MultiPICo: Multilingual Perspectivist Irony Corpus

Silvia Casola[⋆], **Simona Frenda**[⋆⊙], **Soda Marem Lo**[⋆], **Erhan Sezerer**[◇], **Antonio Uva**[◇], **Valerio Basile**[⋆], **Cristina Bosco**[⋆], **Alessandro Pedrani**[◇], **Chiara Rubagotti**[◇], **Viviana Patti**[⋆], **Davide Bernardi**[◇]

[⋆] Computer Science Department, University of Turin, Turin, Italy

[⊙] aequa-tech, Turin, Italy

[◇] Alexa AI, Amazon, Amazon Development Centre Italy, Turin, Italy

{silvia.casola|simona.frenda|sodamarem.lo|valerio.basile|cristina.bosco|viviana.patti}@unito.it

{erhanszr|antonuva|pedrana|crubagot|dvdbe}@amazon.it

Abstract

Recently, several scholars have contributed to the growth of a new theoretical framework in NLP called perspectivism. This approach aims to leverage data annotated by different individuals to model diverse perspectives that affect their opinions on subjective phenomena such as irony. In this context, we propose MultiPICo, a multilingual perspectivist corpus of ironic short conversations in different languages and linguistic varieties extracted from Twitter and Reddit. The corpus includes sociodemographic information about its annotators. Our analysis of the annotated corpus shows how different demographic cohorts may significantly disagree on their annotation of irony and how certain cultural factors influence the perception of the phenomenon and the agreement on the annotation. Moreover, we show how disaggregated annotations and rich annotator metadata can be exploited to benchmark the ability of large language models to recognize irony, their positionality with respect to sociodemographic groups, and the efficacy of perspective-taking prompting for irony detection in multiple languages.

1 Introduction

The pervasiveness of AI-based technologies has renewed the interest in making Artificial Intelligence more inclusive and attentive to users' needs. AI-based technology mirrors the quality and problems of data that feed it (Dignum, 2023). Therefore, the way corpora are created and the design biases that are — consciously or unconsciously — included in datasets and models are of central importance. *Perspectivist* corpora allow turning bias from an undesirable criticality into a measurable trait by reporting the perception of multiple individuals with different social and cultural traits on pragmatic phenomena (e.g., hate speech or irony).

In this work, we present MultiPICo (Multilingual Perspectivist Irony Corpus), a multilingual corpus of short conversations (Post-Reply) extracted from Twitter and Reddit and annotated as ironic or not ironic by crowdsourcing workers with different social backgrounds[1].

MultiPICo covers 9 languages and 25 varieties, ranging from high to low resources. Specifically, we include 5 varieties of English (Australian, British, Indian, Irish and US English), 5 varieties of Spanish (Argentinean, Castilian, Colombian, Mexican, and US Spanish), 5 varieties of Arabic (Egyptian, Iraqi, Moroccan, Saudi Arabian, and Yemen), 2 varieties of French (Canadian and French), 3 varieties of German (Austrian, German, and Swiss), 2 varieties of Portuguese (Brazilian and Portuguese), and Italian, Dutch, and Hindi. Selected annotators' nationality is related to these linguistic varieties, the annotators' gender is balanced, and a rich set of other sociodemographic information (age, ethnicity, student and employment status) is provided.

Starting from these metadata, we analyze the perception of irony in different sociodemographic groups and explore which demographic dimensions primarily explain the annotation. Furthermore, we also show how to exploit this multifaceted corpus as a benchmark for AI-based technologies, including Large Language Models (LLMs). These models are becoming pervasive, but the lack of information on their training data and methods makes their intrinsic bias hard to measure.

In short, our contributions are the following:

- We present MultiPICo, a disaggregated multilingual dataset annotated with the presence of irony in social media short conversations (Section 3).
- We analyze the impact of sociodemographic

S. Casola, S. Frenda, and S. M. Lo contributed equally to this work.

[1] MultiPICo is available at https://huggingface.co/datasets/Multilingual-Perspectivist-NLU/MultiPICo with a CC-BY 4.0 license.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16008–16021

dimensions and cultural background in the perception of irony in multiple languages (Section 4).

- We explore the performance of three LLMs on the task of irony detection (Section 5).

- We propose a framework to evaluate LLMs' positionality and the efficacy of sociodemographic prompting (Section 5).

2 Related Work

The NLP community is increasingly questioning how annotators' background can influence their choices and perceptions of pragmatic linguistic phenomena, such as irony or hate speech.

Bender and Friedman (2018) proposed a set of best practices for NLP technicians to avoid ethical issues such as exclusion and misrepresentation of users. Their recommendations include reporting annotators' demographic data, as they can impact how annotators use and interpret language.

Numerous works have highlighted human annotation is subjective (Aroyo and Welty, 2015), especially for semantic and pragmatic phenomena — as the point of view might differ in relation to one's social background, beliefs, and demographics. An increasing number of works emphasize abandoning a single ground truth and preserving the disagreement among annotators (Leonardelli et al., 2021; Uma et al., 2021a; Leonardelli et al., 2023). Dealing with human label variation has consequences on the whole Machine Learning pipeline (Plank, 2022), from data to modeling (Mostafazadeh Davani et al., 2022) and evaluation (Uma et al., 2021c; Basile et al., 2021).

The *perspectivist* approach aims at leveraging disagreement and modelling annotators' perspectives (Basile et al., 2021). To do so, scholars have exploited individual annotators' decisions (Mostafazadeh Davani et al., 2022) or grouped their annotations based on their beliefs (Akhtar et al., 2019) or demographic traits (Frenda et al., 2023b). To work in this direction, disaggregated corpora with demographic information become of fundamental importance (Cabitza et al., 2023). In fact, aggregated data tend to reflect a minority of perspectives, under-representing others (Prabhakaran et al., 2021; Frenda et al., 2023b).

