Which Demographics do LLMs Default to During Annotation?

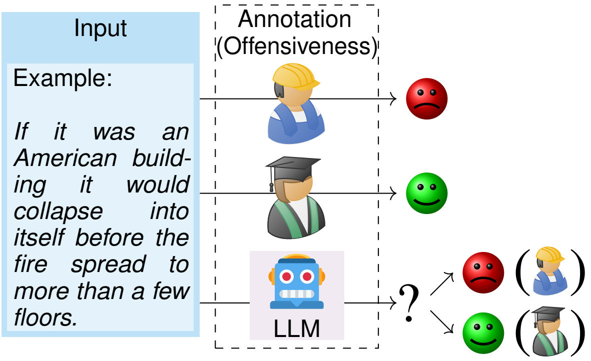
Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li,
Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz,
Aswathy Velutharambath, Amelie Wührl, Sabine Weber, and Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
roman.klinger@uni-bamberg.de

Abstract

Demographics and cultural background of annotators influence the labels they assign in text annotation – for instance, an elderly woman might find it offensive to read a message addressed to a "bro", but a male teenager might find it appropriate. It is therefore important to acknowledge label variations to not underrepresent members of a society. Two research directions developed out of this observation in the context of using large language models (LLM) for data annotations, namely (1) studying biases and inherent knowledge of LLMs and (2) injecting diversity in the output by manipulating the prompt with demographic information. We combine these two strands of research and ask the question to which demographics an LLM resorts when no demographics is given. To answer this question, we evaluate which attributes of human annotators LLMs inherently mimic. Furthermore, we compare non-demographic conditioned prompts and placebo-conditioned prompts (e.g., "you are an annotator who lives in house number 5") to demographics-conditioned prompts ("You are a 45 year old man and an expert on politeness annotation. How do you rate {instance}"). We study these questions for politeness and offensiveness annotations on the POPQUORN data set, a corpus created in a controlled manner to investigate human label variations based on demographics which has not been used for LLM-based analyses so far. We observe notable influences related to gender, race, and age in demographic prompting, which contrasts with previous studies that found no such effects.

1 Introduction

In some text[1] annotation tasks, it is feasible to obtain an aggregated ground truth label, for instance in named entity annotation (Yadav and Bethard, 2018) or semantic role labeling (Shi et al., 2020). In other tasks, perhaps even in the majority of tasks, it is more obvious that annotators' traits influence the label assignments, for instance in sentiment annotation (Liu, 2012), emotion annotation (Klinger, 2023; Plaza-del Arco et al., 2024), or personality profiling (Neuman, 2015). The diversity of annotations, conditioned on annotators' profiles, has been recognized as an important variable to consider, instead of aggregating all labels to an adjudicated score, which might not correspond to any of the annotators (Plepi et al., 2022).



If it was an American building it would collapse into itself before the fire spread to more than a few floors.

Figure 1: Our objective is to identify which human demographic groups are mimicked by LLMs during subjective annotation tasks on text data.

With the advent of (instruction-tuned) large language models (LLMs), automatic data annotation and zero or few shot predictions became more popular (Brown et al., 2020). However, language models do not provide the same diversity as human annotators do in a simple zero-shot setup, which can lead to a lower performance of the labeling process (Bagdon et al., 2024). To mitigate this problem, it is important to understand whether large language models exhibit a default persona when acting as annotators and how this can be controlled to ensure greater diversity in annotations. One concern is that large language models may reflect biases present in their training data, which can disproportionately

[1] We provide our code and model predictions on https://www.uni-bamberg.de/en/nlproc/resources/llms-default-demographics/.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17331–17348
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

emphasize certain viewpoints. This raises the issue of potential marginalization of minority perspectives in the outputs generated by these models, making it crucial to address this bias for equitable representation in annotations. One idea to address this limitation is socio-demographic prompting (Muscato et al., 2024), where we guide the model by specifying characteristics in the prompt, like age, gender, ethnicity or socio-economic status.

Previous research has explored the effects of using prompts informed by demographic or cultural contexts on model predictions, but did not find consistent patterns (Beck et al., 2024; Mukherjee et al., 2024). We build on top of this work and particularly contribute in two directions of socio-demographic prompting. We study if large language models default to a particular demographic, i.e., we evaluate if the LLM-based predictions more closely align with those of people of a particular demographic when not being conditioned. Our general objective is depicted in Figure 1. Additionally, we utilize placebos as suggested by Mukherjee et al. (2024) – irrelevant information that should not affect the model's output – to compare the impact of demographic prompts and evaluate the stability of the model's predictions. More concretely, we answer the following questions.

RQ1 What default demographic values can we infer from the annotation behavior of large language models?

RQ2 Are the changes to the models' annotations more pronounced for demographic prompting than with non-relevant additional information to the prompt?

RQ3 How do task properties of offensiveness rating vs. politeness rating influence the role of demographic information in prompts?

RQ4 Are observed patterns consistent across different large language models?

The remainder of this paper is structured as follows: We review related work in Section 2 and explain our data set and task choices as well as prompt setup in Section 3. In Section 4, we present the results of our experiments and conclude including a discussion of possible future work in Section 5.

2 Related Work

This paper relates to various areas which we review in the following, namely biases in large language models, perspectivism, and (socio-demographic) prompting of LLMs.

2.1 Inherent Knowledge and Biases in LLMs

Large language models may be understood as databases that store information encoded in the training data (Petroni et al., 2019). These data are stored in a probabilistic manner, which allows models to show sometimes unexpected generalization capabilities beyond the original instructions available in the training data, i.e., emergent abilities (Wei et al., 2022). It has also been shown that prompting models can adapt parameters during inference, similar to how backpropagation works (von Oswald et al., 2023). Therefore, it is reasonable to view prompting as a form of programming, as it involves manipulating model behavior by crafting specific inputs to achieve desired outputs (Beurer-Kellner et al., 2023).

There are cases in which the information requested from a large language model is not available in the training data "as is" and the generalization abilities reach its limits. In such cases, the model might output text that is not correct. Such cases are sometimes considered "hallucinations". The confidence of hallucinations is often lower than for correct information (Farquhar et al., 2024).

In cases in which subjective properties are requested, the models therefore rely on "emergent" abilities. There is some research in understanding the "stances" or "worldviews" encoded in language models. For instance, Motoki et al. (2023); Feng et al. (2023) study political biases and how these lead to unfair models. Ceron et al. (2024) study the reliability and robustness of such stances. Wright et al. (2024) extend such analysis to more fine-grained opinions. The opinions expressed by models correspond more to particular subsets of a society than to others (Santurkar et al., 2023).

