PERSEVAL: A Framework for Perspectivist Classification Evaluation

Soda Marem Lo[1]*, Silvia Casola[2]*, Valerio Basile[1],
Erhan Sezerer[3], Franco Sansonetti[1], Antonio Uva[3], Davide Bernardi[3]

[1]Computer Science Department, University of Turin, Turin, Italy
[2]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany and Munich Center for Machine Learning (MCML)
[3]Alexa AI, Amazon, Amazon Development Centre Italy, Turin, Italy

{sodamarem.lo|valerio.basile}@unito.it s.casola@lmu.de
franco.sansonetti@edu.unito.it {erhanszr|antonuva|dvdbe}@amazon.it

Abstract

Data perspectivism goes beyond majority vote label aggregation by recognizing various perspectives as legitimate ground truths. However, current evaluation practices remain fragmented, making it difficult to compare perspectivist approaches and analyze their impact on different users and demographic subgroups. To address this gap, we introduce PERSEVAL, the first unified framework for evaluating perspectivist models in NLP. A key innovation is its evaluation at the individual annotator level and its treatment of annotators and users as distinct entities, consistently with real-world scenarios. We demonstrate PERSEVAL's capabilities through experiments with both Encoder-based and Decoder-based approaches, as well as an analysis of the effect of sociodemographic prompting. By considering global, text-, trait- and user-level evaluation metrics, we show that PERSEVAL is a powerful tool for examining how models are influenced by user-specific information and identifying the biases this information may introduce.

1 Introduction

Recently, part of the Natural Language Processing (NLP) community has seen what Cabitza et al. (2023) called a *perspectivist turn*. Researchers have increasingly questioned data harmonization techniques such as majority vote in favor of considering multiple perspectives as legitimate ground truths (Basile, 2020; Plank, 2022a). Perspectivist models thus leverage annotator disagreement to better account for user diversity (Prabhakaran et al., 2021) and adopt evaluation strategies capable of embracing disagreement (Uma et al., 2021b).

This framework assumes that part of the disagreement observed in annotation can be explained by the background and beliefs of the annotators, who might have different perspectives on the phenomena under study. For example, in a sensitive and difficult (Röttger et al., 2021) task such as hate speech detection, annotator specific modeling improved classification performance (Mostafazadeh Davani et al., 2022).

Evaluation practices in perspectivism vary widely. Inspired by early work on understanding and predicting annotator disagreement, one popular approach treats annotators' judgments separately, training with their individual labels (Fleisig et al., 2024). Another common practice is to consider all annotators as known at training time (Mostafazadeh Davani et al., 2022). While this represents a reasonable research scenario, it remains unclear how it would translate to real-world applications, where users are unknown during training, and adaptation occurs through limited interactions or feedback. To this end, some works in the perspectivist realm also account for unseen annotators (Deng et al., 2023; Orlikowski et al., 2025).

Given the diversity of assumptions and approaches, the developed models are not directly comparable, and quantifying their performance on different tasks remains hard. Moreover, the impact of using perspectivist models on new users and texts, together with the possible biases introduced, remains underexplored.

With the overarching goal of rationalizing perspectivist evaluation and quantifying its impact, this paper presents PERSEVAL, a framework for Perspectivist Evaluation. To mirror real-world scenarios, we consider annotators, who provide the bulk of the annotation for training models, as disjoint from system users, for which performance is tested. Relaxing our working hypothesis, we also define two scenarios for which minimal test users' annotations are available, for example from user feedback or human-in-the-loop approaches. The first, inspired by Kocoń et al. (2021b), accounts for cases in which only a little information about test users' preferences is available during training;

*Soda Marem Lo and Silvia Casola contributed equally to this work.

the model can thus use this information to learn a user-specific bias. The second scenario assumes a system has been already trained and deployed, and allows using test user information for adaptation. All the variants of PERSEVAL are explained in Section 3.

Moreover, we consider two scenarios depending on the availability of explicitly-defined annotator and user characteristics: they can either be known by their identifier only, or they can be represented as a set of metadata, for example, describing their sociodemographic information or declared preferences.

Evaluation within PERSEVAL occurs at the individual annotation level and incorporates both global and fine-grained metrics—evaluating at the user, text, and trait levels (Section 5). This enables a comprehensive comparison and analysis across different perspectivist models.

We showcase our evaluation framework on encoder- and decoder-based models; we primarily focus on presenting a comprehensive framework for evaluating perspectivist models rather than testing an extensive range of models. We consider five disaggregated datasets focused on phenomena such as irony and offensive speech detection or AI safety and with diverse designs concerning the number of annotators and the provided demographics. By performing evaluation at the individual annotation level, on the one hand, we can compare multiple perspectivist systems and the impact of explicitly modeling perspectives on their performance; on the other, we measure which point of view is privileged, e.g., by taking into account demographic data when available.

In summary, our contributions are the following: (1) We present PERSEVAL, an evaluation framework for perspective systems. We rationalize the user representation, the user splitting, and the evaluation functions. (2) We collect and harmonize five disaggregated datasets with diverse domains, classification tasks, and user representations. (3) We test several models and compare their performance in the proposed settings. (4) We carefully analyze the evaluation results, focusing on whether the models can bias their prediction following annotator-specific sociodemographic information and whether this introduced bias improves models' performance. (5) We develop and share a user-friendly library providing functionalities facilitating the comparison and analysis of different perspectivist approaches.

To support the reader's understanding, we provide a glossary that briefly explains key concepts of the Perspectivist approach and PERSEVAL in Appendix A.

2 Related works

Perspectivism *Data perspectivism* aims at leveraging disaggregated annotations to model human perspectives in NLP (Frenda et al., 2024). The traditional approach towards annotators' disagreement "solves" it through data harmonization. However, aggregating annotations may result in an increasing bias toward specific groups, often minorities (Prabhakaran et al., 2021; Goyal et al., 2022). Perspectivist approaches, instead, challenge the assumption of a single ground truth (Aroyo and Welty, 2015) and consider the coexistence of different standpoints. These perspectives are defined differently depending on the task and the data: tied to cultural backgrounds (Akhtar et al., 2021), demographic information (Frenda et al., 2023; Casola et al., 2024), a combination of attitudes and behavior (Chulvi et al., 2023), a set of psychological characteristics (Mieleszczenko-Kowszewicz et al., 2023) or beliefs (Kazienko et al., 2023), moving in a continuum from a group-based to an individual perspective (Kocoń et al., 2021a).

Demographic data as perspectives A considerable body of work is exploring the influence of demographics in annotators' choices (Al Kuwatly et al., 2020; Larimore et al., 2021; Sap et al., 2022; Biester et al., 2022; Kumar et al., 2021; Davani et al., 2023; Jaggi et al., 2024). Encoding users' explicit traits has been an effective strategy when working with individual annotations (Milkowski et al., 2021; Wan et al., 2023), and it is increasingly investigated to evaluate generative models' cultural alignment (Cao et al., 2023; Tao et al., 2024; Casola et al., 2024). Sociodemographic prompting has been explored also to simulate human responses, showing mixed results (Argyle et al., 2023; Wang et al., 2025). Recent works have systematically studied both the potential and limitations of this approach, in zero-shot settings (Beck et al., 2024) as well as with fine-tuned generative models (Orlikowski et al., 2025).

