Quantifying the Persona Effect in LLM Simulations

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown remarkable promise in simulating human language use and behavior. In this study, we delve into the intersection of persona variables and the capability of LLMs to simulate different perspectives. We find that persona variables can explain <10% variance in annotations in existing subjective NLP datasets. Nonetheless, incorporating them via prompting in LLMs provides modest improvement. Persona prompting is most effective on data samples where disagreements among annotators are frequent yet confined to a limited range. A linear correlation exists: the more persona variables influence human annotations, the better LLMs predictions are using persona prompting. However, when the utility of persona variables is low (i.e., explaining <10% of human annotations), persona prompting has little effect. Most subjective NLP datasets fall into this category, casting doubt on simulating diverse perspectives in the current NLP landscape.

1 Introduction

Annotation questions such as "how do you feel emotionally after reading this text" are subjective - there are rarely definitive right or wrong answers (Ovesdotter Alm, 2011). This subjectivity is increasingly being recognized within the NLP community. Subjective NLP tasks are typically characterized by low inter-annotator agreement, making label aggregation inappropriate (Ovesdotter Alm, 2011; Plank, 2022; Cabitza et al., 2023). Previous research has established the significant influence of sociodemographic variables on the annotations of these tasks (Sap et al., 2022; Santy et al., 2023; Pei and Jurgens, 2023, *inter alia*).

One approach to model these persona variables[1] is to use LLMs. LLMs have been effectively utilized for role-playing and simulating human behavior, primarily by defining the persona of interest within the prompt (Aher et al., 2023; Horton, 2023; Kovač et al., 2023; Argyle et al., 2023). Their success has even spurred debates on whether LLMs could replace human subjects (Dillion et al., 2023; Grossmann et al., 2023). However, there are also concerns about such "persona prompting" methodology (Figure 1) (Beck et al., 2023), citing ecological fallacy (Orlikowski et al., 2023), and LLMs' susceptibility to caricatures (Cheng et al., 2023), misportrayal and erasure of subgroup heterogeneity (Wang et al., 2024).

Existing studies have often sought to measure the effects of individual persona variables, overlooking a holistic analysis of the potential explanatory power of persona variables on annotation variance. It is then hard to contextualize the models' ability to utilize persona information. Furthermore, the influence of *persona variables* is often conflated with that of *text samples* (Figure 1), making it difficult to understand the true capacity of LLMs to simulate personas. To address these issues, our research explores the following questions:

RQ1: How much variance in human annotation could persona variables explain? Understanding this will help us assess the overall influence of persona variables on human annotation, providing context to our subsequent investigations. We find that persona variables explain relatively little variance (<10%) for many NLP tasks (Section 3).

RQ2: Can incorporating persona variables via prompting improve LLMs' predictions? Building on our findings from RQ1, we assess how much the explained variance by persona variables translates into prediction gains in LLMs. We find that in three of four datasets, incorporating persona variables provides a modest improvement (Section 4).

RQ3: For what types of samples is persona prompt-

[1] In our work, we adopt a broad definition of *persona variables* to include not only demographic and social variables but also other variables that could help describe a persona, such as variables relating to attitudes, behaviors, lived experiences, and values. It is worth noting that most NLP datasets have no information of any kind available about the annotators.

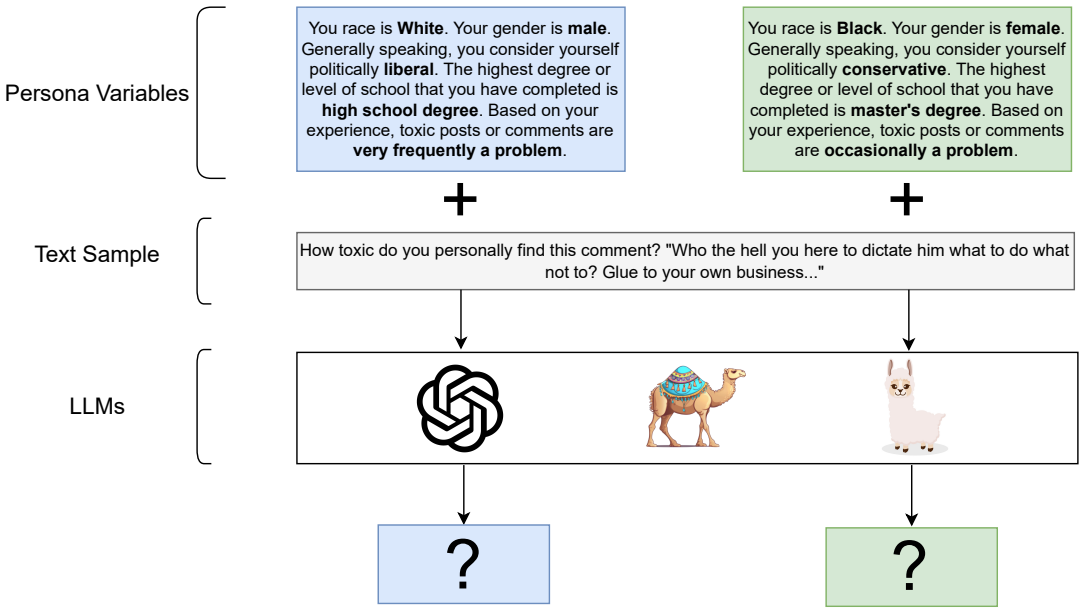


Figure 1: Illustration of persona prompting. We prepend the persona information of an annotator before the text sample and task description to investigate the capacity of LLMs to simulate diverse perspectives in subjective NLP tasks.

ing most useful? To delve deeper into the utility of persona prompting, we examine its impact across sample types. We identify that most gains occur in samples characterized by frequent annotator disagreements within a relatively narrow range, suggesting that models can adjust their annotation to suit the persona, though not drastically (Section 5).

RQ4: How well can LLMs simulate personas with controlled text randomness and varied persona utilities? This setting isolates the evaluation of persona simulation evaluation from the variance in text, allowing us to fully understand the capacity of LLMs in simulating personas with when persona variables matter to a varying degree. We find a linear relationship: the more persona factors influence human response, the better LLMs predictions are using persona prompting. Larger, fine-tuned models perform best and can explain up to 81% of variance found in human responses. However, when the utility of persona variables is low, persona prompting has little effect. Regrettably, most subjective NLP datasets fall into this category, casting doubt on the efficacy of persona prompting in the current NLP context (Section 6).

2 Related Work

2.1 The Relationship between Persona Variables and Annotation Outcome

The role of persona variables, such as demographics and lived experiences, in influencing annotations in NLP tasks is well established. Many studies have highlighted how persona variables affect tasks like hate speech detection (Kumar et al., 2021; Sap et al., 2022; Pei and Jurgens, 2023; Santy et al., 2023; Hettiachchi et al., 2023; Lee et al., 2023), sentiment analysis (Ding et al., 2022; Biester et al.), and irony detection (Frenda et al., 2023). While these studies shed light on the subjectivity of NLP annotations in many tasks, they often stop short of a holistic account of the explanatory power of persona variables on annotation variance. By contrast, in social science, the impact of persona variables on attitude are long studied and quantified (Bobo and Licari, 1989; Bartels, 2002). In our work, we analyze the utility of the persona variables in explaining annotation outcomes across subjective NLP tasks.