2.2 Perspectivism

Natural language processing for a long time focused on seemingly objective tasks such as parsing or named entity recognition. Therefore, it has been a standard approach to adjudicate annotations into a single gold standard (Stubbs, 2011). To do so, a set of methods for aggregation that considers the distributions and disagreements have been developed (Paun et al., 2018). More recently, research moved towards acknowledging the importance of disagreements. Plank (2022) has been one of the first to discuss challenges and potential approaches to this problem. Recent work looked into methods to consider disagreements (Fleisig et al., 2024) and discusses ethical considerations (Valette, 2024). Nowadays, this field perspectivism developed dedicated shared tasks (Uma et al., 2021) and workshops (Abercrombie et al., 2024).

Disagreement is related to confidence of annotators (and models), an aspect that has received some attention. Baumler et al. (2023) made use of this property in the context of active learning, to understand which instances require annotations by multiple people in order to understand the disagreements. Troiano et al. (2021) and Baan et al. (2024) showed that annotators' own confidence predicts inter-annotator agreement scores.

To understand disagreement in labeling better, more and more corpora are being published with more detailed information about the annotators. Troiano et al. (2023) annotated event reports for emotions and appraisals and collected demographic information, personality, and current emotional state of both the person who lived through the event and multiple annotators that read the event description. Plepi et al. (2022) studied the role of demographics, automatically extracted from the data, on the perception of social norms. Bizzoni et al. (2022) discuss the role of individual differences on the judgement of literary quality. Romberg (2022) integrates perspectivism in argument mining, by making explicit the subjective nature of argument interpretation. May et al. (2024) study the effect of demographics on the role of numbers in social judgements. Frenda et al. (2023) study how the perception of irony varies by nationality, employment status, student status, ethnicity, age, and gender. Sachdeva et al. (2022) measure different aspects of hate speech which include sentiment, disrespect, insult, attacking/defending, humiliation, inferior/superior status, dehumanization, violence, genocide, and a 3-valued hate speech benchmark label. They study these variables under the condition of identity target groups and annotator demographics. Xu et al. (2023) look into disagreement of legal case outcome differences. Next to subjective tasks, perspectivism has also been considered in seemingly objective tasks, for instance in named entity recognition (Peng et al., 2024) and natural language inference (Gruber et al., 2024).

The large POPQUORN corpus (Pei and Jurgens, 2023) has been created specifically to study perspectivism of annotators and the relationship between demographics and annotations in the tasks of offensiveness detection, question answering, text rewriting and style transfer, and politeness rating. We use this corpus because it has been created for the study of perspectivism, but it has not yet been used to analyze large language models.

2.3 Prompting for Automatic Data Set Annotation or Zero-Shot Predictions

While fine-tuning models is currently still the state-of-the-art approach to obtain the best possible performance for a variety of natural language processing tasks, prompting language models for a zero-shot of few-shot prediction gained popularity recently. This is due to the possibility of efficiently adapting model outputs by prompt engineering, without fine-tuning the model – in fact, prompt optimization (by a human or automatically) can be seen as parameter-efficient model adaptation.

This field builds on top of instance-based classification methods and became popular with the work by Palatucci et al. (2009) who suggest to perform dataless classification by semantically encoding output concepts. Another non-NLP example is Banerjee et al. (2022) who embed emotion concepts for zero-shot classification of body gestures. A milestone in the natural language processing community is the work by Yin et al. (2019) who show how natural language inference can be applied across multiple classification tasks and Brown et al. (2020) who show that auto-regressive language models are zero-shot learners.

Since then, a set of studies have been proposed, including work that focuses on cross-linguality (Bareiß et al., 2024), data augmentation (Chen and Shu, 2023), emotions (Plaza-del Arco et al., 2022; Bagdon et al., 2024), named entity recognition (Shen et al., 2023), and sentiment classification (Fei et al., 2023; Ma et al., 2022). A more comprehensive survey has been provided by Li (2023). Prompts can also be learned, but this setup is out of scope for our work in this paper (Liu et al., 2023).

2.4 Socio-Demographic Prompting

Automatic annotation with language models is not a replacement for human annotation. Humans have previous world knowledge, experiences, and perspectives on a matter that differ individually. LLMs have difficulties replicating these differences, causing issues in annotation. For instance, Bagdon et al. (2024) show that the diversity of annotations that is beneficial in comparative annotations by multiple people cannot be straightforwardly replaced by multiple runs of a language model. Lee et al. (2023) focus on this aspect in particular and find poor alignment of the distribution of labels predicted by LLMs with human annotations on natural language inference tasks.

Most relevant for our work are the following studies. Beck et al. (2024) study the impact of demographic information on subjective annotation tasks. They find that model variation is larger across other parameters such as prompt formulation techniques than demographic information (e.g., age and gender) in the prompt. Mukherjee et al. (2024) examine cultural aspects like food preferences and find that most language models exhibit significant response variability, casting doubt on the reliability of socio-demographic prompting. To understand if the variables influence the annotation in a systematic way, they compare the predictions to what they call "placebos" – information that should not influence the prediction but looks like relevant parameters (favorite planet or house number).

Additionally, Sun et al. (2025) highlight that most LLMs show demographic biases in subjective judgment tasks, favoring perceptions from White participants over those from Asian or Black participants. Hu and Collier (2024) find that incorporating persona variables in LLM prompting improves model predictions slightly, especially in conditions of significant annotator disagreement. Movva et al. (2024) reveal that while GPT-4 shows reasonable alignment with human assessments of safety, there are significant demographic disparities in how well it correlates with different annotator groups.

Finally, there is a set of studies that investigate biases on large language models (Cheng et al., 2023; Santy et al., 2023, i.a.). Their findings suggest that these models may generate outputs reflecting racial stereotypes and exhibit some performance disparities across different demographic groups, indicating potential biases in their design and outputs.

Our work combines aspects of previous research, namely on *biases*, *placebos*, and *demographics*.

3 Experimental Setting

This section presents the methodology and resources utilized in our experiments.

3.1 Data Sets

We chose the POPQUORN data for our experiments (Pei and Jurgens, 2023). In contrast to data used by other previous research we are aware of, these data have been particularly sampled for the study of annotators' properties and their impact on annotation tasks. Therefore, these data render themselves as a straightforward choice also for an LLM-based analysis. The original data stem from the Ruddit corpus (Hada et al., 2021) which was originally annotated via best–worst scaling. This may lead to different label frequencies than rating scale annotations and is considered as not appropriate for skewed distributions (Finn and Louviere, 1992; Louviere et al., 2015). The authors of POPQUORN reannotated the data with 1–5 rating scales, which we adopt for our annotation setup. Further, it is noteworthy that the creators of the data set removed annotators with particularly low annotation competency according to MACE (Hovy et al., 2013), to filter out potential random noise.

The POPQUORN data set consists of 45,000 annotations from 1,484 annotators with information on their gender, race, age, occupation and education. We use data from the subjective partitions for the offensiveness and politeness rating tasks. To equalize the number of annotations per instance, we randomly sample three annotations for each instance.[2] When sampling, we omit the annotations where the annotator preferred not to disclose demographic information. We exclude data instances with fewer than three annotations remaining. The two subsets of the sample which we use in our experiments consist of 4,500 annotations of 1,500 instances for offensiveness ratings and 11,151 annotations of 3,717 instances for politeness ratings.