Disaggregated datasets have become more widespread in recent years, as listed in the Perspectivist Data Manifesto[2] and by Plank (2022)[3]. Some of these corpora account for annotators' demographic. For example, Sachdeva et al. (2022) built a perspectivist corpus for hate speech detection, asking annotators to self-identify their race, gender, sexuality, religion, education, and income. Almanea and Poesio (2022) collected a dataset about misogyny and sexism in Arabic and had it annotated by three Muslim annotators (2 female, 1 male), who self-identified their religious belief as liberal, moderate, and conservative. They also provided annotations by a larger cohort on a small set of their dataset. The work showed a correlation between annotators' beliefs and their perception of misogyny and sexism in the texts.

Similarly, Akhtar et al. (2021) developed a dataset for hate speech detection involving migrants as victims of offensive texts and demonstrating that members of the same group tend to agree more. Moreover, they implemented perspective-aware models by creating a gold standard for each group, training two separate models, and finally ensembled by majority vote. Perspective-based ensembling methods have also been explored by Casola et al. (2023) for irony and hate speech detection in English. In their work, perspectives are extracted using sociodemographic information of annotators and mined from annotations.

The use of demographic data when working on highly subjective language phenomena has shown improvement in NLP research. Sap et al. (2022) investigated how annotators with different identities and beliefs perceive toxic content considering the characteristics of texts along three dimensions: racial prejudices, African American English vernacular and dialects, and vulgar words. The authors demonstrated that identities and beliefs impact annotators' judgments. Sociodemographic information has also been used to build a disagreement predictor by Wan et al. (2023), who tried to quantify how controversial an utterance is. They show that adding demographic information to the input improves results and returns insights on which instances need a more diverse group of annotators. Finally, Santy et al. (2023) developed a framework to identify datasets and model design biases, highlighting their positionality. Results show available datasets strongly align with the young WEIRD population (Western, Educated, Industrialized, Rich,

[2] https://pdai.info/

[3] www.github.com/mainlp/awesome-human-label-variation

Democratic).

Few disaggregated datasets for irony detection exist. Simpson et al. (2019) released a corpus about humor detection in English, used as a benchmark in the first edition of the Learning With Disagreement shared task (Uma et al., 2021b). No annotators' metadata, however, are included. Frenda et al. (2023b) proposed a dataset for irony detection and investigated the influence of the annotators' characteristics on their perception (Frenda et al., 2023a); the dataset, however, contains English text only.

To the best of our knowledge, no multilingual disaggregated dataset exists. We fill this gap by proposing a perspectivist dataset for irony in 9 languages and a total of 25 high and low-resourced varieties. We also provide demographic information on all the annotators involved.

3 MultiPICo

In this section, we describe MultiPICo, a corpus of 18,778 short conversations collected from Reddit (8,956) and Twitter (9,822) in 9 languages, and a total of 25 varieties. Data has been collected reproducing the structure of short conversations. Annotators were asked to read a set of Post and Reply pairs and answer whether the text of the reply was ironic or not, given the context.

We collected Reddit comments (first comments in the vast majority of cases and second-level comments in a few cases) with their direct replies and conversation-starting messages. For Twitter, the Post could be both a conversation starter or a direct reply to it. Reddit data were retrieved using the Pushshift repository[4] from January 2020 to June 2021. To collect data in several linguistic varieties, we picked 26 subreddits, as reported in Table 10. Thus, we inferred the linguistic variety of subreddit users from the subreddit. We filtered out pairs having at least one deleted or removed comment and further analyzed the target languages using the Python library for language identification LangID[5]. We collected Twitter data via Twitter Stream API, using the geolocation service and excluding quotes and retweets. Then, we retrieved the full conversation and retained tweets that directly replied to the starting ones. We tried to collect the same number of Post-Reply pairs from the two selected sources. However, there was not enough data on Reddit for some linguistic varieties. This was the case of US-American Spanish, Swiss-German, Iraqi, and Yemen Arabic (see Table 10). The data collection resulted in 18,778 instances, together with their metadata, consisting of Post-Reply original IDs, subreddits, and geolocation information.

The human annotation of the collected data was performed on the crowdsourcing platform Prolific[6], through an integrated custom-built annotation interface designed to collect a diverse and balanced set of annotators. The interface mimicked a message conversation, having the Post as context and asking whether the Reply was Ironic or Not ironic[7] . We prevented an instance from being annotated more than a predefined number of times, with a mean of 5.02 annotations per instance (see Table 1).

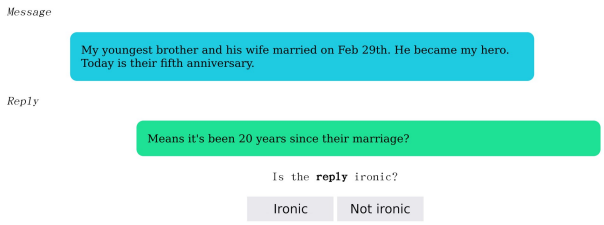

Figure 1: Screenshot of the annotation interface.

Annotators were selected based on three factors: i) their completion rate had to be at least 99%; ii) they had to be native speakers of the considered language; iii) the set of annotators needed to be balanced both across genders (230 Female, 274 Male, 1 Prefer not to say, 1 Null[8]) and nationality (we hired around 24 annotators for all linguistic varieties, except for the English dataset with around 15). Considering both annotators' mother tongue and nationality, we inferred that annotators from a specific country speak the corresponding linguistic variety. Because of the lack of annotators, we were not able to preserve balance for the Arabic subset of the dataset[9].

