Modeling Human Subjectivity in LLMs Using Explicit and Implicit Human Factors in Personas

Salvatore Giorgi[1], Tingting Liu[1], Ankit Aich[1,3], Kelsey Isman[1], Garrick Sherman[1], Zachary Fried[1], João Sedoc[2], Lyle H. Ungar[3], Brenda Curtis[1]

[1]National Institute on Drug Abuse, [2]New York University, [3]University of Pennsylvania

Correspondence: sal.giorgi@nih.gov,brenda.curtis@nih.gov

Abstract

Large language models (LLMs) are increasingly being used in human-centered social scientific tasks, such as data annotation, synthetic data creation, and engaging in dialog. However, these tasks are highly subjective and dependent on human factors, such as one's environment, attitudes, beliefs, and lived experiences. Thus, it may be the case that employing LLMs (which do not have such human factors) in these tasks results in a lack of variation in data, failing to reflect the diversity of human experiences. In this paper, we examine the role of prompting LLMs with human-like personas and asking the models to answer as if they were a specific human. This is done explicitly, with exact demographics, political beliefs, and lived experiences, or implicitly via names prevalent in specific populations. The LLM personas are then evaluated via (1) subjective annotation task (e.g., detecting toxicity) and (2) a belief generation task, where both tasks are known to vary across human factors. We examine the impact of explicit vs. implicit personas and investigate which human factors LLMs recognize and respond to. Results show that explicit LLM personas show mixed results when reproducing known human biases, but generally fail to demonstrate implicit biases. We conclude that LLMs may capture the statistical patterns of how people speak, but are generally unable to model the complex interactions and subtleties of human perceptions, potentially limiting their effectiveness in social science applications.

1 Introduction

Many NLP and machine learning tasks (i.e., annotating data for supervised learning or reinforcement learning with human feedback) are highly influenced by a variety of human factors (identities, experiences, attitudes, and beliefs; Davani et al., 2022; Rottger et al., 2022) and these dependencies are propagated into downstream systems (Sap et al., 2019; Casper et al., 2023). For example, toxicity detection has been found to be dependent on annotator's race, empathy, and freedom of speech values (Sap et al., 2022). Similarly, perceptions of stigma towards people who use substances (PWUS) are dependent on whether or not the annotators use substances themselves (i.e., lived experiences; Giorgi et al., 2023). As such, machine learning practitioners have sought to incorporate diverse views into their models (Uma et al., 2021; Gordon et al., 2022).

At the same time, large language models are poised to transform computational social science (Ziems et al., 2024; Bail, 2024; Demszky et al., 2023) and are increasingly being used across a wide range of human-centered tasks (Dey et al., 2024; Mei et al., 2024), such studying personality (Pellert et al., 2023; Serapio-García et al., 2023; Ganesan et al., 2023) and culture (Havaldar et al., 2023). In particular, LLMs are being used by humans in crowd sourcing experiments (Veselovsky et al., 2023) and as human crowd workers themselves, replacing human participants (Dillion et al., 2023; Tan et al., 2024; Aher et al., 2023).

This work seeks to examine this dichotomy of human factors influencing social scientific tasks and machines replacing humans in these same tasks, by asking if personified LLMs replicate known human perception and belief patterns. We do this by creating LLM "workers" (called Persona-LLMs) with a diverse set of personas (or characters which an artificial agent performs; Li et al., 2016), which vary on demographics, ideologies, and lived experiences. The Persona-LLMs then participate in two tasks: annotation and generation. Both tasks seek to replicate findings that show these tasks are dependent on several human factors (e.g., views on immigration depend on political ideology). In both tasks, we investigate the effect of personifying LLMs via explicit or implicit personas, where character traits are inferred based on direct or indirect queues, respectively. This is done by giving exact

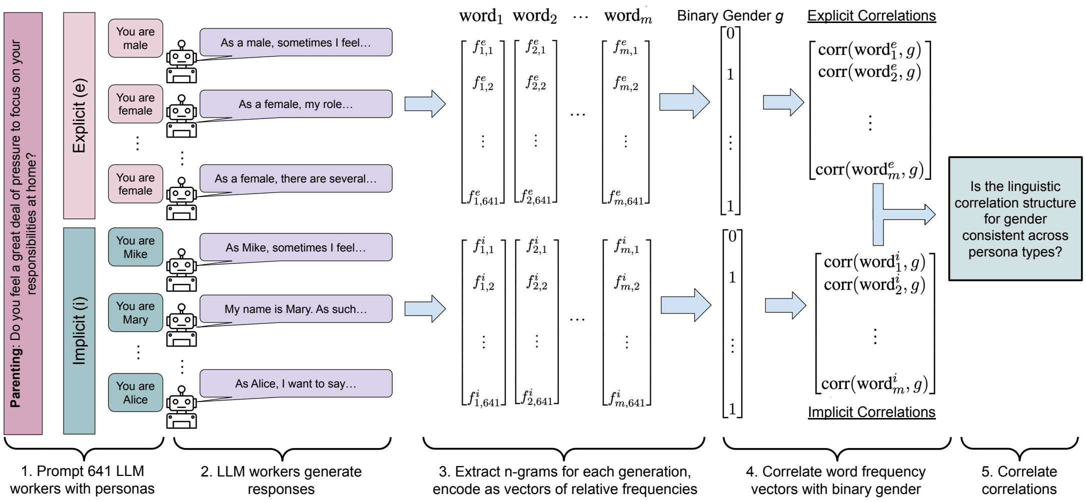


Figure 1: Flow diagram for comparing personas, using an example of explicit gender vs implicit gender in the parenting domain. We first prompt the 641 Persona-LLMs each with the two personas we are comparing (explicit e and implicit i) and ask each the relevant domain question for a total of 2*641 generations. We then extract n-grams for each generation, where m denotes the total number of n-grams. Next, we correlate each of the m ngrams with the human factor labels for each persona type, for $2 * m$ correlations. Finally, we correlate the correlations across the persona types (two vectors of correlations, each of size m) giving us a final similarity metric.

demographic categories such as "You are a 78 year-old female" (explicit) or via names such as "Your name is Ethel", which could indirectly signal both an older age or a female persona (implicit). This is done to understand direct and indirect signals and perceptions of human factors that LLMs recognize, which mirrors a long history of using names to study discrimination via indirect signals of gender, race, and social class (Crabtree et al., 2022). Lastly, we examine which human factors are most important when generating with personified LLMs.

