Personas with Attitudes: Controlling LLMs for Diverse Data Annotation

Anonymous ACL submission

Abstract

We present a novel approach for enhancing diversity and control in data annotation tasks by personalizing large language models (LLMs). We investigate the impact of injecting diverse persona descriptions into LLM prompts across two studies, exploring whether personas increase annotation diversity and whether the impacts of individual personas on the resulting annotations are consistent and controllable. Our results indicate that persona-prompted LLMs generate more diverse annotations than LLMs prompted without personas, and that the effects of personas on LLM annotations align with subjective differences in human annotations. These effects are both controllable and repeatable, making our approach a valuable tool for enhancing data annotation in subjective NLP tasks such as toxicity detection.

Content Warning: This document shows content that some may find disturbing, including content that is hateful towards protected groups.

1 Introduction

Many diverse NLP tasks rely on data acquired through human annotations, often collected via crowdsourcing studies. Röttger et al. (2022) recently offered a detailed introduction and discussion of the two different paradigms available to handle label differences resulting from diverse beliefs and backgrounds of individual annotators. The *prescriptive* paradigm discourages annotator subjectivity and strives towards a single label for each instance in a dataset, effectively enabling the training of models that consistently apply a single understanding of the construct. The ease with which the resulting data may be used for model training makes this paradigm popular in NLP, where a simple majority vote is oftentimes applied to the annotations offered by a diverse pool of crowdworkers to derive a single groundtruth label for the annotated instances. Thereby, this paradigm is akin to the "wisdom of the crowd" approach, assuming that the "average opinion" of multiple, diverse annotators would sufficiently approximate the desired groundtruth label.

The second and alternative paradigm discussed by Röttger et al. (2022) is the *descriptive* paradigm, in which annotator subjectivity is explicitly encouraged and used to gain insights into diverse beliefs and to improve model training and evaluation via the consideration of annotator disagreement. Röttger et al. (2022) conclude that neither of the two paradigms is inherently superior, but that both serve different purposes and applications - for example, the prescriptive paradigm makes it easier to train classification models on the resulting data, while only the descriptive paradigm allows researchers to study and understand differences in perceptions across annotators with diverse beliefs and backgrounds.

In this work, we propose to combine the idea of using LLMs as annotators with the large pool of personas offered by the Persona Hub (Ge et al., 2024) to both increase and control the diversity of the generated annotations.

To explore the feasibility of injecting personas into LLM prompts to diversify and steer the models' zero-shot annotations in all its facets, we organize this work into two studies. Study 1 covers the *prescriptive* paradigm towards diverse annotations, assuming the existence of a single label per instance and evaluating our approach's annotation diversity by comparing the persona-prompted LLM annotations to it. Study 2, in contrast, is in line with the *descriptive* paradigm, exploring the approach's ability to reconstruct the diversity found in human annotations and to controllably replicate the observed effects of human subjectivity. While both studies individually test the suitability of our LLM persona-prompt approach for the corresponding paradigm, they - taken together - are designed

to establish that our approach is suitable to increase the diversity in LLM annotations (Study 1) and that the persona impact is not entirely random, but that it follows predictable and controllable patterns in line with patterns of subjectivity observed for human annotations (Study 2). In this paper, therefore, we set out to answer the following two research questions:

- **RQ Study 1**: Does the inclusion of persona descriptions in LLM-prompts consistently increase the diversity of the resulting LLM annotations?
- **RQ Study 2**: Does the inclusion of persona descriptions in LLM-prompts lead to controllable annotation patterns, and do these patterns align with effects of subjectivity observed for human annotators?

2 Related Work

Researchers have explored the abilities and performance of LLMs in annotating datasets for different types of constructs. Ziems et al. (2024), Faggioli et al. (2023), Gilardi et al. (2023) and Pavlovic and Poesio (2024) provide overviews of scenarios for which the idea of using LLMs as annotators has already been tested, including tasks as diverse as determining the relevance of texts for specific issues, the detection of humor, or the extraction of medical information. Other researchers argued that LLMs are particularly well suited for the annotation of subjective constructs like hate speech, offensive language and toxicity (Li et al., 2024). They argue that using LLMs can counter the instability in annotations that often arises from the varying social backgrounds of human annotators. Even more recently, researchers have started to explore the performance impacts of aggregating annotations generated with different LLMs (Del Arco et al., 2024; Schoenegger et al., 2024), acting upon the assumption that an increased diversity of the crowd of annotators - be they human or LLM - would lead to gains in performance. He et al. (2024) explore yet another angle of LLM annotations by comparing the annotation performance of GPT-4 with the annotations resulting from a carefully designed and conducted crowdworker annotation pipeline.

Yet another line of research is moving beyond the use of LLMs to predict the groundtruth label only, proposing to *personalize* LLMs via the inclusion of socio-demographic information in order to steer the LLM annotations towards those provided by human annotators. Among the most prominent approaches are Argyle et al. (2023), Bisbee et al. (2023) and Santurkar et al. (2023), who explore the ability of LLMs to predict survey responses of individual participants, as well as Beck et al. (2024), Pei and Jurgens (2023), Sun et al. (2023) and Orlikowski et al. (2023), who evaluate the performance of personalized LLMs in predicting the annotations of individual annotators as well as the resulting majority vote labels.