Perspectivist evaluation Given the differences in disaggregated dataset design, number of annotators, available metadata, and corpus size (Plank,

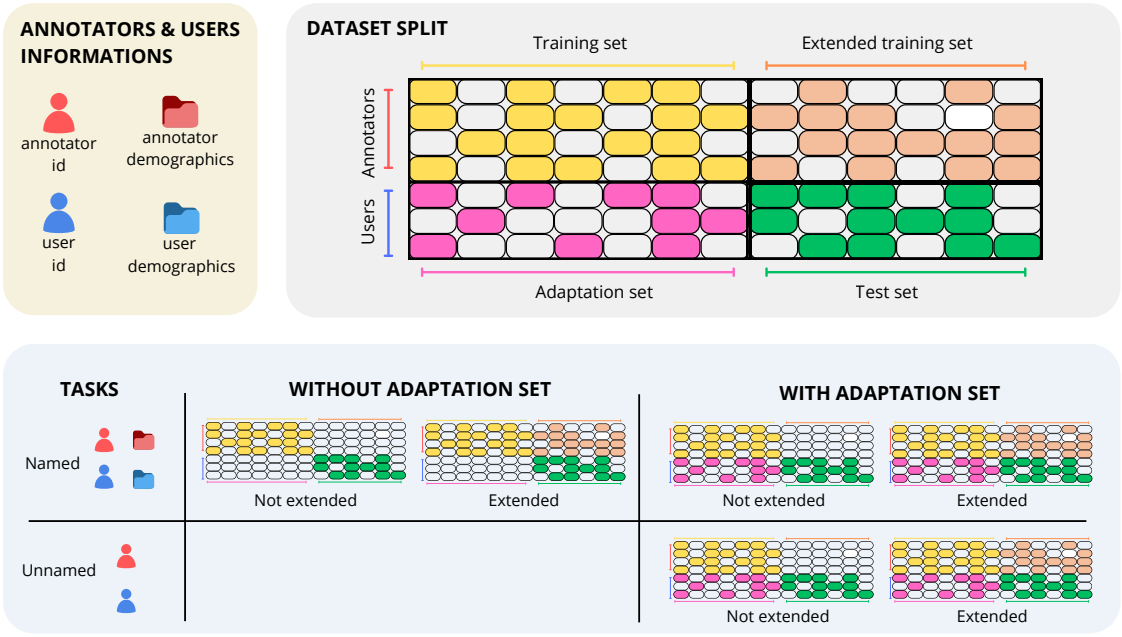

Figure 1: Data split in PERSEVAL.

2022b)[1,2], researchers have developed different ways to modeling and evaluating annotator perspectives. Works in the middle of the continuum between data- and human-centric approaches focus on modeling and evaluating groups of annotators (Akhtar et al., 2021; Frenda et al., 2023; Casola et al., 2023; Lo and Basile, 2023; Mostafazadeh Davani et al., 2024). Moving towards individuals, Mostafazadeh Davani et al. (2022) propose a multi-task approach, where the goal is to predict each annotator's label. Recent studies increasingly follow this line of research, experimenting in active-learning settings (Wang and Plank, 2023), proposing personalized methods. Personalization techniques have been used for modeling annotators (Plepi et al., 2022), often inspired by recommender systems methods (Kazienko et al., 2023; Heinisch et al., 2023), also being particularly attentive to annotators' demographics (Gordon et al., 2022) and minority voices (Mokhberian et al., 2024). A structured framework of evaluation comes from the LeWiDi shared task (Uma et al., 2021a; Leonardelli et al., 2023, 2025), which evaluates the impact of disagreement via cross-entropy. Nevertheless, this approach does not address the challenge of evaluating the models' ability to capture human perspectives. To the best of our knowledge, the only previous benchmark in this field of research is The Inherent Disagreement 8 dataset (TID-8) by Deng et al. (2023), a collection of 8 language-understanding disaggregated datasets with a varying number of annotators. In this work, we cover a larger diversity of approaches with a systematization of the possible splits, reporting a set of metrics to evaluate models also at the user, text and metadata level, and benchmarking encoder- and decoder-based models for perspectives classifications.

[1] https://github.com/mainlp/awesome-human-label-variation

[2] https://pdai.info/

3 PERSEVAL: the Framework

We propose a conceptual framework for the evaluation of perspectivist text classification models. According to our framework, each instance is a <*text, annotator*> (or <*text, user*>) pair. We make this choice to better understand the impact of explicitly modeling perspectives (e.g., in terms of sociodemographic traits) on individual users. This choice is fundamentally different from traditional evaluation methodologies in NLP where instances are typically just textual. This difference has implications for several aspects that we discuss in the following sections.

3.1 Data Split

Previous research on disaggregated datasets has taken different approaches to data splitting. Most perspectivist evaluation practices rely on a fixed set of annotators, who typically annotate every text

in the corpus; a standard text-based split is then adopted. This approach is useful from a theoretical standpoint, but does not reflect real-world scenarios and implications. In practice, a system is typically trained on data provided by one group of individuals (the *annotators*), while its inference is run on a set of instances encoding the perspectives of a distinct set of individuals (the *users*).

We conceptualize the data split in PERSEVAL under the assumption that annotators, who provide the training annotations, are disjoint from the test users. When explicit knowledge about the users is available — for example, in the form of sociodemographic information or preferences —, a model can attempt to learn biases toward such characteristics. When no such information is available, however, inferring preferences for completely unknown users is unfeasible. As a consequence, we define two adaptation scenarios:

- **Adaptation at training time (T)**: we assume minimal annotation from users has been obtained before training the system. A few annotations from test users are thus included in the training split.
- **Adaptation at inference time (I)**: we assume an already trained system has to be adapted to new users. A few test users instances can thus be used to adapt an existing model.

In both cases, we assume that a minimal amount of annotations from users in the test set (e.g., collected through user interaction with the system or human-in-the-loop approaches) is available.

3.2 User representation

The degree of user information available varies across datasets. When metadata are available, we represent the annotators and users through a set of traits, which may include sociodemographic or other explicit information. This representation enables models to learn annotator-specific perspectives based on these traits. We refer to this as the ***Named*** representation. With our proposed data split separating training annotators and test users, the challenge is to learn annotator perspectives and generalize them to unseen users using only their traits. This setting is motivated by a well-established body of research examining the influence of annotators' demographic backgrounds on their choices, yielding both positive and contentious findings (Section 2).

To address scenarios where user metadata is unavailable, we also define a setting where annotators and users are represented solely by unique identifiers. While this restricts the model's ability to personalize predictions, it is a common scenario in many real-world applications. We call this representation ***Unnamed***. In the *Unnamed* perspectives task, the model must classify perspectives without any explicit knowledge of the annotators' and users' characteristics. This variant necessitates adaptation to infer user perspectives from the available annotations: the strict hypothesis for which test users are completely disjoint from training annotators must be relaxed (Section 3.1). As a consequence, the *Named* classification task can be performed with and without adaptation, while the *Unnamed* task requires some form of adaptation. Table 1 summarized the available variants.

Task	Adaptation	Adapt. Phase
Named	No adaptation	Never
	Adaptation-T	At training time
	Adaptation-I	At inference time
Unnamed	Adaptation-T	At training time
	Adaptation-I	At inference time

Table 1: Task variants proposed in PERSEVAL.

3.3 Extended training set

Since instances in PERSEVAL are <*text, annotators*> pairs, training and test instances could, in principle, share the same text, associated with labels from different annotators. This is the approach adopted in previous work when an annotator-level split was performed, for example by Orlikowski et al. (2025). However, this behavior is not always desirable, as the knowledge learned by a model from a training text may affect the inference on an instance with the same text in unpredictable ways.

To ensure fair evaluation and avoid data leakage, we follow standard practice and exclude any text instances in the test split that have been annotated by the training annotators. However, we also explore a variant where texts that appear in both training and test sets but are annotated by different people, are allowed in the training data (*extended*). This variant tests the model's ability to learn from systematic disagreements among users who annotate the same text differently, capturing the diversity of perspectives inherent in the data.

All task variants can use the extended training set. While they differ in training splits—and in some cases include additional sets for adaptation

at inference time—the test set remains consistent to ensure fair performance comparison.

4 Datasets

PERSEVAL incorporates a diverse range of datasets, varying in task, domain and annotator information. Table 2 summarizes the dataset characteristics. A description of each dataset and the available metadata are available in Appendix B and C.

5 Evaluation metrics

Our evaluation setting is inspired by previous work in personalization. Given predicted labels and true annotations for each <*text, user*> pair, we compute standard classification metrics, i.e., precision, recall, and F1-score (referred to as *global metrics)*.

Moreover, the annotator-based characteristic of the disaggregated labels allows us to gain further insights into the models' capability to learn from diverse human perspectives. Inspired by Mokhberian et al. (2024), we report *user-level* metrics. These metrics are computed individually for each test user and then averaged; they provide a fairer evaluation regardless of the contribution in terms of annotations of each user to the dataset. We also report *text-level* metrics, computed individually for each text and averaged. The analysis of these metrics helps understand whether some texts are easier to classify for a given model and whether having instances with the same textual content (but different users, and thus, different annotations, in the extended version of the dataset), helps the model in the classification. Finally, for the named task, we also report *trait-level* metrics. These metrics, computed for each trait and then averaged for each dimension, are meant to describe if the preference of all groups of people is fairly learned by the model or if the model underperforms when considering users with certain characteristics.