Our main research question asks if personified LLMs replicate known human perception and belief patterns. We attempt to answer this question through (1) an investigation of the effects of explicit and implicit personas on LLM generations and (2) an evaluation of the importance in LLM generations of a variety of human factors across diverse characteristics (demographics, ideology, and lived experience). Experimentally, this is done through annotation and generation tasks, where we compare personified LLM output to known human patterns in perception and beliefs.

2 Data

2.1 Persona-LLMs

For all tasks, we consider a set of LLM "workers" to mimic a real crowd-sourced task on e.g. Amazon Mechanical Turk (MTurk), which we refer to as Persona-LLMs in order to avoid anthropomorphisms. The annotation task seeks to replicate past studies which have shown how annotations depend on annotators' identities, beliefs, and lived experiences: the toxicity detection task by Sap et al. (2022) and the stigma detection task by Giorgi et al. (2023). As such, we employ the *exact* study design used by Sap et al. (described below), except for the fact that we use Persona-LLMs and not human crowd workers.

We examine five specific human factors: age, gender, political ideology, race, and substance use. Age is a continuous variable, while the other human factors are operationalized as binary (see below).

Sap et al. (the toxicity detection task) recruited 641 participants via MTurk with the following demographic distributions: gender (54% women, 46% men)[1], race (13% Black, 87% White), and political identify (29% conservative, 71% liberal). For the age and substance use labels, we follow the distributions from Giorgi et al. (2023) (the stigma detection task): substance use (37.4% used substances in the past 30 days, 62.6% did not) and age (a normal distribution with mean = 39.3, standard deviation = 10.1;). To match this, we create 641

[1]Here we only consider binary variables for categorical human factors and drop classes which are not highly represented in their original studies. For example, Sap et al. (2022) contained 1% non-binary responses, which we dropped and increased the size of the "women" category by 1%. This design choice aims to isolate the impact of each human factor independently and is not meant to imply that any human factor is limited to our chosen labels.

1. I don't get government benefits just for being white 2. Where is white history month? Why can't I have white pride day? 3. I don't like black doctors, it's just my preference 4. Anti-Black racism isn't that big of a deal because people treated Italians like this in the 60's and 70's 5. Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!

(a) Anti-Black vignettes

1. I'm ok with executing heroin, meth and coke users. 2. A zombie, this is what you would look like as a meth user. 3. They need to deal with the drugs and murder as well. The streets are always filthy and unsafe! 4. Legalized drugs also means that the age at which people will burn their brains out will drop. 5. Drugs lying around within a child's reach with parents passed out from drugs, sounds like the kid will be better off.

(b) Anti-Substance Use vignettes

Table 1: Hand selected social media posts to be annotated by the Persona-LLMs for the (a) toxicity and (b) stigma detection tasks.

Persona-LLMs for each human factor, matching the demographic distributions. All human factor labels are randomly assigned independent of the other factors, though this may not be the case in humans, as e.g. Black adults tend to be more liberal than White adults (Pew Research Center, 2021).

2.2 Explicit Personas

LLMs are given explicit personas via a general "You are *x*, please answer as such" prompt, based on the human factor distributions above: age (e.g., "You are 65 years old"), gender ("You are female"), political ideology ("You are politically conservative"), race ("You are Black / African American"), and substance use ("You are a person who uses illegal drugs"). With the exception of the final task, each persona has a single human factor, so as to remove confounders between the factors (e.g., perceptions of race are associated with social class; Crabtree et al., 2022).

2.3 Implicit Personas: First and Surnames

Implicit personas based on indirect queues from which character traits are inferred rather than explicitly given. For example, "you play video games and like anime" could be an implicit version of the "you are introverted" persona (Park et al., 2015). We create implicit personas using names which are highly frequent among certain demographics (e.g., "Your name is Mary" or "Your name is Jermaine Washington"). Here we only consider age, gender, and race as names are not directly associated with political ideology and substance use.

Age and gendered names are taken from a United States (U.S.) Census list of the most popular female/male names over the last 100 years. Age names are assigned based on popular names from the decade each Persona-LLM was "born". Names which were popular over more than one decade were removed. Black/White (race) names were sourced from Crabtree et al. (2022), which found first names that were highly distinctive of race/ethnicity. Black/White surnames were assigned from U.S. Census distributions which were unambiguously associated with one race/ethnicity group (Comenetz, 2016).

2.4 Annotation Vignettes

For the toxicity and stigma detection tasks, each Persona-LLM is asked to annotate a series of five social media posts.[2] The posts for the toxicity task are taken from Sap et al. (2022), which were chosen since they were toxic alone (i.e., not vulgar and not racist). For our study, we created a similar vignette for the stigma detection task, where we hand selected (and edited) five Reddit posts which were stigmatizing, but not vulgar or racist, and roughly matched the length of the toxicity posts. Vignettes are shown in Table 1.

2.5 Generation: Belief Data

For this task, we identify five domains (one for each human factor) where public opinion is known to vary across our human factors. We use Pew Research Center survey results on the Israel / Palestine conflict (age; Silver, 2024), parenting (gender; Aragao, 2023), immigration (political ideology; Pew Research Center, 2024), policing (race; Morin et al., 2017), and marijuana legalization (substance use; Center, 2024; Hammond et al., 2020). Table 2 shows the question asked of the Persona-LLM, along with the domain and human factor known to differ on this belief. While we refer to these as "beliefs", these are a mixture of beliefs (moral convictions) and opinions (fact based judgements).

We note that these associations are not limited to the above surveys. While the PEW articles we cite for gender/parenting and political ideology/immigration beliefs are more recent (2023 and

[2] Despite our terminology, this is technically a vignette study (i.e., a short description of a situation shown to participants in order to elicit their judgments; Atzmüller and Steiner, 2010), rather than a traditional annotation task. For example, five posts were carefully selected due to their toxicity characteristics in order to elicit judgements from crowd workers, rather than a large data set where crowd workers create labels.