While the generation of dataset labels via LLMs might be interpreted as a form of synthetic training data generation, this description is usually reserved for efforts that synthetically create the instances to annotate, not (only) the corresponding labels. Timpone and Yang (2024) offer an extensive review of the state of the literature together with a detailed discussion of associated challenges and opportunities. Fundamental for this work, Ge et al. (2024) introduce the Persona Hub, a collection of 1,000,000,000 diverse persona descriptions, as a way to diversify the synthetic instances that LLMs generate, and show that their persona descriptions - when included in the prompts used to synthesize, e.g., novel math or logical reasoning problems - are successful in increasing the diversity of the resulting datasets and thereby also the generalizability of the models trained on the datasets' tasks.

3 Data

To systematically test the impact of including persona descriptions in the prompts used to collect toxicity annotations from different LLMs, we rely on two external sources of data described next.

3.1 Persona Descriptions

Central to our proposal to increase the diversity of LLM-generated annotations via the injection of personas into the prompt is the collection of personas introduced by Ge et al. (2024) via their Persona Hub. While the personas themselves are just a brief, natural language description of an - ideally - human individual, the scale and diversity of the collection is what makes the Persona Hub such an ideal resource for our approach. Ge et al. (2024) developed the Persona Hub as part of a novel paradigm for the creation of synthetic data, not driven by seed datasets or manual prompt-design, but by a large number of personas to be automatically injected into LLM prompts.

ID	Persona
189476	An experienced biomedical engineer who has successfully brought cognitive rehabilitation devices to market
11276	A project manager who is skeptical about the practicality and cost-effectiveness of containerization
123381	A Muslim immigrant seeking legal assistance in defending their right to religious expression
81722	a French postgraduate student focusing on multidisciplinary research
73592	A retiree living in the Auckland suburbs who often has Sunday barbecues with her neighbors and enjoys discussing local issues with them.

Table 1: Random sample of five persona descriptions used in our experiments.

Their persona collection features brief descriptions of more than 1 billion different personas, created by asking different LLMs for a shown webtext instance: "who is likely to [read|write|like|dislike|...] this text". Depending on the prompt as well as the nature and level of detail of the webtext instance, the LLM will come up with different persona descriptions in response, varying both in content and complexity. Table 1 shows a selection of different personas included in the Persona Hub and the Appendix provides additional detail on the approach used by Ge et al. (2024) to create the persona descriptions as well as our efforts to clean them.

3.2 Toxicity Datasets

As the basis for our experiments, we use the datasets prepared and shared by Sap et al. (2022). To study the impact of human annotators' beliefs and identities on their annotations, Sap et al. (2022) curate datasets of potentially toxic texts with different characteristics. From the datasets shared by Founta et al. (2018) and Vidgen et al. (2021), they select those posts that are either vulgar, anti-Black, or detected to be African-American English.

For their breadth of posts study, Sap et al. (2022) collect 571 posts that are annotated by 173 Amazon Mechanical Turk (MTurk) crowdworkers for their levels of toxicity, amounting to six annotations per post. While annotators indicated the level of toxicity on a 5-point Likert scale, we first calculate the average level over the six annotations and then map this to a binary toxic/not toxic outcome, considering all instances with an average toxicity level larger than 2.5 as toxic. We use this dataset of 571 instances with binary toxicity labels resulting from a majority vote of six individual annotators as data basis for our Study 1 and refer to it as the SINGLE-LABEL DATASET.

For their breadth of workers study, Sap et al. (2022) carefully select 12 posts that are then annotated for their levels of toxicity by a large population of 641 annotators. The annotators have been recruited from MTurk to be diverse along racial, political, and gender identities. Each annotator annotated every post, (subjectively) answering the question "How toxic/hateful/disrespectful or offensive does this post seem to you?" [1]. For our Study 2, we use the non-aggregated responses provided by the individual annotators, with responses for each instance ranging from "Not at all" (1) to "Very much so" (5) on a 5-point Likert scale. We refer to this dataset of 12 posts annotated by 641 annotators with labels from 1 to 5 as the MULTI-LABEL DATASET.

4 Methodology

Experiments in all studies were conducted with two different LLMs from two popular families of open-source models - the Mistral-7B-Instruct-v0.1 model and the Qwen2-7B-Instruct model. Further justification for our choice of models and details on model deployment can be found in Appendix A.3.

4.1 Study 1

By answering RQ1, we want to establish that the inclusion of persona descriptions into the prompts used to generate LLM annotations consistently increases the diversity of the models' annotation decisions, especially when compared to a baseline in which no persona description is added to the same prompt.

Diversity We randomly sample 1,000 personas from the Persona Hub and collect their annotations on the SINGLE-LABEL DATASET. We collect annotations by injecting the persona description directly into the prompt and asking for a binary label response using *Prompt Template 1* shown in Appendix A.2. To compare these persona-prompted LLM annotations against LLM annotations with-

[1] In early experiments, we also tried the alternative question asked by Sap et al. (2022), inquiring about the perceived toxicity of the post "to anyone" instead of "to you". Since early results did not differ much, we decided to focus on the "to you" variant of the question.

out any persona-influence, we run the same models 1,000 times without any personas included in the prompt using *Prompt Template 2*. We refer to the generated annotations without persona-influence as baseline LLM annotations. The variation in this setup is expected to originate exclusively from the randomness of the sampling process and a temperature-setting of 1 across different annotation runs of the same model.

Consistency To further establish that the effect of including personas in LLM prompts are not randomly fluctuating, but that the inclusion of specific personas in prompts has a consistent effect on the resulting annotations, we select the 30 personas with the highest, median and lowest alignment to the SINGLE-LABEL DATASET labels (as measured via the macro-average F1 score) and let each of them annotate the dataset 30 additional times.

4.2 Study 2

Through RQ2, we want to show that the diversity introduced through persona prompting follows controllable patterns that align with those found in human annotations.