2.2 Modeling Persona Variables and LLM for Simulation

Several works in NLP have sought to account for the differences between individual annotators or the group-level attributes of annotators through adding individual (group) specific lay-

ers (Mostafazadeh Davani et al., 2022; Gordon et al., 2022; Fleisig et al., 2023; Orlikowski et al., 2023, *inter alia*), or via prompting (Beck et al., 2023). Results from these studies have been mixed, with some work indicating success using group-level persona variables (Gordon et al., 2022; Fleisig et al., 2023), while others cast doubt on the effectiveness of such methods (Orlikowski et al., 2023; Cheng et al., 2023; Beck et al., 2023). Simultaneously, in the social sciences, a multitude of studies have been employing use persona prompts in LLMs to simulate human behavior (Horton, 2023; Argyle et al., 2023; Kim and Lee, 2023; Törnberg et al., 2023), while others have pointed out the lack of fidelity and diversity in such simulations (Bisbee et al., 2023; Park et al., 2024; Wang et al., 2024; Taubenfeld et al., 2024).

Our work builds on the uncertainty raised by these mixed results, focusing on the potential of persona prompting with LLMs for simulating different perspectives in NLP tasks, which is currently understudied. Furthermore, our work aims to isolate the evaluation of *persona* prompting from the impact of *text samples* in the modeling process, a separation that has not been much explored in previous studies.

2.3 Persona Prompting and AI Alignment

Apart from the research focused on incorporating demographic factors into NLP models and using LLMs for simulations, another line of studies has examined persona prompting in the context of AI alignment (Santurkar et al., 2023; Durmus et al., 2023). These studies have employed LLMs to answer multiple-choice survey questions concerning societal values and attitudes, comparing the LLM-generated answer distribution with actual human response distribution derived from survey data representing diverse demographic groups. In contrast to these studies, our work aims to explore the efficacy of LLMs in leveraging persona variables to inform task predictions, rather than the degree to which LLM responses mirror those of specific demographic groups.

3 RQ1: How much variance in human annotation could persona variables explain?

Methodology Given the relative gap in literature in a holistic understanding of the impact of persona variables on annotation variance, we investigate to what extent persona variables explain human annotation variance. This analysis would provide valuable context to any modeling exercise of incorporating persona variables.

We employ a mixed-effect linear regression model[2] to assess how much variance in annotation can be explained by persona variables (fixed effect), while controlling for the text-specific variability in the text sample (random effect) by fitting a random intercept for each text. Using a mixed-effect linear regression allows us to separate the impact of persona variables from the inherent variation of the text being annotated. We evaluate 10 subjective NLP datasets which provide unaggregated annotations and persona variables of their annotators. We also consider the presidential vote question in the ANES 2012 public opinion survey (ANES), in which every human subject answers the same question and therefore does not require a text random effect, for comparison.

Results We show a comparison of the tasks, sources of data, annotation methods, sizes, types of persona information included, and the regression R^2 values in Table 1.

We observe that the datasets mostly come from social media sources and annotations are collected through crowd-sourcing. They vary substantially in size, persona variables provided and R^2 values. While persona variables (fixed effect) do significantly explain some variance in annotation outcomes, they account for just **1.4%-10.6%** of the total variance (Marginal R^2), even when controlling for text variation. Conversely, variability inherent to individual texts (random effect) can explain up to **70%** of the total variance, i.e. $\sim$ (Conditional R^2 − Marginal R^2). For comparison, in the ANES dataset, persona variables explain more than **70%** human response variance.

The marginal R^2 values provide a baseline indication of the variance in annotations that persona variables could explain. The regression model assumes a linear relationship between persona variables and annotation and does not consider any interaction between the persona variables. Therefore, while it is straightforward and interpretable, for LLMs, it should be considered a **weak baseline**.

Acknowledging that a substantial portion of variance remains unexplained (**25%-70%**) by either the text or persona variables across all tasks consid-

[2] In R notation, `annotation ~ persona variables + (1 | text_id)`

Task	Dataset	Data Source	Annotation	Size	Persona Variables	$R^2_{Cond.}$	$R^2_{Marg.}$
Toxicity Detection	annWithAttitudes (Sap et al., 2022)	Twitter	5-point MTurk	N=626 A=5.5[a] U.S.	Basic Attitude	0.611	0.045
Offensiveness Rating	POPQUORN (Pei and Jurgens, 2023)	Reddit	5-point Prolific	N=1,500 A=8.7 U.S.	Basic	0.319	0.029
Politeness Rating	POPQUORN (Pei and Jurgens, 2023)	Email	5-point Prolific	N=3,718 A=6.7 U.S.	Basic	0.454	0.014
Toxicity Detection	Kumar et al. (2021)	Twitter Reddit 4chan	5-point MTurk	N=106,035 A=5.1 U.S.	Basic Attitude Behavior	0.349	0.106
Sentiment Analysis	Diaz et al. (2018)	Twitter	5-point	N=14,071 A=4.2 U.S.	Basic Attitude	0.329	0.036
Social Acceptability	Social-Chem-101 (Forbes et al., 2020)	Reddit	5-point MTurk	N=9,740 A=6.1 Mostly U.S.	Basic	0.432	0.097
Social Acceptability	NLPositionality[b] (Santy et al., 2023)	Reddit	5-point Opt-in volunteer	N=291 A=50.2 87 countries	Basic	0.513	0.005
Toxicity Detection	NLPositionality[b] (Santy et al., 2023)	Twitter	3-point Opt-in volunteer	N=299 A=29.6 87 countries	Basic	0.432	0.017
Social Bias	SBIC (Sap et al., 2020)	Twitter Reddit Gab Stormfront	3-point MTurk	N=35,504 A=3.2 U.S. and Canada	Basic	0.758	0.031
Irony Detection	EPIC (Frenda et al., 2023)	Twitter Reddit	Binary Prolific	N=2,994 A=4.7 IE, UK, US, IN, AU	Basic	0.289	0.091
Presidential Vote	ANES 2012	Survey	Binary Face-to-face	A=2,728[c] U.S.	Basic Attitude Behavior	-	0.719

[a] Another phase of this dataset has 600+ annotators labeling a total of 15 tweets.
[b] We consider the action acceptability, to be in line with the NLPositionality dataset. As it is a volunteer-annotated dataset, substantial persona information is unavailable.
[c] After filtering out participants with missing attributes.