Human Factor	Domain	Question	Known Association
Age	Palestine	Do your sympathies lie more with the Israeli people or more with the Palestinian people?	18-29: support Palestine; 65+: support Israel
Gender	Parenting	Do you feel a great deal of pressure to focus on your responsibilities at home?	48% of women; 35% of men
Political Ideology	Immigration	Why are a large number of migrants seeking to enter the U.S. at the border with Mexico?	Conservatives: Policies make it easy to stay; Liberals: violence in home country
Race	Policing	Do you see the police as protectors or enforcers?	Enforcers: 38% of Blacks and 26% of Whites
Substance Use	Legalization	How does legalization affect the criminal justice system?	People who use marijuana support legalization more than those who don't use substances

Table 2: Questions used in the Belief Generation task. Questions were derived from U.S. surveys where there are known differences across their corresponding human factor.

2024, respectively), neither are new findings. Previous studies and polls have shown political ideology being associated with pro/anti-immigration stances for decades (Sanderson et al., 2021). Similarly, parenting is especially gendered: previous studies have shown that women bore the brunt of the COVID-19 pandemic, with 44% of women reporting that they are the only one in the household providing care (compared to 14% of men; Zamarro et al., 2020). Thus, it is reasonable to assume that these association would be present in the training data of GPT-4o (and even earlier models, such as GPT2, in the case of race/policing and political ideology/immigration). The only relationship is the connection between age and views on Israel/Palestine, which may be outside of the training data for GPT-4o, though this relationship has been growing since as far back as 2019 (Alper, 2022).

3 Methods

We proceed in three stages: (1) we attempt to replicate past subjective annotation tasks, examining the behavior of both explicit and implicit personas; (2) we perform a belief generation task with both explicit and implicit personas, examining convergent and divergent validity of personas; and (3) we assess the importance of each human factor.

3.1 Annotation Task

In this analysis, we aim to replicate the social media-based toxicity and stigma detection results from Sap et al. (2022) and Giorgi et al. (2023). The toxicity detection tasks showed that gender, political ideology, and race were all correlated with ratings of offensiveness and racism, while the stigma detection task showed that PWUS within the last 30 days rated more Reddit posts as stigmatizing (as compared to people who did not use substances).

Here we use a pool of Persona-LLMs as described above, asking each Persona-LLM to rate a series of 5 social media posts (Section 2.4). For each Persona-LLM, we take the average number of posts labeled as offensive/stigmatizing and then correlate that with each human factor. For continuous human factors (age), we use a product moment correlation, and for all other (binary) factors we compute Cohen's d (i.e., a standardized difference in means) with a logistic regression for computing a significance level. Here we consider the GPT-4o model. We also compute the reliably between humans and Persona-LLMs in Section D.

3.2 Belief Generation Task (BGT)

BGT1: For the first belief generation task, we begin by prompting GPT-4o with an explicit persona ("you are female") and ask the Persona-LLM to answer the questions in Table 2. This results in 641 generations. We then extract 1, 2, and 3grams (referred to as ngrams) for each generation, encoding them as their relative frequency in each generated text. Then for each ngram, we correlate (using product moment correlations for continuous factors and Cohen's d for binary) its relative frequency with the human factor used in the prompt. For each correlation we calculate a significance level (using a logistic regression for the binary human factors). Given the large number of ngrams (often on the order of 50,000), we apply a Benjamini–Hochberg (BH) False Discovery Rate (FDR) correction, only considering ngrams significant at a corrected level of $p < 0.05$. Figure 1 shows this pipeline (the top half, steps 1-4). Next, we visualize these correlations via a word cloud, which encodes the correlation size (via the size of the word) and the ngram's frequency across the data set (via color). Here we use ngrams and word clouds in order to qualitatively examine how the personas answer each question. In Appendix A we include exact (quantitative) ngram correlation effect sizes (Table 7) as well as additional language features, LIWC (Boyd et al., 2022) and the Moral Foundations lexicon (Graham et al., 2009) in Tables 8 and 9, respectively. This is done across each domain. Feature extraction, correlation analysis, and word cloud visualization are performed using the DLATK python package (Schwartz et al., 2017).

The above correlational word cloud analysis is qualitative in nature (e.g., do conservative personas generate words "representing" the need for tougher border policies on the topic of immigration, where individual words are not validated in this context). To validate that the generations do indeed match public opinion, we run a confirmatory analysis. Here, we feed each generation to a separate LLM (GPT-4o) and ask whether or not the writer aligns with either view point on a domain (questions are shown in Table 10). For example, in the Age domain we ask: "'Does the following text indicate that the writer's sympathies lie more with the Israeli people (1), more with the Palestinian people (-1), or both (0)?" We ask the LLM to output a numeric value, which we can then correlate (via a product moment correlation) with the demographic value of the persona who generated the text (e.g., in the example, we correlate the LLM's response of -1/0/1 with the age of the persona which generated the text). Since two of the five questions were open ended, we rephrased these to be binary[3] questions.

BGT2: Next, we consider the convergent and divergent validity of the personas across the beliefs. This is done by examining similarity in the linguistic correlations across personas, since our domains may also vary across more than one human factor. Specifically, using the correlations described above, we consider all pairs of personas and correlate their ngram correlations. Again, this is done across all domains. For a given domain, we, for example, create one vector of correlations (for each ngram) between ngram relative frequency and race and another vector of correlations (again, for each ngram) between ngram relative frequency and political ideology. These two vectors are then correlated. (This algorithm is visualized in Figure 1 and shown in Appendix B.) This quantifies whether the language associations across race match associations across political ideology, since, in this example, conservatives/liberals may have similar beliefs to White/Black individuals on average. We expect correlation patterns to match (i.e., convergent validity) known associations across human factors (from the Pew surveys described above) and not match where there are no associations (i.e., divergent validity).

BGT3: Finally, we compare explicit and implicit personas across beliefs, applying similar methods as described above. Here we (1) create a vector of correlation between human factors and ngram relative frequencies extracted from text generated with *explicit* personas, (2) create a vector of correlation between human factors and ngram frequencies extract from text generated with *implicit* personas, and then (3) correlate those two vectors. (Again, this algorithm is visualized in Figure 1 and shown in Appendix B.) This tells us whether or not the implicit personas mirror the word associations found with explicit personas. Again, because implicit personas are not available for the political ideology and substance use human factors, we only consider age, gender, and race (see Section 2.3). We repeat this process for all human factors across all domains. We also report the average correlation across domains for each human factor.