The quality of the annotation has been further assured using attention check questions in the form of *"Please answer X to this question"*. Annotators had 1% probability of receiving these special questions. We set a threshold of 50% correct answers; only 13 annotators failed the test, and their annotations

[4] https://redditsearch.io/

[5] https://github.com/saffsd/langid.py

[6] https://www.prolific.com/

[7] The instructions and an example of annotation is presented in Appendix A.1.

[8] This annotator did not share the information.

[9] The Arabic portion of MultiPICo has 46 self-assessed males, 21 self-assessed females, and 1 annotator who prefers not to reveal their gender. Since we did not find enough annotators from the 5 considered countries, we removed this filter, collecting annotations from all mother-tongue annotators.

Language	#Annotators	#Annotations	Label rate		#Texts	Sources		Annotation mean
			%not	%iro		#Reddit	#Twitter	
Arabic	68	10,609	68	32	2,181	949	1,232	4.86
Dutch	25	4,991	73	27	1,000	500	500	4.99
English	74	14,171	69	31	2,999	1,499	1,500	4.73
French	50	8,770	70	30	1,760	1,000	760	4.98
German	70	12,510	68	32	2,375	1,042	1,333	5.27
Hindi	24	4,711	65	35	786	286	500	5.99
Italian	24	4,790	69	31	1,000	500	500	4.79
Portuguese	49	9,754	62	38	1,994	997	997	4.89
Spanish	122	24,036	67	33	4,683	2,183	2,500	5.13
Total	506	94,342	68	32	18,778	8,956	9,822	5.02

Table 1: Number of annotators, annotations, texts per source, and annotation means for each language.

were excluded from the final corpus, leading to a total of 506 annotators.

Together with the annotators' gender and nationality, we collected other demographic information, specifically: *Age Group* (13 Baby Boomer, 66 Gen-X, 260 Gen-Y, 162 Gen-Z, 5 Null), *Ethnicity* (315 White, 64 Mixed, 44 Asian, 13 Black, 66 Other, 4 Null), *Student status* (260 No, 165 Yes, 81 Null), *Employment status* (178 Full-time, 74 Part-time, 74 Unemployed and job seeking, 24 Not in paid work, 11 Due to start a new job within the next month, 36 Other, 109 Null)[10].

4 Annotators' polarization

In this section, we present an exploratory analysis of MultiPICo. We employ primarily the *polarization index* (P-index) proposed by Akhtar et al. (2019), which measures the instance-based polarization for k groups. For each instance i, the P-index is defined as:

$$P(i) = \frac{1}{k} \sum_{1 \leq w \leq k} a(G_w)(1 - a(G)) \quad (1)$$

where k is the number of different groups of annotators, $a(G_w)$ indicates the internal agreement for group G_w, and $a(G)$ indicates the overall agreement on the instance. We compute the instance-level agreement as:

$$a(G) = 1 - \frac{\chi^2(G)}{|M|} \quad (2)$$

where $\chi^2(G)$ is the chi-square statistics and M is the complete set of annotators. A high P-index (close to 1) indicates that the annotation of an instance is highly divergent (or *polarized*) across groups, while each group is internally consistent.

[10]Details per language are in Appendix A.2.

[11]The values of Ethnicity for French and Dutch are uncertain because of the unbalance amount of annotators for each trait of this dimension (see Table 8 in Appendix A.2).

For instance, looking at table 3, Example #1 is annotated as ironic regardless of the annotators' sociodemographic traits, and its P-index is 0; Example #2 has a P-index of 1 because it is annotated as ironic by annotators who self-identified as male and not ironic by the one belonging to the other group (female).

We applied the P-index for two analyses to understand the relevance of sociodemographic information in the interpretation of irony and to investigate its cultural basis:

1. analysis of dimensions: we compared the P-indices obtained by dividing the population based on gender, generation, nationality, ethnicity, student, and employment status (Section 4.1);

2. analysis of cultural basis: we examined the possible connection between the linguistic variety of the instances and the nationality of annotators, observing the mean of P-index scores of the instances belonging to the same linguistic variety when annotated as ironic by workers from the same country (Section 4.2).

To evaluate the significance of the obtained P-index scores, we computed random P-index values for each instance using random partitions with the same number of annotators for dimension and trait (see Table 8 in Appendix A). Thus, we could compare the mean of real P-index scores per group of annotators (henceforth 'real P-index') and the mean of their corresponding random P-index values (henceforth simply 'random P-index'). Finally, specifically for analyses 1 and 2, we also calculated the difference in percentage (Δ) between real and random P-index.

	Gender			Generation			Nationality			Ethnicity			Student status			Employment status		
	real P	random P	%Δ	real P	random P	%Δ	real P	random P	%Δ	real P	random P	%Δ	real P	random P	%Δ	real P	random P	%Δ
ar	.18	.16	15.7	.19	.17	14.5	–	–	–	.24	.22	9.1	.18	.17	3.3	.26	.23	13.8
nl	.17	.17	-.4	.18	.18	-2.1	–	–	–	.22	.15	*44.2*	.19	.17	12.3	.23	.19	21.0
en	.22	.20	12.4	.25	.21	16.9	.26	.25	4.1	.25	.21	21.5	.23	.21	11.6	.27	.24	12.2
fr	.19	.16	13.7	.21	.19	12.2	.20	.17	19.5	.24	.17	*44.0*[11]	.19	.17	11.5	.24	.19	27.3
de	.19	.17	13.5	.22	.19	12.1	.21	.20	3.7	.21	.19	8.5	.20	.19	3.4	.25	.23	9.5
hi	.26	.23	15.4	.25	.24	1.0	–	–	–	–	–	–	.26	.23	12.9	.30	.27	9.4
it	.17	.16	3.8	.18	.16	14.1	–	–	–	–	–	–	.20	.15	27.2	.18	.19	-5.0
pt	.22	.20	7.7	.24	.21	15.3	.23	.20	13.9	.25	.20	23.2	.24	.20	20.4	.26	.24	9.0
es	.21	.19	7.0	.23	.21	12.3	.25	.24	4.0	.24	.21	15.9	.23	.21	12.1	.25	.24	3.1

Table 2: Polarization per dimensions and language. The table shows the percentage of delta (%Δ) between the mean of real P-index (real P) for dimension and its corresponding random P-index (random P).