Table 1: An overview of datasets with unaggregated annotations and persona information. This table compares the tasks, sources of data, annotation methods, sizes, types of persona information included, and to what degree the persona variables can explain the variance of annotations in each dataset. The "Size" column specifies the number of text samples (N) and the average number of annotators per sample (A), alongside the geographical location of the annotators. The "Persona Variables" column indicates the available persona categories: "Basic" for standard demographics like gender and age, "Attitude" for annotators' personal views, and "Behavior" for actions such as media consumption habits. The conditional ($R^2_{Cond.}$) and marginal ($R^2_{Marg.}$) R-squared values are reported from regression models that predict the annotations based on persona variables, while accounting for text-specific variability (using a random effect for each text).

ered is crucial. This unexplained variance could be attributed to theoretically measurable persona factors such as personality traits and complex moral and political beliefs, which are not currently collected in existing datasets. Additionally, it could be due to hard-to-measure factors like the annotators' lived experiences, interpersonal dynamics, and other personal variables.

The elevated R^2 value in the ANES dataset may be attributed to the escalating degree of polarization in U.S. politics in recent years. This rise in polarization has lead to more predictable voting patterns (Pew Research Center, 2014) and the increasing tendency of U.S. voters to behave in a manner consistent with their in-groups (Graham and Haidt, 2010).

In contrast, tasks such as assessing the hatefulness of a tweet offer more room for personal interpretation, leading to diverse opinions. Thus, persona factors may account for a lesser portion of the variance in annotation for such tasks.

4 RQ2: Can incorporating persona variables via prompting improve LLMs' predictions?

Methodology Since persona variables can explain a small but significant amount of human annotation variations, we then explore whether persona prompting would improve LLM's predictions.

As depicted in Figure 1, we prepend each *text sample* with *persona variables* in a zero-shot prompting setup. We prompt the LLMs twice: once with persona variables, and once without, to zero-shot predict individual annotations on AnnotatorwithAttitude (Sap et al., 2022), Kumar et al. (2021), EPIC (Frenda et al., 2023) and the politeness rating task in POPQUORN (Pei and Jurgens, 2023). We preserve the original language of the persona descriptions to the extent possible, adopt a multiple-choice format, include a description of the question and the answer choices, and predict only the next token as the model's response, as done in prior work (Santurkar et al., 2023; Durmus et al., 2023). Due to cost constraints, we sample 600 instances from each dataset. The details of the prompt format are provided in the Section B.

We additionally perform a set of robustness experiments by swapping the order of persona variables in the prompt or paragraphing the language used to describe each persona variables and repeat the experiments on Kumar et al. (2021). The detailed setting can be found in Section C.

To evaluate, we compare model predictions with individual human annotations using R^2 value, Cohen's Kappa (Cohen, 1960), mean absolute error (MAE) for multi-class classification or F1 score for binary classification. Our focus is on observing the performance change before and after persona prompting, rather than the absolute performance of each model.

Result We show the results in Table 2. The first row shows the "Target" R^2 values, which refer to the conditional (and marginal) R^2 value of the mixed-effect regression on the sampled data computed as in Table 1, while the R^2 in subsequent rows are from a fixed-effect linear regression predicting the human annotation with model predictions[3]. While these two R^2 values cannot be compared directly, the "Target" R^2 gives context to the fixed-effect R^2 values. As the 7b and 13b models exhibit much weaker performance, we only feature results from 70b models in the main text, while the results from smaller models are included in Table 3.

On average, among the 6 models considered, persona prompting shows varying levels of improvement on 3/4 datasets. However, the improvement in terms of R^2 is small compared to the target R^2. For instance, in EPIC, where persona variables could explain up to 9% of annotation variance, persona prompting only provides 1% gain on average. The effectiveness of persona prompting also varies across models: for each dataset, persona prompting improves the performance of some models but not others, echoing the results in Beck et al. (2023).

We note that overall, with and without persona prompting, GPT-4 consistently outperforms all other models in every task. Tulu-2 models outperform Llama-2 with performance on par with GPT-3.5. The Llama-2 models are, on the other hand, much more sensitive to persona variables, arguably to an excessive degree. For example, on AnnotatorwithAttitudes, persona prompting improves the R^2 by as much as 0.23 even though persona variables only has a marginal Target R^2 of 0.03. We show the robustness experiment result in Table 5. The model performances are consistent across variations in the ordering and language use of the persona variables.

5 RQ3: For what types of samples is persona prompting most useful?

Methodology To better understand persona prompting as a technique, we aim to investigate its effectiveness on data samples with varying degrees of annotation *entropy* and *standard deviation*. We focus on Kumar et al. (2021), as persona variables play a relatively more important role in explaining annotation variances in this dataset.

We create a new subsample of the dataset based on four categories: low entropy-low standard deviation (most annotators agree with one another and the magnitude of the disagreement is small, e.g. 1,1,1,1,2); low entropy-high standard deviation (e.g. 0,4,4,4,0), high entropy-low standard deviation (e.g. 1,1,2,2,3) and high entropy-high standard deviation (e.g. 0,1,2,3,4). The low/high division is based on the medians for entropy and standard deviation.

Then, we further stratify samples from each category into four bins according to their average annotation value. We then randomly sample 150 from

[3] in R notation, `annotation ~ prediction`

Model	annwAttitudes			Kumar et al. (2021)			EPIC			POPQUORN-P		
	$R^2\uparrow$	$\kappa\uparrow$	MAE $\downarrow$	$R^2\uparrow$	$\kappa\uparrow$	MAE $\downarrow$	$R^2\uparrow$	$\kappa\uparrow$	F1 $\uparrow$	$R^2\uparrow$	$\kappa\uparrow$	MAE $\downarrow$
Target	0.64 (0.03)	-	-	0.42 (0.20)	-	-	0.28 (0.09)	-	-	0.47 (0.03)	-	-
GPT-4-0613	0.56	0.42	0.70	0.16	0.24	0.87	0.03	0.12	0.52	0.34	0.22	0.89
+Persona	0.53	0.40	0.74	0.12	0.20	0.90	0.05	0.20	0.58	0.33	0.22	0.90
GPT-3.5-Turbo-0613	0.53	0.29	0.80	0.12	0.17	1.12	0.04	0.18	0.59	0.28	0.09	1.07
+Persona	0.49	0.31	0.82	0.12	0.15	0.97	0.03	0.14	0.54	0.28	0.14	1.14
Llama-2-70b	0.17	0.14	1.70	0.01	0.04	1.51	0.00	0.00	0.24	0.24	0.13	1.42
+Persona	0.40	0.30	0.91	0.03	0.05	1.01	0.00	0.00	0.24	0.21	0.17	1.10
Llama-2-70b-chat	0.39	0.13	1.33	0.11	0.07	1.70	0.00	0.05	0.49	0.32	0.15	1.00
+Persona	0.42	0.15	1.22	0.10	-0.01	1.45	0.02	0.14	0.56	0.31	0.14	0.90
Tulu-2-70b	0.49	0.29	0.90	0.16	0.13	1.09	0.05	0.20	0.59	0.34	0.20	0.89
+Persona	0.49	0.26	0.88	0.14	0.16	0.90	0.07	0.27	0.63	0.31	0.16	0.92
Tulu-2-dpo-70b	0.51	0.35	0.84	0.15	0.15	1.16	0.03	0.14	0.54	0.35	0.21	0.83
+Persona	0.51	0.30	0.84	0.15	0.20	0.92	0.04	0.18	0.58	0.33	0.19	0.87
Avg. Δ	0.03	0.02	-0.14	-0.01	-0.01	-0.23	0.01	0.04	0.03	-0.02	0.00	-0.04

Table 2: Comparison of performance across LLMs in estimating individual annotations, with and without the inclusion of persona variables. Performance is measured using R^2, Cohen's Kappa (κ), Mean Absolute Error (MAE) and F1 score.

each bin, culminating in a total of 600 samples per category. This approach is implemented to mitigate extreme class imbalances within certain categories. For instance, the low entropy-low standard deviation category would predominantly include samples with a rating of 0 (Not at all toxic). We then run the LLMs twice, once with persona prompting, once without, in the same setting as described in Section 4, on Llama-2-70b, Llama-2-70b-chat, Tulu-2-70b, and Tulu-2-dpo-70b.