6 Models

We benchmarked a series of approaches for perspectivist classification, using Encoder- and Decoder-based models, covering all task variants proposed in Section 3.

Due to the different settings supported by each approach, we test a subset of settings with each model. In particular, when working with the Encoder-based model, we did not include inference-time adaptation since this architecture does not support it. On the other hand, performing zero- and few-shot learning by prompting the Decoder-based model, we did not cover the *Adaptation-T*variant for the *Named* or the *Unnamed* Task.

6.1 Encoder-based Model

We fine-tuned RoBERTa (Zhuang et al., 2021)[3], customized implementing Focal Loss (Lin et al., 2017) to prevent overfitting in case of unbalanced datasets. All splitting and training parameters are reported in Appendix D. Inspired by the personalized User-ID model from Ferdinan and Kocoń (2023), we added identifiers and traits of the annotators to the text embedding as a special token. The input thus concatenates the annotator ID, a special token for each of the annotator's traits, and the input text to classify. The special tokens explicitly encode the annotator's identity and characteristics and are used by the model to learn annotator- and trait-specific features in the classification. The model is then trained with a classification head to predict the label. We also computed a baseline without any additional special tokens.

6.2 Decoder-based Models

For the Decoder-based model, we focus on open-source models and benchmark the performance of Mixtral-8 7B[4] and Llama-3.1 8B,[5] both instruction tuned. We consider several settings:

Base-zero We prompt the models to classify the test set examples, with no additional information.

Perspective Inspired by work on role-based sociodemographic prompting (Beck et al., 2023), we ask the models to impersonate each user's trait. To do so, we prepend the given trait to the prompt (for example *You are a person from Generation X.*). We use this variant to test models without adaptation with a named user representation. We prompt the model for each available user trait.

In-Prompt Augmentation (IPA) We reproduced Salemi et al. (2024)'s approach, using In-Prompt Augmentation (*IPA*). It consists of prompting the model with user-specific input selected via retrieval augmentation, a framework which extracts pertinent texts, relevant to the classification of the unseen test case. Using the authors' terminology, given a sample (x_i, y_i) and a user u, a query generation function ϕ_q transforms the input x_i into a

[3] FacebookAI/roberta-base
[4] Mixtral-8x7B-Instruct-v0.1
[5] meta-llama/Llama-3.1-8B

Dataset	Reference	Task	#Annot.	#Texts	#Inst.	Source	Label	Positive Class	Metadata
BREXIT	Akhtar et al. (2021)	Abusive Language	6	1,120	3,872	Twitter	Binary	Offensiveness	Target and control group
EPIC	Frenda et al. (2023)	Irony	74	3,000	14,172	Twitter, Reddit	Binary	Irony	Gender, Nationality, Age/Generation
MHS	Sachdeva et al. (2022)	Hate Speech	7,912	39,565	135,556	YouTube, Twitter, Reddit	Binary	Hate Speech	Gender, Age/Generation, Education, Income
MD-Agreement	Leonardelli et al. (2023)	Offensiveness	819	10,753	53,765	Twitter	Binary	Offensiveness	—
DICES	Aroyo et al. (2024)	AI Safety	123	350	43,050	Human-chatbot conversations	Non-binary	Harmful	Gender, Age/Generation, Education, Ethnicity

Table 2: Overview of the datasets used in social media text classification tasks.

query q for retrieving the user profile P_u (i.e. the user's historical data) from the Adaptation set. To do so, we used the Contriever model (Izacard and Grave, 2021), a pre-trained dense retrieval model $\mathcal{R}(q, P_u, k)$ that retrieves the k most pertinent entries. Finally, the prompt construction function ϕ_p assembles the personalized prompt. Specifically, we selected 5 examples per user. We used this approach both giving information about the user's trait value (*Named* with *Adaptation-T*) and without providing demographic information.

When some outputs could not be properly parsed — such as when the model refused to provide an answer, particularly for datasets related to hate speech — we assigned an additional uncorrect label.

In the *Named* task, we prompt the model separately for each available user trait and determine the final label through a majority vote across the outputs of the trait-specific models. The prompts used for each setting are detailed in Appendix E.

7 Results

In this section, we present the results for the Encoder-based and Decoder-based Models. In all cases, we report metrics related to the positive class, with the exception of DICES, the only multi-class dataset, for which we present the macro-averaged metrics.

The performance of encoder-based models is generally better than the Decoder-based models. This is expected since the former are fine-tuned while the latter are used in zero- or few-shot mode. However, we note how Decoder-based models are much more sensitive to the injection of annotator metadata, paving the way for more sophisticated decoder-based models for perspectivist classification. Section 7.1 and 7.2 present the results separately. To gain more insights into the effect of sociodemographic prompting, we performed an in-depth analysis (Section 7.3).

7.1 Encoder-based Model

Table 3 shows the results on the datasets with binary labels. We notice that when considering the *non-extended training set* — i.e., the case in which the text to be annotated has not been seen by the model at training time — the baseline tends to have higher scores in terms of global F1. With the *extended training set*, instead, providing information about the user (both in terms of sociodemographic traits or IDs) leads to improved results over the baseline. This indicates a mild tendency of the model to learn the relation between latent features of the text and the annotator labeling them, although marginal.

The user and text-based F1 scores highlight the benefit of including demographic traits at training time, especially in the setting without adaptation set (*Adaptation-None*). When demographics are not available, such as for MD-Agreement, providing the user ID still results in being beneficial.

The trait-based F1 scores show that some traits are more informative than others, e.g., *Nationality* for EPIC, coherently with its focus on differences in the perception of irony across language varieties. This pattern is consistent in all settings. As for MHS, the model tends to be fairer to annotators grouped based on their generation.

Results for DICES are in Table 4. Providing demographics confirms a positive impact, with improved results in all settings in terms of global, user- and text-level F1 scores. Moreover, adaptation helps the performance across all the metrics.

In all settings, the model presents a higher trait-based F1 score on *Generation*, showing its influence during training, which aligns with intuition for a task related to human-AI conversations.

	Dataset		Adapt	Global F1	User F1	Text F1	Traits F1s				
Not extended	EPIC	baseline	-	**.555**	**.538**	**.376**	Gender -	Nationality -	Generation -	-	-
		Named	None	.542	.527	.364	.520	.547	.527		
			Train	.550	.534	.371	.531	.556	.538		
		Unnamed	Train	.534	.518	.352	-	-	-		
	BREXIT	baseline	-	**.567**	.519	.403	Group -	-	-	-	-
		Named	None	.558	**.524**	**.416**	.524				
			Train	.544	.512	.405	.512				
		Unnamed	Train	.512	.484	.378	-				
	MHS	baseline	-	.688	.642	.515	Gender -	Generation -	Education -	Income -	Ideology -
		Named	None	.691	.640	.518	.666	.692	.690	.691	.687
			Train	.689	.641	.516	.662	.690	.686	.689	.685
		Unnamed	Train	**.692**	**.643**	**.521**	-	-	-	-	-
	MD	baseline	-	**.665**	.591	**.500**					
		Unnamed	Train	**.665**	**.597**	.499					
Extended	EPIC	baseline	-	.579	.559	**.405**	Gender -	Nationality -	Generation -	-	-
		Named	None	.575	.560	.392	.555	.591	.560		
			Train	.578	.564	.594	.560	.589	.565		
		Unnamed	Train	**.586**	**.571**	**.405**	-	-	-		
	BREXIT	baseline	-	**.592**	.543	.427	Group -	-	-	-	-
		Named	None	.587	**.557**	**.455**	.557				
			Train	.540	.509	.424	.509				
		Unnamed	Train	.554	.524	.436	-				
	MHS	baseline	-	.696	.647	.526	Gender -	Generation -	Education -	Income -	Ideology -
		Named	None	**.700**	**.651**	.530	.663	.699	.698	.700	.696
			Train	**.700**	.650	**.532**	.674	.702	.699	.700	.696
		Unnamed	Train	.697	.648	.527	-	-	-	-	-
	MD	baseline	-	.667	.603	.495					
		Unnamed	Train	**.681**	**.620**	**.518**					

Table 3: Encoder model's global F1 score, and user-, text-, trait- level F1 for the positive class for binary datasets.