Exploratory Analysis of Annotation Patterns To explore the patterns that drive differences in annotations, we create two different embedding spaces in which the personas' descriptions and labels are projected, allowing us to calculate distances between the different personas.

First, we use a pre-trained sentence-transformer model to project our persona descriptions into an embedding vector space. [2] We refer to this embedding space as the *persona space*.

Second, we construct an embedding space from the personas' annotations for the MULTI-LABEL DATASET. Each of the twelve instances in the dataset is represented as a dimension in the label embedding space, with values for each dimension ranging from 1 to 5 and corresponding to the possible toxicity labels. We use the persona's annotations for the MULTI-LABEL DATASET to project each persona into the label embedding space. The annotations are collected using *Prompt Template 3* shown in Appendix A.2, soliciting toxicity levels on a 5-point Likert scale. We refer to this embedding space as the *label space*.

We start our analysis of the embedding spaces by using k-means to find clusters in the *persona space*, i.e., persona descriptions that are similar to each other. We then calculate the average distances in the *label space* between the persona clusters. This results in a symmetric matrix with dimensionality equal to the number of *persona space* clusters, where each entry represents the average distance between two persona clusters in the *label space*. Based on that matrix, we can identify persona clusters that annotate alike as well as those that annotate very differently from each other.

Additionally, we test the assumption that similar personas (i.e., small distance in the *persona space*) annotate alike (i.e., small distance in the *label space*). For each persona, we calculate the pairwise distances to every other persona in both spaces and measure the correlation between these distances.

Alignment with Human Annotators To test whether the annotation patterns we find for the persona-prompted LLM annotations are in line with the annotation patterns displayed by human annotators, we first formulate expectations of subjectivity effects based on the findings of Sap et al. (2022). For their human annotators, they showed the following effects:

- **Effect 1**: Conservative annotators are less likely to rate anti-Black posts as toxic,
- **Effect 2**: Conservative annotators are more likely to rate African-American English (AAE) posts as toxic, and
- **Effect 3**: Black annotators rate anti-Black posts as more toxic than White annotators.

They derive a further effect from theory, for which they fail to find conclusive evidence in their collected annotations:

- **Effect 4**: Black annotators rate AAE posts as less toxic than White annotators.

We propose to test whether persona-prompting replicates these effects by comparing the annotations collected from personas that are explicitly marked as conservative and Black. We do this by first identifying neutral personas that are not explicitly assigned to an ethnicity or an ideology. We then create variants by injecting explicit markers (the terms "black" and "conservative") at manually selected, adequate positions in the persona descriptions. This results in three different groups,

[2] We use all-MiniLM-L12-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

all based on the same set of neutral persona descriptions - ethnically and ideologically neutral personas, personas manually changed to be identifiable as Black and personas manually changed to be identifiable as conservative (see Appendix Table A.1 for example personas of each group). We then use these persona groups to annotate subsets of the SINGLE-LABEL DATASET that are identified as anti-Black or as AAE by Sap et al. (2022), again soliciting annotations on the 5-point Likert scale introduced above.

5 Results

The following sections present the results of the experiments described above. All results can be reproduced using the code made available upon publication together with the publicly available datasets shared by Sap et al. (2022) and Ge et al. (2024). [3]

5.1 Study 1

In Study 1, we start with establishing the increased diversity and the consistency of LLM annotations in the persona-prompting approach.

Diversity When we examine the alignment of annotation runs with and without persona descriptions used in the prompt with the human majority vote labels represented by the SINGLE-LABEL DATASET, the first thing to notice is the great increase in the diversity of alignment levels between the persona-prompted LLMs and human annotators. Panels a) and c) in Figure 1 show boxplots of the distributions of the annotation performances (measured via macro-average F1 scores) resulting from prompting with the 1,000 sampled personas and from running the model 1,000 times without personas.

For Mistral, the baseline LLM annotations are generally better aligned with the majority vote human annotations than the persona-prompted LLM annotations. More importantly, however, persona-prompted annotation runs exhibit significantly more fluctuation in the resulting levels of alignment to the labels in the SINGLE-LABEL DATASET, indicating a higher opinion diversity introduced by the persona descriptions.

For Qwen, we observe that the median persona-prompted LLM annotation runs align slightly better with the human majority vote label than the median baseline LLM runs. Nonetheless and parallel to Mistral, we observe a much higher variance

[3]For purposes of the review, we make the code available as part of the submission's supplementary materials.

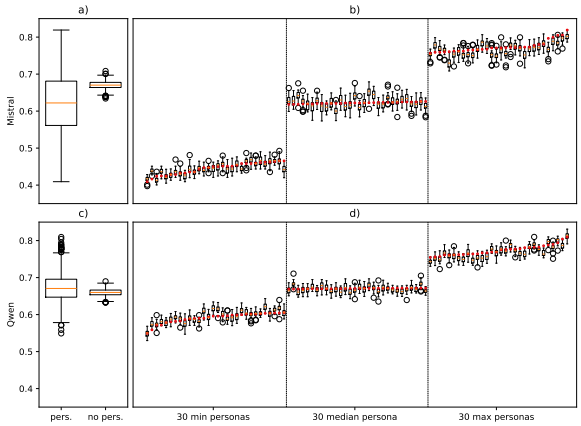

Figure 1: Boxplots of macro-average F1 scores achieved in 1,000 different persona-based LLM annotation (*pers.*) and 1,000 baseline LLM annotation runs (*no pers.*) for a) Mistral and c) Qwen, showing the increased diversity introduced by personas. Boxplots of macro-average F1 scores achieved in 30 additional annotation runs for the 30 personas with min, median and max alignment to the human majority vote label for b) Mistral and d) Qwen, showing the consistency of the persona-prompting.

in annotation alignment for the persona-prompted annotations.