Result We show in Figure 2a the mean improvement in MAE between models with and without persona prompting, averaged across the four models, in each of the four categories, with darker color indicating a greater degree of improvement in predictions when persona prompting is used. To reduce the possibility of finding a dataset-specific effect, we also repeat the same experiment on POPQUORN-Politeness dataset (Pei and Jurgens, 2023), and show the same plot Figure 2b.

Our findings indicate that including persona information leads to only slight changes in the model's predictions for data with low entropy. This is as expected - with or without persona prompting, a capable LLM should already capture the consensus among annotators if there is one, thus only necessitating minor adjustments to individual predictions.

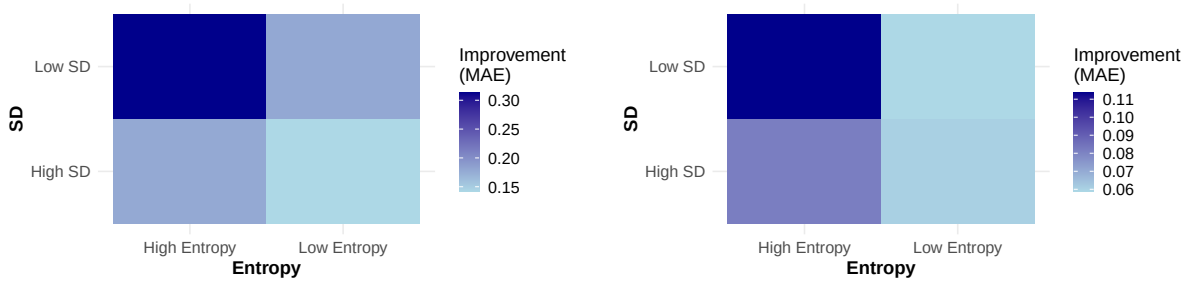
On the contrary, we observe larger shifts in prediction when annotations have high entropy but low standard deviation. These instances often involve substantial disagreement among individuals, though within a small margin. The integration of persona variables may then enhance the model's ability to refine its predictions. An example in this would be a prediction transition from 3 (without persona variables) to 4 (with persona variables).

However, when both entropy and standard deviation are high, the task of adjusting predictions based on persona information becomes considerably more challenging, as this would require significant shifts in the predicted values from the "mean" level, when no persona variables are provided. For instance, imagine a case where a prediction needs to change from 0 (without persona variables) to 4 (with persona variables).

6 RQ4: How well can LLMs simulate personas with controlled text randomness and varied persona utilities?

Motivation Within the context of NLP annotation, both the *text sample* and the persona variables may vary across instances (Figure 1). Both factors, along with their interactions, could potentially influence model predictions. To understand the models' capacity for simulating different perspectives with persona prompting, we designed a case study that minimizes the impact of the *text sample*.

Methodology We use the ANES dataset (ANES), a comprehensive U.S. national-level election survey, as a data source for this section. This dataset offers a wealth of persona variables from a large sample of survey respondents. From the perspective of NLP annotation, surveys can be seen as having a large number of individuals (typically >1,000)



(a) Kumar et al. (2021)

(b) POPQUORN-P

Figure 2: Mean improvement in MAE with persona prompting across four 70b models in annotations characterized by low/high entropy and standard deviation, with darker colors denoting more substantial improvement in predictions.

annotating a small number of sentences, each representing a survey question. One key difference is that the survey questions, carefully crafted and tested by seasoned professionals, are designed to eliminate ambiguity common in social media-based NLP text annotation datasets. Therefore, by running experiments on the ANES dataset, we can minimize the impact of the randomness in the text samples.

We select a number of questions from ANES 2012 as the *text sample*, or the questions to be predicted, using a fixed set of persona variables. We ensure that these questions have varying predictability from persona variables, indicated by R^2 values. Further details of the dependent and independent variables considered are included in Section D. After filtering out respondents who did not answer some of the questions of interest, we arrive at a sample size of 2,372 human respondents and 42 questions. We then run the LLMs with persona prompting.

We perform a robustness check with the presidential vote prediction question from ANES by swapping the order of persona variables in the prompt or paragraphing the language used to describe each persona variables. The detailed setting can be found in Section C.

Result We visualize the relationship between the predicted and target R^2 values in Figure 3 of Tulu-2-70b-dpo and Llama-2-7b-chat. The results for other models are provided in the Figure 4. Each point in the scatter plot represents an experiment result, where the x-coordinate signifies the target R^2 and the y-coordinate denotes the predicted R^2. The line $Y = X$ is also included to represent the maximum possible performance, where predicted R^2 equals target R^2. We additionally fit a linear regression line to the data points and show the fitted equation and R^2 in the figure.

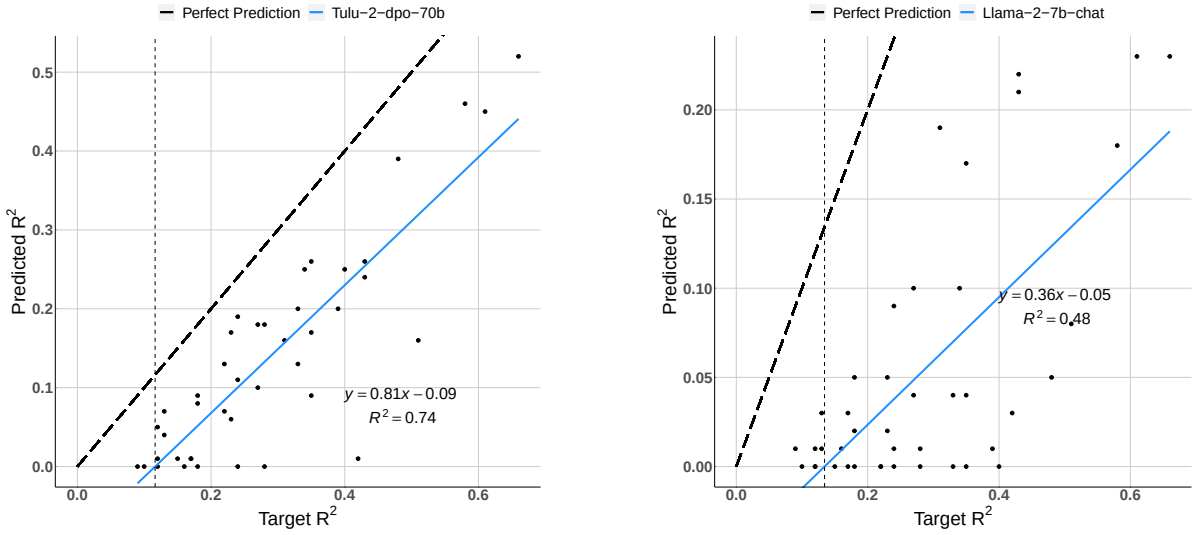
Our results show a positive correlation between the target and predicted R^2 values - the higher the target R^2 value, the higher the predicted R^2. Tulu-2-70b-dpo, one of the best-performing models on the 70b scale, can capture 81% of the target R^2. However, it still fails to utilize the persona information effectively when target R^2 is low, especially when $R^2 < 0.1$. The other 70b models, except for the base model Llama-2-70b (Figure 4), have largely the same simulation capabilities, while the smaller models (7b and 13b) do much worse.