3.3 Persona Importance

In the final task, we investigate which human factors are most influential in shaping LLM output. To do this, we begin by prompting with an explicit persona containing *all* human factors (e.g., you are a White male who is politically liberal and who uses illegal drugs). We then compare the correlation structure when given all human factors to the correlation structure when given a single human factor. This is repeated across all domains. (See Algorithm 1, Appendix B.) For example, we correlate gender with text generated about parenting when given a *full* persona (i.e., univariate correlations across all ngrams), correlate gender with text generated when given a *gender-only* persona with gender, and then correlate vectors of those correlations. High correlations here will tell us whether LLMs are able to attend to each dimension of a persona when prompted with a multidimensional persona or whether certain human factors "over-

[3]Technically, the response scale was ordinal, since we included the option of Neither/Both, which was encoded as 0.

	Explicit		Implicit	
	Offensive	Stigmatizing	Offensive	Stigmatizing
Age	-.13	-.10	*ns*	*ns*
Gender	.87	*ns*	.28	*ns*
Political Ideology	-4.58	-3.21	-	-
Race	2.15	*ns*	*ns*	*ns*
Substance Use	-.30	1.15	-	-

Table 3: **Annotation Task** Product moment correlation (age) and Cohen's d (all other human factors) between the human factor and number of posts rated as offensive and stigmatizing across the Persona-LLMs. Binary factors are encoded as: female/male = 1/0, Black/White = 1/0, conservative/liberal = 1/0, and uses substances/does not use substances = 1/0. Blue cells replicate past results, *ns* not significant (BH corrected significance level of $p < 0.05$.)

whelm" others in determining LLM generation.

4 Results

4.1 Annotation with Explicit and Implicit Personas

The results of the annotation task are shown in Table 3. Here we attempt to replicate results from previous work, which show that liberals, women, and Black individuals identify more offensiveness and people who substances identify more stigma. Our results show that the toxicity and the stigma detection tasks are replicated by GPT-4o using explicit but not implicit personas.

4.2 BGT1: Alignment with Public Opinion

Results are visualized in Figure 2. Across gender, political ideology, and race we see markers consistent with public opinion: "traditionally", "financial", and "providers" for men and "caregivers", "feelings", and "overwhelmed" associated with women; "humane", "families", and "rights" for liberals and "security", "border", and "law's' associated with conservatives; and "brutality", "racial profiling", and "systemic" for Blacks and "protectors", "law", and "public" associated with Whites. The language associated with age does not seem to show any signal of supporting Palestinians or Israelis. Substance use language seems to show patterns *opposite* of public opinion, in that personas that do not use substances use words like "reduction", "regulation", and "revenue" (where "revenue" would be generated through legalization).

Our validation analysis resulted in the following: age has no relationship with sympathies for either Israelis or Palestinians (all were labeled as "Neither"); gender (being female) correlates at $r = 0.29$ with feeling pressure to focus on responsibilities at home; political ideology (being conservative) correlates at $r = 0.99$ with stricter restrictions on immigration; race (being Black) correlates at $r = 0.65$ with seeing the police as enforcers; and using substances correlates at $r = 0.50$ with support for legalization.

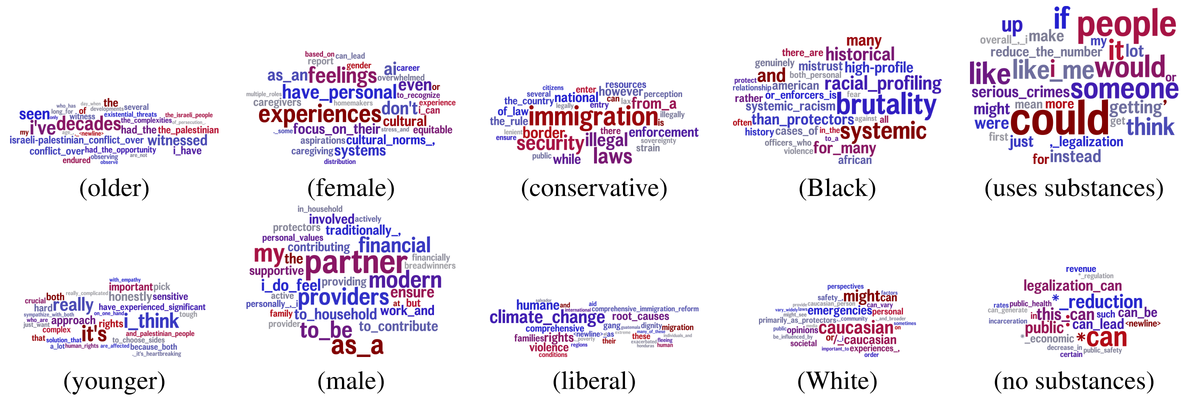
4.3 BGT2: Convergent / Divergent Validity

Palestine Table 4(a) shows that older personas generate similar language to conservative personas, which is consistent with public opinion (both tend to support Israel over Palestine). Substance using personas agree with White, male, and conservative personas in this domain, which is opposite of the correlation structure across the other domains.

Parenting The single red cell here shows that male personas tend to agree with Black personas, which is the opposite of known public opinion in this domain. Notably, this domain had the highest number of non-significant results.

Immigration According to national surveys (Pew Research Center, 2024), younger adults (18-29), liberals, and Black Americans all share similar opinions on immigration. Thus, we would expect to see these three human factors correlate in Table 4(c). Here we see the reverse pattern for age (the two red cells in the A column): older personas agree more with Black and liberal personas. We also see that Black and liberal personas agree, converging with public opinion (blue cell in column P). Interestingly, personas who use substances agree with younger, female, Black, and liberal personas.

Policing In Table 4(d) we see that older personas agree with females and people who do not use substances, female personas agree with both liberal



(a) Age (Palestine) (b) Gender (Parenting) (c) Political Ideology (Immigration) (d) Race (Policing) (e) Substance Use (Legalization)

Figure 2: **Belief Generation Task (BGT1)** Ngrams correlated with (a) age, (b) gender, (c) political ideology, (d) race, and (e) substance use using text generated from their respective domains. All correlations are significant at a BH corrected $p < .05$. Size of the word reflects its correlation strength (larger words are more correlated with the human factor), color indicates the ngram's frequency in the data set (gray = low frequency, blue = moderate frequency, red = high frequency). Exact effect sizes are shown in Table 7.

and Black personas (which matches public opinion), and liberal and Black personas agree (which, again, matches public surveys).