#	Post	Reply	ann_1	ann_2	ann_3	ann_4	ann_5	ann_6	P-index
			female	female	female	male	male	male	
1	This man is so completely focused towards engineering riots.	Gotta stick to your strengths	ironic	ironic	ironic	ironic	ironic	ironic	0
#	Post	Reply	ann_1	ann_2	ann_3	ann_4			P-index
			female	female	male	male			
2	Curse Zeus for making me a mortal	Damn you Zeus! Damnnnn yooouuu!!!!! ... Every morning.	not ironic	not ironic	ironic	ironic			1

Table 3: Example of polarized and non-polarized annotations.

4.1 Analysis of dimensions

As the first exploratory analysis of MultiPICo, we examined the general level of polarization of all the dimensions available in this dataset: gender, generation, nationality, ethnicity, student, and employment status. In this case, annotators have been separated into various groups according to their trait (for instance, Boomers), and the P-index is computed among all the groups of the same demographic dimension (thus, e.g., Boomers, GenX, GenY, and GexZ for generation)[12].

Table 2 shows that all the differences between real and random P-index are positive (see column of %Δ), except in very few cases (i.e., two cases in Dutch and one in Italian), likely due to the unbalance of information available about annotators. Moreover, for specific dimensions, such as generation and student status, %Δ is higher than 10% for the majority of languages. This observation suggests that age and student status are the most polarizing dimensions when detecting irony, i.e., people from the same age group tend to agree on the annotation while often disagreeing with those in another age range.

[12]In this case k depends on the number of traits per dimension (i.e., $k = 2$ in gender, $k = 4$ in generation, etc.) (Akhtar et al., 2019).

4.2 Analysis of cultural basis

In this second analysis, to examine if irony is affected by the cultural background, we investigated the connection between linguistic varieties and nationality, observing the P-index of instances annotated as ironic by the majority of contributors coming from the corresponding nationality. In this case, we used the P-index in a binary fashion, i.e., *one-vs-all*. For each nationality, we divided the annotators into two groups, i.e., those with that nationality vs. those with other nationalities, and computed the P-index with $k = 2$ accordingly.

In Table 4, the agreement in the interpretation of irony appears to be country-based for specific varieties. A clear example is in German, where Austrian and German texts report the highest P-index when the texts are considered ironic by annotators coming from Austria and Germany respectively.

The case of English is also interesting: the identification of irony in Indian texts polarizes specifically when the annotators come from India. Similar results are also observed, for instance, between American and British English and the annotators from their corresponding countries. The same tendency is visible in Spanish, where higher values of the P-index are reported when Mexican annotators detect irony in Mexican texts.

Nationality	Language variety				
English	US	GB	AU	IE	IN
USA	***.39***	.43	.21	.38	.32
India	.35	.37	.38	.39	**.50**
UK	.34	***.39***	.32	.36	.45
Ireland	.27	.32	.44	.36	.42
Australia	.42	.41	.33	.40	.46
French	CA	FR			
Canada	.31	.38	–	–	–
France	.19	**.27**			
German	CH	DE	AT		
Austria	.20	.28	**.29**	–	–
Switzerland	***.27***	.23	.28	–	–
Germany	.23	**.27**	.19	–	–
Portuguese	PT	BZ			
Portugal	.30	.32	–	–	–
Brazil	.28	**.32**			
Spanish	ES	US	MX	CO	AR
USA	.26	.32	.37	.45	.42
Colombia	.35	.31	.33	.24	.24
Spain	.29	.32	.39	.36	.35
Mexico	.31	.38	**.40**	.34	.35
Argentina	.33	.31	.31	.39	.33

Table 4: Connection between the nationality of annotators and the language varieties per language through the P-index.

The situation in French and Portuguese is less crisp, showing specific varieties (French from France and Brazilian, respectively) that induce polarization regardless of the provenience of annotators. Looking at French texts, we noticed that most of the Post-Reply pairs annotated as ironic also by Canadians are humorous:

(3) [Post] Direction Amsterdam ?

[Reply] @user Salon de l'herbe?[13]

In general, results in Table 4 suggest that the interpretation of irony is similar in speakers of the same language. However, for specific varieties, the sensibility of annotators is affected by their cultural background.

5 MultiPICo as a benchmarking tool

Given its fine-grained metadata, both at the annotation and the annotator level, MultiPICo allows for detailed performance evaluation in a perspectivist setting. In this section, we will first analyze LLMs zero-shot performance for irony detection (Section 5.1) and then show some of the evaluation possibilities in exploring the LLMs' positionality (Section B.3) and the effectiveness of socio-demographic prompting (Section 5.3).

5.1 Baseline performance

This section analyzes the irony detection performance of multilingual Large Language Models on MultiPICo. In particular, we will consider ChatGPT (gpt-3.5-turbo), PolyLM (Polyglot Large Language Model) (Wei et al., 2023)[14] in its instruction-tuned 13B version and mT0 (Scao et al., 2023) with 13B parameters[15].