Considering that most existing NLP datasets, as discussed in Section 1, have marginal $R^2 < 0.1$, we argue that **persona prompting cannot reliably simulate different perspectives within existing NLP tasks**. This finding may explain the modest or non-existent gain of persona prompting observed earlier in Section 2 and in Beck et al. (2023).

We propose two potential explanations as to why LLMs, however powerful they are in other tasks, may be deficient in simulating diverse perspectives:

1) The persona variables typically accessible to researchers are group-level, while people form their identity based on both individual and group-level characteristics (Marsden and Pröbster, 2019). Therefore, there could be an inherent mismatch between the group-level variables we provide and individual perspectives we aim to simulate.

2) LLM generations can be understood as simulating the medium of a group, rather than individuals (West et al., 2023). Therefore, LLMs can have the tendency to represent a group as a monolith in simulation (Wang et al., 2024). While using more fine-grained group-level persona variables may in



(a) Tulu-2-dpo-70b

(b) Llama-2-7b-chat

Figure 3: Comparison of predicted R^2 and target R^2. Each point in the X-Y plane represents an experimental result with persona prompting, where the x-coordinate signifies the target R^2 and the y-coordinate denotes the predicted R^2. We then fit a linear regression line and also plot the theoretical maximum performance line $y = x$ in the same figure.

theory bring us closer to individual ratings, it remains to be seen whether this could lead to true individualization in practice.

We show the robustness experiment result in Table 5. The model performances are consistent across variations in the ordering and language use of the persona variables.

7 Conclusion and Recommendation

Our study reveals that persona variables account for less than 10% of variance in human annotations across most NLP datasets we considered. The use of persona prompting offers modest and inconsistent improvements across different tasks. The improvement is most produced in cases where the annotators largely disagree but only by a small margin (high entropy, low standard deviation). By running a case study with opinion survey data, we uncovered a linear relationship between target and predicted R^2 values. Alarmingly, when the target R^2 value falls below 0.1, the predicted R^2 often drops to zero. This could explain the small and inconsistent improvements observed in NLP tasks with persona prompting, as existing datasets often have R^2 values smaller than 0.1.

Given these insights, we have the following recommendations:

1) **Exercise Caution in LLM Simulation Work in NLP Settings** In light of our findings, we advise caution for researchers intending to use LLMs to simulate text annotations from different perspectives. Unvalidated, zero-shot simulations with LLMs may not yield reliable results.

2) **Implement More Strategic Dataset Design:** As persona variables in existing datasets account for only 10% of human annotation variation, deliberate and strategic dataset design is imperative for advancing human-centric NLP research. One approach could be to be more conscientious about the persona variables collected in future datasets and include more nuanced and target questions that probe individual-level characteristics such as attitudes and behaviors, as exemplified by Kumar et al. (2021), which as a result has a relatively high target R^2. However, this approach is not without challenges, including ethical concerns as well as receiving intentionally inaccurate responses to sensitive demographic questions in crowdsourcing (Huang et al., 2023). Furthermore, we call for the expansion of dataset collection efforts to include more diverse cultural perspectives, noting the scarcity of datasets that include annotator persona information from non-U.S. contexts, not to mention a multilingual one.

8 Limitations

While we exerted considerable effort to include a diverse range of datasets, the vast majority of

available datasets with persona information from annotators have been collected in the U.S., featuring persona questions primarily relevant to this particular context. Consequently, we can only speculate about the effectiveness of persona prompting for questions that are specifically tailored to other countries. Furthermore, to the best of our knowledge, we have not identified any datasets that include annotator persona variables in a language other than English. Considering that even the most sophisticated LLMs still exhibit significant performance disparities between English and non-English languages (Ahuja et al., 2023), it is highly probable that the ability of LLMs to simulate different perspectives based on persona information is considerably weaker in non-English languages. Additionally, many terms used to denote identities are deeply rooted in specific cultural and societal contexts, which cannot be readily translated into other languages. Thus, it is crucial to evaluate the simulation capabilities of an LLM independently for each language, without translation.

The zero-shot simulation ability of LLMs largely depends on their extensive training data, essentially a compressed digital snapshot of the internet. However, previous studies have indicated that the pretraining corpora used by LLMs are riddled with various social biases (Gao et al., 2020; Dodge et al., 2021; Bailey et al., 2022; Hu et al., 2023, *inter alia*). Consequently, LLM simulations could potentially be tainted by biases and stereotypes, among other issues.

We did not carry out extensive prompt engineering due to computational limitations and the targeted scope of our study. Instead, we presented the same prompts with persona information using language that closely mirrors how questions were asked of human participants. We believe this constitutes a fair setting for comparing LLMs. Additionally, we conducted a robustness check and found little variation for different persona variable orders and the exact wordings used to describe each variable (Section C).

9 Ethical Considerations

We utilize persona variables from publicly available datasets, which have been anonymized prior to their release. Therefore, no human participants were involved or personal data collected in this study. The research acknowledges the potential risks associated with the use of LLMs for simulation purposes, including issues such as identity fraud and manipulation. We sternly denounce such nefarious applications of this technology. We also acknowledge the concerns related to categorizing individuals into different demographic groups. However, we argue that our study merely utilizes existing datasets and does not involve any original data collection. Furthermore, the categorizations employed within these datasets adhere to established best practices, such as those used by the U.S. Census Bureau, thereby ensuring their appropriateness. In addition, the use of these demographic categories is only aimed at understanding and demonstrating the potential for LLMs to simulate diverse perspectives.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. ArXiv:2208.10264 [cs].

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

ANES. The american national election studies 2012 time series study.

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 - Supplementary Material. *Political Analysis*, 31(3):337–351.

April H. Bailey, Adina Williams, and Andrei Cimpian. 2022. Based on billions of words on the internet, PEOPLE = MEN. *Science Advances*, 8(13):eabm2463.

Larry M. Bartels. 2002. Beyond the running tally: Partisan bias in political perceptions. *Political Behavior*, 24(2):117–150.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (Not) to Use Sociodemographic Information for Subjective NLP Tasks. ArXiv:2309.07034 [cs].

