Annotating Dimensions of Social Perception in Text: A Sentence-Level Dataset of Warmth and Competence

Mutaz Ayesh[1], **Saif M. Mohammad**[2], **Nedjma Ousidhoum**[1]
[1] Cardiff University, [2] National Research Council Canada
Correspondence: OusidhoumN@cardiff.ac.uk

Abstract

Warmth (W) (often further broken down into *Trust* (T) and *Sociability* (S)) and *Competence* (C) are central dimensions along which people evaluate individuals and social groups (Fiske, 2018). While these constructs are well established in social psychology, they are only starting to get attention in NLP research through word-level lexicons, which do not fully capture their contextual expression in larger text units and discourse. In this work, we introduce *Warmth and Competence Sentences (W&C-Sent)*, the first sentence-level dataset annotated for warmth and competence. The dataset includes over 1,600 English sentence–target pairs annotated along three dimensions: *trust* and *sociability* (components of *warmth*), and *competence*[1]. The sentences in W&C-Sent are social media posts that express attitudes and opinions about specific individuals or social groups (the targets of our annotations). We describe the data collection, annotation, and quality-control procedures in detail, and evaluate a range of large language models (LLMs) on their ability to identify trust, sociability, and competence in text. W&C-Sent provides a new resource for analyzing warmth and competence in language and supports future research at the intersection of NLP and computational social science.

Content warning: This paper contains potentially offensive examples.

1 Introduction

Language is a powerful medium through which people express their emotions and opinions about topics and individuals. In social psychology, interpersonal evaluation is largely structured around warmth and competence, the two primary dimensions along which people form impressions of, and make judgments about, individuals and social groups (Fiske et al., 2002; Fiske, 2018). To explain how such evaluations manifest, the Stereotype Content Model (SCM), proposed by Fiske et al. (2002), examines social perception, bias, and stereotyping. Central to the SCM is the idea that humans rapidly infer whether others are well-intentioned and socially oriented—even in the case of strangers—before assessing their ability to act on those intentions. Accordingly, **warmth (W)** reflects perceptions of **trust (T)** and **sociability (S)**, capturing whether an individual is seen as benevolent, cooperative, and inclined toward positive social interaction. In contrast, **competence (C)** reflects perceived capability to act on those intentions. These dimensions are fundamental to the study of social interaction, emotional responses, stereotypes, and human behavior more broadly. As such, understanding them is crucial for building human-centered NLP and AI systems.

However, while the study of warmth and competence is well established in psychology, they have only recently begun to receive attention in NLP. For example, Mohammad (2025b) introduced a lexicon of over 42k words and multiword expressions associated with warmth, sociability, trust, and competence. Lexical approaches to language analysis are simple yet powerful, and are especially useful for identifying aggregate trends (e.g., across large collections of social media posts over time) (Teodorescu and Mohammad, 2023). Nevertheless, for applications that require instance-level understanding (e.g., determining the warmth communicated in a sentence), machine learning approaches trained on sentence-level annotated datasets are markedly more accurate. One reason is that meaning does not correspond to the simple sum of individual words' senses (Szabó, 2024). Instead, expressions of trust, sociability, and competence are often shaped by syntax, compositional semantics, and pragmatic cues, which cannot be fully captured at the word level. Consequently, word-based resources provide

[1]Available at https://github.com/nedjmaou/W_C_Sent

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6374–6412
July 2-7, 2026

limited insight into how these social traits are conveyed in context.

To address this gap, we build W&C-Sent—a dataset designed to capture contextualized expressions of trust, sociability, and competence toward specific targets at the sentence level. We collect instances from the SemEval-2016 stance dataset (Mohammad et al., 2016)—chosen for its focus on opinions about specific targets—and augment them with instances from the Affect, Body, Cognition, Demographic, and Emotion (ABCDE) dataset (Wahle et al., 2026). Each instance is then independently annotated by 4 to 7 fluent English speakers for each dimension (T, S, and C) with respect to seven distinct targets: individual politicians (Hillary Clinton, Barack Obama, and Donald Trump) and social groups (women, religious people, atheists, and climate change activists). Annotations are recorded on a 7-point scale ranging from -3 (very low) to +3 (very high), with 0 representing neutrality (e.g., see Figure 1).

We provide a detailed description of our data collection, annotation, and quality-control procedures and make the dataset publicly available. Then, we evaluate a range of LLMs on their ability to detect trust, sociability, and competence in text. We find that LLMs struggle to reliably assess these social dimensions, leading to suboptimal performance in downstream applications and their deployment in real-world systems, such as chatbots, content moderation, and machine translation. W&C-Sent supports a broad range of applications in NLP and social science including analytics, annotation, discourse analysis, and bias studies. It also serves as a benchmark for evaluating whether models accurately capture trust, sociability, and competence in language.