This initial analysis of the various annotation runs leads to an important conclusion: The introduction of personas into LLM prompts broadens the distribution of performances across annotation runs with both models. We confirm this finding through a Levene test for equality of variances, which for both LLMs rejects the null hypothesis of equal variances at significance levels of $\alpha = 0.001$. In other words, the personalization shifts the LLMs further away from their baseline performance than the typical randomness introduced by the sampling procedure does, showing that the inclusion of personas indeed increases the diversity of LLM annotations.

Consistency Next, we test whether the effects of persona descriptions are consistent and stable across multiple runs and thus controllable, or whether the personas impact annotations randomly. Each boxplot in panels b) and d) of Figure 1 represents 30 annotation runs with the same persona. For both models, we see how the order of the achieved F1-scores is almost perfectly restored when running each persona multiple times. We take this as confirmation that the annotation differences associated with different personas are not purely contingent, but that the same personas consistently push the models into the same perspectives.

This is another important finding for the proposed persona-based annotation approach, as it establishes not just the consistency and stability of the persona-based annotation runs, but indicates also a degree of control that is required to steer the models towards specific annotation perspectives.

Qualitative Analysis of Annotation Patterns In an attempt to identify characteristics of personas that lead to particularly weak and strong alignment between persona-prompted LLM annotations and the human majority vote labels, we manually search for themes and patterns in the descriptions of the personas with the minimum and maximum alignment levels.

For Mistral, we find that personas with high alignment to the human majority vote labels are described as "appreciative", as "interested in" different questions and topics, as well as "offering" or "seeking advice". In contrast, for personas with low alignment, the term "competitive" occurs most frequently in the persona descriptions, together with expressions of "being against" something.

Interestingly, the tendency that personas described as more open and outreaching achieve higher alignment than personas defined as fundamentally in opposition to something or someone is pretty much inverted for Qwen. There, we find that the personas described as "being critical", as "skeptical" or "questioning" of something have higher alignment, while the personas that "share" things or "seek" and "offer" advice have lower alignment.

The opposing directions of the effect for the two models makes it inherently difficult to meaningfully interpret, but one certain conclusion is that character traits and psychological attributes seem to be more important for annotation diversity than socio-demographic attributes, at least for the extreme ends of the widened persona distributions.

5.2 Study 2

After having established in Study 1 that different personas consistently lead to different levels of alignment with the human majority opinion, we are now taking a systematic look at the label patterns associated with different personas.

Exploratory Analysis of Annotation Patterns We select a clustering solution in the *persona space* with 2,180 different clusters, using a similarity threshold of 0.6 for cluster formation (see Appendix for justification and a basic evaluation of our clustering). Figure 2 shows the intra- and inter-cluster cosine distances in the *label space* for the annotations of the persona clusters with Qwen, and Appendix Figure A.2 shows the same for the persona cluster annotations using Mistral.

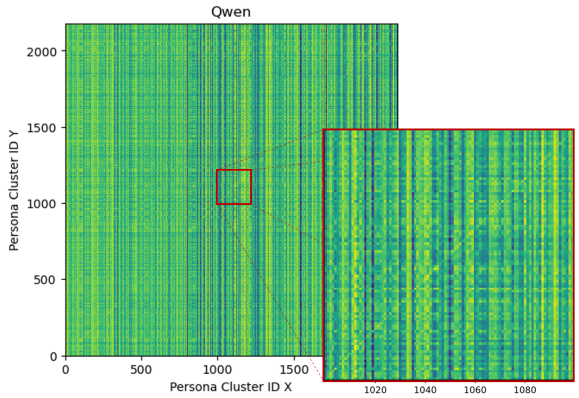

Figure 2: Intra- and inter-cluster cosine distances of persona-space clusters measured in *label space* resulting from Qwen annotations. Values are normalized per row. Lighter-colored cells represent lower average distances between the respective clusters, the lighter colors along the diagonal thus indicate that similar personas annotate alike. The inset zooms in on clusters with IDs from 1,000 to 1,100.

For both models, we see that the clusters along the diagonal are lighter in color, indicating that personas that ended up in the same cluster based on their descriptions are also relatively close to each other in the *label space*, i.e., personas with similar descriptions tend to annotate alike.

These first indications of a positive association between distances in the *persona space* and the *label space* are further confirmed by the pairwise cor-

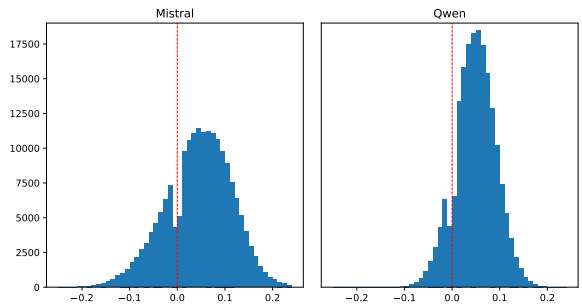

Figure 3: Histograms of Spearman correlation coefficients for pairwise distances measured in the *persona* and the *label space*. A single correlation coefficient represents the correlation between distances from a specific persona to every other persona in both spaces. The mostly positive correlations between distances in both spaces confirms that similar personas annotate alike.

relation results shown in Figure 3. For both models, more than 95% of pairwise Spearman correlation coefficients between inter-persona distances measured in both spaces are significantly different from zero, with 75.5% of these significant correlation coefficients being positive for Mistral. For Qwen, the number of significantly positive correlation coefficients for pairwise distances in the two spaces is with 88.3% even higher.