		Adapt	Global F1	User F1	Text F1	Traits F1s			
						Gender	Generation	Education	Race
Not ext.	baseline	-	.340	.311	.245				
	Named	None	.400	.391	.361	.401	.400	.388	.397
		Train	.420	.407	.373	.419	.422	.408	.414
	Unnamed	Train	**.434**	**.424**	**.389**	-	-	-	-
Ext.	baseline	-	.440	.448	.335				
	Named	None	.453	.439	.378	.452	.454	.436	.447
		Train	**.457**	.445	**.389**	.457	.457	.439	.454
	Unnamed	Train	.456	**.446**	.388	-	-	-	-

Table 4: Macro-averaged global F1 score, and user-, text-, trait- level F1 for the Encoder model with DICES.

7.2 Generative Models

Table 5 presents the results for the Decoder-based models using the ensembling strategy described in Section 6. Across all datasets, all the approaches outperform the baseline, except for Mixtral on MHS. Focusing on *Named* tasks, we notice that adding annotators' demographics consistently helps the model. As expected, the *Unnamed* task is harder, however the user-based selection of few-shot examples of *IPA* significantly outperforms the baseline. Indeed, *IPA* is the most effective strategy for perspective classification with generative models, except on BREXIT. We speculate this is due to the high polarization of the annotations in this dataset, and the narrow characterization of the users. The same pattern can be observed on DICES (Table 6), where *IPA* is the best approach with both models. The positive influence of sociodemographic information is also confirmed using IPA-Llama. *IPA-Mixtral* presents higher scores in the *Unnamed* setting, demonstrating the effectiveness of providing user-specific examples alone.

Examining the trait-based F1 scores, for both the *Perspective* and *IPA* approaches, the most informative traits are *Nationality* in EPIC (consistently with results on the Encoder). This is consistent when using LLama and Mixtral and aligns with intuition, given that the dataset focuses on various linguistic varieties. A similar pattern is observed for *Generation* in MHS when using Llama. With Mixtral, *IPA*

Model	Dataset	Approach		Adapt	Global F1	User F1	Text F1	Traits F1s				
								Gender	Nationality	Generation		
	EPIC	Base-zero	Baseline	-	.529	.511	.363	-	-	-		
		Perspective	Named	None	.484	.467	.322	.465	.492	.463		
		IPA	Named	Test	**.547**	.528	**.387**	.515	.543	.532		
			Unnamed	Test	.546	**.530**	.386	-	-	-		
								Group				
Llama	BREXIT	Base-zero	Baseline	-	.502	.476	.340	-				
		Perspective	Named	None	**.527**	**.502**	**.371**	.502				
		IPA	Named	Test	.364	.362	.238	.362				
			Unnamed	Test	.330	.319	.231	-				
								Gender	Generation	Education	Income	Ideology
	MHS	Base-zero	Baseline	-	.593	.543	.425	-	-	-	-	-
		Perspective	Named	None	.515	.454	.354	.419	.522	.522	.515	.512
		IPA	Named	Test	**.637**	**.573**	**.467**	.587	.649	.640	.637	.638
			Unnamed	Test	.626	.570	.456	-	-	-	-	-
	MD	Base-zero	Baseline	-	.556	.515	.381					
		IPA	Unnamed	Test	**.613**	**.535**	**.451**					
								Gender	Nationality	Generation		
	EPIC	Base-zero	Baseline	-	.487	.477	.305	-	-	-		
		Perspective	Named	None	.507	.494	.328	.501	.515	.493		
		IPA	Named	Test	**.554**	.521	.380	.528	.551	.543		
			Unnamed	Test	.553	**.528**	**.384**	-	-	-		
								Group				
Mixtral	BREXIT	Base-zero	Baseline	-	.255	.235	.128	-				
		Perspective	Named	None	.344	.323	.193	.323				
		IPA	Named	Test	.406	.382	.263	.382				
			Unnamed	Test	**.448**	**.410**	**.313**	-				
								Gender	Generation	Education	Income	Ideology
	MHS	Base-zero	Baseline	-	**.648**	**.599**	**.483**	-	-	-	-	-
		Perspective	Named	None	.644	.594	.480	.655	.648	.649	.644	.649
		IPA	Named	Test	.634	.569	.459	.621	.639	.643	.633	.634
			Unnamed	Test	.632	.571	.457	-	-	-	-	-
	MD	Base-zero	Baseline	-	**.538**	**.495**	**.678**					
		IPA	Unnamed	Test	.531	.398	.643					

Table 5: Decoder-based approach global F1 score, and user-, text- and trait- level F1 scores for the positive class.

Models	Dataset	Approach		Adapt	Global F1	User F1	Text F1	Traits F1s			
								Gender	Generation	Education	Race
Llama	DICES	Base-zero	Baseline	-	.290	.282	.310	-	-	-	-
		Perspective	Named	None	.298	.290	.289	.297	.295	.300	.297
		IPA	Named	Test	**.365**	**.354**	**.428**	.367	.363	.352	.362
			Unnamed	Test	.355	.340	.425	-	-	-	-
Mixtral	DICES	Base-zero	Baseline	-	.232	.228	.402	-	-	-	-
		Perspective	Named	None	.256	.323	.412	.297	.311	.310	.304
		IPA	Named	Test	.303	.350	**.448**	.302	.303	.309	.300
			Unnamed	Test	**.306**	**.356**	.443	-	-	-	-

Table 6: Macro-averaged global F1, user-, text- and trait- level F1 scores for the Decoder-based models with DICES.

exhibits greater fairness toward annotators grouped by education. This result is consistent with the encoder, suggesting that these traits are particularly influential on the models' predictions.

7.3 Analysis

Representing users through their sociodemographic traits could lead to the risk of stereotype propagation. Previous research showed that considering demographic traits only can be limiting (Jiang et al., 2024; Biester et al., 2022), since they do not necessarily align with annotations (Orlikowski et al., 2023; Lo and Basile, 2023; Vitsakis et al., 2024). Thus, to evaluate the effect of sociodemographic prompting, we investigate whether the provided annotator metadata inform the models and whether the learned biases align with those observed in the datasets.

Q1: What is the contribution of each trait when ensembling the model's outputs? We conducted an ablation study by ensembling model outputs across all possible combinations of traits. Since BREXIT has only one trait, and MD lacks information about the annotators, we computed the results for EPIC, MHS and DICES, reported in Appendix F.1 On EPIC, users' *Generation* is the most informative trait in *Perspective* settings on both Llama and Mixtral; combining it with *Nationality* also shows a positive impact. These results are consistent with the dataset design and discussion by Frenda et al. (2024), where annotators' generation is one of the most polarizing demographic dimensions. The same pattern can be found for *IPA-Llama*. In DICES, *Education* is an important factor in *Perspective* settings with both models, an interesting result considering the focus on AI-safety. On the other hand, when models can see examples,

Generation and *Gender* are more positively influential for Llama and Mixtral respectively. MHS is the only dataset where ensembling more traits is beneficial in the *IPA* setting with both models, suggesting a less clear-cut influence of traits than in other settings and datasets. In fact, the *IPA* approach shows smaller F1 score differences across traits compared to the *Perspective* setting, as models benefit from personalized examples rather than relying solely on demographics.

Q2: Which demographic trait difference most significantly impacts models' labels? Here, we focused on a single demographic feature (e.g., gender, age, etc). We filtered out texts annotated by only one subgroup. Then, we computed the impact of changing the demographic variable in the prompt on the models' label (Appendix F.2). Results in the *Perspective* approach tend to be consistent across the two models. For EPIC, *Nationality* and *Generation* are the most influential traits, while in MHS *Ideology* tends to make the model change the label. These are the same traits resulting in being most influential in the ablation study (Q1), consistently with the dataset task. On the other hand, *IPA* shows a higher percentage of cases where the model changes the label, confirming the positive effect of providing user-specific examples. Finally, DICES presents a very low label change compared to the other datasets.

Q3: How similar is the distribution of models' predictions to that of the annotators' chosen labels? Leveraging soft evaluation metrics (Rizzi et al., 2024), we measured the alignment between models and annotators' labels using Jensen-Shannon Divergence (JSD) (Uma, 2021).[6] Taking each trait separately and filtering texts annotated by only one demographic group, we calculated the similarity of the distributions of each demographic variable by text and averaged. The lower the score, the higher the similarity (see also Appendix F.3). Results show that in BREXIT, the models asymmetrically present a higher alignment with the target group in all cases.