	Random	ChatGPT	PolyLM	Bloom
ar	.348	.567	0	–
nl	.350	.152	.347	.319
en	.341	.481	.386	.023
fr	.353	.477	0	0
de	.336	.559	.426	.268
hi	.349	.475	–	–
it	.335	.533	.405	.334
pt	.387	.374	.498	.385
es	.382	.474	.440	.376

Table 5: Models' f1 results for positive class per language, when prompted zero-shot; we use the annotators' majority vote as the gold standard. PolyLM does not support the Hindi language.

As a first baseline, we consider the models' performance on the aggregated dataset[16]. For all languages, we prompt the models to classify examples as ironic or not ironic. All prompts are written by native speakers of the target language. An example of a prompt for English is in Appendix B.1.

Table 5 reports the f1-score for the positive class for all languages. Open models tend to perform lower than ChatGPT and are often close to random when considering irony. All models tend to output some non-standard labels, i.e., different than *irony* or *not irony* as requested in the prompt; for ChatGPT (always) and PolyLM (in the majority of cases); however, the labels are semantically meaningful and can be mapped to the correct class in a semi-automatic way. In contrast, mT0 tends to prepend the prompt to the labels and produces many meaningless labels; we map the output to the intended label[17] when possible and consider meaningless outputs incorrect by default. Results show that irony detection is still an open problem in a zero-shot scenario, particularly in non-mainstream languages.

[13] [Post] Direction Amsterdam ? [Reply] Herbal salon?

[14] huggingface.co/DAMO-NLP-MT/polylm-chat-13b

[15] Following BLOOM's recommendations, we use *mT0-xxl* for English and *mT0-xxl-mt* for other languages.

[16] For each instance, we consider the majority vote among all annotators.

[17] The process was semi-automatic. We checked non-standard outputs for patterns that refer to the ironic or not ironic label (e.g., "irony", "ironic", "Output: ironic", "[prompt] ironic", etc.). We were not able to perform this operation for non-Latin scripts, and thus do not report results for these languages.

language	Young	Old
ar	**.398**	.364
nl	**.172**	.045
en	**.272**	.219
fr	**.324**	.260
de	**.377**	.332
hi	.252	**.263**
it	**.415**	.380
es	**.317**	.190

Table 6: ChatGPT's positionality with respect to age. We report the correlation between the model's output and the annotators' mean label.

5.2 Positionality

Santy et al. (2023) have exploited annotators' sociodemographic metadata to investigate design bias and the intrinsic positionality of datasets and models. Given the rich set of metadata available in MultiPICo, we can explore LLMs' positionality in the perception of irony for multiple languages.

Following Santy et al. (2023), we grouped annotators according to their demographic traits and computed the mean label for annotators having a certain trait only. We use an LLM to label the instance in a zero-shot approach, using the base prompt in Appendix B.1. For all instances, we take the label produced by the model and compute their Pearson's correlation coefficient with the group-specific mean labels.

We find that, for all languages, ChatGPT tends to align with the perspective of young annotators (see Table 6). This is consistent with OpenAI's report on InstructGPT (Ouyang et al., 2022), where the annotators are described as «quite young (75% less than 35 years old)». Other dimensions — other than student status — show a weaker signal (see Table 11). Characterizing models' positionality is particularly important with closed-source systems, for which the design choices, training data, and algorithms are unknown and can only be characterized by their observed behaviors (Gallegos et al., 2023; Li et al., 2023; Kotek et al., 2023).

5.3 Perspective-taking prompting

To explicitly model the differences in human views toward given phenomena, recent work has explored ad-hoc prompting, impersonating a specific user (Deshpande et al., 2023; Cheng et al., 2023) or exploiting relevant opinions previously expressed by the user (Hwang et al., 2023). In this context, Beck et al. (2023) have shown that sociodemographic prompting lacks robustness and propose to use it to identify ambiguous instances. However, its effect is hard to evaluate.

MultiPICo provides a novel benchmark to evaluate sociodemographic prompting for irony detection in multiple languages. To showcase this, we construct perspective-based datasets per language. For each sociodemographic trait, we consider labels from annotators with that trait only and aggregate them using majority voting. Then, we divide each dataset into a training (80%) and a test (20%) set. We perform the following experiments:

- No additional information (base): we prompt the model with a Post-Reply pair as discussed in Section 5.1.
- Trait-based perspective-taking (trait): we ask the model to impersonate a person from a specific socio-demographic group by prepending the information, e.g., "You are a self-identified male/female/. . . /Indian/. . . ".
- Data-driven perspective-taking (data): we extract the most representative examples for each trait. We split the annotators into two groups: one composed of people with the trait and its complement. Among instances with a high agreement in the target group (> .60), we then rank the instances by their P-index computed according to this binary split and select the top 3 ironic and not ironic instances according to the target group annotation. We prompt the model in a few-shot setting, providing the extracted examples and the related labels. Note that this approach is completely data-driven and can be exploited to perform perspective-taking prompting even when the nature of the target group is difficult to describe.

An example of all prompts is in Appendix B.1. We use PolyLM for all experiments. Table 7 shows a widely variable performance across languages. For English, base results show the best performance in many cases, particularly for gender and generation. For Spanish and German, perspective-taking prompting works well, with the few-shot approach consistently superior to the prompt-based one. The data-driven approach also obtains good results for the Dutch language, where the fixed prompts do not improve over the baseline. For French, the few-shot approach is the only one that works; otherwise, the model never predicts irony. Finally, results for Portuguese and Italian are mixed, while the model