This is another central finding for our proposed approach, as it establishes that similar persona descriptions lead to similar annotation outcomes - yet another indication that the persona descriptions allow for control of the annotation perspectives taken by the model beyond purely random differences.

Alignment with Human Annotators Next, we test whether personas marked as Black and conservative lead to LLM annotations that replicate the effects of subjectivity observed in annotations produced by real humans as described above. Figure 4 shows the mean toxicity level shifts for the Black and conservative personas relative to their neutral versions across all instances in the AAE and anti-Black datasets.

Note that the observed shifts are relatively small, with Wilcoxon rank-sum tests for the theorized effects revealing that only the effects for Qwen on the anti-Black instances are statistically significant at significance levels of $\alpha = 0.05$. This, however, is not too surprising, given that even for human annotations the significance of demographic effects on annotation patterns is difficult to establish.

However, we observe for the AAE instances in both models shifts in toxicity levels that are in line with **Effects 2** and **4**, i.e., that personas marked as conservative tend to perceive these instances as more toxic and personas marked as Black as less toxic. These effects hold both when compared to the neutral personas as well as when comparing Black and conservative annotations directly. For Qwen, the mean (absolute) toxicity level assigned to AAE instances across Black personas is $\mu_{Q;B} = 3.39$ and $\mu_{Q;C} = 3.43$ across conservative personas. For Mistral, the values are $\mu_{M;B} = 2.67$ and $\mu_{M;C} = 2.76$.

According to **Effect 3**, we would expect Black personas to rate anti-Black instances as more toxic. This effect is to a degree confirmed for Qwen, where Black personas tend to annotate anti-Black posts as slightly more toxic than their neutral counterparts. Additionally and in line with **Effect 1**,

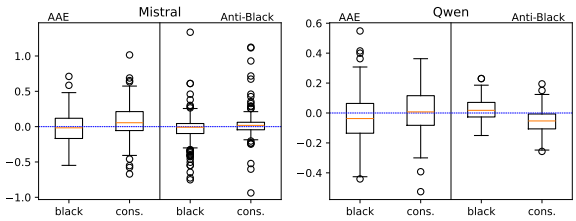

Figure 4: Boxplots of shifts in average toxicity labels assigned to instances in the AAE and anti-Black datasets. The shifts are on a persona-level and are calculated as the difference in average toxicity label of the manually changed black and conservative personas relative to the original, neutral persona. We see first evidence that the human annotation effects are replicated through the personas.

conservative personas tend to annotate the same posts as less toxic than the neutral personas. For Qwen, the mean (absolute) toxicity level assigned to anti-Black instances across Black personas is $\mu_{Q;B} = 4.69$ and $\mu_{Q;C} = 4.61$ across conservative personas.

For Mistral, the shifts caused by the inclusion of markers for Blackness and conservatism are far less pronounced on the anti-Black posts and are, if different from zero at all, shifted in a direction that is not in line with **Effects 1** and **3**. For Mistral, the mean toxicity level for anti-Black posts from Black personas is $\mu_{M;B} = 4.51$ and $\mu_{M;C} = 4.58$ from conservative personas.

Qualitative Analysis of Annotation Patterns While the quantitative shifts are relatively small and not in all settings entirely conclusive, a closer look at some of the instances for which Black and conservative annotators differ in their toxicity annotations most strongly is instructive. Table 2 shows the three instances in the AAE dataset for which the differences between the absolute average toxicity level assigned by conservative and Black personas via the Qwen model are the largest.

All of the shown instances for which Black annotators assigned a (much) lower toxicity level than their conservative counterparts are examples of a reclaimed usage of the n-word and thus examples of an explicitly non-toxic usage of a term usually used as a slur. This finding perfectly mirrors what Sap et al. (2022) observe for human annotators, where "raters who are more conservative tend to score those posts [containing the n-word] as significantly more racist". This indicates that the inclusion of the Black marker in the persona prompt triggers an awareness in the model for the possible use of the

AAE Instance	Diff.	Anti-Black Instance	Diff.
whats up my *n****	-2.31	I was raised in Idaho, amazing place. There is a reason they call it whiteaho. There was 1 single black in my town growing up. He was a halfie and wasn't too bad. Though still did the whole poor black me whine. [...]	1.09
ma brotha ma *n****	-1.36	I always hated visiting *n**** York, but if I'm ever forced to I'll be sure to pack all my Confederate clothes and flags and to dance around Central Park! (I may need a little protection)	0.69
I don't even wanna get married but you know if my *n**** propose to me imma say yes plz	-1.16	Side note, can anyone explain why racism is necessarily bad? [...] Only on the last 60 years has it been an issue for the SOME PEOPLE.	0.53

Table 2: AAE and anti-Black instances with largest difference in average toxicity ratings assigned by Black and conservative personas. Differences calculated as Diff $= \mu_{Q;B} - \mu_{Q;C}$ based on Qwen annotations. We replace all mentions of the n-word and its variants with *n**** and mark omissions due to length constraints with [...].

n-word in a reclaimed, colloquial manner, a usage that should not be annotated as toxic.

Table 2 further shows the three instances in the anti-Black dataset for which the differences between the absolute toxicity level assigned by Black and conservative personas were the largest - i.e., those instances, for which Black personas on average assigned a higher absolute toxicity level than conservative personas did. These instances here are blatantly racist, thus confirming that the LLM prompted with Black personas has a higher sensitivity for racist contents and accordingly rates it as more toxic than the non-Black personas do. Both observations are also true for annotations generated via Mistral, as shown in Appendix Table A.4.