3.2 Prompt Setup

We use three different prompt setups, namely socio-demographically informed (SD) prompts, placebo-informed (P) prompts, and prompts without additional information (N prompts). We show the

[2] The decision to retain three annotators per instance was a deliberate choice to standardize our methodology, as the original dataset has varying numbers of annotators. We aimed to ensure consistency across our analysis and view the balance we achieve as a necessity for the comparison to non-demographic prompting, which always produces one prediction per instance. With our method, we avoid a skewed distribution towards instances with a larger number of annotators in demographic prompting. We have checked whether the distribution is skewed due to our sampling: For the offensiveness data, our sample consists of 35% of the overall annotations from POPQUORN. As for the distribution of each demographic attribute in our sample: It comprises 31%-38% of the annotations per attribute (standard deviation: 1%). For the politeness data, our sample consists of 45% of the overall annotations. As for the distribution of each demographic attribute in our sample: It comprises 34%-56% of the annotations per attribute (standard deviation: 4%; the wide range is due to the fact that some attributes being extremely rare in the original dataset, leading to outliers). Our sample maintains a certain degree of representativeness of the demographics in POPQUORN.

Prompt Type	Prompt 1 Template	Prompt 2 Template
Socio-Demographic (SD)	You are a person of gender [gender], race [race], age [age], occupation [occupation], education [education]. Given a text, how do you rate the [task] of the following text? Possible values are [task-values]. In your answer, only provide the label you chose.	Imagine [Prompt 1 with initial character in lower case]
Placebo (P)	You are a person of height [height], Zodiac sign [zodiac sign], living in a house with the number [house number], enjoying [enjoying], and whose favorite color is [favorite colour]. Given a text, how do you rate the [task] of the following text? Possible values are [task-values]. In your answer, only provide the label you choose.	Imagine [Prompt 1 with initial character in lower case]
Non-Demographic (N)	You are a person of any gender, any race, any age, with any occupation, and any education. Given a text, how do you rate the [task] of the following text? Possible values are [task-values]. In your answer, only provide the label you choose.	Given a text, how do you rate the [task] of the following text? Possible values are [task-values]. In your answer, only provide the label you chose.

Table 1: Prompt types and Templates for system messages. For the [task] offensiveness, the [task-values] are ‘not offensive’, ‘slightly offensive’, ‘moderately offensive’, ‘very offensive’ or ‘extremely offensive’. For the [task] politeness, the [task-values] are ‘not polite’, ‘slightly polite’, ‘moderately polite’, ‘very polite’ or ‘extremely polite’.

Placebo Attribute	Possible Values
Height	140 cm, 150 cm, 160 cm, 170 cm, 180 cm, 190 cm, 200 cm, 210 cm.
Zodiac sign	Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, Sagittarius, Capricorn, Aquarius, Pisces.
House number	6, 12, 13, 24, 45, 68, 98, 122, 234, 1265.
Enjoying	food, sleep, friends.
Favorite colour	red, green, blue, yellow, purple, turquoise, orange, pink, black, white, brown.

Table 2: Sets of values for placebo attributes used in P prompts in our experiments.

templates used for these prompts in Table 1. The input to the LLM consist of a concatenation of the system message with the respective text to be classified. The values for the demographic attributes are taken from the demographic data of the annotators included in the sample. The sets of values for the placebo attributes are displayed in Table 2. During prompting, each placebo attribute value is randomly sampled from the respective set.

3.3 Model Choice and Access

We use two LLMs, namely GPT-4o (OpenAI, 2024b) and Claude (Anthropic, 2024). We access both models via their respective APIs with default hyperparameters. The total cost for using GPT-4o was $54 and the cost for using Claude was $60.

3.4 LLM Output Parsing

We transform the texts generated by the LLMs for the annotation tasks into a defined label set of five categories, corresponding to the 1–5 rating scale for each of the two tasks. We use the Langchain StrOutputParser[3] to interpret the output from GPT-4o as one of the designated labels, while we use the output from Claude directly.

This process does not successfully parse the output in all cases. We encountered 22 error cases with GPT-4o and three with Claude. Of the failed cases with GPT-4o, 14 were due to the input text from the original data set being in Polish instead of English. This resulted in the model predicting labels in Polish, despite the prompt specifying an English-only label set. The remaining eight cases involved instances with very short text or questions, which the model misinterpreted. In these cases, GPT-4o asked for further input instead of performing the intended classification task.

For Claude, the three instances of failure involved the names of famous actors in the text. This triggered the model’s copyright protection protocols, preventing it from reproducing or paraphrasing potentially copyrighted content.

Given the total number of instances analyzed in our experiments, this failure rate is insubstantial. Consequently, we disregard the output in the cases where parsing was unsuccessful and assign the labels “not offensive” or “not polite”.[4]

[3] https://api.python.langchain.com/en/latest/output_parsers/langchain_core.output_parsers.string.StrOutputParser.html

[4] Our decision to label such instances with the negative class, rather than removing them, was made to maintain comparability across different experimental settings. Removing instances due to a single failed label would distort the consistency of the number of variants of prompts per text instance.

3.5 Evaluation Settings

We conduct various analyses of the labels generated by the LLMs, utilizing the different prompt setups described in Section 3.2. First, we examine the extent to which LLM predictions align with the judgments of different human annotators by comparing the outputs generated from non-demographic (N) prompts to the human annotations available for each instance. Second, we examine the impact of socio-demographic prompting by comparing the models' automatic annotations produced with SD prompts to those produced with N prompts. In this analysis, we assess the differences in the LLMs' annotations when demographic data is included versus when it is omitted. Thirdly, we investigate the effect of placebo prompting in a similar manner by comparing the automatic annotations generated with P prompts to those created with N prompts.

4 Results

In this section, we present the results of our experiments based on the annotations generated by the LLMs for two labeling tasks. Each of our research questions is addressed with particular results derived from the various experiments conducted. Detailed results for individual prompts show that the two N prompts behave similarly (see Appendix A.2). We also observed this for the N and SD prompt templates. Thus, we here report all results as an average over the two respective prompt templates for the different prompt types.

4.1 RQ1: What default demographic values can we infer from the annotation behavior of large language models?

We approach the identification of the default persona of the models in two different ways following the first two settings as described in Section 3.5. The following sections present and discuss the results of these analyses.

