracy of 90.82% and Cohen’s Kappa κ of 0.83, successfully reproducing the results of (Argyle et al., 2022).

²Since supervised classifiers need labeled data to train, we split the dataset into 80%/10%/10% as training, validation, and test sets. The supervised setting offers models more information and eases the tasks.

³Cohen’s Kappa κ has values ranging from 0 to 1, where 1 indicates stronger agreement and 0 indicates almost no agreement.

However, after removing two shortcuts, the performance of both methods drops dramatically: the performance of GPT 3.5 drops to the accuracy of 61.60% and κ of 0.43; similarly, Random Forest drops from 90.29%, 0.78 to 69.22%, 0.23. On the other hand, in the **Full** setting, even the simple Random Forest achieves results as almost good as GPT, and also outperforms GPT in **w/o shortcut** setting. Given the nearly half-half distribution of voting decision, the performance without shortcut features is considerably unsatisfactory.

Our preliminary study suggests LLMs’ promising social prediction performance reported by prior works (Argyle et al., 2022) possibly derives from the existence of shortcut features. This finding motivates us to question if LLMs are *really* powerful in social prediction, or if the startling results are merely because of the shortcut features. To explore it, we propose a set of tasks that avoid shortcut features as inputs and resemble realistic scenarios.

3 Social Profile Prediction

In this section, we introduce a social prediction task evaluating LLMs’ predictive power without shortcuts. First, we coin a social prediction dataset based on survey data and methods to eliminate shortcut features. Then we introduce three settings to simulate real-world scenarios. Finally, we demonstrate and discuss the performance of LLMs’ prediction in new settings.

3.1 Task and Dataset

As illustrated in Sec. 2, the inclusion of shortcut features can affect the evaluation of the authentic social prediction power of LLMs. To address it, we design Soc-PRF Prediction as shown below.

The dataset derives from Gallup World Poll (Gallup, 2009), one of the most prestigious social surveys that guarantees reliability and diversity of features⁴. In this paper, we construct our dataset on its data from the USA and primarily between 2016 and 2020. To ensure information completeness, sample individuals with missing demographic features are removed. After careful data cleaning, the dataset includes 4,941 profiles of American individuals (samples). From feature views, we pick a

⁴Initialized in 2006, Gallup World Poll is conducted in over 150 countries and follows strict random sampling. Questions are designed by political scientists, measuring key indicators of social life, such as law, finance, civic engagement, etc., along with individual demographic data.

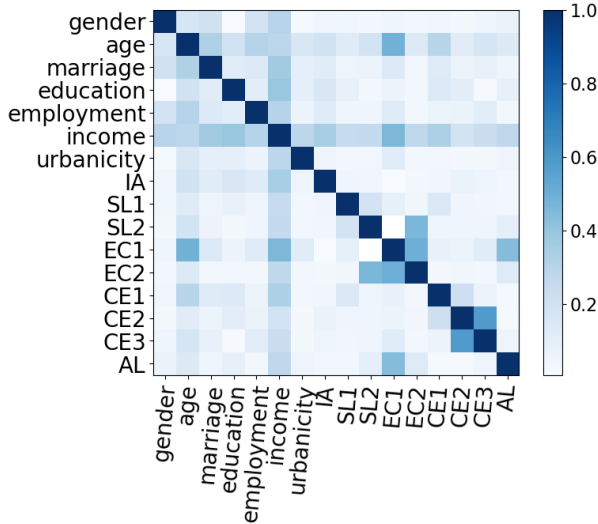


Figure 2: Correlation between features. The metric is Cramer’s $V \uparrow$. The labels IA, SL, EC, CE, AL stand for features of Internet Access, Social Life, Economic Confidence, Civic Engagement, and Approval of Leadership.

subset of features to construct the profiles of individuals, encompassing 16 social features reflecting various socio-demographic characteristics. Details are shown in Sec. 3.2.

3.2 General Settings stemming from Realistic Scenarios

In social studies, social datasets mainly derive from two methodologies: either traditional surveys or online data collection (Couper, 2017; Diaz et al., 2016; Callegaro et al., 2014). Following the statement in Sec. 1, features are not always available in realistic scenarios. To simulate this situation, we first retrieved the works of social studies and selected the concept *mutability*⁵ to classify the social features bipartite. Most of the time features with high mutability (like viewpoint, ideology, social behavior, etc.) and low mutability (like age, gender, profession, etc.) can hardly be collected simultaneously. For example, online data collection, such as crawling posts from social networking platforms, has the advantage of collecting high-mutability features by analyzing real-time attitudes and opinions with natural language processing (NLP) tools (Alghamdi and Alfalqi, 2015; Vayansky and Kumar, 2020; Hussein, 2018; Yue

⁵Mutability measures the features’ propensity to change or be influenced by social context. For more details please refer to social studies as (Bailey, 1992; Brensinger and Eyal, 2021; Sen and Wasow, 2016; Halley, 2017)

et al., 2019). In contrast, low mutability features (e.g. demographic features) often remain inaccessible unless the users reveal them online. Survey data from in-person interviews is complementary, capturing low-mutability features precisely, while the capture of highly mutable features is constrained to limited topics/years/individuals and inevitably missing data.