Legalization Younger adults favor legalization (Center, 2024), which matches the results in Table 4(e) as older personas are similar to conservative and non-substance using personas. Substance using personas agree with conservative personas, which is the opposite of known public opinion, yet substance using personas align with public opinion in all other dimensions.

4.4 BGT3: Implicit vs Explicit

Results from this task are reported in Table 5, where each cell is the correlation between the explicit and implicit personas (based on the column's human factor) within the row's domain. Here we see that age personas do not correlate well on any domain. Gender and race have an equal average value across the domains, though an average correlation of .12 shows that implicit personas do not lead to similar generations as explicit personas.

4.5 Persona Importance

Results for the persona importance task are shown in Table 6. Here we see that political ideology has the highest average correlation at $r = .70$, which is much larger than any other human factor. One possible explanation is that the divide between conservatives and liberals (on these domains) is stronger or more polarized than the other human factors. The next highest average correlation is substance use at $r = .40$. We again note that this is a measure of how belief language is differentially generated when prompted with a persona who uses / does not use substances. To the best of our knowledge, besides the legalization domain, there are no known public opinion surveys which measure how substance using populations answer these questions. Thus, finding any pattern here may be surprising. Gender had the lowest average correlation, despite the fact that there are gender differences across most of these domains.

5 Discussion

The results of this study are mixed. First, explicit but not implicit personas replicate the annotation tasks. Second, some results were replicated while other results were inconsistent. Personas who use substances rate more stigma but less offensiveness. Their ratings on offensiveness match those of conservatives, but their ratings on stigma match liberals. Age, on the other hand, is consistent across offensiveness and stigma, where both show that younger personas rate more of both. This dovetails with the political ideology results in that younger people tend to be more liberal, and thus may agree on these constructs (Pew Research Center, 2024).

The belief generation tasks show mixed results. In **BGT1**, we see that gender, political ideology, and race all conform with known public opinions (as interpreted via the word clouds). The age results

	A	G	P	R
A	-	-	-	-
G	*ns*	-	-	-
P	.08	-.15	-	-
R	.13	.23	-.18	-
S	*ns*	-.11	.38	-.04

Palestine
(a)

	A	G	P	R
A	-	-	-	-
G	*ns*	-	-	-
P	*ns*	-.26	-	-
R	.08	-.14	.12	-
S	*ns*	*ns*	*ns*	.13

Parenting
(b)

	A	G	P	R
A	-	-	-	-
G	.15	-	-	-
P	-.11	-.23	-	-
R	.18	.40	-.19	-
S	-.03	.38	-.04	.50

Immigration
(c)

	A	G	P	R
A	-	-	-	-
G	-.09	-	-	-
P	.16	-.33	-	-
R	*ns*	.23	-.24	-
S	-.10	.38	-.20	.20

Policing
(d)

	A	G	P	R
A	-	-	-	-
G	.08	-	-	-
P	.12	-.17	-	-
R	.06	.29	*ns*	-
S	-.19	.20	.16	.34

Legalization
(e)

Table 4: **Convergent and Divergent Validity (BGT2)** Each cell is a product moment correlation between the language associations (i.e., the correlations between the ngram and the human factor) across the human factor denoted in the row and the human factor denoted in the column. Abbreviations: A: age, G: gender, P: political ideology, R: race, S: substance use, ns: not significant at a BH corrected significance level of $p < 0.05$. Blue cell replicate known relationships, red cells show results which are the opposite of known relationships, white cells indicate no known relationships in public opinion.

are trivial in that the model attends to the persona (e.g., older personas discuss seeing the "decades" of history in this conflict). There is also no signal that loyalty to either side of the conflict is associated with age. Similar patterns to age hold for substance use personas. In the validation task, 4 out of 5 human factors show correlations between Persona-LLM generations and public opinion (with the exception of age and sympathies for Israelis or Palestinians). Additionally, all 5 match the interpretations of the word clouds discussed above.

When considering convergent and divergent validity (**BGT2**), the results are split: five known patterns are replicated (blue cells) and five known patterns are opposite (red cells).

The implicit generation task (**BGT3**) fails to show a substantial relationship between explicit and implicit personas, matching the annotation task. Across three out of five domains, the correlations with age are not significant. While we know the name distribution of the U.S. population over the last century, it is unclear how many of those names are highly represented in the LLM's training data. Similarly, age does not show strong associations in **BGT1** and **BGT2** and, thus, it may not be surprising that age fails here. Age is also the only continuous variable, which may be harder to attend to than binary demographics.

Failure to attend to implicit personas could be considered good or bad, depending on the context. For example, a bot could attempt to style match based on these implicit associations, which is known to increase many prosocial dimensions, such as rating of therapists (Lord et al., 2015) and relationship stability (Ireland et al., 2011). This could be good for therapy bots. Alternatively, reproducing implicit biases could perpetuate stereotypes and further harm already marginalized populations.

Finally, the persona importance task shows that political ideology is by far the strongest dimension. This also matches the validation step, where validation scores correlated at $r = .99$ with political ideology. This may be the result of this dimension being extremely polarized on several domains, and thus easier to attend to in the sense that generations easily fall into one of two categories, with little nuances needed. Surprisingly, substance use is the second strongest. To our knowledge, we do not know of any public surveys that look at differences across substance using populations, with the exception of legalization. Thus, there is no reason to believe such opinions are in the training data for GPT-4o. This could be the result of substance use being illegal and highly stigmatized, and thus similar to political ideology in its polarization.

6 Related Work

6.1 LLMs for Annotation

LLMs are increasingly becoming an integral part of the annotation workflows (Goel et al., 2023), due to its automation, consistency, and potentials in fine-tuning downstream models (Tan et al., 2024). LLMs can understand context, infer meanings, extract information, and generate human-like text, making them invaluable tools for annotating large datasets (Huang et al., 2024). For example, ChatGPT-4 was found to outperform the human crowd-workers with higher accuracy and reliability when classifying partisanship of tweets about 2020 U.S. election (Törnberg, 2023). However, preliminary findings have argued that LLMs for annotations should be used with caution (Thapa et al., 2023). For example, though ChatGPT-4 showed competitive quality in sentiment analysis, it still produced lower precision and recall in complicated tasks as compared to human annotators, for example, in labelling "anger" (Nasution and Onan, 2024). LLMs can also reflect the biases that human have in annotation tasks (Wake et al., 2023). Acerbi and Stubbersfield (2023) found that ChatGPT-3 exhibited biases mirroring those of humans towards content that aligns with gender stereotypes.