Which socio-demographic attributes of human annotators does an LLM inherently mimic in the absence of explicit information (N prompts)? Table 3 shows the results for this analysis which corresponds to the first setting described in Section 3.5. We combine socio-demographic attributes that are represented in only a few cases. The table specifies the reference categories (ref.) used for each of the categorical variables.[5]

	Offensiveness		Politeness	
Socio-Demographic Attribute	GPT-4o	Claude	GPT-4o	Claude
Age	**0.01	**0.01	0.00	0.00
Gender (ref.: Male)				
Female	0.00	−0.03	−0.05	−0.05
Non-binary	−0.06	−0.01	−0.05	−0.05
Race (ref.: White)				
Asian	0.09	0.03	−0.08	0.00
Black/Afri. Am.	***0.22	**0.19	**0.14	**0.15
Hispanic or Latino	−0.11	−0.05	0.09	0.12
Other race	−0.14	−0.26	−0.17	−0.15
Occupation (ref.: Employed)				
Unemployed	0.04	−0.08	−0.08	−0.08
Homemaker	−0.07	−0.02	−0.03	−0.06
Retired	−0.11	−0.13	0.03	0.03
Self-employed	0.04	0.02	−0.03	−0.03
Student	0.13	0.13	−0.11	−0.13
Other occupation	−0.05	0.02	0.08	0.11
Education (ref.: Less than high school)				
High school dipl.	−0.01	−0.08	−0.34	*−0.37
College degree	0.05	−0.09	*−0.43	*−0.48
Graduate degree	0.06	−0.01	*−0.36	*−0.44
Other education	−0.02	0.01	*−0.50	**−0.57

Table 3: Coefficients indicating the effect of particular human demographic categories on the distance between human and LLM annotations, calculated using mixed-effects regression models with random intercepts for annotators and instances. Statistical significance is calculated using standard error (see Appendix A.1 for these values) and is here marked by asterisks: * corresponds to $P \leq 0.05$, ** to $P \leq 0.01$, and *** to $P \leq 0.001$.

Table 3 shows the coefficients of regression models predicting absolute values of the distance between LLM annotations (N prompts) and annotations provided by human annotators for the same instance. Independent variables are the socio-demographic characteristics of the human annotators. We report results separately for the two models (GPT-4o and Claude) and two classification tasks (offensiveness and politeness rating). The reported coefficients can be interpreted as the effect of particular human demographic categories on the distance between human and LLM annotations (change in distance between the category in question and the reference category). Positive coefficients indicate that the respective LLM is further from the human annotators in that demographic category than human annotators in the reference category, i.e., less accurate. From this we can follow that the categories with positive, statistically significant coefficients are those categories which

[5] The reference categories for the categorical variables occupation and education are chosen according to the recommendations of Johfre and Freese (2021). For race and gender, their scheme does not result in unambiguous recommendations, so we choose the categories we expect to lie closest to the models' defaults: male for gender and white for race.

the respective model does not default to. In contrast, negative coefficients indicate that the LLM is closer to the human annotators in the given demographic category than it is to the human annotators in the reference category; i.e., more accurate. Categories with negative and statistically significant coefficients are those that are nearer to the model's default than the reference category.

The analysis of coefficients (see Table 3) reveals significant biases in the LLMs' predictions based on demographic factors, particularly concerning race and age. Specifically, the models demonstrate a tendency to align more closely with annotations provided by persons identifying as White as opposed to those identifying as Black or African American. The distance between their annotations measures .19 to .22 Likert scale points for offensiveness rating and .14 to .15 points for politeness rating. Additionally, the LLMs are progressively less accurate in reflecting the views of older individuals at offensiveness rating, with a .01 point increase in distance for each year of age. In terms of educational background, the models exhibit a greater discrepancy from those with less than a high school education compared to those with higher educational levels, ranging from .34 to .57 points for politeness rating. One possible explanation for these discrepancies by annotator sociodemographics is that the models' training data may lack representation for these demographic groups.

Notably, this analysis does not show any significant effects related to gender or occupational status. This could be because the LLMs do not consistently favor any particular group of people on these dimensions, or it could be because there are in fact no systematic differences in human annotation along these dimensions. Overall, while the statistically significant differences in annotation distances suggest some demographic biases, we emphasize that these effects are small compared to the full Likert scale range. In addition, the R2 Marginal values for these models (reported in Appendix A.1) are quite low, indicating that the sociodemographic categories explain little of the variability in human–LLM annotation differences. This complexity makes it challenging to establish a clear default persona for the LLMs. Therefore, we perform a second analysis in the following.

The Effect of Demographic Prompting. In a second approach, we examine how the inclusion of demographic information in the prompt (cf. second setting described in Section 3.5) influences the automatic annotation outputs of the models. We compare the outputs generated with `SD` prompts to those with `N` prompts. We assess the differences in the LLMs' annotations with and without the inclusion of demographic data, which enables us to infer the demographics closest to the models' default.

Table 4 shows the results as differences in the prediction scores (Δ_μ).[6] For some demographic attributes, there are relatively few samples ("count" columns). This implies that there may not be sufficient statistical evidence to support the observed average distance values for certain categories. We focus our discussion on cases with count ≥ 100.

GPT-4o shows substantial differences in the scores for the offensiveness task across different gender attribute values: The prediction differences for Non-Binary (.29) are substantially larger than those for Male (.18) and Female (.20). This suggests that prompting the LLM to act as a non-binary individual has a more pronounced effect on its predictions, indicating that the male or female attributes are more aligned with its default persona. This observation is not consistent across tasks.

Claude shows a notable pattern for the politeness task concerning the age socio-demographic attribute. With an increasing age, the prediction difference increases (from .16 to .26). Similarly, the occupation attribute reflects the highest difference for the value Retired. This indicates that when Claude is prompted to act as an older individual, it exhibits an increased sensitivity to politeness.

The discrepancy between the results from the analysis displayed in Table 3 to those from the analysis displayed in Table 4 are not a contradiction. These analyses are based on distinct interpretations of the default persona of the models. Table 3 indicates a poor representation of some demographics in the models' responses. Conversely, Table 3 highlights that, when prompted with specific demographics, the output was marginally closer to the default, suggesting that the models' interpretation of these demographic attributes influences its behavior only slightly. Thus, the model exhibits a measurable lack of alignment with certain demographics while it simultaneously demonstrates minimal variation in its behavior when prompted to act as a person with those demographics.

[6] Note that there is an overlap in the age ranges (50–54, 54–59), a consequence of the survey question design in the original annotation task for the POPQUORN dataset. We have preserved these original categories, despite this issue, in order to maintain the fine-grained distinctions they provide.