	EN			ES			FR			DE		
	base	trait	data	base	trait	data	base	trait	data	base	trait	data
Sex												
Female	.408	.352	.380	.495	.513	.368	0	0	.439	.364	.404	.414
Male	.395	.315	.377	.426	.416	.430	0	0	.373	.401	.407	.451
Generation												
Boomer	.519	.333	.469	.538	.552	–	0	0	–	.429	.512	.520
GenX	.368	.276	.294	.520	.524	.527	0	0	.424	.394	.421	.463
GenY	.441	.462	.381	.443	.447	.479	0	0	.404	.369	.410	.448
GenZ	.439	.423	.359	.472	.483	.493	0	0	.393	.382	.444	.425
Old	.378	.354	.352	.519	.517	.544	0	0	.453	.417	.432	.441
Young	.410	.435	.408	.448	.449	.458	0	0	.386	.379	.393	.432
Student												
No	.387	.386	.341	.442	.463	.431	0	0	.396	.413	.404	.412
Yes	.495	.517	.482	.516	.514	.510	0	0	.373	.402	.484	.498
Employed												
No	.389	.407	.374	.464	.469	.413	0	.034	.372	.356	.412	.412
Yes	.412	.429	.365	.485	.479	.484	0	0	.425	.439	.455	.453
Nationality												
	Australia			Argentina				Canada		Austria		
	.459	.456	.293	.474	.426	.433	0	0	.473	.424	.457	.478
	India			Colombia				France		Germany		
	.435	.452	.405	.453	.460	.445	0	0	.337	.376	.404	.419
	Ireland			Mexico						Switzerland		
	.410	.429	.374	.562	.538	.556				.342	.392	.419
	UK			Spain								
	.466	.454	.396	.435	.455	.452						
	US			US								
	.400	.417	.401	.521	.520	.521						

	PT			IT			NL			AR		
	base	trait	data	base	trait	data	base	trait	data	base	trait	data
Sex												
Female	.491	.487	.499	.400	.400	.405	.430	.430	.425	0	0	0
Male	.520	.523	.524	.438	.438	.443	.279	.279	.278	0	0	0
Generation												
Boomer										0	0	0
GenX	.836	.836	–	.558	.558	–	.389	.389	.416	0	0	–
GenY	.436	.436	.436	.389	.389	.391	.352	.359	.397	0	.059	0
GenZ	.540	.545	.512	.476	.476	.477	.362	.362	.397	0	0	0
Old	.836	.836	.776	.558	.558	.483	.389	.389	.420	0	0	0
Young	.491	.490	.494	.398	.398	–	.339	.339	.368	0	0	0
Student												
No	.469	.469	.473	.389	.389	.388	.333	.333	.372	0	0	0
Yes	.538	.536	.549	.512	.512	.535	.367	.367	.374	0	0	0
Employed												
No	.487	.489	.511	.500	.500	.415	.429	.429	.417	0	0	0
Yes	.536	.533	.555	.374	.374	.358	.308	.308	0	0	0	0
Nationality												
	Brazil											
	.471	.483	.465									
	Portugal											
	.537	.535	.528									

Table 7: PolyLM's f1 results for the positive class when prompting without sociodemographic information (base), with trait-based perspectives (trait) and with data-driven perspective examples (data) for all languages. The best performance is underlined. Data for some sociodemographic groups might be missing due to the lack of or the very low number of annotations.

does not seem to have a notion of irony for the Arabic language.

6 Conclusions and Future Work

We have presented MultiPICo, a multilingual perspectivist corpus for irony. The corpus consists of Post-Reply pairs from social media platforms, and it is released with disaggregated annotations and annotators' metadata.

Our analysis of the corpus shows a high variability in the perception of irony across sociodemographic groups, confirming the importance of considering the annotators' characteristics and background for irony analysis. Some demographic dimensions like age and student status particularly induce a strong polarization of annotations. In some languages, such as German, English, and Spanish, we notice a connection between annotators' nationality and their sensitivity to irony across their linguistic varieties. It suggests that annotators with a background that is in line with a specific language variety tend to have higher agreement.

Furthermore, we showcase the utility of MultiPICo as a benchmark for perspectivist modeling of irony through experiments of perspective-taking prompting with LLMs. In particular, we compare prompts without sociodemographic information, with trait-based perspectives, and with data-driven perspective examples, showing that it is possible to evaluate models in a perspectivist setting. Moreover, MultiPICo allowed us to perform a positionality analysis of ChatGPT, discovering that tends to be aligned with the youngest annotators.

In the future, we plan to expand the analysis of linguistic patterns that characterize irony across languages and language varieties, looking also at the different types of irony (i.e., situational and verbal irony) involved in the Post-Reply pairs. Furthermore, we plan to mine meaningful clusters of annotators based on their annotation rather than on their demographic traits and to analyze whether a meaningful relationship exists between these groups and those based on sociodemographic information.

7 Limitations

While our proposed resource is the first multilingual corpus of irony distributed with disaggregated annotations, we acknowledge that the choice of languages is still limited and somewhat arbitrary. The sociodemographic information about the annotators is also partial, bound to what was available from the crowdsourcing platform, and following a discretization of human personal traits that could be perceived as forces (e.g., representing self-identified gender as a single binary label).

Similarly to Sachdeva et al. (2022); Sap et al. (2022); Forbes et al. (2020), we noticed the ethnicity of annotators was unbalanced. We point out this issue to highlight the limitation of paid crowdsourcing (Santy et al., 2023). Furthermore, as shown by Orlikowski et al. (2023), annotators' sociodemographics do not always align with the most relevant grouping of annotators according to the language phenomenon under study. However, approaches based on mining perspectives (Lo and Basile, 2023), as opposed to strictly categorizing annotators, may alleviate this issue.

In the vast majority ($\sim$90%) of cases, we downloaded the conversation-starting messages and their direct replies to capture the full conversational context. In a few cases, the downloaded reply was not direct but rather a second-level reply (a reply to a direct reply); thus, some conversational context might be missing.

7.1 Ethical considerations

Our work stresses the necessity to consider and include 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 corpus, moreover, can be used as a starting point also for investigating and evaluating LLMs across a multilingual spectrum, to make them apt to final users.