6 Discussion and Conclusion

In this work we explored the potential of personalizing LLMs through persona-based prompts to enhance diversity and control in data annotation tasks. By injecting persona descriptions into LLM prompts, we observed an increase in the variability of model annotations compared to annotation runs that did not include persona descriptions, demonstrating through various experiments that personas can influence model outputs in a consistent and controllable manner. We show that our persona-based approach to LLM data annotations offers a novel way to simulate human subjectivity in annotations, which can be particularly useful in tasks that require diverse and subjective perspectives, such as the detection of toxicity.

Our findings suggest that personas not only introduce desirable diversity in annotations, but that they also enable researchers to guide LLMs toward specific annotation behaviors, making them - under certain conditions - more aligned with groups of human annotators and being successful in replicating effects of annotation subjectivity also found in human annotations.

7 Limitations

Our study is not without limitations. First, we restricted our analysis to two open-source LLMs. While we intended to include other models, such as LLaMA 3.1 and Falcon, different challenges unconnected to our proposed approach made their use for our purpose impossible - for various LLaMA models, the guardrails stopped the model from consistently complying with the toxicity annotation task, and for Falcon, the model's general ability to comply with the prompt instructions was insufficient for producing meaningful annotations. While our study establishes that the injection of personas into LLM prompt leads to the same effect of widening the annotation performance distribution across different models, future research could still investigate less complex subjective constructs, constructs that do not depend on potentially harmful language (e.g., sentiment detection), as well as additional model families and sizes, including those with (strong) guardrails.

Additionally, there are several limitations that we inherit from our use of the Persona Hub (Ge et al., 2024) dataset. Importantly, our experimental study was conducted on a sample provided by the authors rather than the full dataset. This limitation may introduce sampling biases when certain demographic groups are captured in the sample while others are excluded, thereby potentially reducing the diversity effects observable in our analyses. Furthermore, we cannot guarantee that all persona descriptions included in the sample represent individual humans (rather than groups of individuals or non-human characters like animals or even objects) and are written in English, although we took measures to

filter out any persona descriptions written in languages other than English. Importantly, we do not have any control over the focus and make up of the persona descriptions. While this is not a necessary condition for our goal of showing that persona descriptions increase annotation diversity, we speculate that control over the information included in the descriptions would probably even lead to more significant effects than what we observed. We note that many of the personas have professions or hobbies as the most important descriptor, which are probably less important dimensions along which perceptions of toxicity differ than, e.g., dimensions such as race, gender or political ideology. Future research could explore the annotation effects caused by personas that differ along dimensions that are known from theory to be important factors for the annotation task at hand.

8 Ethical Considerations

The successful personalization of LLM annotations and the output control that comes with it is not without risks. First, there is the risk that bad actors could exploit the approach to identify personas corresponding to destructive or abusive perspectives and abuse them for the generation of harmful content. Second, the control over LLMs that allows to tailor their outputs to the preferences of individuals includes the risk of abusing this ability, potentially allowing bad actors to persuade them into actions and reactions possibly harmful to themselves.

References

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

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.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2023. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, pages 1–16.

Flor Miriam Plaza Del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds. exploring model label variation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 19–30.

Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. *arXiv preprint arXiv:2405.01299*.

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.

Paul Röttger, 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.

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, 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.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, and Philip E Tetlock. 2024. Wisdom of the silicon crowd: Llm ensemble prediction capabilities match human crowd accuracy. *arXiv preprint arXiv:2402.19379*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Richard Timpone and Yongwei Yang. 2024. Artificial data, real insights: Evaluating opportunities and risks of expanding the data ecosystem with synthetic data. *arXiv preprint arXiv:2408.15260*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682.

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 Appendix

This Appendix is organized in sections that provide additional details on the Persona Hub dataset we use, the LLM prompts designed to collect (personalized) annotations from the models, the model deployment, as well as the results of the two experimental studies.

A.1 Personas

Ge et al. (2024) use two different approaches to automatically create personas from webtext (i.e., large-scale collections of text supposed to represent *all text on the web*); text-to-persona, as described in the main part above, as well as persona-to-persona, an approach designed to complete the persona collection by leading the persona-generating LLM to consider personas beyond those visible and represented in the web, e.g., children, via their relations to the personas obtained from the text-to-persona approach. Their persona-to-persona prompt asks for any already created persona "who is in close relationship with the given persona" for up to six iterations, thereby enriching and diversifying the initial persona collection. Personas are then deduplicated based on embedding proximity as well as ngram-overlaps.

In our experiments on crowd size and annotation diversity, we use the 200,000 personas that are publicly available (as of 03.07.2024). However, we noticed that some of these personas were written in languages other than English. Since we design our experiments assuming that the persona to be injected into the prompt is in English and since we do not (explicitly) leverage a multilingual LLM for the annotations, we make an effort to filter out any persona descriptions that are not written in English. To do so, we use the langdetect package [4] to get a list of persona descriptions that the classifier labels as primarily written in a language other than English. Since we noticed that the classifier showed a slight tendency of producing false positives, i.e., flagging descriptions written in English as non-English, we manually went through this list and only selected those for removal that were clearly written in a language other than English. This cleaning process left us with a dataset of 198,039 different personas. During our experiments with the persona collection, we also noticed that some of the descriptions did not describe individual human personas, but groups of individuals or institutions. In one case, we even came across a persona description corresponding to a Brazilian spider. However, given the extremely low prevalence of these non-personas, the difficulty of finding and removing them, as well as the assumption that groups of individuals or institutions diversify annotations in a similar manner and direction as the single individuals they represent would do, we accept and acknowledge the inclusion of some non-personas as a limitation of our work.