The 16 selected social features are assigned to low-mutability and high-mutability groups respectively. The low-mutability features are socio-demographic features, including age, gender, marriage, education, employment, income, and urbanicity of residence. The high-mutability features are attitudes or behaviors of social life, with topics of Internet Access, Social Life, Economic Confidence, Civic Engagement, and Approval of Leadership. In save of space, we denote features of them as IA, SL1, SL2, EC1, EC2, CE1, CE2, CE3, AL. Please note mutability is continuous; features even in the same group could have different degrees of mutability. For example, employment status is more mutable than gender, while Civic Engagement is more mutable than Internet Access. The details of the features are shown in Appendix A.1.

According to features’ mutability, we design three settings to assess the social prediction capability of LLMs, which simulate real-world scenarios for social data: giving low-mutability features to predict high-mutability features; giving high-mutability features to predict high-mutability or low-mutability features. Following the prior works especially (Argyle et al., 2022), we employ the same zero-shot prompt template without taking in any labeled data.

3.3 Details of 3 Settings

low2high. This setting takes in low mutability features to predict high mutability features, resembling traditional survey datasets mentioned in Sec. 3.2. One example of the prompt is:

- I am a male in the USA. I am 42 years old. My current marital status is married. My highest completed level of education is middle level. My current employment status is employed. My Annual Household Income is \$12600. I am from a suburb of a large city.
- When I’m asked “Do you have access

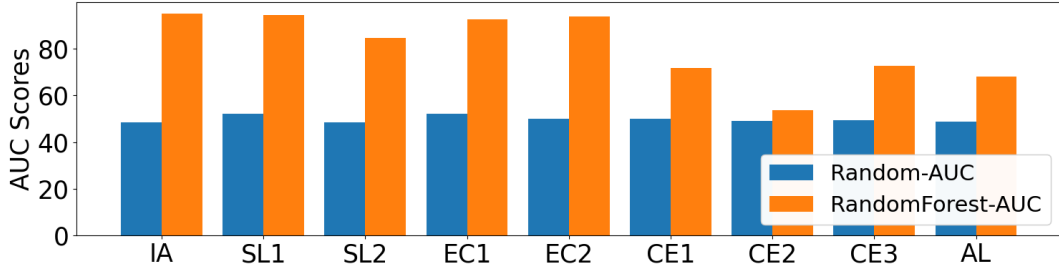


Figure 3: Performance of Random Forest and Random Guessing. The metric is AUC.

326 *to the Internet in any way, whether on a*
 327 *mobile phone, a computer, or some other*
 328 *device?"*, my answer is _____

329 Here the underlined text indicates the values of
 330 individual features, and italicized text presents the
 331 question to elicit final responses. For the subse-
 332 quent settings, we utilize prompts with similar tem-
 333 plates.

334 **high2low**: This setting denotes the prediction from
 335 high-mutability to low-mutability features. The in-
 336 puts include values of all 9 high mutability features,
 337 followed by the question about one low mutability
 338 feature. Serving as the inverse setting of low2high,
 339 this setting is designed for profile construction us-
 340 ing online data: with the in-time individual atti-
 341 tudes extracted from online posts, the demographic
 342 features are inaccessible.

343 **high2high**. High-mutability features are utilized
 344 as input to predict other high-mutability features.
 345 Different from high2low setting, to avoid short-
 346 cuts, the high mutability features of the same topic
 347 with the response feature are excluded from the
 348 input prompts. This setting simulates a specific
 349 real-world scenario, where individuals' attitudes to-
 350 ward one topic are collected, but opinions on other
 351 topics remain unexpressed.

352 **Evaluation Metrics**. Most features in the dataset
 353 have imbalanced distributions. For example, the
 354 feature IA has 91.82% samples with "yes" labels,
 355 while only 8.18% samples with "no". In this sit-
 356 uation, accuracy is not a proper metric (Gu et al.,
 357 2009). Thus we employ AUC as the metric.