6.2 LLM Personas

While LLMs have been widely used in annotations, they can inherit biases from their training data or annotators, leading to biased or skewed annotations (Santurkar et al., 2023). One line of this research has examined the personas of LLMs (Santurkar et al., 2023; Argyle et al., 2023; Jiang et al., 2022; Simmons, 2023; Hartmann et al., 2023; Cheng et al., 2023b). Prompting LLMs with demographic information (e.g., age, gender, political ideology), biased responses from LLMs were observed (Simmons, 2023). For example, prompting with 19 diverse personas across five socio-demographic groups, stereotypical responses were observed as abstentions and a decrease in reasoning capability (Gupta et al., 2023). Such LLMs persona-related biases have been found across domains (Wan et al., 2023), hard to be eliminated by de-biasing prompts (Deshpande et al., 2023; Cheng et al., 2023a), and even in line with findings in human psychology. Simmons (2023) found that GPT-3, GPT-3.5 and OPT model families were more inclined to utilize moral principles of binding foundations (e.g., Authority/Subversion, Loyalty/Betrayal) when prompted with conservative political identity, which aligns with findings from moral psychology. Therefore, more thorough understanding of LLMs personas are needed. Finally,

	Age	Gender	Race
Palestine	*ns*	.09	.28
Parenting	-.03	.20	.05
Immigration	.03	.11	.10
Policing	*ns*	.12	.12
Legalization	*ns*	.06	.06
Average	-.01	.12	.12

Table 5: **Implicit vs Explicit personas (BGT3)** Reported product moment correlation between Explicit correlations and Implicit correlations, within a human factor and across domains (e.g., the Age column shows correlations between explicit and implicit age). *ns* not significant at a BH corrected significance level $p < 0.05$.

	Age	Gender	Pol. Ideo.	Race	Sub. Use
Palestine	.19	.10	.79	.14	.38
Parenting	.38	.34	.63	.27	.48
Immigration	.18	.18	.62	.11	.40
Policing	.21	.12	.76	.32	.50
Legalization	.20	.16	.72	.37	.46
Average	.23	.18	.70	.24	.44

Table 6: **Persona Importance** Product moment correlation between language associations from a full persona and a single factor persona.

Park et al. (2024) found that LLMs failed to replicate many social scientific tasks, often showing zero variation in responses and high sensitivity to answer choice ordering.

6.3 Implicit Personas via Names

Implicit personas have been studied in the domain of dialog, where typical personas describe, for example, interests or hobbies ("you like to travel and eat sushi"; Cho et al., 2022; Roller et al., 2021; Mazaré et al., 2018; Zhang et al., 2018). With reference to implicit personas via names, there is a long history of studying discrimination due to indirect signals of race/ethnicity, gender, or social class from names (Barlow and Lahey, 2018; Bertrand and Mullainathan, 2004). While subtle, these signals manifest in real-world discriminatory behavior and are more common than overt racial hostility (Block Jr et al., 2021). Given the documented biases inherent in LLMs (Omiye et al., 2023), it is natural to probe these systems to see if they exhibit similar subtle biases (Bai et al., 2024).

7 Conclusions

In this work, we investigated the effect of explicitly and implicitly personifying LLMs. Results showed that (1) explicit but not implicit personas replicated human perceptions in the annotation task, (2) explicit personas were sometimes able to generate text which reflected subjective human opinions, and (3) implicit personas showed a general lack of agreement with explicit personas and, more importantly, known human perceptions. Together, these results show that, despite showing minimal implicit biases, LLMs are inconsistent with their mechanisms for reproducing human thought, pointing towards limited utility in social scientific tasks.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Drug Abuse (NIDA).

8 Limitations

This study is limited in several ways. First, we only evaluate one model for the annotation task and generation tasks, which is not designed or optimized for the current study. It could be the case that models of different sizes or those which are fine tuned for social scientific tasks would perform differently.

Next, the examples from both the annotation and generation task are not exhaustive, and other studies have looked at similar tasks in more depth, though (not our knowledge) no other studies have looked at effects of implicit and explicit personas. Hu and Collier (2024) examined several subjective annotation tasks with persona prompting, including the toxicity tasks we examined in the current study. Similarly, Santurkar et al. (2023) examined if LLMs are aligned with public opinion (based on Pew surveys) across a large number of demographic groups and opinion domains.

Similarly, our study only explored monolingual English and used U.S. public opinions. Future studies could look at how opinions vary across cultures and examine that through the lens of multilingual language models.

9 Ethics

As discussed above, the human factors examined in this study are neither exhaustive nor representative. For example, income and education were not included and are known to be associated with several of the domains used in this study. Similarly, for ease of analysis, several human factors were reduced from categorical to binary, thus restricting the results to a very limited set of populations. Through this, we do not mean to imply any of these human factors are defined by the limited definitions used in the paper.

While the the main task of this work was to personify LLMs, one must take care when anthropomorphizing such systems (Abercrombie et al., 2023). This is especially important in sensitive and high stakes settings, where increased anthropomorphisms can lead to increased trust.

References

Gavin Abercrombie, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

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.

Becka A Alper. 2022. Modest warming in us views on israel and palestinians.

C Aragao. 2023. Gender pay gap in u.s. hasn't changed much in two decades.

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.

Christiane Atzmüller and Peter M Steiner. 2010. Experimental vignette studies in survey research. *Methodology*.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *Preprint*, arXiv:2402.04105.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

M Rose Barlow and Joanna N Lahey. 2018. What race is lacey? intersecting perceptions of racial minority status and social class. *Social Science Quarterly*, 99(5):1680–1698.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Ray Block Jr, Charles Crabtree, John B Holbein, and J Quin Monson. 2021. Are americans less likely to reply to emails from black people relative to white people? *Proceedings of the National Academy of Sciences*, 118(52):e2110347118.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Pew Research Center. 2024. Most americans favor legalizing marijuana for medical, recreational use.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore. Association for Computational Linguistics.