On DICES, *Race* tends to be the most aligned trait. DICES and MHS present higher alignment on the *IPA* approach than *Perspective*, the opposite for BREXIT and EPIC. Overall, while in the previous analysis we saw that *IPA* ensures a higher label variability in the predictions at the text level, this does not systematically correlate with a higher alignment with the label distributions.

[6]For cases where the model produced an uninterpretable or invalid output, we assigned an additional label. Consequently, in some settings, the model's label distribution includes this extra category, whereas the human distribution does not. Jensen–Shannon Divergence was therefore computed over the union of all categories.

8 Conclusion

We introduced PERSEVAL, the first unified framework for the evaluation of perspectivist text classification. We assume train annotators and test users are different, and design a *Named* perspectivist classification task where users are represented by their explicit traits and an *Unnamed* task where only their identifier is available. We included five datasets and implemented three baseline models, presenting a robust benchmark for complex real-world applications. Results show that the fine-tuned Encoder benefits more from learning latent annotator-specific biases. For Decoder-based models, within the *Perspective* approach, models appear particularly sensitive to specific demographic cues, which vary according to the dataset and its task. Conversely, providing user-specific examples increases label variability but does not always lead to greater alignment between the models' and annotators' label distributions.

PERSEVAL is implemented in a Python library available on Github at the following link: `https://github.com/valeriobasile/PersEval`. All details in Appendix G.

Limitations

In this paper, we primarily focus on presenting a comprehensive framework for evaluating perspectivist models. Our goal is not to test an extensive range of models; instead, we conducted experiments on just three baseline models. We believe that the framework and library introduced here will serve as a valuable resource for future research in evaluating real-world systems within similar contexts. While we considered multiple datasets, all are in English and most feature binary labels. In future work, we plan to expand this work by incorporating disaggregated datasets in various languages.

While PERSEVAL covers a framework of evaluation, it is supposed to be applied in those contexts where is not possible to assess a single ground truth. Thus, this framework does not cover tasks having a single possible correct answer.

When developing annotation guidelines, the ability to provide annotations about a linguistic phe-

nomenon is considered as something that can be taught and refined. Although some of the selected datasets, such as EPIC, were not designed to provide guidelines, it is essential to note that labelling data is a skill informed by culture, rather than being determined by it.

The presented framework provides tools to study user- and trait-specific bias learned by classification models; however, guidelines for practically limiting the impact of such biases depend on the specific context under study and remain a responsibility of users leveraging the tool.

Ethical statement

The work presented in this paper is in the context of a broader initiative to consider the subjectivity of the annotators in NLP applications, encouraging reflection on the different perspectives encoded in annotated datasets to minimize the amplification of biases. The proposed benchmark can be used as a basis for evaluating a wide range of NLP models, including LLMs, according to their capability of representing the variability of human perspectives.

However, as discussed by Fortuna et al. (2022), working with grouped or individual annotators may represent a risk if it is not clearly defined which perspectives are warranted in the real-world usage of models and resources. For example, as the authors note, while understanding how white supremacists view hate speech could be informative, training models on their annotations would result in systems that would hurt marginalized communities.

Since perspectivist research is recently proposing annotator- and personalization-based approaches, analyzing models' biases becomes fundamental. PERSEVAL is conceived as a tool to systematically evaluate perspectivist classification—accounting for the risk of stereotype propagation in models that encode user metadata or treat annotators as isolated sources—while aiming to prevent harm to targeted groups and minorities. We believe this is a necessary step in the NLP community interested in considering annotators' subjectivity, especially for monitoring the possible drawbacks associated with using these approaches in tasks such as offensive and hate speech detection.

As regards the language resources included in the benchmark, they were built adopting measures to protect the privacy of annotators and data handling protocols designed to safeguard personal information. Some of the material could contain racist, sexist, stereotypical, violent, or generally disturbing content.

Acknowledgments

This work was funded by the 'Multilingual personalization through perspective-aware Language Modeling' partnership with Amazon Science.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

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.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2024. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, page 441–453, Berlin, Heidelberg. Springer-Verlag.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective NLP tasks. *CoRR*, abs/2309.07034.

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.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Silvia Casola, Soda Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, Paolo Rosso, et al. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *CEUR WORKSHOP PROCEEDINGS*, volume 3494.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Teddy Ferdinan and Jan Kocoń. 2023. Personalized models resistant to malicious attacks for human-centered trusted ai. *Emotion*, 40000:50000.

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.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

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.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*,

pages 874–880, Online. Association for Computational Linguistics.

Harbani Jaggi, Kashyap Coimbatore Murali, Eve Fleisig, and Erdem Biyik. 2024. Accurate and data-efficient toxicity prediction when annotators disagree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21910–21917, Miami, Florida, USA. Association for Computational Linguistics.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021a. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021b. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA. USENIX Association.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. Lewidi-2025 at nlperspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP*.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanislaw Wozniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP*.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online. Association for Computational Linguistics.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. 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 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

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

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Be-

yond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

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

Barbara Plank. 2022a. 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.

Barbara Plank. 2022b. 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*, Abu Dhabi. Association for 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.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, Cristina Bosco, et al. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *Ceur workshop proceedings*, volume 2006, pages 1–6. CEUR-WS.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for 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.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. 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.

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

Alexandra Nnemamaka Uma. 2021. *Making the Most of Crowd Information: Learning and Evaluation in AI tasks with Disagreements*. Queen Mary University of London.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of

unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: quantifying annotation disagreement using demographic information. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Glossary

The glossary in Table 7 provides definitions of central concepts of the Perspectivist approach and PERSEVAL framework, deepening into the terminology used throughout the paper.

B Datasets

We present all the datasets incorporated in PERSEVAL. For each dataset, the authors provided details about the instructions given to annotators, except in the case of Brexit, where the proposed guidelines are described in (Sanguinetti et al., 2018; Poletto et al., 2017).

BREXIT is a dataset for abusive language detection, consisting of 1,120 English tweets. The dataset is annotated by 6 annotators from 2 groups: 3 Muslim immigrants in the UK (target group), and 3 researchers (control group). Each annotator annotated the entire corpus with binary label for multiple aspects: hate speech, aggressiveness, offensiveness, and stereotype. The only available trait is the group each annotator belongs to i.e., target or control. In PERSEVAL the positive class is *hate speech*, however the dataset is highly unbalanced toward the negative class.

EPIC (English Perspectivist Irony Corpus) consists of 3,000 texts collected from Twitter and Reddit in 5 English-speaking countries and annotated by 74 crowd workers. Each annotator labeled around 200 texts, for a total of 14,172 annotations. The authors also released annotators' demographic information (Appendix C), balanced across gender and nationality. The target class is *irony*.

MHS (Measuring Hate Speech) contains 39,565 English comments extracted from YouTube, Twitter, and Reddit. It has been annotated by 7,912 people, resulting in 135,556 annotations with both a specific label and multiple hate-informative labels to capture the degree of hatefulness in a continuum. The annotators shared their demographics, reported in Appendix C.The positive class is *hate speech*.

MD-Agreement (Multidomain Agreement), recently used in LeWiDi (Leonardelli et al., 2023), comprises 10,753 English tweets from three domains associated with the hashtags #BlackLivesMatter, #Election2020 and #Covid-19. Each text has been annotated 5 times by 819 annotators, for a total of 53,765 annotations. No demographic trait is available. The positive class is *offensiveness*.

DICES (Diversity in Conversational AI Evaluation for Safety) focuses on AI safety. It is a multi-turn conversation corpus generated by humans interacting with an AI-chatbot, provoking it to respond with an undesirable or unsafe answer. For PERSEVAL we opted for DICES-350, designed to study in-depth cross-demographic differences within the US. Specifically, it consists of 350 multi-turn conversations (within a maximum of 5 turns), fully annotated by 123 people, having a total of 43,050 annotations. This is the only dataset with a non-binary label (with values harmful, not harmful and unsure). The author released annotators' traits, reported in Appendix C.

C Sociodemographic traits

Table 8 shows the traits in the BREXIT, DICES, EPIC and MHS datasets.

D Training parameters

Table 9 presents the training parameters for the Encoder model.

E Prompt example

In this section, we report the prompts used in each setting, taking as examples EPIC and BREXIT datasets, the first with both a context and a text, the second with text only.
Warning: Examples could contain racist, sexist, violent, and generally offensive content.