358 3.4 Feature Analysis

359 Remind that our study is motivated by the shortcut
 360 features which are closely *related* to the response
 361 feature, and thus *inaccessible* in realistic scenar-
 362 ios. Mutability only guarantees the features are
 363 of different *accessibility*, but says little about re-

364 latedness between features. To prevent the emer-
 365 gence of shortcut features, we check the Cramer's
 366 V between all feature pairs. As shown in Fig. 2,
 367 most Cramer's V scores lie at low levels that are
 368 lower than 0.5. The only large values (such as
 369 0.58 between CE2 and CE3) come from the high
 370 mutability features with the same topics (i.e., Civic
 371 Engagement); however, all our settings do not in-
 372 clude this kind of cases.

373 Then we evaluate the predictive power of the se-
 374 lected features by a traditional supervised method.
 375 Take the low2high setting as the example, we
 376 train Random Forest Classifier with dataset split
 377 by 80%/10%/10% as training/validation/test sets.
 378 Then we compare its results with random guessing
 379 as the baseline. Fig. 3 shows Random Forest Clas-
 380 sifier outperforms the random guessing baseline
 381 by a considerable margin. For example, the AUC
 382 score of IA is 95.07, compared to 48.34 of random
 383 guessing.

384 In conclusion, all selected features are not short-
 385 cuts, and they are still powerful enough in predic-
 386 tion tasks.

387 3.5 LLMs as the Predictor

388 In this section, we leverage LLMs for the Soc-PRF
 389 Prediction task in the three aforementioned set-
 390 tings. The results of the three settings are illustrat-
 391 ed in Table 1, Table 2, and Fig. 4, respectively. In the
 392 tables, "Random" indicates the random guessing
 393 baseline. Note that for the settings high2high,
 394 we only show part of the results because the ob-
 395 servations are similar. We note that the perfor-
 396 mance of LLMs is closely similar to the random
 397 guessing and is far from satisfactory. The poor
 398 results appear consistently in all the settings and
 399 with all the LLMs. These observations indicate that
 400 LLMs struggle to predict individual features with
 401 the given information in the proposed settings.

Table 1: Performance of LLMs of setting low2high. The IA, SL, EC, CE, AL stand for indexes of Internet Access, Social Life, Economic Confidence, Civic Engagement, and Approval of Leadership.

Model	IA	SL1	SL2	EC1	EC2	CE1	CE2	CE3	AL
Random	48.34	52.09	48.47	52.12	50.07	49.89	49.16	49.32	48.60
Llama-7B	50.00	50.00	50.00	48.75	55.41	50.00	50.00	50.00	50.00
Llama-7B-chat	50.00	50.00	50.00	50.95	51.80	50.00	50.00	50.00	50.00
Mistral-7B	50.00	50.00	50.00	53.12	56.89	50.00	50.00	50.00	50.00
Gemini Pro	50.00	50.00	50.00	50.76	60.93	50.00	50.00	50.00	50.00
GPT-3.5	50.00	50.00	50.00	52.63	58.20	50.00	50.00	50.00	50.00
GPT-4	50.00	50.00	50.00	53.82	56.57	50.00	50.00	50.00	50.00

Table 2: Performance of LLMs (GPT 3.5 and GPT 4) of setting low2high.

Model	age	gender	marriage	education	employment	income	urbanicity
Random	49.50	49.62	49.45	49.99	50.54	48.14	50.22
Llama-7B	33.50	49.81	50.00	55.15	50.00	50.05	49.85
Llama-7B-chat	40.00	50.00	50.00	35.21	50.33	51.18	50.09
Mistral-7B	33.55	49.81	50.00	55.15	50.00	50.05	49.85
Gemini Pro	38.80	51.14	50.00	66.70	50.00	50.10	49.75
GPT-3.5	41.35	50.00	51.29	57.76	49.59	50.95	50.94
GPT-4	40.75	50.00	50.88	65.65	52.01	53.80	52.09

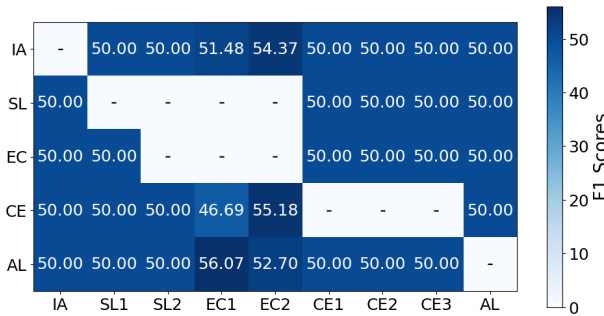


Figure 4: Performance of GPT 3.5 of setting high2high. The metric is AUC score. The sign "-" indicates no valid data, either because the input features (Y-axis) and output features (X-axis) share the same topic, or they are not conducted simultaneously in the survey.