	Offensiveness			Politeness		
Socio-Demographic Attribute	Count	Δ_μ (GPT-4o)	Δ_μ (Claude)	Count	Δ_μ (GPT-4o)	Δ_μ (Claude)
Gender						
Male	2,157	0.18	0.17	5,195	0.26	0.20
Female	2,219	0.20	0.15	5,623	0.24	0.22
Non-binary	124	0.29	0.17	333	0.24	0.17
Race						
White	3,396	0.18	0.16	8,163	0.25	0.20
Hispanic or Latino	95	0.21	0.11	790	0.22	0.21
Native American	97	0.25	0.12	0	–	–
Arab American	17	0.32	0.06	0	–	–
Native Hawaiian or Pacific Islander	0	–	–	36	0.31	0.15
American Indian or Alaska Native	0	–	–	28	0.18	0.23
Black or African American	559	0.22	0.15	1,386	0.25	0.21
Asian	336	0.21	0.15	731	0.24	0.24
Hebrew	0	–	–	17	0.18	0.26
Age						
18-24	499	0.21	0.16	1,241	0.25	0.16
25-29	444	0.20	0.16	894	0.25	0.18
30-34	540	0.16	0.15	1,190	0.24	0.17
35-39	532	0.20	0.19	834	0.26	0.18
40-44	418	0.19	0.18	1,176	0.23	0.21
45-49	388	0.19	0.15	957	0.22	0.22
50-54	314	0.17	0.17	995	0.27	0.21
54-59	627	0.18	0.15	1,083	0.26	0.22
60-64	251	0.22	0.15	1,193	0.24	0.22
>65	487	0.20	0.15	1,588	0.25	0.26
Occupation						
Unemployed	571	0.19	0.14	1,328	0.28	0.19
Employed	2,189	0.19	0.17	4,944	0.24	0.20
Homemaker	199	0.24	0.14	784	0.23	0.21
Retired	500	0.19	0.15	1,783	0.26	0.25
Other	86	0.16	0.22	268	0.28	0.19
Self-employed	617	0.17	0.17	1,395	0.23	0.21
Student	338	0.23	0.14	649	0.23	0.16
Education						
Less than a high school diploma	84	0.18	0.17	76	0.30	0.22
High school diploma or equivalent	1,379	0.19	0.14	3,312	0.26	0.17
Graduate degree	846	0.21	0.16	2,160	0.23	0.25
College degree	2,098	0.18	0.17	5,352	0.24	0.21
Other	93	0.25	0.11	251	0.24	0.23

Table 4: Sample sizes (count) and mean distance scores of demographic-prompting (SD prompts) predictions in comparison to predictions with N prompts for models GPT-4o and Claude at two rating tasks. Variables with very few cases (count $\leq$ 100) are shown in gray to indicate a lack or reliability of these numbers.

4.2 RQ2: Are the changes to the models' annotations more pronounced for demographic prompting than with non-relevant additional information to the prompt?

We investigate whether the modifications to the model are more substantial for socio-demographic prompting compared to non-relevant additional information provided in the prompts. Specifically, we analyze how the models' predictions for the studied tasks are affected when presented with placebo information (cf. third setting from Section 3.5).

We evaluate if these results are as pronounced as the observations based on Table 4 as described above. Overall, the results of the placebo prompting (P prompts) in comparison to N prompts do not reveal any notable differences for specific attribute values (see Appendix A.3). The scores remain consistently stable across these comparisons. Consequently, we conclude that the changes in model behavior are indeed more systematic for prompting with specific socio-demographic attributes than when irrelevant additional information is included.

The changes resulting from placebo prompting in comparison to N prompts appear to be substantial, with some discrepancies even surpassing those related to socio-demographic prompting. This pattern is more pronounced within the offensiveness task. The discrepancies arise from the analysis that focuses on the absolute values of the differences, regardless of their direction, thus capturing notable fluctuations that may not reflect a consistent trend.

4.3 RQ3: How do task properties of offensiveness rating vs. politeness rating influence the role of demographic information in prompts?

We investigate how task properties influence the role of demographic information in prompting and whether patterns remain consistent across the tasks. Our general conclusion indicates that there is no clearly distinguished default persona of the models for both tasks. Table 3 shows that results are in general consistent across the two tasks. However, the analysis regarding RQ1 also highlights some task-specific tendencies. Notably, there are also differences in the prediction distances when comparing demographic prompting to no-information prompting across the two tasks. The average prediction difference associated with socio-demographic prompting of GPT-4o for offensiveness rating is calculated to be 0.19. In contrast, the average for politeness rating is substantially higher at 0.25. Similar results are evident for the model Claude (0.16 for offensiveness rating and 0.21 for politeness rating). This suggests that, in general, these LLMs are more influenced by demographic information at rating politeness than rating offensiveness.

4.4 RQ4: Are observed patterns consistent across different large language models?

The patterns identified in the analysis presented in Table 3 demonstrate consistency across the two models examined, particularly in their ability to replicate human annotations. Notably, Table 4 highlights specific behaviors of the LLMs in response to socio-demographic prompting. However, each of these distinct behaviors is only observed in one of the models. Overall, both tested LLMs do not display clear default personas. Thus, the models remain consistent in this aspect.

5 Conclusion and Future Work

In this paper, we present an analysis of the effect of socio-demographic prompting on tasks in the POPQUORN data set. Our findings show that demographic prompting exerts measurable effects on the annotation behaviors of large language models. We contrast this to placebo prompting, which elicits no consistent changes across various attributes. Specifically, our analyses reveal that LLMs show variations in annotation based on demographic attributes, particularly for gender, race, and age. While we cannot infer one concrete, unique default persona, we conclude that large language models do not represent all members of a society alike, and that socio-demographic prompting does influence the result in a structured manner. This stands in contrast to the results from previous studies, such as those by Beck et al. (2024) and Mukherjee et al. (2024), which report no consistent patterns.

Furthermore, our results echo some of the findings from Sun et al. (2025) regarding gender influences and that predictions tend to align more closely with the perceptions of White individuals. However, our analysis includes a broader range of demographic attributes. Hu and Collier (2024) observe that while persona variables produce modest improvements, their limited explanatory power aligns with our findings on demographic prompting. Similarly, Movva et al. (2024) highlight challenges in aligning LLMs with safety evaluations across demographics, underscoring the need to understand these impacts on model behavior. Together, these insights confirm the necessity of exploring demographic influences in LLM outputs.

Our study highlights biases in the annotation behavior of LLMs regarding demographics. These biases raise critical concerns about the perpetuation of racial inequities in applications of these models. The discrepancies we observe regarding different demographic groups suggest that LLMs may entrench existing biases rather than mitigate them. This reinforces societal norms that marginalize diversity, demonstrating that models struggle to accurately represent certain demographic groups.

Not all our findings are consistent across models. While this could be considered an issue that is inherent to LLMs, it also presents itself with a substantial challenge: depending on the model that is used by an end-user, the impact of socio-demographic information varies, and different models default to different demographic information.

Consequently, we advocate for future research to concentrate on developing reliable tools to measure the influence of both implicitly and explicitly provided demographic information. Making demographic information explicit and ensuring its accurate interpretation by models is crucial not only for addressing hidden biases but also for creating models that genuinely reflect the diversity and variance among individuals. Ultimately, our work underscores the necessity for bias-aware design in training models to provide equitable representation across demographic groups and to mitigate the risks of reinforcing existing societal inequities.

Limitations

Our study considers only a limited set of models and variables. Particularly limiting is the set of attributes available in the data set, which are also culturally biased towards the USA. Another considerable limitation is the unequal representation of demographic subgroups. We additionally only analyze a sample of the human raters to represent the text instances equally, which might lead to a skewed representation of certain demographics. However, our sample maintains a certain degree of representativeness of the demographics in POPQUORN. Using a larger sample would increase robustness of our results.