E.1 Base-zero

Base prompt with EPIC:

> *Instruction: You are provided in input (Input) a pair of sentences (post, reply) extracted from social media conversations. Your task is to determine if the reply is ironic in the context of the post. Please provide in output (Output) a single label among the following array ["irony", "not irony"].*
>
> *Input:*
>
> *- post: YouTube told me to.*
>
> *- reply: Lol*
>
> *Output:*

Base prompt with BREXIT:

	Definition
Perspectivist concepts	
Perspectivism	A theoretical framework and a family of methods in AI that aim at **modeling different human perspectives** in predictive models. It typically involves the use of labeled data where the disaggregated labels are available (Frenda et al., 2024; Basile, 2020).
Perspectivist classification	Classification tasks **performed on individual labels provided by the annotator** (i.e. disaggregated data), which are leveraged to capture a plethora of perspectives. In PersEval, we evaluate each <text, annotator, label> input tuple separately. The output of the task is a label like in standard classification.
Perspectivist model	NLP model that leverages disaggregated data to **learn human perspectives** on language data. PersEval supports models that output a label for each <text, annotator> pair.
PERSEVAL glossary	
Instance and text	An instance is a **<text, annotator> pair**, where text is the text to be annotated and user is the annotator ID.
Annotators	Individuals who provide the annotation to *train* the model.
Users	Individuals who **use and interact** with the system. In Perseval, annotators and users are distinct. The annotators are a group of individuals whose annotations are used to train the system. Its inference is then run on a distinct set of individuals, i.e. the **users**.

Table 7: Definitions of key concepts of the Perspectivist approach, and the proposed evaluation framework PERSEVAL.

Instruction: You are provided in input (Input) a sentence (tweet) extracted from Twitter. Your task is to determine if the tweet is hateful. Please provide in output (Output) a single label among the following array ["hate speech", "not hate speech"].

Input: tweet: #brexit is going to effect other countries economically and they are most likely going to pass laws against the refugees that come in from

Output:

E.2 Perspective

Prompting in Perspective-setting with EPIC:

Instruction: You are Irish. You are provided in input (Input) a pair of sentences (post, reply) extracted from social media conversations. Your task is to determine if the reply is ironic in the context of the post. Please provide in output (Output) a single label among the following array ["irony", "not irony"].

Input:

- post: YouTube told me to.

- reply: Lol

Output:

Prompting in Perspective-setting with BREXIT:

Instruction: You are a researcher. You are provided in input (Input) a sentence (tweet) extracted from Twitter. Your task is to determine if the tweet is hateful. Please provide in output (Output) a single label among the following array ["hate speech", "not hate speech"].

Input:

tweet: #brexit is going to effect other countries economically and they are most likely going to pass laws against the refugees that come in from

Output:

E.3 IPA Named

Prompting in IPA Named setting with EPIC:

Instruction: You are Irish. You are provided in input (Input) a pair of sentences (post, reply) extracted from social media conversations. Your task is to determine if the reply is ironic in the context of the post. Please provide in output (Output) a single label among the following array ["irony", "not irony"].

{User-specific Example 1}

{User-specific Example 2}

{User-specific Example 3}

{User-specific Example 4}

{User-specific Example 5}

{Example to label}

Input:

- post: YouTube told me to.

- reply: Lol

Output:

Prompting in IPA Named setting with BREXIT:

Instruction: You are a researcher. You are provided in input (Input) a sentence (tweet) extracted from Twitter. Your task is to determine if the tweet is hateful. Please provide in output (Output) a single label among the following array ["hate speech", "not hate speech"].

{User-specific Example 1}

{User-specific Example 2}

{User-specific Example 3}

{User-specific Example 4}

{User-specific Example 5}

{Example to label}

Input:

tweet: #brexit is going to effect other countries economically and they are most likely going to pass laws against the refugees that come in from

Output:

E.4 IPA Unnamed

Prompting in IPA Unnamed setting with EPIC:

Instruction: You are provided in input (Input) a pair of sentences (post, reply) extracted from social media conversations. Your task is to determine if the reply is ironic in the context of the post. Please provide in output (Output) a single label among the following array ["irony", "not irony"].

{User-specific Example 1}

{User-specific Example 2}

{User-specific Example 3}

{User-specific Example 4}

{User-specific Example 5}

{Example to label}

Input:

- post: YouTube told me to.

- reply: Lol

Output:

Prompting in IPA Unnamed setting with BREXIT:

Instruction: You are provided in input (Input) a sentence (tweet) extracted from Twitter. Your task is to determine if the tweet is hateful. Please provide in output (Output) a single label among the following array ["hate speech", "not hate speech"].

{User-specific Example 1}

{User-specific Example 2}

{User-specific Example 3}

{User-specific Example 4}

{User-specific Example 5}

{Example to label}

Input:

tweet: #brexit is going to effect other countries economically and they are most likely going to pass laws against the refugees that come in from

Output:

F Error analysis

We present the complete results of the error analysis.

F.1 Q1: What is the contribution of each trait when ensembling the model's outputs?

Results from the ablation study are presented separately for each dataset in Table 10 (EPIC), Table 11 (MHS on *Perspective setting*), Table 12(MHS on *IPA setting*), and Table 13 (DICES).

F.2 Q2: Which demographic trait most significantly impacts the model's label predictions in the presence of varying annotator characteristics?

Table 14 illustrates the extent to which changing the value for each trait in the prompt influences the label assigned to the text.

F.3 Q3: How similar is the distribution of models' predictions to that of the annotators' chosen labels?

This third question has been assessed by computing the Jensen-Shannon Divergence between models and annotators' label distributions. Scores are presented separately for each dataset in Table 15 (EPIC), 16 (MHS), Table 17 (DICES), and Table 18 (BREXIT).

G The PERSEVAL Python Library

PERSEVAL is implemented as a Python library to facilitate access to the data, the different splits related to task variants, and the evaluation metrics.[7] The main interaction starts by instantiating a dataset from the `data` submodule. The user can then request the training, test, and optionally adaptation data splits with the `get_splits()` `method`, indicating whether the adaptation data (`user_adaptation`) is absent (`False`), available at training time (`train`) or at inference time (`test`). Additionally, the user chooses whether to extend the training split including texts also in test instances (`extended=True`) or to exclude them (`extended=False`). The dataset object contains a series of metadata about the dataset, such as its name, label names, and a dictionary of the annotator traits. Moreover, it contains the three splits, instantiated as objects of the same `PerspectivistSplit` class. These objects, called `training_set`, `test_set`, and `adaptation_set`, contain the list of users, texts, and the annotations, for the respective split. The `User` objects contain a unique identifier and a dictionary of traits. The `Text` objects contain a dictionary with the textual content of an instance, depending on the structure of the dataset. The `annotation` property is a dictionary where the keys are a pair (*User id*, *Text id*), and the value is a dictionary containing a value for each annotated label.

Besides providing access to the datasets and appropriate splits of the data for each task variant, the PERSEVAL library facilitates the automatic evaluation of models. The library implements the class `Evaluator`, which can be instantiated by passing the path of a file containing the predictions, a test set, and a target label name. The `Evaluator` object implements the functions to calculate the evaluation metrics described in Section 5. The output of the global, annotator-, text-, and trait-level metrics can be visualized in their aggregated forms and can be accessed (also at the level of each individual annotator, text, and trait) programmatically for a deeper analysis.

[7] The code will be released upon acceptance.

Dataset	Traits	Values
BREXIT	Group	Target, Control
DICES	Gender	Male, Female
	Age	GenX+, GenY, GenZ
	Education	College degree or higher, High school or below
	Ethnicity	Asian, Black, Latinx, White
EPIC	Gender	Male, Female
	Age	19-64 y/o, grouped in Boomer, GenX, GenY and GenZ
	Nationality	Australia, India, Ireland, United Kingdom, United States
MHS	Gender	Male, Female
	Age	18-81 y/o, grouped in Boomer, GenX, GenY, GenZ
	Education	College degree or higher, High school or below
	Income	less than 50k annual income, more than 50k annual income

Table 8: The sets of user traits included in PersEval for the BREXIT, DICES, EPIC and MHS datasets.

Parameter	Value
eval_strategy	epoch
greater_is_better	False
learning_rate	$5e^{-6}$
load_best_model_at_end	True
metric_for_best_model	eval_loss
num_train_epochs	5
per_device_eval_batch_size	32
per_device_train_batch_size	16

Table 9: Model parameters for the Encoder-based model.