4 Discussions

Some may wonder if the degraded results are caused by suboptimal or even trivial prompts: *are there other prompts that can make good predictions?* We admit there is a possibility, but this goes beyond the range of our paper. The prompts can be augmented by better-crafted prompts or examples of labeled data (so-called few-shot), but the settings will be incomparable with the prior works, and also converting the focus from LLMs' authentic predictive ability to advanced prompt engineering. Similar to the critical work of self-correction (Huang

et al., 2023), we are not devoted to addressing questions like *"what are the better social prediction prompt templates to induce better performance?"*. Rather, with the overwhelming evidence that several representative LLMs fail on social prediction with the popular and straightforward prompts, we wonder *"do LLMs really have social prediction ability at the individual level, without the help of other external resources?"* Below we propose deeper analysis and potentially helpful methods for social prediction.

4.1 Population v.s. Individual

As shown in the previous section, even advanced LLMs like GPT 4 encounter challenges in accurately predicting social features, only to yield outcomes similar to random guessing. To explore the underlying causes for such phenomena, we take the distribution comparisons between predicted response features and true counterparts. The case study is conducted in low2high setting and the results are shown in Fig. 5. We have the following observations:

(1) Although all the response features are high-mutable, LLM's predictions of relatively less mutable features (such as IA and SL, first 2 sub-figures of Fig. 5) are prone to have smaller discrepancies with true distributions. This fact indicates LLMs

do contain global knowledge about these social features, but they are only well-aligned at the population level. To validate this claim, we check the distributions of predictions in the true positive group and the whole dataset: among all the individuals, 88.60% are predicted as positive; however, among the true positive group (all the predictions are expected to be positive), the proportion of positive prediction is 90.52%. The gap is only 1.92%, which means the input features of individuals have few impacts on the prediction. We speculate that even conditioned by individual-level features, the population-level pre-trained knowledge prevails over that of individuals from prompts, leading LLMs to predict by simple sampling from the population distribution, rather than making case-specific predictions.

(2) The patterns of highly mutable features, such as CE1 and CE2 (last 2 sub-figures of Fig. 5), are not captured by existing LLMs even at the population level. Rather, LLMs prefer to predict more negative responses to these features. This fact indicates building accurate predictors with LLMs for highly mutable features is more challenging, requiring LLMs to be well-aligned not only to individual information but also to population-level knowledge.

4.2 Incorporating Labeled Data

We try several popular methods below, only to find social prediction is still a challenging task without further advancement of LLMs.

The strong performance of the random forest classifier in Fig. 3 indicates that our proposed prediction task is reasonable if sufficient labeled data is considered. Based on this finding, we explore the effectiveness of incorporating supervision signals to LLMs based on the low2high setting as the example. Following prior studies (Brown et al., 2020; Song et al., 2023), we leverage the in-context learning ability of LLMs (Dong et al., 2022; Zhang et al., 2023) to incorporate a few labeled samples as demonstrations. Specifically, for each individual profile, we sample some other individual profiles from the dataset as the reference. In addition to the vanilla prompts introduced in Sec. 3.2, we append full information (including the input and response features) of these reference samples to the prompts. One example of such prompts is:

- Here are self-descriptions of two people:

"I am a male in the USA ... my answer is yes"; "I am a female in the USA ... my answer is no"; ...

- I am a male in the USA. I am 42 years old ...;

- When I'm asked "*Do you have access to the Internet in any way, whether on a mobile phone, a computer, or some other device?*", my answer is _____

To improve the efficiency of demonstrations, we select reference samples with tricks. (1) We choose 2 or 4 samples with the same *year* and *marriage* features with the predicted sample, and the positive and negative labels are balanced within the reference samples. (2) In addition, we adopt two more sets of in-context learning with Active Learning algorithms (Margatina et al., 2023). Among the demonstration selection methods, we select the most powerful Diversity (Yu et al., 2022) and Similarity (Liu et al., 2021) variants. The samples with the most distinct or similar representations are selected as context. Like the supervised methods, these demonstrations allow LLMs to make predictions aided by supervision signals from ground truth.

The results of experiments are shown in Table 3. It's plausible that augmented prompts with demonstrations help LLMs achieve better prediction performance, but it's worth noting the performance gain is unstable and sometimes minimal. Besides, incorporating 4 demonstrations (column 3) only has marginal or no improvement compared to incorporating 2 demonstrations (column 2). The active learning methods (columns 4-5) showcase similar results. This observation suggests solely searching for optimal prompts takes intensive effort, given the search space of prompts is infinite and the outputs are sensitive to prompts.