Target: Hillary Clinton
Text: Would you wanna be in a long term relationship with some bitch that hides her emails, & lies to your face? Then #Dontvote

High Distrust -3
Moderate Unsociability -2
Moderate Competence +2

Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.

Neutral trust 0
High Unsociability -3
High Incompetence -3

Figure 1: Examples of sentence–target pairs with different annotations across trust, sociability, and competence.

2 Related work

Prior work in social perception has established warmth and competence as core dimensions underlying social judgments. The Stereotype Content Model (Fiske et al., 2002) showed that different combinations of warmth and competence elicit distinct emotional responses (e.g., pity or envy), explaining why groups are associated with varied affective reactions rather than uniform prejudice (Fiske, 2018). Subsequent work by Abele et al. (2016) challenged a unidimensional view of warmth and proposed a multifaceted conceptualization of warmth (communion), distinguishing between morality-related traits (e.g., honesty, fairness) and sociability-related traits (e.g., friendliness, empathy). They further demonstrated that morality plays a stronger role in group judgments. This distinction was further supported by Koch et al. (2024), who showed that these facets cluster differently and predict several behavioral outcomes, motivating more fine-grained modeling of warmth and competence. This framework is particularly helpful for assessing bias and stereotypes in NLP (Fraser et al., 2024), given the growing interest in bias and stereotype evaluation in NLP models (Kiritchenko and Mohammad, 2018; Blodgett et al., 2020; Nadeem et al., 2021; Ousidhoum et al., 2021), and in LLMs in particular (Cheng et al., 2023; Siddique et al., 2024; Plaza-del Arco et al., 2024a,b).

Current computational approaches to warmth and competence are largely lexicon-based. The Valence–Arousal–Dominance (VAD) lexicon (Mohammad, 2018, 2025a) provides large-scale word-level scores for valence—previously argued to be the largest component of warmth (Cuddy et al., 2007)—and dominance (also referred to as competence), offering broad coverage but treating warmth as a single dimension. Other work has developed small dictionaries specifically for warmth and competence text analysis (Nicolas et al., 2021). Mohammad (2025b) introduced the first large-scale lexicon, Words of Warmth, providing reliable human ratings of word–warmth, word–sociability, word–trust, and word–competence associations for over

42k words and multiword expressions.[2] Mohammad (2025b) used the lexicon to study: (1) the rate at which children acquire warmth-, sociability-, and trust-related words in their vocabulary with age; and (2) the average degree of warmth and competence of words that co-occur with mentions of social groups (such as "immigrant", "doctor", "girl", etc.) in social media. The Words of Warmth lexicon provides a foundation for more nuanced modeling of social traits in language and motivates extending such analyses beyond the word level, which we address in this paper.

3 Dataset

We construct a dataset of 1,633 sentences annotated for warmth (W)—broken down into trust (T) and sociability (S)—and competence (C). We describe the data sources, annotation procedures, and quality-control processes in the following.

3.1 Data Sources

Stance Data We use the SemEval-2016 Stance dataset (Mohammad et al., 2016) as the primary source of textual instances (1,490 sentences; 90.8%). It consists of social media posts manually annotated for stance toward predefined targets. Specifically, each post is labeled according to whether the expressed stance toward a given target is *in favor of*, *against*, or *neutral* (more precisely, neither favor nor against stance can be inferred). The text may express stance overtly or implicitly. The dataset also includes labels for the overall sentiment of the post independent of stance. The six predefined targets in the dataset are *Hillary Clinton*, *Donald Trump*, *Atheism*, *Feminism*, *Climate Change*, and *Abortion*.

The ABCDE Dataset We additionally use the ABCDE dataset (Wahle et al., 2026), a large-scale corpus of over 400 million text utterances from diverse sources, including social media, blogs, and books. The dataset contains self-reported features that capture affective, cognitive, linguistic, bodily, and demographic information. We selected 151 sentences from the ABCDE dataset, constituting 9.2% of the total sentences in *W&C-Sent*, using a simple keyword-based search targeting the seven predefined targets. The search keywords can be found in Appendix A.

[2] https://saifmohammad.com/WebPages/warmth.html

Target	SemEval	ABCDE	Total (%)
Hillary Clinton	610	7	617 (37.7)
Donald Trump	301	108	409 (25.0)
Women	292	21	313 (19.0)
Barack Obama	109	4	113 (7.0)
Religious P.	98	7	105 (6.5)
Environ.	37	3	40 (2.5)
Nonreligious P.	35	1	36 (2.3)
Total	**1,482**	**151**	**1,633 (100)**

Table 1: **Distribution of targets in W&C-Sent,** showing frequency counts by source dataset and overall percentages for the seven targets (Hillary Clinton, Donald Trump, women, Barack Obama, religious people (Religious P.), climate change activists (Environ.), and atheists (Non-religious P.).