Traits combination	Model	Precision	Recall	F1
Gender	Perspective-Llama	.473	.489	.481
Generation		.473	.553	.510
Nationality		.463	.473	.468
Generation-Nationality		.471	.518	.494
Gender-Generation		.471	.513	.491
Gender-Nationality		.470	.481	.475
Gender-Generation-Nationality		.472	.498	.484
Gender	Perspective-Mixtral	.485	.525	.504
Generation		.495	.592	.539
Nationality		.497	.501	.499
Generation-Nationality		.501	.557	.528
Gender-Generation		.484	.551	.515
Gender-Nationality		.489	.512	.500
Gender-Generation-Nationality		.485	.530	.507
Gender	IPA-Llama	.404	.829	.544
Generation		.401	.857	.546
Nationality		.389	.880	.539
Generation-Nationality		.397	.865	.544
Gender-Generation		.402	.841	.544
Gender-Nationality		.397	.850	.541
Gender-Generation-Nationality		.402	.857	.547
Gender	IPA-Mixtral	.456	.688	.548
Generation		.458	.696	.552
Nationality		.453	.706	.552
Generation-Nationality		.451	.691	.546
Gender-Generation		.455	.693	.549
Gender-Nationality		.455	.699	.551
Gender-Generation-Nationality		.460	.696	.554

Table 10: Precision, Recall, and F1 scores of the positive class on each trait, and their ensembled combinations on EPIC.

Traits combination	Model	Precision	Recall	F1
Age		.448	.580	.505
Education		.471	.818	.598
Income		.404	.270	.324
Gender		.474	.719	.571
Ideology		.460	.413	.435
Gender-Income		.452	.492	.471
Gender-Ideology		.468	.564	.512
Gender-Education		.471	.766	.583
Gender-Age		.458	.644	.536
Age-Education		.464	.709	.561
Age-Ideology		.450	.493	.471
Age-Income		.436	.426	.431
Education-Income		.453	.542	.493
Education-Ideology		.466	.612	.529
Income-Ideology		.443	.348	.389
Gender-Age-Education	Perspective-Llama	.467	.724	.568
Gender-Age-Income		.454	.567	.504
Gender-Age-Ideology		.456	.590	.514
Gender-Education-Income		.471	.703	.564
Gender-Education-Ideology		.472	.710	.567
Gender-Income-Ideology		.446	.451	.448
Age-Education-Income		.449	.581	.506
Age-Education-Ideology		.452	.604	.517
Age-Income-Ideology		.443	.432	.437
Education-Income-Ideology		.444	.457	.450
Gender-Age-Education-Income		.460	.640	.536
Gender-Age-Education-Ideology		.461	.652	.540
Gender-Age-Income-Ideology		.445	.503	.473
Gender-Education-Income-Ideology		.462	.582	.515
Age-Education-Income-Ideology		.448	.518	.480
Gender-Age-Education-Income-Ideology		.457	.591	.515
Age		.505	.875	.640
Education		.511	.870	.644
Income		.514	.866	.645
Gender		.506	.873	.641
Ideology		.525	.853	.650
Gender-Income		.511	.871	.644
Gender-Ideology		.516	.863	.646
Gender-Education		.508	.872	.642
Gender-Age		.506	.874	.641
Age-Education		.509	.874	.644
Age-Ideology		.515	.862	.645
Age-Income		.510	.871	.644
Education-Income		.513	.868	.645
Education-Ideology		.518	.861	.647
Income-Ideology		.520	.858	.648
Gender-Age-Education	Perspective-Mixtral	.505	.876	.641
Gender-Age-Income		.505	.876	.641
Gender-Age-Ideology		.509	.875	.643
Gender-Education-Income		.508	.873	.642
Gender-Education-Ideology		.511	.870	.644
Gender-Income-Ideology		.515	.869	.647
Age-Education-Income		.509	.875	.644
Age-Education-Ideology		.511	.873	.645
Age-Income-Ideology		.514	.870	.646
Education-Income-Ideology		.515	.868	.646
Gender-Age-Education-Income		.507	.875	.642
Gender-Age-Education-Ideology		.509	.874	.643
Gender-Age-Income-Ideology		.510	.872	.643
Gender-Education-Income-Ideology		.512	.871	.645
Age-Education-Income-Ideology		.512	.871	.645
Gender-Age-Education-Income-Ideology		.509	.874	.644

Table 11: Precision, Recall, and F1 scores of the positive class on each trait, and their ensembled combinations on MHS in *Perspective* setting.

Traits combination	Model	Precision	Recall	F1
Age		.546	.763	.636
Education		.534	.778	.633
Income		.545	.756	.633
Gender		.540	.772	.635
Ideology		.534	.778	.633
Gender-Income		.542	.767	.635
Gender-Ideology		.537	.775	.634
Gender-Education		.537	.775	.634
Gender-Age		.543	.767	.636
Age-Education		.541	.772	.636
Age-Ideology		.539	.768	.634
Age-Income		.545	.760	.635
Education-Income		.540	.767	.634
Education-Ideology		.533	.777	.632
Income-Ideology		.540	.766	.633
Gender-Age-Education	IPA-Llama	.540	.774	.636
Gender-Age-Income		.544	.766	.636
Gender-Age-Ideology		.541	.775	.638
Gender-Education-Income		.540	.772	.636
Gender-Education-Ideology		.537	.780	.636
Gender-Income-Ideology		.542	.773	.637
Age-Education-Income		.542	.769	.636
Age-Education-Ideology		.540	.778	.637
Age-Income-Ideology		.542	.770	.636
Education-Income-Ideology		.539	.775	.636
Gender-Age-Education-Income		.541	.770	.636
Gender-Age-Education-Ideology		.539	.776	.636
Gender-Age-Income-Ideology		.542	.772	.637
Gender-Education-Income-Ideology		.541	.777	.638
Age-Education-Income-Ideology		.541	.772	.636
Gender-Age-Education-Income-Ideology		.542	.773	.637
Age		.579	.688	.629
Education		.587	.681	.631
Income		.579	.687	.628
Gender		.572	.708	.633
Ideology		.607	.645	.626
Gender-Income		.575	.697	.630
Gender-Ideology		.589	.676	.630
Gender-Education		.580	.693	.631
Gender-Age		.576	.700	.632
Age-Education		.583	.682	.629
Age-Ideology		.594	.668	.629
Age-Income		.578	.688	.628
Education-Income		.583	.684	.629
Education-Ideology		.597	.664	.629
Income-Ideology		.593	.668	.628
Gender-Age-Education	IPA-Mixtral	.581	.695	.633
Gender-Age-Income		.578	.698	.632
Gender-Age-Ideology		.586	.686	.632
Gender-Education-Income		.581	.695	.633
Gender-Education-Ideology		.592	.683	.634
Gender-Income-Ideology		.587	.688	.633
Age-Education-Income		.581	.688	.630
Age-Education-Ideology		.593	.677	.632
Age-Income-Ideology		.589	.681	.631
Education-Income-Ideology		.592	.677	.632
Gender-Age-Education-Income		.580	.695	.632
Gender-Age-Education-Ideology		.588	.687	.634
Gender-Age-Income-Ideology		.584	.688	.632
Gender-Education-Income-Ideology		.589	.687	.635
Age-Education-Income-Ideology		.590	.682	.633
Gender-Age-Education-Income-Ideology		.586	.689	.634

Table 12: Precision, Recall, and F1 scores of the positive class on each trait, and their ensembled combinations on MHS in *IPA* setting.