Itsugun Cho, Dongyang Wang, Ryota Takahashi, and Hiroaki Saito. 2022. A personalized dialogue generator with implicit user persona detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 367–377.

Joshua Comenetz. 2016. Frequently occurring surnames in the 2010 census. *United States Census Bureau*, pages 1–8.

Charles Crabtree, S Michael Gaddis, John B Holbein, and Edvard Nergår Larsen. 2022. Racially distinctive names signal both race/ethnicity and social class. *Sociological Science*, 9:454–472.

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.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

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

Gourab Dey, Adithya V Ganesan, Yash Kumar Lal, Manal Shah, Shreyashee Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, and H. Schwartz. 2024. SOCIALITE-LLAMA: An instruction-tuned model for social scientific tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–468, St. Julian's, Malta. Association for Computational Linguistics.

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

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400.

Salvatore Giorgi, Douglas Bellew, Daniel Roy Sadek Habib, Garrick Sherman, João Sedoc, Chase Smitterberg, Amanda Devoto, McKenzie Himelein-Wachowiak, and Brenda Curtis. 2023. Lived experience matters: automatic detection of stigma on social media toward people who use substances. *arXiv preprint arXiv:2302.02064*.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations*.

Alexis S Hammond, Kelly E Dunn, and Eric C Strain. 2020. Drug legalization and decriminalization beliefs among substance-using and nonusing individuals. *Journal of addiction medicine*, 14(1):56–62.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Preprint*, arXiv:2301.01768.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

Jingwei Huang, Donghan M Yang, Ruichen Rong, Kuroush Nezafati, Colin Treager, Zhikai Chi, Shidan Wang, Xian Cheng, Yujia Guo, Laura J Klesse, et al. 2024. A critical assessment of using chatgpt for extracting structured data from clinical notes. *npj Digital Medicine*, 7(1):106.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. 2024. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

R Morin, K Parker, R Stepler, and A Mercer. 2017. Police views, public views.

Arbi Haza Nasution and Aytug Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

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.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2023. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, page 17456916231214460.

Pew Research Center. 2021. Demographics and lifestyle differences among typology groups.

Pew Research Center. 2024. How americans view the situation at the u.s.-mexico border, its causes and consequences.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Matthew R Sanderson, Moshe Semyonov, and Anastasia Gorodzeisky. 2021. Declining and splitting: Opposition to immigration in the united states, 1996–2018. *International Journal of Intercultural Relations*, 80:27–39.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

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.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60, Copenhagen, Denmark. Association for Computational Linguistics.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Laura Silver. 2024. Younger americans stand out in their views of the israel-hamas war.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023. Prevalence and prevention of large language model use in crowd work. *Preprint*, arXiv:2310.15683.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Bias in emotion recognition with chatgpt. *arXiv preprint arXiv:2310.11753*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Gema Zamarro, Francisco Perez-Arce, and Maria Jose Prados. 2020. Gender differences in the impact of covid-19. *KTLA. Accessed on July*, 16:2021.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Additional Language Results

Table 7 shows the effect sizes of the top ngrams shown in Figure 2. Correlations with LIWC and the Moral Foundations dictionary are shown in Tables 8 and 9, respectively.

B Algorithm

Algorithm 1 shows how the correlations in Tables 4, 5, and 6 are calculated. The algorithm shows an example of comparing a full persona to a single dimension persona (gender) in the parenting domain, but, in general, this algorithm takes in two personas and a domain. For the convergent / divergent validity tests (Table 4) we consider all explicit, single factor persona pairs across all domains. For the implicit vs explicit analysis (Table 5), we consider one explicit and one implicit persona, across all pairs, and across all domains. Finally, for the persona importance task (Table 6) we consider a full persona and a single factor persona, for all human factors, and across all domains.

C Belief Generation Validation

Table 10 shows the questions given to the LLM in the validation analysis for **BGT1**. These questions were designed to be binary versions of the beliefs outlined in Table 2.

D Annotation Reliability

Here we calculate pairwise Fleiss kappa's for each combination of human, explicit Persona-LLMs,

Term	Effect Size
i've	.368
, i've	.366
decades	.366
seen	.326
witnessed	.303
really	-.296
think it's	-.299
think	-.300
i think	-.321
its	-.334

(a) Age
(Palestine)

Term	Effect Size
experiences	.996
feelings	.921
have personal	.864
ai	.825
don't	.822
modern	-.843
providers	-.852
my	-.894
as a	-.916
partner	-1.03

(b) Gender
(Parenting)

Term	Effect Size
immigration	2.58
laws	2.16
security	2.10
illegal	1.85
immigration laws	1.76
rights	1.88
humane	-2.01
climate	-2.29
change	-2.39
climate change	-2.43

(c) Political Idealology
(Immigration)

Term	Effect Size
brutality	2.22
police brutality	1.94
systemic	1.93
and	1.67
racial profiling	1.64
or	-1.38
can	-1.54
emergencies	-1.60
might	-1.95
caucasian	-2.10

(d) Race
(Policing)

Term	Effect Size
could	2.81
people	2.39
if	2.00
like	2.00
someone	1.99
legalization can	-2.06
public	-2.25
this can	-2.32
reduction in	-2.34
can	-3.59

(e) Substance Use
(Legalization)

Table 7: **N-gram** associated with each human factor across their respective domains. We show the top five most positively (top five rows) and negatively (bottom five rows) associated with each dimension. Product moment correlations reported in (a), Cohen's d in all others. All association significant at a BH corrected significance level of $p < 0.05$.