For building the proposed resource, we adopted measures to protect the privacy of annotators, and our data handling protocols are designed to safeguard personal information (like anonymization of users' mentions). Although our attention during the collection of data was focused on ironic content spread online, we acknowledge that some of the material to annotate could contain racist, sexist, stereotypical, violent, or generally disturbing content.

Regarding the annotation process, we aimed to pay annotators fairly, estimating an average rate of 9£ per hour. Moreover, we tried to balance annotators through their nationality and self-identified gender. However, we are aware that considering gender in a binary form is limited. We plan to adopt a more inclusive approach toward non-binary annotators in future work.

Some sociodemographic information had not been managed by us before the annotation process, and, when examining them in this work, we noticed a substantial unbalance for some dimensions, like the self-identified ethnicities. This pattern suggests the need to interact differently with annotators or social communities if we want a diversity of annotators and perspectives in terms of social background.

Acknowledgements

This work was funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Alexa. The work of Cristina Bosco is also supported by Compagnia di San Paolo - Bando ex-post 2020 - SteroetypHate. The work of Valerio Basile is also supported by Compagnia di San Paolo - Bando ex-post 2020 - "Toxic Language Understanding in Online Communication - BREAKhateDOWN".
We thank all the people involved in providing and checking the prompt translations.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 – Advances in Artificial Intelligence*, pages 588–603, Cham. Springer International Publishing.

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*.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

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, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

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

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

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

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.

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

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 1236–1270.

Virginia Dignum. 2023. Responsible Artificial Intelligence: Recommendations and Lessons Learned. In *Responsible AI in Africa: Challenges and Opportunities*, pages 195–214. Springer International Publishing Cham.

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

Simona Frenda, Soda Marem Lo, Silvia Casola, Bianca Scarlini, Cristina Marco, Valerio Basile, and Davide Bernardi. 2023a. Does anyone see the irony here? Analysis of perspective-aware model predictions in irony detection. In *ECAI 2023 Workshop on Perspectivist Approaches to NLP*.

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. 2023b. 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.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 10528–10539.

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

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models.

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 colocated with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023*, volume 3494 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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, 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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

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

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*, page 133–138.

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.

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

Maarten Sap, 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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

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.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021b. 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. 2021c. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence - AAAI Special Track on AI for Social Impact*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. PolyLM: An open source polyglot large language model.

A Appendix

A.1 Annotation process

Instructions provided to the annotators before starting the annotation task:

> **Is it ironic?** In this study, we ask the participants to read a message and a reply, and judge **if the reply is ironic**.
>
> **Irony** is a figurative language device that conveys the opposite of literal meaning, profiling intentionally a secondary or extended meaning.
>
> For instance:
>
> `message`: *If ur homeless u probably wouldn't have a phone.*
>
> `reply`: *Yes, and all your belongings would be in a handkerchief tied at the end of a stick.* –> `irony: yes`
>
> `message`: *If ur homeless u probably wouldn't have a phone.*
>
> `reply`: *Yes, you're right.* –> `irony: no`
>
> This annotation consists of 200 small conversations of tweets and Reddit's posts and will take more or less 2 hours. There are no requirements for taking part in this study, simply annotate the presence of irony as you perceive it. Thank you for your interest in this research!

Example of task:

> *Message* This man is so completely focused towards engineering riots.
> *Reply* Gotta stick to your strengths
> Is the **reply** ironic?
> Ironic or Not ironic

A.2 Sociodemographic information

In Table 9, we report the characteristics of each sociodemographic cohort.

In Tables 10 and 8, we report the details of the sets of data and the annotators' sociodemographic information per language, respectively.

B Appendix

B.1 Prompt example

In this section, we report the prompts written by native speakers and used for our experiments.

> *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 "irony" or "not irony".*
>
> *Input:*
>
> *- Post: Because the last generation was Gen. Z and there's no where to go past that, so we gotta start from the beginning.*
>
> *- Reply: but we should have just named the first generation "Alpha" instead of doing it now*
>
> *Output:*

Base prompt for English

> *Instruction: You are British. 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 "irony" or "not irony".*
>
> *Input:*
>
> *- Post: Because the last generation was Gen. Z and there's no where to go past that, so we gotta start from the beginning.*
>
> *- Reply: but we should have just named the first generation "Alpha" instead of doing it now*
>
> *Output:*

Demographics		Languages								
		English	Spanish	Italian	French	Dutch	German	Hindi	Arabic	Portoguese
Age group	Boomer	3	2	–	2	–	5	–	1	–
	GenX	22	17	1	7	4	7	3	4	1
	GenY	38	66	15	23	10	36	13	36	23
	GenZ	10	37	8	17	11	20	8	26	25
Ethnicity	White	47	60	23	40	22	66	–	20	37
	Mixed	1	31	1	3	2	3	–	13	10
	Asian	18	1	–	1	1	–	22	1	–
	Black	3	2	–	5	–	–	–	2	1
	Other	3	27	–	1	–	1	8	31	1
Student	Yes	13	39	14	16	7	14	8	29	30
	No	46	60	9	30	16	39	14	25	16
Employment	Full-time	25	41	9	24	10	24	10	20	15
	Unemployed	11	24	7	5	4	3	1	11	8
	Part-time	11	17	5	5	3	10	4	13	6
	Not in paid work	4	4	1	5	4	5	–	1	–
	Due to start	–	3	1	1	–	2	2	–	2
	Other	1	6		6	–	3	1	5	14

Table 8: Sociodemographic information about annotators per language.