Again, as this is not the focus of our work, we leave the improvement of in-context prompts as a further direction.

5 Related Work

With the advent of LLMs, predicting social features with LLMs has been studied by numerous works (Ziems et al., 2023; Veselovsky et al., 2023). Among social studies, LLMs have been deployed to predict the potential responses or outcomes with

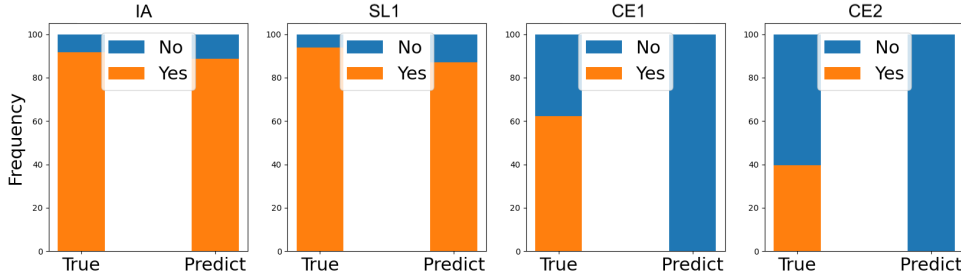


Figure 5: Distributions of Social Features. Note that the last two features (civic engagement behaviors) are more mutable than the first two.

Table 3: Performance of LLMs (GPT-3.5) with demonstrations. 2-Demos and 4-Demos indicate label-balanced demonstrations; AL-Sim and AL-Div stand for active learning with similar or diverse demonstrations.

	Zero Shot	2-Demos	4-Demos	AL-Sim	AL-Div
IA	50.00	71.61	82.67	60.46	54.19
SL2	50.00	50.60	50.04	48.54	50.65
EC1	52.63	50.52	53.47	49.11	56.64
CE1	50.00	60.17	55.34	54.15	54.13
CE2	50.00	53.22	52.79	55.90	60.34
AL	50.00	52.03	50.80	46.89	51.22

ease, especially in scenarios where traditional methods are constrained by cost or ethical concerns. In economics, Phelps and Russell (2023) studied game theory by examining cooperative and competitive behaviors with LLMs. Within political science, Wu et al. (2023) deployed LLMs to predict the ideological views of politicians. For communication studies, LLMs are used to simulate and predict the potential outcomes of toxic discourse (Törnberg et al., 2023), the political affiliation of Twitter posts (Törnberg, 2023), etc.

Additionally, there are growing interests in leveraging LLMs with social survey and interview, aiming to replicate human-like responses to certain questions or attributes of individuals. For example, Argyle et al., 2022 proposed "silicon samples" that deploy LLMs to simulate the people in a survey or interview and predict their partisan views and voting decisions. Dillion et al., 2023 examined the LLMs response to psychological tests, comparing the decisions and judgements from LLMs and humans. Aher et al., 2023 proposed sets of experiments to check LLMs response to interview and games. Besides, fine-tuning LLMs is a promising method for better prediction of social attitudes across years of surveys (Kim and Lee, 2023). At

the same time, discussions (Jansen et al., 2023) are hold about the potential and risks of deploying LLMs in social survey studies.

6 Conclusion

In this study, we introduce a survey-based social prediction task to assess the LLMs' predictive ability using general features. Through the replication of experiments and ablation studies of voting prediction tasks, we reveal a significant performance gap between input prompts with and without shortcut features. To further study the LLMs' predictive ability, we propose a real-world survey dataset with rigorously selected features. Based on it, we demonstrate the inability of LLMs to predict social features only with general features. Furthermore, our empirical studies further showcase the potential reasons that constrain the LLMs' predictive power. In our future research, we aim to explore the efficient methods of providing supervision signals and reference information to improve LLMs prediction performance. Moreover, with the abundant social survey and online data, we plan to use fine-tuning methods to fit the LLMs knowledge with social prediction tasks.

587 **7 Limitations**

588 As not the focus of this paper, we do not propose
589 methods to address the poor performance issue of
590 social prediction, nor provide experiments with
591 better results to validate our suggestions. Second,
592 the LLMs are not further fine-tuned and the op-
593 timal prompts are not searched. Tailoring LLMs
594 to advance social prediction abilities and finding
595 optimal prompts are potential directions to ex-
596 plore. Besides, we merely deploy large-scale close-
597 sourced LLMs and less powerful open-sourced
598 LLMs. However, large-scale open-sourced LLMs,
599 such as the Llama-70B, have both access to fine-
600 tuning and enormous language capabilities. For
601 researchers with enough computing resources, we
602 encourage further experiments and tuning on large-
603 scale open-sourced LLMs.