For Study 2, we select and modify neutral personas to isolate and measure the effects associated with different annotator characteristics. Table A.1 shows three example personas manually selected

[4] https://pypi.org/project/langdetect/

Persona ID	Neutral Persona with Replacement Token
130831	[ATOKEN] political science professor writing their first book about democracy
164597	[ATOKEN] receptionist at a boutique hotel who hates fake news
82521	An internationally recognized [TOKEN] car restoration expert with a web-based reality show

Table A.1: Persona descriptions selected for their undefined ethnicity and ideology. These descriptions are changed into Black and conservative personas by replacing [ATOKEN] with "a black" and "a conservative" and [TOKEN] with "black" and "conservative", respectively.

for being ethnically and ideologically undefined, or neutral, as well as versions of the same personas that are explicitly marked as Black and conservative.

A.2 LLM Prompts

This section shows the prompts used for soliciting the toxicity annotations from the LLMs. In each prompt, the [text]-token marks where the text to be annotated is inserted into the prompt. In prompts that are persona-injected, the [persona]-token marks where the persona description is inserted. In Study 1, we were interested in binary toxicity labels, using the following prompts for annotations with (Prompt Template 1) and without (Prompt Template 2) personas:

> *Prompt Template 1*
> <TASK> Your task is to annotate the post shown below for toxicity, taking on the perspective of the following persona: [persona]. Respond with 'TRUE' if this post seems toxic/hateful/disrespectful to you and 'FALSE' if this post does not seem toxic/hateful/disrespectful to you. <POST> [text]. <RESPONSE>

> *Prompt Template 2*
> <TASK> Your task is to annotate the post shown below for toxicity. Respond with 'TRUE' if this post seems toxic/hateful/disrespectful to you and 'FALSE' if this post does not seem toxic/hateful/disrespectful to you. <POST> [text]. <RESPONSE>

In Study 3, we were interested in levels of toxicity on a 5-point Likert scale, using the following prompts for annotations with personas:

> *Prompt Template 3*
> <TASK> Your task is to annotate the post shown below, taking on the perspective of the following persona: [persona]. Answer the following question on a scale from 1 to 5, where 1 stands for 'Not at all' and 5 stands for 'Very much so'. How toxic/hateful/disrespectful or offensive does this post seem to you? <POST> [text]. <RESPONSE>

A.3 Model Deployment

We used the Mistral-7B-Instruct-v0.1 model [5] and the Qwen2-7B-Instruct model [6] via the Python bindings [7] of the `llama.cpp` framework [8], running them on two separate 40GB partitions of a NVIDIA A100 GPU. We choose to run all experiments on two separate models to establish that our approach is not dependent on a specific model, and selected the two models for their good zero-shot performance in the toxicity annotation task as well as their comparably good efficiency and usability, fitting on a single 40GB GPU partition.To further decrease the compute workload and streamline model generations, we make use of the outlines framework [9], effectively restricting the LLM generations to a provided set of response options. Since we are not interested in the LLMs' abilities to generate open ended responses but in their preferred alternative from a restricted set of options (either binary toxicity labels or the five ordinal options from the 5-point Likert scale for the level of toxicity), we consider this to be a sensible choice that does not impact the validity of our results. We use the multinomial sampler implemented in Outlines with a temperature of 1 and max_new_tokens of 1 across all generations in our studies, in line with our need to only generate integers.

A.4 Study 2 Results

We set the similarity threshold for cluster formation to 0.6 - in combination with a minimum personas per cluster threshold of 25 - based on a comparison of the resulting clustering solutions using different similarity thresholds (Table A.2). We settle on

[5] https://huggingface.co/mistralai/Mistral-7B-v0.1
[6] https://huggingface.co/Qwen/Qwen2-7B-Instruct
[7] https://github.com/abetlen/llama-cpp-python
[8] https://github.com/ggerganov/llama.cpp
[9] https://github.com/outlines-dev/outlines

Threshold	# Clusters	# Personas
0.50	1,627	184,761
0.55	2,065	169,382
0.60	**2,180**	**138,519**
0.65	1,676	87,653
0.70	702	30,954
0.75	102	3,613

Table A.2: Similarity thresholds resulting in different clustering solutions in the *persona space*, together with the number of resulting clusters and the number of personas they include. Resulting cluster solution printed in bold.

this threshold value as it affords a high number of different clusters, which we think is necessary to account for the heterogeneity of persona descriptions, while still including a sufficiently high number of personas (70% of all personas).

Table A.3 provides examples for the three largest and three of the smallest clusters of our resulting cluster solution.

As further confirmed in Figure A.1, the resulting clusters are internally homogeneous (i.e., small average intra-cluster distance in the *persona space*), as indicated by the light colors along the diagonal of Figure A.1, as well as heterogeneous across different clusters (i.e., high(er) average inter-cluster distance in the *persona space*), as indicated by the dark colors everywhere but on the diagonal of Figure A.1.