Female	Annotators who self-identify as female
Male	Annotators who self-identify as male
Boomers	Annotators whose age is ≥ 58
GenX	Annotators whose age is ≥ 42 and < 58
GenY	Annotators whose age is ≥ 26 and < 42
GenZ	Annotators whose age is ≤ 26
Old	Annotators whose age is ≥ 42
Young	Annotators whose age is < 42
Student status: yes	Annotators who self-declare to be students
Student status: no	Annotators who self-declare not to be students
Employment status: yes	Annotators who are in paid work (part-time, full-time)
Employment status: no	Annotators who are not in paid work (unemployed, not in paid work, due to start)

Table 9: Characteristics of each demographic cohort

	Varieties	Annotations			Subreddits
		#Reddit	#Twitter	Total	
en	Australian	1,403	1,384	2,787	r/australia
	British	1,406	1,439	2,845	r/CasualUK, r/britishproblems
	Irish	1,432	1,409	2,841	r/ireland
	Indian	1,426	1,429	2,855	r/india
	US English	1,443	1,400	2,843	r/AskReddit
es	Argentinean	2,399	2,407	4,806	r/argentina
	Colombian	2,388	2,417	4,805	r/Colombia
	Spanish	2,402	2,402	4,804	r/spain
	Mexican	2,403	2,409	4,812	r/mexico
	US Spanish	1,323	3,486	4,809	r/estadosunidos
it	Italian	2,400	2,390	4,790	r/italy
fr	Canadian	2,488	1,300	3,788	r/Quebec
	French	2,487	2,495	4,982	r/france
nl	Dutch	2,500	2,491	4,991	r/nederlands
de	Austrian	2,295	2,312	4,607	r/Austria
	Swiss	363	2,870	3,233	r/schwiiz
	German	2,335	2,335	4,670	r/de
hi	Indian	1,700	3,011	4,711	r/Hindi
ar	Egyptian	1,217	1,239	2,456	r/Egypt
	Iraqi	443	2,197	2,640	r/Iraq
	Moroccan	1,587	1,043	2,630	r/Morocco
	Saudi Arabian	1,216	1,239	2,455	r/saudiarabia
	Yemen	15	413	428	r/Yemen
pt	Brazilian	2,454	2,427	4,881	r/brasil
	Portuguese	2,439	2,434	4,873	r/portugal

Table 10: Varieties for each language, number of annotated text per language variety, subreddits.

Prompt-based perspective-taking prompt for English, for the British perspective.

Instruction: You are British. 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 "irony" or "not irony".

Example 1:

Input:

- Post: I went there about 10 years ago. Costs about £20 to go in and it's just a greenhouse.

- Reply: You go to look at the plants not the greenhouse. It's like saying a restaurant is just a bunch of tables.

Output: irony

Example 2:

Input:

- Post: casually joins you. Every time I log out of my bank they're there. And ad with them in is on repeat in my local branch. How I'm a Zelebrity is still going after all these years is beyond me. And please, God, why do people pretend they can't tell them apart?

- Reply: I think I'd have to switch banks

Output: irony

Example 3:

Input:

	Generation				Age		Gender		Nationality					Working		Studying	
	Boomer	GenX	GenY	GenZ	Young	Old	Male	Female						Y	N	Y	N
AR	.345	.362	.384	.271	.398	.364	.372	.384						.378	.333	.326	.384
NL	-	.045	.136	.149	.172	.045	.144	.131						.182	.124	.129	.178
EN	.237	.207	.240	.234	.272	.219	.259	.223	UK .249	IR .203	US .218	AU .215	HI .196	.235	.237	.239	.247
FR	.162	.277	.267	.278	.324	.260	.291	.293	FR .308	CA .283				.320	.248	.298	.297
DE	.278	.337	.345	.272	.377	.332	.337	.353	DE .319	AU .299	CH .336			.345	.322	.319	.354
HI	-	.263	.204	.168	.252	.263	.166	.258						.237	.223	.137	.292
IT	-	.380	.359	.259	.415	.380	.350	.372						.389	.306	.365	.374
ES	.208	.181	.282	.235	.317	.190	.267	.277	ES .261	US .190	MX .242	AR .229	CO .226	.265	.255	.269	.256

Table 11: ChatGPT positionality.

- Post: Can we just stop with this now?

- Reply: I suspect it's coming to an end....

Output: irony

Example 4:

Input:

- Post: We just outraised Greg Abbott — again. Now we're going to defeat him. That's how we overcome his extremism and move Texas forward.

- Reply: Money won't get you elected

Output: not irony

Example 5:

Input:

- Post: When you're young, work to learn don't work to earn. You should prioritise study over work. Go full time uni and part time work.

- Reply: >work to learn don't work to earn

Output: not irony

Example 6:

Input:

- Post: How none of my reddit posts ever get guilded.

- Reply: I'd guild you if I wasn't poor, my friend

Output: not irony

Example to label:

- Post: Because the last generation was Gen. Z and there's no where to go past that, so we gotta start from the beginning.

- Reply: but we should have just named the first generation "Alpha" instead of doing it now

Your output:

Data-driven perspective-taking prompt for English, for the British perspective.

B.2 Computational resources

Baseline performance (Sections 5 and B.3) ChatGPT (gpt-3.5-turbo) has a cost of $0.001/1000 tokens. We run PolyLM-13B on 5 v100 GPUs with 16GB VRAM; the experiment took approximately 7 hours. We ran the BLOOM zero-shot experiments on two NVIDIA-A40 GPUs; experiments were completed in around 200 hours. The long processing time for BLOOM is likely because the model tends to generate unnecessarily long outputs (e.g., repeating the prompt) rather than the actual label.

Perspective-taking prompts We run PolyLM-13B on 5 v100 GPUs with 16GB VRAM; the experiments took approximately 200 hours in total.

B.3 ChatGPT positionality

Table 11 reports the results for ChatGPT's positionality.