604 **8 Ethics Statement**

605 As far as we know, there are no major ethical
606 concerns thanks to our use of publicly available
607 and anonymized data. However, it's important to
608 acknowledge potential ethical issues when using
609 LLMs to mimic human responses in surveys. If
610 the privacy features are easy to infer, there's a
611 risk of privacy leakage. Moreover, addressing bias
612 and ensuring fairness is another significant ethical
613 challenge. LLMs may perpetuate societal biases
614 present in training data, resulting in social predic-
615 tion responses that reinforce harmful stereotypes or
616 discrimination. Thoroughly testing for bias across
617 different demographics is vital to mitigate these
618 risks and promote fairness.

619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672

References

Josh Achiam et al. 2023. [Gpt-4 technical report](#).

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, 6(1).

ANES. User’s guide and codebook for the anes 2012 time series study.

Rohan Anil et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Lisa P. Argyle, E. Busby, Nancy Fulda, Joshua Ronald Gubler, Christopher Michael Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31:337 – 351.

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.

Susan Athey. 2018. The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*, pages 507–547. University of Chicago Press.

Kenneth D Bailey. 1992. Globals, mutables, and immutable: a new look at the micro-macro link. *Quality and Quantity*, 26(3):259–276.

Kenneth D Bailey. 1998. A theory of mutable and immutable characteristics: Their impact on allocation and structural positions. *Quality and Quantity*, 32(4):383–398.

Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74.

Jordan Brensinger and Gil Eyal. 2021. The sociology of personal identification. *Sociological Theory*, 39(4):265–292.

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.

Mario Callegaro, Reginald P Baker, Jelke Bethlehem, Anja S Göritz, Jon A Krosnick, and Paul J Lavrakas. 2014. *Online panel research: A data quality perspective*. John Wiley & Sons.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yunsong Chen, Xiaogang Wu, Anning Hu, Guangye He, and Guodong Ju. 2021b. Social prediction: a new research paradigm based on machine learning. *The Journal of Chinese Sociology*, 8:1–21.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936.

Mick P Couper. 2017. New developments in survey data collection. *Annual review of sociology*, 43:121–145.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Russell J Dalton. 2016. Party identification and its implications. *Oxford research encyclopedia of politics*.

Fernando Diaz, Michael Gamon, Jake M Hofman, Emre Kıcıman, and David Rothschild. 2016. Online and social media data as an imperfect continuous panel survey. *PloS one*, 11(1):e0145406.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Jéssica Soares dos Santos, Flávia Cristina Bernardini, and A. Paes. 2021. [A survey on the use of data and opinion mining in social media to political electoral outcomes prediction](#). *Social Network Analysis and Mining*, 11.

G Gallup. 2009. World poll methodology. Technical report, Technical Report, Washington, DC.

Qiong Gu, Li Zhu, and Zhihua Cai. 2009. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer.

Janet E Halley. 2017. Sexual orientation and the politics of biology: A critique of the argument from immutability. In *Sexual orientation and rights*, pages 3–68. Routledge.

Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science*, 355(6324):486–488.