While the decision to further analyze only demographics with over 100 annotated instances aimed to enhance the manageability and interpretability of our results, it may introduce concerns regarding the statistical confidence of findings, particularly for the non-binary group, which marginally meets this threshold. Fluctuations in the observed effects may also partially arise from the group's under-representation in this evaluation.

A limitation of our analysis is the lack of post-hoc statistical testing, which could strengthen our findings. Although such tests are often omitted in regression model comparisons with few predictors, examining them could analyze whether assumptions like homoscedasticity and collinearity might have been violated. It is unlikely that the results or the conclusion of our work is influenced, however, this can be a possible factor that might have undermined the overall process.

Our experiments are conducted with only two large language models and each experiment was run just once, for pragmatic reasons stemming from limited financial resources. A large scale analysis across larger set of tasks would mitigate this issue.

Finally, the reported differences in annotation outputs are relatively small compared to the overall Likert scale range of 1–5. This raises questions about the practical significance of some of the observed effects, necessitating further exploration to understand their impact in real-world applications.

Ethical Considerations

Our work has the goal to make challenges transparent which the use of large language models has. A considerable limitation from an ethical perspective is that the variables we consider are not relevant across cultures in the world. The response options of the "race" variable, in particular, are specific to US American society. Other variables might be relevant in other cultures in the world that we are not aware of. We therefore suggest to conduct future studies that consider more open sets of variables that describe the diversity of users of language model-based systems.

Beyond the limitations of variable selection, our findings also have implications for issues of inequality in the context of large language models. If models effectively mimic particular demographic groups, this can lead to the privileging of the viewpoints and experiences of those groups, potentially marginalizing individuals who are not adequately represented. Furthermore, if the responses are influenced by the inclusion of certain demographic information over others, it suggests that some demographic categories are represented as more "normal" or typical than others. This pattern mirrors historical inequities that persist in new technologies, revealing an unsettling continuity in the ways that cultural biases may be perpetuated.

ChatGPT (OpenAI, 2024a) was used to gain inspiration for formulations of our initial notes for the text of some sections of this paper, as well as to find typos.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the projects KL2869/1–2 (CEAT, project number 380093645), KL2869/5–1 (FIBISS, 438135827), KL2869/11–1 (ITEM, 513384754), KL2869/12–1 (EMCONA, 516512112), KL2869/13–1 (INPROMPT, 521755488).

References

Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Anthropic. 2024. Claude 3.5 Sonnet. Large language model.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Abhishek Banerjee, Uttaran Bhattacharya, and Aniket Bera. 2022. Learning unseen emotions from gestures via semantically-conditioned zero-shot perception with adversarial autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence, 36*.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(Issue PLDI):1946–1969.

Yuri Bizzoni, Ida Marie Lassen, Telma Peura, Mads Rosendahl Thomsen, and Kristoffer Nielbo. 2022. Predicting literary quality how perspectivist should we be? In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 20–25, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *Preprint*, arXiv:2402.17649.

Canyu Chen and Kai Shu. 2023. PromptDA: Label-guided data augmentation for prompt-based few shot learners. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 562–574, Dubrovnik, Croatia. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Adam Finn and Jordan J. Louviere. 1992. Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing*, 11(2):12–25.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. 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.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717, Online. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Sasha Shen Johfre and Jeremy Freese. 2021. Reconsidering the reference category. *Sociological Methodology*, 51(2):253–269.

Roman Klinger. 2023. Where are we in event-centric emotion analysis? bridging emotion role labeling and appraisal-based approaches. In *Proceedings of the Big Picture Workshop*, pages 1–17, Singapore. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 641–647, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bing Liu. 2012. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9).

Jordan J. Louviere, Terry N. Flyn, and A.A.J. Marley. 2015. *Best-Worst Scaling – Theory, Methods and Applications*. Cambridge University Press.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland. Association for Computational Linguistics.

Marlon May, Lucie Flek, and Charles Welch. 2024. A perspectivist corpus of numbers in social judgements. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 42–48, Torino, Italia. ELRA and ICCL.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *Preprint*, arXiv:2406.11661.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, and Fosca Giannotti. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 49–55, Torino, Italia. ELRA and ICCL.

Yair Neuman. 2015. Personality research for NLP. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Lisbon, Portugal. Association for Computational Linguistics.

OpenAI. 2024a. ChatGPT (model GPT4o-mini). Large language model.

OpenAI. 2024b. GPT-4o. Large language model.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

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.

Siyao Peng, Zihang Sun, Sebastian Loftus, and Barbara Plank. 2024. Different tastes of entities: Investigating human label variation in named entity annotations. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 73–81, Malta. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

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.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Tianze Shi, Igor Malioutov, and Ozan Irsoy. 2020. Semantic role labeling as syntactic dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7551–7571, Online. Association for Computational Linguistics.

Amber Stubbs. 2011. MAE and MAI: Lightweight annotation and adjudication tools. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 129–133, Portland, Oregon, USA. Association for Computational Linguistics.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 111–115, Torino, Italia. ELRA and ICCL.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. Revealing fine-grained values and opinions in large language models. *Preprint*, arXiv:2406.19238.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A Additional Results

In this section we provide additional details regarding the results of our experiments.

A.1 Full Results for LLMs Mimicing Annotators

Table 5 provides additional statistics complementing those presented in Table 3. Here, we describe these additional statistics in more detail. At the bottom of Table 5, we present regression model fit statistics. The numbers in parentheses represent the standard error of the estimates, which indicates the uncertainty associated with these estimates and is used for calculating statistical significance.

"Intercept" reflects the overall intercept of the model, representing the expected distance between the LLM's predictions and those of a human, assuming all reference categories and an average age of 0 with an average annotation skill on a prompt of average ambiguity. The other coefficients serve as adjustments to this baseline value. The statistical significance of the intercept itself is not meaningful, and it is the effect sizes that are of primary interest rather than the absolute values of the predicted distances. A "+" sign indicates marginal significance ($0.05 \leq P \leq 0.1$). SD (Standard Deviation) refers to the standard deviations of the mixed-effect model's random intercepts for instances and annotators. The model assigns a unique intercept to each instance and annotator to account for uninteresting idiosyncrasies, with a mean of 0. The SD value indicates the expected variability of LLM-human differences across instances and annotators. "Num. Obs." denotes the number of rows (annotations) on which the regression model was conducted. R2 Marg. and R2 Cond. are measures of