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. Analyzing the Effects of Annotator Gender Across NLP Tasks. Technical report.

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. Artificially precise extremism: How internet-trained llms exaggerate our differences.

Lawrence Bobo and Frederick C Licari. 1989. Education and political tolerance: Testing the effects of cognitive sophistication and target group affect. *Public Opinion Quarterly*, 53(3):285–308.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868. Number: 6.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. ArXiv:2310.11501 [cs].

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA. Association for Computing Machinery.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600. Epub 2023 May 10.

Yi Ding, Jacob You, Tonja-Katrin Machulla, Jennifer Jacobs, Pradeep Sen, and Tobias Höllerer. 2022. Impact of Annotator Demographics on Sentiment Dataset Labeling. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):519:1–519:22.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. ArXiv:2306.16388 [cs].

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. ArXiv:2305.06626 [cs].

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. ArXiv:2101.00027 [cs].

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. ArXiv: 2202.02950.

Jesse Graham and Jonathan Haidt. 2010. Beyond beliefs: Religions bind individuals into moral communities. *Personality and social psychology review*, 14(1):140–150.

Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. How Crowd Worker Factors Influence Subjective Annotations: A Study of Tagging Misogynistic Hate Speech in Tweets. ArXiv:2309.01288 [cs].

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2023. Generative language models exhibit social identity biases. *arXiv preprint arXiv:2310.15819*.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Junsol Kim and Byungkyu Lee. 2023. AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys. ArXiv:2305.09620 [cs].

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large Language Models as Superpositions of Cultural Perspectives. ArXiv:2307.07870 [cs].

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. ArXiv:2106.04511 [cs].

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. Crehate: Cross-cultural re-annotation of english hate speech dataset. *arXiv preprint arXiv:2308.16705*.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139.

Nicola Marsden and Monika Pröbster. 2019. Personas and identity: Looking at multiple identities to inform the construction of personas. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Peter S Park, Philipp Schoenegger, and Chongyang Zhu. 2024. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, pages 1–17.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Pew Research Center. 2014. Political polarization in the american public. Pew Research Center. Accessed: Dec 19, 2023.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? ArXiv:2303.17548 [cs].

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models cannot replace human participants because they cannot portray identity groups. *CoRR*, abs/2402.01908.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. 2023. The generative ai paradox:" what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv:1910.03771 [cs].

A Implementation Detail

We run all experiments on a NVIDIA A100-SXM-80GB cluster. The OpenAI models are called through the OpenAI API in Python[4]. Due to resource constraint, inference with open-source models are with 4-bit quantization with QLora (Dettmers et al., 2023) implemented in bitsandbytes[5] in the Transformers library (Wolf et al., 2020). For the Llama-2 models, we directly prepend the persona prompts. For Llama-2-chat, Tulu-2, and Tulu-2-dpo models, we apply the persona prompts in the system variable. All regressions are performed in R. For the mixed-effect regressions, we use the performance library (Lüdecke et al., 2021) which implements the method proposed by Nakagawa and Schielzeth (2013) to obtain R^2 values.

B Prompt Template

AnnwAttitudes

```
**Your Profile**: You race/ethnicity is
    %s. Your gender identity is %s. You
    are %s years old. In terms of your
    political orientation, on a five-
    point Likert scale of -1 (Strongly
    liberal) to 1 (Strongly conservative
    ), you are %s.
**Question**: How toxic/hateful/
    disrespectful or offensive does this
     post seem to you? (Note: The
    purpose is to understand how
    disrespect/offense can show up in
    language, not making statements
    about the actual content of the
    posts.) "%s"
(A) not offensive at all
(B) mildly inappropriate
(C) somewhat offensive
(D) moderately offensive
(E) very offensive
**Answer**: (
```

Kumar et al. (2021)

```
**Your Profile**: You race is %s. Your
    gender is %s. You %s describe
    yourself as transgender. You
    consider yourself to be %s. You are
    %s years old. The highest degree or
    level of school that you have
    completed is %s. Religion is %s in
    your life. Generally speaking, you
    consider yourself politically %s.
    You %s a parent. You %s use social
    networking sites (Facebook, Twitter)
    . You %s use video sites (YouTube,
    Twitch). You %s use news sites (CNN,
     Fox, NYT, WSJ). You %s use
    community forums (Reddit, Craigslist
    , 4chan). In general, you rate the
    impact of technology on peoples
    lives as %s. Based on your
    experience, toxic posts or comments
    are %s.
**Question**: How toxic do you
    personally find this comment? "%s"
(A) Extremely toxic
(B) Very toxic
(C) Moderately toxic
(D) Slightly toxic
(E) Not at all toxic
**Answer**: (
```

EPIC

```
**Your Profile**: You ethnicity is %s.
    Your gender is %s. You are %s years
    old. Your country of birth is %s.
    Your country of residence is %s. You
     are a national of %s. You %s a
    student. You are %s.
Irony is a figurative language device
    that conveys the opposite of literal
     meaning, profiling intentionally a
    secondary or extended meaning.
For instance,
message: "if ur homeless u probably
    wouldn't have a phone."
reply: "Yes, and all your belongings
    would be in a handkerchief tied at
    the end of a stick." --> irony: yes
message: "if ur homeless u probably
    wouldn't have a phone."
reply: "Yes, you're right."--> irony:
    yes
**Question**: Is the reply ironic in the
     following message and reply pair?
message: "%s"
```

[4]https://github.com/openai/openai-python

[5]https://github.com/TimDettmers/bitsandbytes

```
reply: "%s"
(A) Ironic
(B) Not ironic
**Answer**: (
```

POPQUORN-P

```
**Your Profile**: In terms of race or
    ethnicity, you are %s. You are a %s.
     You are %s years old. Occupation-
    wise, you are %s. Your education
    level is %s.
**Question**: Consider you read this
    email from a colleague, how polite
    do you think it is?
**Email:**: "%s"
(A) not polite at all
(B) barely polite
(C) somewhat polite
(D) moderately polite
(E) very polite
**Answer**: (
```

C Robustness Test

For Kumar et al. (2021) and ANES, we perform a set of robustness checks. Specifically, we swap the order of the persona variables in the prompt five times (Order 1-5) or use GPT-4 to come up with five different paraphrases of the prompt template that are meant to have the same semantics (Semantics 1-5). The results are shown in Table 5. While there are variations between each Order or Semantics setting, the variations are very small.