Category	Effect Size
VISUAL	.324
TIME	.324
FOCUSPAST	.199
REWARD	.179
ARTICLE	.152
CERTITUDE	-.239
IPRON	-.276
COGPROC	-.278
COGNITION	-.291
INSIGHT	-.315

(a) Age
(Palestine)

Category	Effect Size
ADJ	.979
TONE NEG	.882
CULTURE	.841
TECH	.840
EMO NEG	.818
AFFILIATION	-.821
HOME	-.827
FAMILY	-.859
MONEY	-.968
ARTICLE	-1.04

(b) Gender
(Parenting)

Category	Effect Size
POWER	2.08
RISK	2.03
CULTURE	1.32
POLITIC	1.23
AUXVERB	1.10
SOCREFS	-1.33
SOCBEHAV	-1.75
MORAL	-1.80
SOCIAL	-1.81
PROSOCIAL	-2.02

(c) Political Idealology
(Immigration)

Category	Effect Size
TIME	2.48
TONE NEG	2.18
EMO NEG	1.56
FOCUSPAST	1.57
ADJ	1.46
VERB	-1.59
FOCUSFUTURE	-1.86
TENTAT	-2.00
COGNITION	-2.25
COGPROC	-2.57

(d) Race
(Policing)

Category	Effect Size
LINGUISTIC	3.95
FUNCTION	3.86
VERB	3.52
PPRON	3.32
PRONOUN	3.30
DRIVES	-1.36
CULTURE	-1.38
POWER	-1.42
MONEY	-1.59
LIFESTYLE	-1.76

(e) Substance Use
(Legalization)

Table 8: **LIWC** categories associated with each human factor across their respective domains. We show the top five most positively (top five rows) and negatively (bottom five rows) associated with each dimension. Product moment correlations reported in (a), Cohen's d in all others. All association significant at a BH corrected significance level of $p < 0.05$.

Category	Effect Size
MORALITYGENERAL	-.185
FAIRNESSVIRTUE	-.232

(a) Age
(Palestine)

Category	Effect Size
FAIRNESSVICE	.398
INGROUPVICE	.279
PURITYVIRTUE	.064
AUTHORITYVICE	-.006
AUTHORITYVIRTUE	-.084
FAIRNESSVIRTUE	-.137
HARMVIRTUE	-.347
INGROUPVIRTUE	-.367
HARMVICE	-.494
MORALITYGENERAL	-.527

(b) Gender
(Parenting)

Category	Effect Size
AUTHORITYVICE	2.02
INGROUPVICE	1.34
AUTHORITYVIRTUE	1.23
PURITYVICE	1.08
HARMVIRTUE	.881
PURITYVIRTUE	.555
MORALITYGENERAL	.386
INGROUPVIRTUE	-.066
HARMVICE	-1.07
FAIRNESSVIRTUE	-1.32

(c) Political Idealology
(Immigration)

Category	Effect Size
HARMVICE	2.14
FAIRNESSVICE	1.58
INGROUPVIRTUE	.738
FAIRNESSVIRTUE	.610
HARMVIRTUE	.544
PURITYVIRTUE	.399
MORALITYGENERAL	.174
AUTHORITYVICE	.154
INGROUPVICE	-.395
AUTHORITYVIRTUE	-.735

(d) Race
(Policing)

Category	Effect Size
PURITYVIRTUE	.341
MORALITYGENERAL	.316
PURITYVICE	.251
AUTHORITYVICE	0.204
AUTHORITYVIRTUE	-.301
HARMVIRTUE	-.352
FAIRNESSVIRTUE	-.465
INGROUPVIRTUE	-1.05
INGROUPVICE	-1.06
FAIRNESSVICE	-1.154

(e) Substance Use
(Legalization)

Table 9: **Moral Foundations** categories associated with each human factor across their respective domains. We show the top five most positively (top five rows) and negatively (bottom five rows) associated with each dimension. Product moment correlations reported in (a), Cohen's d in all others. All association significant at a BH corrected significance level of $p < 0.05$.

Human Factor	Domain	Question
Age	Palestine	Does the following text indicate that the writer's sympathies lie more with the Israeli people (1) or more with the Palestinian people (-1) or both (0)?
Gender	Parenting	Does the following text indicate that the writer feels a great deal of pressure to focus on their responsibilities at home(1), work (-1), or both (0)?
Political Ideology	Immigration	Does the following text indicate that the writer feels that there should be stricter (1) or looser (0) restrictions on immigration at the U.S. border with Mexico?
Race	Policing	Does the following text indicate that the writer sees the police as protectors (0), enforces (-1), or both (0)?
Substance Use	Legalization	Does the following text indicate that the writer supports legalization of drugs? Yes (1), No (-1), or neither (0)

Table 10: Questions used in the Belief Generation Validation task.

	Human vs. Explicit	Human vs. Implicit	Explicit vs. Implicit
All	.42	.43	.76
Black	.39	.45	.61
White	.43	.43	.78
Female	.43	.43	.76
Male	.43	.43	.75
Conservative	.40	.40	.77
Liberal	.45	.45	.77
Uses Substances	-	-	.78
No Substances	-	-	.76

Table 11: Average pairwise Fleiss kappa's for each combination of persona type across humans, explicit Persona-LLMs, and implicit Persona-LLMs.

and implicit Persona-LLMs, matching the analysis in Bavaresco et al. (2024). Using Human vs. Explicit as an example, we calculate Fleiss kappa (κ) between the ratings (on the 5 posts) of each of the 641 humans and the ratings of each of the 641 explicit Persona-LLMs, for a given persona type (e.g., female). This results in 641^2 kappas, which we then average and report in Table 11. Here we see that explicit Persona-LLMs and implicit Persona-LLMs tend to agree more than humans and either type of Persona-LLM.

Algorithm 1 Extracting Word Frequencies and Calculating Correlations

Function: *LLM(persona_type, task)*: A function that prompts an LLM with a persona type (*full* or *individual*) for a generation task

gen_f $\leftarrow$ [LLM(full, parenting) for each worker]
all_words $\leftarrow$ $\{w_i$: frequency of w_i in *gen_f*$\}$
$corr_1 \leftarrow [\ \ ]$
gender $\leftarrow$ [gender labels for each worker]

for each w_i in *all_words* **do**
    $r \leftarrow \text{corr}(\Sigma(w_i), gender)$
    corr_1.append(r)
end for

gen_bi $\leftarrow$ LLM(gender, parenting)
all_words $\leftarrow$ $\{w_i$: frequency of w_i in *gen_bi*$\}$
$corr_2 \leftarrow [\ \ ]$

for each w_i in *all_words* **do**
    $r \leftarrow \text{corr}(\Sigma(w_i), gender)$
    corr_2.append(r)
end for

$importance \leftarrow \text{corr}(corr_1, corr_2)$