	Offensiveness		Politeness	
Socio-Demographic Attribute	GPT-4	Claude	GPT-4	Claude
Intercept	0.45 (0.17)**	0.56 (0.19)**	1.45 (0.19)***	1.53 (0.20)***
Age (years)	0.01 (0.00)**	0.01 (0.00)**	0.00 (0.00)	0.00 (0.00)
Gender (ref: Male)				
Female	0.00 (0.04)	-0.03 (0.04)	-0.05 (0.03)	-0.05 (0.03)
Non-binary	-0.06 (0.12)	-0.01 (0.13)	-0.05 (0.10)	-0.05 (0.10)
Race (ref: White)				
Asian	0.09 (0.08)	0.03 (0.08)	-0.08 (0.07)	0.00 (0.07)
Black or African American	0.22 (0.06)***	0.19 (0.06)**	0.14 (0.05)**	0.15 (0.05)**
Hispanic or Latino	-0.11 (0.14)	-0.05 (0.15)	0.09 (0.06)	0.12 (0.06)+
Other race/ethnicity	-0.14 (0.13)	-0.26 (0.14)+	-0.17 (0.18)	-0.15 (0.19)
Occupation (ref: Employed)				
Unemployed	0.04 (0.07)	-0.08 (0.07)	-0.08 (0.05)	-0.08 (0.06)
Homemaker	-0.07 (0.10)	-0.02 (0.10)	-0.03 (0.07)	-0.06 (0.07)
Retired	-0.11 (0.07)	-0.13 (0.08)	0.03 (0.06)	0.03 (0.06)
Self-employed	0.04 (0.06)	0.02 (0.07)	-0.03 (0.05)	-0.03 (0.05)
Student	0.13 (0.08)	0.13 (0.09)	-0.11 (0.08)	-0.13 (0.08)
Other occupation	-0.05 (0.15)	0.02 (0.16)	0.08 (0.11)	0.11 (0.11)
Education (ref: Less than high school)				
High school diploma or equivalent	-0.01 (0.15)	-0.08 (0.17)	-0.34 (0.18)+	-0.37 (0.19)*
College degree	0.05 (0.15)	-0.09 (0.17)	-0.43 (0.18)*	-0.48 (0.19)*
Graduate degree	0.06 (0.16)	-0.01 (0.17)	-0.36 (0.18)*	-0.44 (0.19)*
Other education	-0.02 (0.21)	0.01 (0.22)	-0.50 (0.21)*	-0.57 (0.22)**
SD (instance intercepts)	0.44	0.42	0.37	0.33
SD (annotator intercepts)	0.23	0.27	0.30	0.32
Num. Obs.	4500	4481	11151	11151
R2 Marg.	0.014	0.016	0.012	0.014
R2 Cond.	0.316	0.321	0.311	0.305
ICC	0.3	0.3	0.3	0.3

Table 5: Detailed analyses and coefficients indicating the effect of particular human demographic categories on the distance between human and LLM annotations.

the model's explanatory power. R2 Marginal indicates how much variation is accounted for by the coefficients alone, which is very low, suggesting that these coefficients contribute little to predictive power despite some being statistically significant. R2 Conditional, on the other hand, indicates the proportion of variance explained by both the coefficients and the random intercepts, which is significantly higher, demonstrating that the random intercepts contribute more substantially to the model's explanatory power.

The ICC (Intraclass Correlation Coefficient) indicates the extent to which the clustering (by instances and annotators) influences the outcomes (distances between human and LLM ratings). It represents the ratio of total variance in the dependent variable attributed to variance between cluster means, as opposed to variance within clusters. A value of 0.3 or higher suggests that the inclusion of random intercepts is necessary to account for clustering. If the ICC were very low (less than 0.1), a simpler model could be justified.

A.2 Results for Individual Prompts

In examining the scores for the individual prompts presented in Table 6 and Table 7, we observe that there are no notable differences between the two prompts. This lack of variation in scores is a positive outcome, as it suggests consistency in the model's performance across different prompts. Such uniformity reinforces the reliability of the results, indicating that the prompts used do not unduly influence the models' annotation behaviors. This consistency allows us to have greater confidence in our findings and their implications regarding the impact of demographic information on model outputs.

A.3 Results for Placebo Prompting

Table 8 presents the results for placebo prompting. It includes sample sizes and mean distance scores (Δ_μ) for predictions generated using placebo prompts compared to those produced with no-info prompts for the models GPT-4o and Claude

	Offensiveness Prompt 1		Offensiveness Prompt 2	
Socio-Demographic Attribute	GPT-4	Claude	GPT-4	Claude
Intercept	0.51 (0.17)**	0.60 (0.18)***	0.39 (0.18)*	0.52 (0.20)**
Age (years)	0.01 (0.00)**	0.00 (0.00)**	0.01 (0.00)**	0.01 (0.00)***
Gender (ref: Male)				
Female	0.00 (0.04)	-0.03 (0.04)	0.00 (0.04)	-0.02 (0.05)
Non-binary	-0.08 (0.12)	0.01 (0.13)	-0.05 (0.13)	-0.02 (0.14)
Race (ref: White)				
Asian	0.12 (0.07)	0.03 (0.08)	0.06 (0.08)	0.02 (0.09)
Black or African American	0.21 (0.06)***	0.19 (0.06)**	0.22 (0.06)***	0.19 (0.07)**
Hispanic or Latino	-0.08 (0.14)	-0.03 (0.15)	-0.15 (0.15)	-0.07 (0.16)
Other race	-0.11 (0.13)	-0.23 (0.14)+	-0.18 (0.13)	-0.29 (0.15)*
Occupation (ref: Employed)				
Unemployed	0.03 (0.07)	-0.08 (0.07)	0.05 (0.07)	-0.07 (0.08)
Homemaker	-0.08 (0.10)	0.01 (0.10)	-0.05 (0.10)	-0.05 (0.11)
Retired	-0.11 (0.07)	-0.12 (0.08)	-0.11 (0.08)	-0.14 (0.09)+
Self-employed	0.02 (0.06)	0.03 (0.06)	0.05 (0.06)	0.01 (0.07)
Student	0.11 (0.08)	0.11 (0.09)	0.15 (0.09)+	0.16 (0.09)+
Other occupation	-0.07 (0.14)	0.04 (0.15)	-0.04 (0.15)	0.01 (0.16)
Education (ref: Less than high school)				
High school diploma or equivalent	-0.04 (0.15)	-0.08 (0.16)	0.02 (0.16)	-0.07 (0.17)
College degree	0.03 (0.15)	-0.09 (0.16)	0.07 (0.16)	-0.09 (0.18)
Graduate degree	0.02 (0.16)	-0.01 (0.17)	0.10 (0.17)	-0.01 (0.18)
Other education	-0.03 (0.20)	0.04 (0.22)	0.00 (0.22)	-0.03 (0.23)
SD (instance intercepts)	0.48	0.43	0.43	0.41
SD (annotator intercepts)	0.22	0.26	0.25	0.29
Num.Obs.	4500	4481	4500	4481
R2 Marg.	0.012	0.014	0.015	0.018
R2 Cond.	0.325	0.314	0.301	0.317
ICC	0.3	0.3	0.3	0.3

Table 6: Results of the analysis of individual prompts for the offensiveness rating task. This table displays coefficients for the two models (GPT-4o and Claude) and indicates the effects of specific socio-demographic attributes of human annotators on the discrepancies between human and LLM annotations. Positive coefficients signify that the LLM is less accurate in mimicking the responses of annotators from certain demographic categories.

across two rating tasks. The results show that placebo prompting (`P` prompts) does not yield any notable differences relative to no-info prompting (`N` prompts) for specific attribute values. Overall, the scores remain stable across these comparisons.