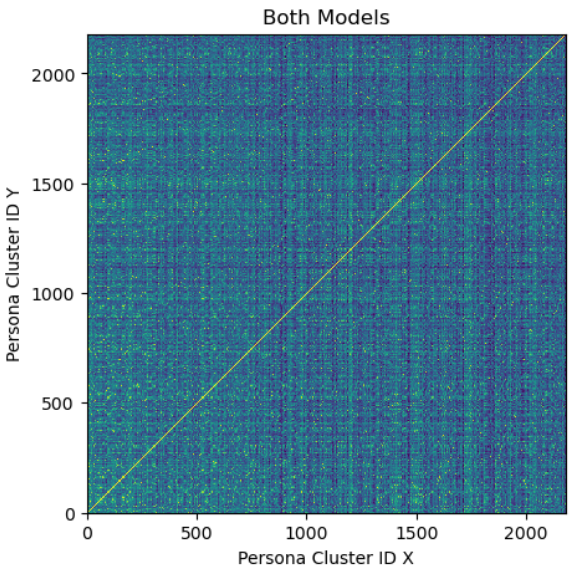

Figure A.1: Intra- and inter-cluster cosine distances of *persona space* clusters measured in the *persona space* shared by both models. Lighter-colored cells represent lower average distances between the respective clusters. The bright diagonal indicates successful clustering, with personas in clusters being more similar to each other than to personas in other clusters.

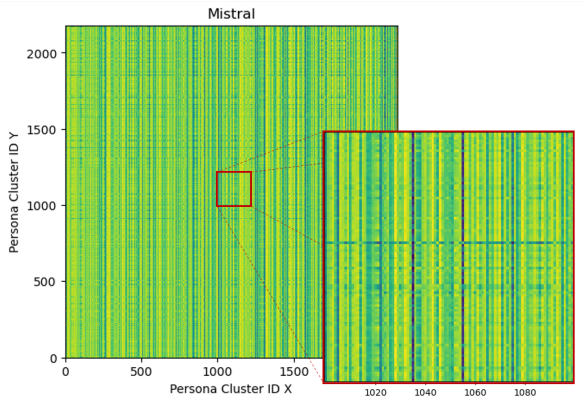

Figure A.2: Intra- and inter-cluster cosine distances of *persona space* clusters measured in *label space* resulting from Mistral annotations. Values are normalized per row. Lighter-colored cells represent lower average distances between the respective clusters, the lighter colors along the diagonal thus indicate that similar personas annotate alike.. The inset zooms in on clusters with IDs from 1,000 to 1,100.

Figure A.2 shows the lighter cell colors along the diagonal as a result of lower inter- than intra-cluster average distances for persona clusters in the label embedding space resulting from Mistral annotations.

Table A.4 provides the instances with the highest absolute difference in average toxicity levels between Black and conservative personas.

Cluster ID	Cluster Size	Top 10 TF-IDF Cluster Terms	3 Random Cluster Personas
0	1,393	sports, athlete, player, basketball, professional, coach, tennis, athletes, sport, football	An athletics coach who focuses on talent development and has been tracking Alemitu's career closely.; a freelance sportswriter; an esports fan who is confused about the appeal of physical sports.
1	1,349	history, professor, historical, teacher, historian, literature, university, political, figures, specializing	A person who is fascinated by elaborate schemes and extraordinary behavior in history.; a literature professor at Lancaster University, United Kingdom.; A professor specializing in the specific subject area for which the content developer is creating material.
2	1,292	journalist, political, news, reporter, politics, journalism, media, politician, commentator, reporting	a sports journalist for a local newspaper in Gloucester; A public relations specialist known for damage control in political controversies; An Orlando City SC player who values the journalist's support and uses their articles as motivation
...	...	...	...
2177	25	fda, clinical, representative, trial, approving, drug, evaluating, responsible, reviewing, efficacy	A pharmaceutical industry regulator responsible for ensuring compliance with clinical trial protocols; A representative from the Food and Drug Administration (FDA) responsible for evaluating the safety and efficacy of the DNA sequencing technology; A representative from the regulatory agency responsible for overseeing and approving the use of CRISPR technology
2178	25	friend, close, neighbor, old, johnson, colleague, dr, dale, long, longtime	an old friend and neighbor of Visanio Johnson.; a proud wife of a long-term friend of Dr. Robert Johnson; An old friend and golf buddy of Peter "Pete" Van Vooren, living in Sioux Falls, SD.
2179	25	spiders, spider, fear, arachnophobic, arachnophobia, person, afraid, arachnophobe, arachnologist, bug	A homeowner who has a severe arachnophobia and wants to ensure a spider-free environment; An arachnophobic librarian; an arachnophobic tour guide in Australia

Table A.3: Cluster size, top ten TF-IDF terms and three randomly selected persona descriptions for the three largest and three of the smallest persona clusters.

AAE Instance	Diff.	Anti-Black Instance	Diff.
All my *n**** wear crowns, all my *n**** bossed up, all my *n**** got bad bitches don't none of us give a fuck.	-1.14	Please save your black weaves for someone who gives a rat's ass	1.78
The way these *n**** hate I hope they get a check for it	-1.13	Is it even proven that we looked like big orangutans? They're most closely related to orangutans but the split happened an extremely long time ago, even before the split between gorillas and humans. Obviousy a large dark section of the population never made this split!	0.73
*N**** rather be known and broke instead of low-key and getting chicken	-1.07	Opps, Under our amazing President, Trump, gang members are gone. Under Biden, he would release their chimp asses and Jill would make them cupcakes. awful ! God Bless President Trump	0.53

Table A.4: AAE and anti-Black instances with largest difference in average toxicity ratings assigned by Black and conservative personas. Differences calculated as Diff $= \mu_{Q;B} - \mu_{Q;C}$ based on Qwen annotations. We replace all mentions of the n-word and its variants with *n****.