Traits combination	Model	Precision	Recall	F1
Gender		.320	.310	.295
Generation		.318	.314	.267
Education		.330	.338	.312
Race		.311	.310	.292
Gender-Generation		.321	.317	.285
Gender-Education		.325	.321	.302
Gender-Race		.315	.305	.291
Generation-Education	Perspective-Llama	.328	.330	.293
Generation-Race		.316	.313	.281
Education-Race		.321	.324	.302
Gender-Generation-Education		.324	.323	.297
Gender-Generation-Race		.317	.308	.289
Gender-Education-Race		.323	.327	.304
Generation-Education-Race		.321	.317	.294
Gender-Generation-Education-Race		.322	.320	.298
Gender		.256	.249	.249
Generation		.262	.252	.251
Education		.279	.256	.257
Race		.246	.242	.241
Gender-Generation		.258	.254	.251
Gender-Education		.264	.248	.251
Gender-Race		.253	.244	.246
Generation-Education	Perspective-Mixtral	.271	.256	.256
Generation-Race		.258	.253	.251
Education-Race		.267	.256	.255
Gender-Generation-Education		.260	.252	.252
Gender-Generation-Race		.260	.256	.254
Gender-Education-Race		.257	.248	.250
Generation-Education-Race		.262	.255	.254
Gender-Generation-Education-Race		.263	.257	.256
Gender		.379	.364	.365
Generation		.392	.385	.385
Education		.390	.370	.375
Race		.374	.359	.363
Gender-Generation		.383	.369	.370
Gender-Education		.386	.368	.372
Gender-Race		.378	.363	.365
Generation-Education	IPA-Llama	.411	.391	.398
Generation-Race		.396	.381	.385
Education-Race		.377	.361	.365
Gender-Generation-Education		.384	.371	.373
Gender-Generation-Race		.372	.362	.364
Gender-Education-Race		.384	.366	.370
Generation-Education-Race		.389	.370	.375
Gender-Generation-Education-Race		.375	.363	.365
Gender		.321	.332	.310
Generation		.293	.327	.298
Education		.299	.329	.299
Race		.294	.323	.300
Gender-Generation		.316	.331	.307
Gender-Education		.331	.334	.311
Gender-Race		.309	.329	.306
Generation-Education	IPA-Mixtral	.282	.326	.295
Generation-Race		.296	.326	.300
Education-Race		.300	.327	.301
Gender-Generation-Education		.311	.331	.302
Gender-Generation-Race		.302	.328	.302
Gender-Education-Race		.320	.332	.305
Generation-Education-Race		.305	.330	.301
Gender-Generation-Education-Race		.310	.331	.303

Table 13: Macro-averaged Precision, Recall, and F1 scores on each trait, and their ensembled combinations for the DICES dataset.

Dataset	Model	Trait	Label change
Epic	Perspective-Llama	Gender	0.011
		Generation	0.028
		Nationality	0.033
	Perspective-Mixtral	Gender	0.027
		Generation	0.037
		Nationality	0.042
	IPA-Llama	Gender	0.122
		Generation	0.117
		Nationality	0.082
	IPA-Mixtral	Gender	0.173
		Generation	0.163
		Nationality	0.161
MHS	Perspective-Llama	Gender	0.037
		Age	0.053
		Education	0.046
		Income	0.045
		Ideology	0.119
	Perspective-Mixtral	Gender	0.034
		Age	0.021
		Education	0.025
		Income	0.025
		Ideology	0.048
	IPA-Llama	Gender	0.086
		Age	0.085
		Education	0.097
		Income	0.100
		Ideology	0.094
	IPA-Mixtral	Gender	0.078
		Age	0.074
		Education	0.070
		Income	0.084
		Ideology	0.076
DICES	Perspective-Llama	Gender	0.001
		Generation	0.006
		Education	0.006
		Race	0.008
	Perspective-Mixtral	Gender	0.003
		Generation	0.008
		Education	0.006
		Race	0.008
	IPA-Llama	Gender	0.004
		Generation	0.005
		Education	0.007
		Race	0.008
	IPA-Mixtral	Gender	0.007
		Generation	0.007
		Education	0.006
		Race	0.010

Table 14: Normalized label change for each trait.

Model	Trait	Value	JSD
Perspective-Llama	Gender	Male	0.380
		Female	0.326
	Generation	GenX	0.352
		GenY	0.316
		GenZ	0.370
	Nationality	Ireland	0.362
		India	0.386
		UK	0.308
		Australia	0.341
		US	0.291
Perspective-Mixtral	Gender	Male	0.346
		Female	0.283
	Generation	GenX	0.290
		GenY	0.291
		GenZ	0.323
	Nationality	Ireland	0.264
		India	0.370
		UK	0.299
		Australia	0.293
		US	0.271
IPA-Llama	Gender	Male	0.393
		Female	0.488
	Generation	GenX	0.464
		GenY	0.426
		GenZ	0.405
	Nationality	Ireland	0.481
		India	0.462
		UK	0.495
		Australia	0.454
		US	0.382
IPA-Mixtral	Gender	Male	0.353
		Female	0.337
	Generation	GenX	0.314
		GenY	0.329
		GenZ	0.338
	Nationality	Ireland	0.301
		India	0.396
		UK	0.320
		Australia	0.287
		US	0.291

Table 15: Similarity of the distributions between models predictions and annotators' labels for each trait, computed through Jensen-Shannon Divergence (JSD) on EPIC.

Model	Trait	Value	JSD
Perspective-Llama	Gender	female	0.332
		male	0.310
		non-binary	0.387
	Generation	Boomer	0.340
		GenX	0.371
		GenY	0.366
		GenZ	0.384
	Education	high	0.322
		low	0.341
	Income	high	0.457
		low	0.471
	Ideology	liberal	0.354
		conservative	0.491
		neutral	0.436
Perspective-Mixtral	Gender	female	0.281
		male	0.295
		non-binary	0.282
	Generation	Boomer	0.251
		GenX	0.276
		GenY	0.269
		GenZ	0.276
	Education	high	0.281
		low	0.270
	Income	high	0.298
		low	0.276
	Ideology	liberal	0.281
		conservative	0.274
		neutral	0.268
IPA-Llama	Gender	female	0.221
		male	0.236
		non-binary	0.208
	Generation	Boomer	0.210
		GenX	0.228
		GenY	0.237
		GenZ	0.208
	Education	high	0.255
		low	0.244
	Income	high	0.237
		low	0.242
	Ideology	liberal	0.238
		conservative	0.234
		neutral	0.253
IPA-Mixtral	Gender	female	0.211
		male	0.212
		non-binary	0.193
	Generation	Boomer	0.217
		GenX	0.201
		GenY	0.224
		GenZ	0.177
	Education	high	0.220
		low	0.236
	Income	high	0.231
		low	0.225
	Ideology	liberal	0.208
		conservative	0.221
		neutral	0.233

Table 16: Similarity of the distributions between models predictions and annotators' labels for each trait, computed through Jensen-Shannon Divergence (JSD) on MHS.

Model	Trait	Value	JSD
Perspective-Llama	Gender	woman	0.460
		man	0.429
	Generation	Millenial	0.531
		GenX	0.506
		GenZ	0.502
	Education	college degree or higher	0.416
		high school or below	0.432
	Race	Black/African American	0.393
		LatinX, Latino, Hispanic or Spanish Origin	0.453
		Asian/Asian subcontinent	0.463
		White	0.413
		Multiracial	0.408
Perspective-Mixtral	Gender	woman	0.312
		man	0.302
	Generation	millenial	0.311
		GenX	0.271
		GenZ	0.323
	Education	college degree or higher	0.253
		high school or below	0.312
	Race	Black/African American	0.260
		LatinX, Latino, Hispanic or Spanish Origin	0.262
		Asian/Asian subcontinent	0.335
		White	0.294
		Multiracial	0.348
IPA-Llama	Gender	woman	0.262
		man	0.265
	Generation	millenial	0.256
		GenX	0.263
		GenZ	0.277
	Education	college degree or higher	0.262
		high school or below	0.296
	Race	Black/African American	0.243
		LatinX, Latino, Hispanic or Spanish Origin	0.242
		Asian/Asian subcontinent	0.271
		White	0.271
		Multiracial	0.332
IPA-Mixtral	Gender	woman	0.246
		man	0.240
	Generation	millenial	0.218
		GenX	0.224
		GenZ	0.252
	Education	college degree or higher	0.239
		high school or below	0.257
	Race	Black/African American	0.223
		LatinX, Latino, Hispanic or Spanish Origin	0.168
		Asian/Asian subcontinent	0.236
		White	0.276
		Multiracial	0.292

Table 17: Similarity of the distributions between models predictions and annotators' labels for each trait, computed through Jensen-Shannon Divergence (JSD) on DICES.

Model	Trait	Value	JSD
Perspective-Llama	Group	Control	0.151
		Target	0.138
Perspective-Mixtral	Group	Control	0.198
		Target	0.193
IPA-Llama	Group	Control	0.518
		Target	0.302
IPA-Mixtral	Group	Control	0.229
		Target	0.174

Table 18: Similarity of the distributions between models predictions and annotators' labels for each trait, computed through Jensen-Shannon Divergence (JSD) on BREXIT.