	Politeness Prompt 1		Politeness Prompt 2	
Socio-Demographic Attribute	GPT-4	Claude	GPT-4	Claude
Intercept	1.40 (0.19)***	1.43 (0.19)***	1.51 (0.19)***	1.65 (0.21)***
Age (years)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Gender (ref: Male)				
Female	-0.05 (0.03)	-0.05 (0.03)	-0.05 (0.03)	-0.05 (0.04)
Non-binary	-0.07 (0.10)	-0.05 (0.10)	-0.04 (0.10)	-0.05 (0.11)
Race (ref: White)				
Asian	-0.08 (0.06)	0.00 (0.07)	-0.08 (0.07)	-0.02 (0.07)
Black or African American	0.13 (0.05)**	0.15 (0.05)**	0.15 (0.05)**	0.16 (0.05)**
Hispanic or Latino	0.07 (0.06)	0.14 (0.06)*	0.10 (0.06)+	0.11 (0.07)
Other race	-0.23 (0.18)	-0.15 (0.18)	-0.11 (0.18)	-0.15 (0.20)
Occupation (ref: Employed)				
Unemployed	-0.08 (0.05)	-0.07 (0.05)	-0.07 (0.05)	-0.10 (0.06)+
Homemaker	-0.03 (0.07)	-0.05 (0.07)	-0.03 (0.07)	-0.06 (0.07)
Retired	0.04 (0.06)	0.04 (0.06)	0.02 (0.06)	0.03 (0.06)
Self-employed	-0.02 (0.05)	-0.02 (0.05)	-0.04 (0.05)	-0.03 (0.06)
Student	-0.10 (0.08)	-0.11 (0.08)	-0.11 (0.08)	-0.14 (0.08)+
Other occupation	0.08 (0.11)	0.10 (0.11)	0.08 (0.11)	0.11 (0.12)
Education (ref: Less than high school)				
High school diploma or equivalent	-0.31 (0.18)+	-0.35 (0.18)+	-0.37 (0.18)*	-0.40 (0.20)*
College degree	-0.41 (0.18)*	-0.45 (0.18)*	-0.46 (0.18)*	-0.52 (0.20)**
Graduate degree	-0.33 (0.18)+	-0.41 (0.18)*	-0.40 (0.19)*	-0.49 (0.20)*
Other education	-0.46 (0.21)*	-0.56 (0.21)**	-0.54 (0.21)*	-0.60 (0.23)**
SD (instance intercepts)	0.41	0.34	0.44	0.40
SD (annotator intercepts)	0.30	0.30	0.30	0.34
Num.Obs.	11151	11151	11151	11151
R2 Marg.	0.011	0.013	0.010	0.013
R2 Cond.	0.322	0.292	0.340	0.340
BIC	28301.7	27294.4	28833.6	28424.2
ICC	0.3	0.3	0.3	0.3
RMSE	0.67	0.66	0.67	0.67

Table 7: Results of the analysis of individual prompts for the politeness rating task. This table presents the coefficients for the two models (GPT-4o and Claude), demonstrating how different socio-demographic characteristics of human annotators influence the distance between human and LLM annotations. Positive coefficients indicate a lower accuracy of the LLM in reflecting the views of annotators from particular demographic categories.

	Offensiveness			Politeness		
Placebo Attributes	Count	Δ_μ (GPT-4o)	Δ_μ (Claude)	Count	Δ_μ (GPT-4o)	Δ_μ (Claude)
Height						
140 cm	601	0.23	0.19	1,380	0.27	0.19
150 cm	555	0.27	0.20	1,393	0.27	0.18
160 cm	579	0.23	0.19	1,417	0.26	0.20
170 cm	536	0.20	0.20	1,407	0.27	0.18
180 cm	574	0.24	0.20	1,354	0.24	0.18
190 cm	564	0.28	0.23	1,385	0.25	0.18
200 cm	572	0.24	0.24	1,446	0.27	0.18
210 cm	519	0.25	0.22	1,369	0.25	0.18
Zodiac sign						
Aries	395	0.24	0.22	889	0.25	0.17
Taurus	382	0.26	0.20	913	0.27	0.18
Gemini	355	0.21	0.23	895	0.26	0.18
Cancer	383	0.29	0.27	956	0.27	0.18
Leo	378	0.24	0.22	928	0.27	0.19
Virgo	360	0.22	0.23	915	0.26	0.19
Libra	398	0.23	0.19	974	0.24	0.19
Scorpio	390	0.25	0.18	975	0.26	0.19
Sagittarius	352	0.23	0.20	919	0.27	0.19
Capricorn	387	0.25	0.21	983	0.27	0.18
Aquarius	365	0.28	0.21	892	0.25	0.17
Pisces	355	0.23	0.17	912	0.24	0.20
House number						
6	460	0.26	0.21	1,130	0.28	0.19
12	446	0.24	0.22	1,090	0.26	0.18
13	447	0.23	0.21	1,107	0.26	0.19
24	424	0.27	0.22	1,103	0.25	0.18
45	455	0.24	0.18	1,123	0.26	0.19
68	438	0.22	0.19	1,098	0.26	0.18
98	456	0.27	0.23	1,190	0.25	0.20
122	465	0.23	0.21	1,116	0.26	0.19
234	466	0.23	0.21	1,118	0.26	0.19
1265	443	0.26	0.22	1,076	0.26	0.17
Enjoying						
food	1,468	0.25	0.22	3,793	0.26	0.19
sleep	1,486	0.23	0.21	3,662	0.26	0.18
friends	1,546	0.25	0.21	3,696	0.26	0.19
Favorite colour						
red	436	0.25	0.22	1,046	0.28	0.16
green	399	0.27	0.24	1,013	0.26	0.18
blue	400	0.23	0.21	997	0.27	0.19
yellow	444	0.24	0.22	984	0.25	0.19
purple	443	0.27	0.20	1,055	0.25	0.20
turquoise	367	0.23	0.20	1,056	0.23	0.21
orange	405	0.23	0.21	1,033	0.26	0.18
pink	371	0.25	0.21	976	0.28	0.16
black	441	0.21	0.20	993	0.25	0.17
white	381	0.27	0.21	991	0.28	0.20
brown	413	0.21	0.21	1,007	0.26	0.19

Table 8: Sample sizes and mean distance scores of predictions for placebo prompting (P prompts) predictions in comparison to predictions with N prompts for models GPT-4o and Claude at two rating tasks.