726	Jie Huang, Xinyun Chen, Swaroop Mishra,	Hannes Rosenbusch, Claire E Stevenson, and Han LJ	779
727	Huaixiu Steven Zheng, Adams Wei Yu, Xiny-	van der Maas. 2023. How accurate are gpt-3's hy-	780
728	ying Song, and Denny Zhou. 2023. Large language	potheses about social science phenomena? <i>Digital</i>	781
729	models cannot self-correct reasoning yet. <i>arXiv</i>	<i>Society</i> , 2(2):26.	782
730	<i>preprint arXiv:2310.01798</i> .		
731	Doaa Mohey El-Din Mohamed Hussein. 2018. A survey	Maya Sen and Omar Wasow. 2016. Race as a bundle of	783
732	on sentiment analysis challenges. <i>Journal of King</i>	sticks: Designs that estimate effects of seemingly im-	784
733	<i>Saud University-Engineering Sciences</i> , 30(4):330–	mutable characteristics. <i>Annual Review of Political</i>	785
734	338.	<i>Science</i> , 19:499–522.	786
735	Bernard J Jansen, Soon-gyo Jung, and Joni Salminen.	Chan Hee Song, Jiaman Wu, Clayton Washington,	787
736	2023. Employing large language models in survey	Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023.	788
737	research. <i>Natural Language Processing Journal</i> .	Llm-planner: Few-shot grounded planning for em-	789
738		bodied agents with large language models. In <i>Pro-</i>	790
739	Albert Q Jiang, Alexandre Sablayrolles, Arthur Men-	<i>ceedings of the IEEE/CVF International Conference</i>	791
740	sch, Chris Bamford, Devendra Singh Chaplot, Diego	<i>on Computer Vision</i> , pages 2998–3009.	792
741	de las Casas, Florian Bressand, Gianna Lengyel, Guil-	Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu,	793
742	laume Lample, Lucile Saulnier, et al. 2023. Mistral	Yongrui Chen, and Guilin Qi. 2023. Can chatgpt	794
743	7b. <i>arXiv preprint arXiv:2310.06825</i> .	replace traditional kbqa models? an in-depth analysis	795
744	Junsol Kim and Byungkyu Lee. 2023. Ai-augmented	of the question answering performance of the gpt llm	796
745	surveys: Leveraging large language models for opin-	family. In <i>International Semantic Web Conference</i> ,	797
746	ion prediction in nationally representative surveys.	pages 348–367. Springer.	798
747	<i>arXiv preprint arXiv:2305.09620</i> .		
748	David Liben-Nowell and Jon Kleinberg. 2003. The	Petter Törnberg. 2023. Chatgpt-4 outperforms experts	799
749	link prediction problem for social networks. In <i>Pro-</i>	and crowd workers in annotating political twitter	800
750	<i>ceedings of the twelfth international conference on</i>	messages with zero-shot learning. <i>arXiv preprint</i>	801
751	<i>Information and knowledge management</i> , pages 556–	<i>arXiv:2304.06588</i> .	802
752	559.		
753	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan,	Petter Törnberg, Diliara Valeeva, Justus Uitermark,	803
754	Lawrence Carin, and Weizhu Chen. 2021. What	and Christopher Bail. 2023. Simulating social me-	804
755	makes good in-context examples for gpt-3? In <i>Work-</i>	dia using large language models to evaluate al-	805
756	<i>shop on Knowledge Extraction and Integration for</i>	ternative news feed algorithms. <i>arXiv preprint</i>	806
757	<i>Deep Learning Architectures; Deep Learning Inside</i>	<i>arXiv:2310.05984</i> .	807
758	<i>Out</i> .		
759	Adrian Mackenzie. 2015. The production of predic-	Hugo Touvron et al. 2023. Llama 2: Open founda-	808
760	tion: What does machine learning want? <i>European</i>	tion and fine-tuned chat models. <i>arXiv preprint</i>	809
761	<i>Journal of Cultural Studies</i> , 18(4-5):429–445.	<i>arXiv:2307.09288</i> .	810
762			
763	Katerina Margatina, Timo Schick, Nikolaos Aletras, and	Gülden Kaya Uyanık and Neşe Güler. 2013. A study on	811
764	Jane Dwivedi-Yu. 2023. Active learning principles	multiple linear regression analysis. <i>Procedia-Social</i>	812
765	for in-context learning with large language models.	<i>and Behavioral Sciences</i> , 106:234–240.	813
766	In <i>Conference on Empirical Methods in Natural Lan-</i>		
767	<i>guage Processing</i> .	Ike Vayansky and Sathish AP Kumar. 2020. A review	814
768		of topic modeling methods. <i>Information Systems</i> ,	815
769	Warren E Miller. 1991. Party identification, realign-	94:101582.	816
770	ment, and party voting: Back to the basics. <i>American</i>	Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil	817
771	<i>Political Science Review</i> , 85(2):557–568.	Arora, Martin Josifoski, Ashton Anderson, and	818
772		Robert West. 2023. Generating faithful synthetic	819
773	Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan	data with large language models: A case study	820
774	Wang, Yingbo Zhou, Silvio Savarese, and Caiming	in computational social science. <i>arXiv preprint</i>	821
775	Xiong. 2022. Codegen: An open large language	<i>arXiv:2305.15041</i> .	822
776	model for code with multi-turn program synthesis.		
777	<i>arXiv preprint arXiv:2203.13474</i> .	Leah von der Heyde, Anna-Carolina Haensch, and	823
778		Alexander Wenz. 2023. Assessing bias in llm-	824
779	OpenAI. 2022. Openai chatgpt .	generated synthetic datasets: The case of german	825
780		voter behavior. Technical report, Center for Open	826
781	Steve Phelps and Yvan I Russell. 2023. Investigating	Science.	827
782	emergent goal-like behaviour in large language mod-		
783	els using experimental economics. <i>arXiv preprint</i>	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	828
784	<i>arXiv:2305.07970</i> .	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	829
785		et al. 2022. Chain-of-thought prompting elicits rea-	830
786		soning in large language models. <i>Advances in Neural</i>	831
787		<i>Information Processing Systems</i> , 35:24824–24837.	832

833 Patrick Y Wu, Joshua A Tucker, Jonathan Nagler, and
834 Solomon Messing. 2023. Large language models
835 can be used to estimate the ideologies of politi-
836 cians in a zero-shot learning setting. *arXiv preprint*
837 *arXiv:2303.12057*.