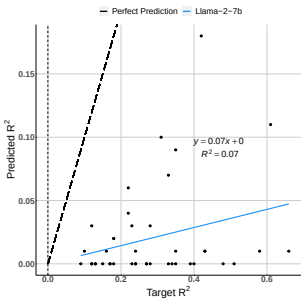
D More Details on Section 6

We include a list of the independent variables we considered in Section 6 as well as the associated R^2 in Table 4. Interested readers could refer to the ANES documentation to find out the exact survey questions asked in these variables. The persona template used is:

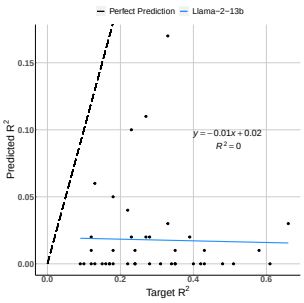
```
**It is 2012. Your Profile**: Racially,
    you are %s. You are a %s. You are %s
     years old. Ideologically, you are %
    s. Politically, you are %s. It makes
     you feel %s when you see the
    American flag flying. You %s. You
    are %s interested in politics and
    public affairs.
```

Model	annwAttitudes			Kumar et al. (2021)			EPIC			POPQUORN-P		
	$R^2\uparrow$	$\kappa\uparrow$	**MAE** $\downarrow$	$R^2\uparrow$	$\kappa\uparrow$	**MAE** $\downarrow$	$R^2\uparrow$	$\kappa\uparrow$	**F1** $\uparrow$	$R^2\uparrow$	$\kappa\uparrow$	**MAE** $\downarrow$
Target	0.64 (0.03)	-	-	0.42 (0.20)	-	-	0.28 (0.09)	-	-	0.47 (0.03)	-	-
GPT-4-0613	0.56	0.42	0.70	0.16	0.24	0.87	0.03	0.12	0.52	0.34	0.22	0.89
+Persona	0.53	0.40	0.74	0.12	0.20	0.90	0.05	0.20	0.58	0.33	0.22	0.90
GPT-3.5-Turbo-0613	0.53	0.29	0.80	0.12	0.17	1.12	0.04	0.18	0.59	0.28	0.09	1.07
+Persona	0.49	0.31	0.82	0.12	0.15	0.97	0.03	0.14	0.54	0.28	0.14	1.14
Llama-2-7b	0.07	-0.02	1.56	0.01	-0.01	2.91	-0.00	0.00	0.25	0.08	-0.04	1.21
+Persona	0.08	0.02	1.64	0.00	-0.01	1.08	0.00	0.02	0.29	0.04	-0.04	1.15
Llama-2-13b	0.11	0.07	1.50	0.00	0.00	2.91	-0.00	0.01	0.44	0.12	0.08	1.51
+Persona	0.02	0.04	1.55	0.00	-0.01	1.78	0.00	0.07	0.53	0.16	0.10	1.35
Llama-2-70b	0.17	0.14	1.70	0.01	0.04	1.51	0.00	0.00	0.24	0.24	0.13	1.42
+Persona	0.40	0.30	0.91	0.03	0.05	1.01	0.00	0.00	0.24	0.21	0.17	1.10
Llama-2-7b-chat	0.25	0.01	1.43	0.00	-0.04	2.03	0.00	0.00	0.41	0.18	0.02	1.07
+Persona	0.32	0.01	1.41	-0.00	-0.00	1.44	-0.00	0.02	0.47	0.10	0.00	1.06
Llama-2-13b-chat	0.29	0.03	1.39	0.07	-0.01	1.84	0.00	0.00	0.41	0.07	0.01	1.06
+Persona	0.17	0.02	1.44	0.03	-0.00	1.46	0.00	0.00	0.41	0.06	0.02	1.01
Llama-2-70b-chat	0.39	0.13	1.33	0.11	0.07	1.70	0.00	0.05	0.49	0.32	0.15	1.00
+Persona	0.42	0.15	1.22	0.10	-0.01	1.45	0.02	0.14	0.56	0.31	0.14	0.90
Tulu-2-7b	0.33	0.04	1.37	0.02	-0.01	2.63	-0.00	0.00	0.25	0.06	0.06	1.10
+Persona	0.35	0.06	1.37	0.01	-0.08	1.31	0.00	0.01	0.27	0.08	0.05	1.07
Tulu-2-13b	0.36	0.12	1.45	0.09	0.05	2.16	0.03	0.15	0.56	0.26	0.07	1.35
+Persona	0.33	0.10	1.34	0.11	0.06	1.42	0.03	0.14	0.52	0.27	0.14	1.02
Tulu-2-70b	0.49	0.29	0.90	0.16	0.13	1.09	0.05	0.20	0.59	0.34	0.20	0.89
+Persona	0.49	0.26	0.88	0.14	0.16	0.90	0.07	0.27	0.63	0.31	0.16	0.92
Tulu-2-dpo-7b	0.38	0.08	1.34	0.04	0.06	1.81	0.00	0.02	0.29	0.08	0.07	1.26
+Persona	0.39	0.09	1.38	0.03	-0.02	1.20	0.01	0.02	0.28	0.08	0.06	1.26
Tulu-2-dpo-13b	0.33	0.13	1.47	0.11	0.07	1.85	0.01	0.11	0.55	0.29	0.11	1.21
+Persona	0.34	0.13	1.28	0.10	0.10	1.32	0.03	0.17	0.57	0.28	0.18	0.93
Tulu-2-dpo-70b	0.51	0.35	0.84	0.15	0.15	1.16	0.03	0.14	0.54	0.35	0.21	0.83
+Persona	0.51	0.30	0.84	0.15	0.20	0.92	0.04	0.18	0.58	0.33	0.19	0.87
Avg. Δ	0.00	0.01	-0.07	-0.01	-0.01	-0.60	0.01	0.03	0.02	-0.01	0.01	-0.08

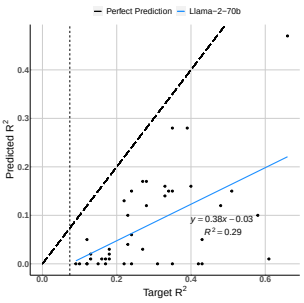
Table 3: Comparison of performance across LLMs in estimating individual annotations, with and without persona prompting. Performance is measured using R^2 for regression annotation prediction, Cohen's Kappa (κ), and Mean Absolute Error (MAE).