838 W. Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingx-
839 uan Ju, Soumya Sanyal, Chenguang Zhu, Michael
840 Zeng, and Meng Jiang. 2022. Generate rather than
841 retrieve: Large language models are strong context
842 generators. *ArXiv*, abs/2209.10063.

843 Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Ming-
844 hao Yin. 2019. A survey of sentiment analysis in
845 social media. *Knowledge and Information Systems*,
846 60:617–663.

847 Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. 2023.
848 Trained transformers learn linear models in-context.
849 *arXiv preprint arXiv:2306.09927*.

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

855 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei,
856 Nathan Scales, Xuezhi Wang, Dale Schuurmans,
857 Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022.
858 Least-to-most prompting enables complex reason-
859 ing in large language models. *arXiv preprint*
860 *arXiv:2205.10625*.

861 Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and
862 Chao Zhang. 2023. Toolqa: A dataset for llm ques-
863 tion answering with external tools. *arXiv preprint*
864 *arXiv:2306.13304*.

865 Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen,
866 Zhehao Zhang, and Diyi Yang. 2023. Can large lan-
867 guage models transform computational social sci-
868 ence? *arXiv preprint arXiv:2305.03514*.

869 **A Appendix**

870 **A.1 Questions for Features**

871 We categorize the selected 16 features into two
872 groups, i.e. high-mutability and low-mutability fea-
873 tures. The details of high-mutability features are
874 shown in Table 4 and those of low-mutability fea-
875 tures are shown in Table 5. The column "Question
876 Abbrev." indicates the abbreviation of the features,
877 which are broadly used in this work. The column
878 "Question Identifiers" indicates the identifier la-
879 bels of the corresponding questions in the original
880 Gallup survey.

881 **A.1.1 Feature Convert Methods**

882 In the main experiments, there are features
883 of integer or several classes, such as income,
884 employment, etc. We convert them into groups
885 (with the number of groups no larger than four).
886 For income, we calculate the 35% and 65% per-
887 centiles of the annual household income. Based
888 on them, we categorize income into three classes:
889 lower level, middle level, and higher level. For fea-
890 tures with more than 4 classes, we combine similar
891 classes to make the number of classes as 2 or 3.

Topic	Question Abbrev.	Question Identifiers	Question	Options
Communication Use	IA	WP16056	Do you have access to the internet in any way, whether on a mobile phone, a computer, or some other device?	yes, no
Social Life	SL1	WP27	If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?	yes, no
	SL2	WP10248	In the city or area where you live, are you satisfied or dissatisfied with the opportunities to meet people and make friends?	satisfied, dissatisfied
Economic Confidence	EC1	WP148	Right now, do you think that economic conditions in this country, as a whole, are getting better or getting worse?	better, worse
	EC2	M30	How would you rate your economic conditions in this country today – as excellent, good, fair, or poor?	excellent, good, fair, poor
Civic Engagement	CE1	WP108	Have you donated money to a charity in the past month?	yes, no
	CE2	WP109	Have you volunteered your time to an organization in the past month?	yes, no
	CE3	WP110	Have you helped a stranger or someone you did not know who needed help?	yes, no
Approval of Leadership	AL	WP150	Do you approve or disapprove of the job performance of the leadership of this country?	approve, disapprove

Table 4: Questions and Options of High-mutability Features of Gallup Dataset.

Immutable Attribute	Question Abbrev.	Question Identifiers.	Options
Age	age	age	-
Gender	gender	WP1219	1. Man, 2. Woman
Marital Status	marriage	WP1223	1. Single/Never been married, 2. Married, 3. Separated, 4. Divorced, 5. Widowed, 6. Domestic Partner;
Highest Completed Level of Education	education	WP3117	1. Completed elementary education or less (up to 8 years of basic education); 2. Secondary - 3 years Tertiary/Secondary education and some education beyond secondary education (9-15 years of education); 3. Completed four years of education beyond high school and/or received a 4-year college degree;
Employment Status	employment	EMP_2010	1. Employed full time for an employer, 2. Out of workforce, 3. Employed part time do not want full time, 4. Employed full time for self, 5. Employed part time want full time, 6. Unemployed;
Annual Household Income	income	INCOME_1	-
Living of Urbanicity	urbanicity	WP14	1. A suburb of a large city, 2. A small town or village, 3. A large city, 4. A rural area or on a farm;

Table 5: Questions and Options of Low-mutability Features of Gallup Dataset.