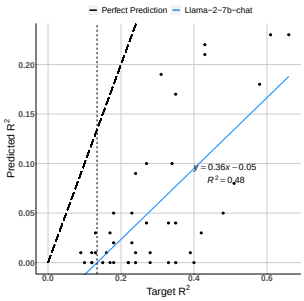
(a) Llama-2-7b



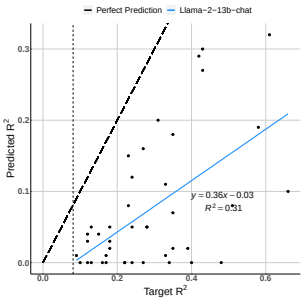
(b) Llama-2-13b



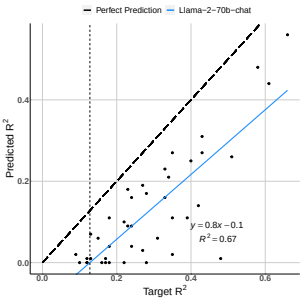
(c) Llama-2-70b



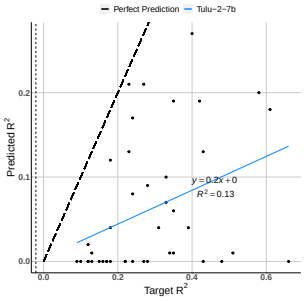
(d) Llama-2-7b-chat



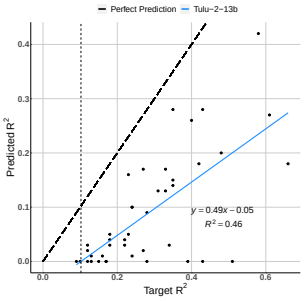
(e) Llama-2-13b-chat



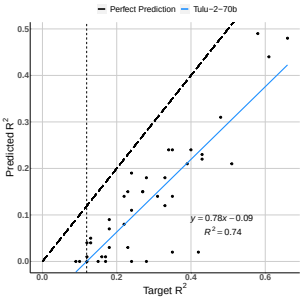
(f) Llama-2-70b-chat



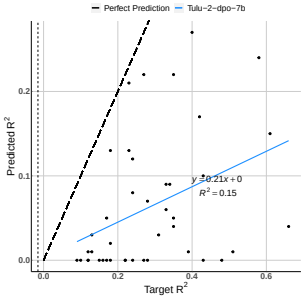
(g) Tulu-2-7b



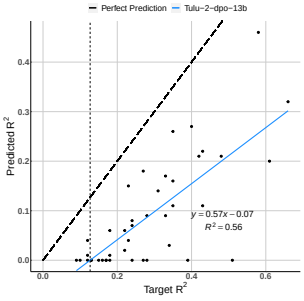
(h) Tulu-2-13b



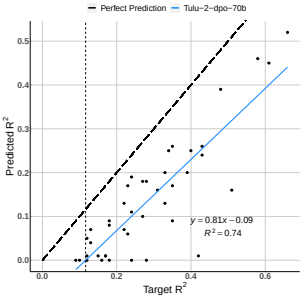
(i) Tulu-2-70b



(j) Tulu-2-dpo-7b



(k) Tulu-2-dpo-13b



(l) Tulu-2-dpo-70b

Figure 4: Comparison of predicted R^2 and target R^2. Each point in the X-Y plane represents an experimental result with persona prompting. We then fit a linear regression line and also plot the theoretical maximum performance line $y = x$ in the same figure.

Variable	Target R^2
aidblack_self	0.35
ecblame_dem	0.43
ecblame_fmpr	0.51
effic_undstd	0.23
ecblame_pres	0.61
egal_toofar	0.34
gayrt_adopt	0.24
gayrt_marry	0.33
govrole_big	0.43
ident_amerid	0.35
immig_checks	0.22
interest_following	0.27
nonmain_bias	0.28
presapp_econ	0.66
presapp_foreign	0.58
prmedia_attvnews	0.28
ptywom_bettrpty	0.42
relig_pray	0.40
resent_deserve	0.39
spsrvpr_ssself	0.48
trad_famval	0.33

Table 4: List of variables considered for the experiment and the associated Target R^2 in Section 6.

Model	**annwAttitudes**			ANES		
	$R^2\uparrow$	$\kappa\uparrow$	**MAE** $\downarrow$	$R^2\uparrow$	$\kappa\uparrow$	**F1**$\uparrow$
Target	0.64 (0.03)	-	-	0.50	-	-
Llama-2-70b	0.01	0.04	1.51	0.00	0.00	0.00
+Persona (Default)	0.03	0.05	1.01	0.33	0.19	0.26
Order-1	0.03	0.03	1.03	0.36	0.19	0.26
Order-2	0.03	0.04	1.01	0.32	0.18	0.26
Order-3	0.04	0.08	1.00	0.29	0.18	0.26
Order-4	0.03	0.05	1.01	0.28	0.18	0.26
Order-5	0.04	0.12	0.97	0.39	0.19	0.26
Semantics-1	0.01	0.00	1.05	0.30	0.19	0.26
Semantics-2	0.01	0.02	1.04	0.36	0.20	0.27
Semantics-3	0.01	0.00	1.05	0.31	0.19	0.26
Semantics-4	0.01	0.01	1.05	0.28	0.18	0.25
Semantics-5	0.03	0.03	1.02	0.29	0.18	0.26
Llama-2-70b-chat	0.11	0.07	1.70	0.00	0.00	0.00
+Persona (Default)	0.10	-0.01	1.45	0.30	0.19	0.26
Order-1	0.11	-0.01	1.49	0.34	0.20	0.26
Order-2	0.10	-0.01	1.46	0.34	0.20	0.26
Order-3	0.12	-0.01	1.45	0.29	0.18	0.26
Order-4	0.10	0.00	1.44	0.28	0.18	0.26
Order-5	0.11	0.00	1.46	0.39	0.21	0.27
Semantics-1	0.10	-0.01	1.45	0.27	0.18	0.25
Semantics-2	0.11	-0.01	1.46	0.28	0.18	0.26
Semantics-3	0.11	-0.01	1.43	0.27	0.18	0.25
Semantics-4	0.10	-0.00	1.40	0.28	0.18	0.25
Semantics-5	0.11	-0.01	1.44	0.30	0.19	0.26
Tulu-2-70b	0.16	0.13	1.09	0.00	0.00	0.00
+Persona (Default)	0.14	0.16	0.90	0.35	0.19	0.26
Order-1	0.13	0.16	0.92	0.33	0.18	0.26
Order-2	0.14	0.17	0.90	0.33	0.18	0.26
Order-3	0.14	0.14	0.94	0.35	0.19	0.26
Order-4	0.12	0.15	0.92	0.38	0.20	0.27
Order-5	0.13	0.13	0.94	0.34	0.18	0.26
Semantics-1	0.12	0.15	0.92	0.39	0.20	0.27
Semantics-2	0.12	0.16	0.93	0.42	0.22	0.27
Semantics-3	0.14	0.15	0.93	0.36	0.19	0.26
Semantics-4	0.13	0.14	0.92	0.38	0.21	0.27
Semantics-5	0.13	0.15	0.92	0.37	0.19	0.26
Tulu-2-dpo-70b	0.15	0.15	1.16	0.00	0.00	0.00
+Persona (Default)	0.15	0.20	0.92	0.36	0.20	0.27
Order-1	0.15	0.21	0.92	0.34	0.19	0.26
Order-2	0.15	0.20	0.92	0.35	0.19	0.26
Order-3	0.16	0.18	0.94	0.36	0.20	0.27
Order-4	0.16	0.22	0.90	0.34	0.20	0.26
Order-5	0.17	0.20	0.94	0.35	0.18	0.26
Semantics-1	0.15	0.20	0.91	0.37	0.21	0.27
Semantics-2	0.16	0.21	0.95	0.38	0.21	0.27
Semantics-3	0.16	0.20	0.94	0.37	0.20	0.27
Semantics-4	0.16	0.20	0.91	0.31	0.19	0.26
Semantics-5	0.15	0.20	0.93	0.38	0.21	0.27

Table 5: Robustness test of LLMs in terms of swapping order of persona variables and paraphrase the text description of persona variables. Performance is measured using R^2 for regression annotation prediction, Cohen's Kappa (κ), Mean Absolute Error (MAE) and F1 score.