Data Selection Table 1 presents the final distribution of the dataset. Human targets (e.g., *Hillary Clinton*, *Donald Trump*, and *women*) were prioritized, as the analysis of warmth and competence requires targets to be human individuals or social groups. Non-human or abstract targets—such as *Atheism*, *Feminism*, *Climate Change*, and *Abortion*—were included only when sentences explicitly referred to human subjects or collectives. Specifically, we retained posts discussing these topics and manually reassigned the targets to relevant groups of people rather than the topics themselves (e.g., *atheism* posts referring to religious or non-religious people; *abortion*-related posts referring to women; and *climate change* posts targeting environmentalists or climate change activists) (see Figure 2). Moreover, several instances in the SemEval stance dataset with *Hillary Clinton* or *Donald Trump* as targets contained references to *Barack Obama*. In such cases, we created additional instances by pairing the same sentence with *Barack Obama* as the target, resulting in a final set of seven targets.

3.2 Annotation Process

We preprocessed the data instances by anonymizing social media posts, replacing most `@mentions` with `@user`, except for mentions of public figures (e.g., *Hillary Clinton*). The instances were then labeled using the annotation platform Prolific.

Annotator Selection Criteria Annotators were required to be fluent English speakers based in English-speaking countries and to have an approval rate above 99% on Prolific. They were compensated at a rate of US $16–22 per hour.

To reduce the likelihood of bot participation or inattentive responses, annotators were required to complete multiple attention checks throughout the study in which they were instructed to assign a

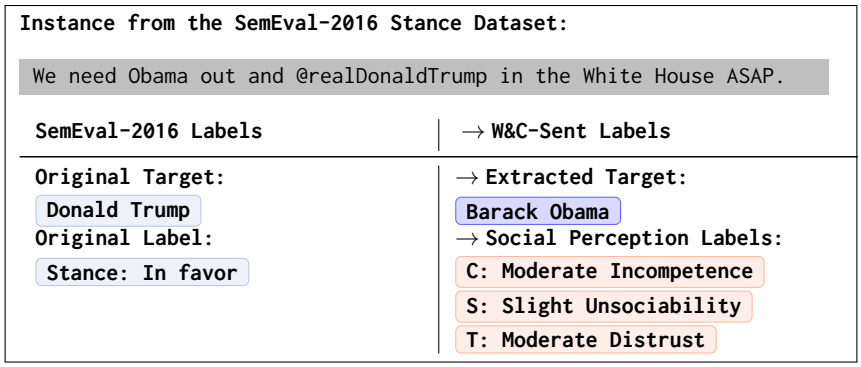


Figure 2: An example illustrating how the SemEval-2016 Stance dataset was used to extract sentence–target pairs for W&C-Sent. Specifically, a social media post is initially labeled as expressing a stance in favor of Donald Trump; however, because it also mentions Barack Obama, we additionally extract Barack Obama as a target and annotate it for competence, sociability, and trust (C, S, and T).

specific, predetermined label (e.g., *"Please assign a +3 score"*). Selecting any other label was treated as a failure to follow the instructions.

In addition, a set of potentially unambiguous sentences ($n = 229$), manually pre-annotated by the authors was used in quality assessment. Annotators were presented with a random subset of these sentences, and if they failed to correctly annotate more than 20% of them, their work was subject to full review and typically discarded.

Instructions To reduce annotator cognitive load, we annotate trust, sociability, and competence independently. We adapted the word-level guidelines proposed by Mohammad (2025b) for sentence–target pair annotations. The full annotation guidelines are provided in Appendix D. Annotators are shown a post together with a target and are asked to rate each of the three dimensions on a 7-point ordinal scale. Scores range from -3—very low trust, very low sociability, or very low competence—to $+3$—very high trust, very high sociability, or very high competence—depending on the dimension being annotated. A score of 0 indicates neutrality, non-applicability, or a lack of (expressed) information.

Agreement and Reliability We compute average split-half reliability (SHR), a commonly used measure for assessing the reliability of ordinal-scale annotations (Kuder and Richardson, 1937; Cronbach, 1951; Weir, 2005). All annotations are randomly split into two halves, and separate aggregate scores are computed for each half. The similarity between the two sets of scores is then measured using a correlation metric. This procedure is repeated 1,000 times, and the resulting correlations are averaged (Mohammad, 2025b). The resulting SHR scores are 0.76 for trust, 0.68 for sociability, and 0.56 for competence, indicating relatively high annotation reliability. For inter-annotator agreement (IAA), we compute Krippendorff's α, and obtain 0.60 for trust, 0.50 for sociability, and 0.30 for competence, which is expected for subjective and fine-grained classification tasks. By contrast, when computing the overall average pairwise agreement (APA) for coarsened labels—mapping $[-3, -0.5]$ to *low*, $[-0.5, 0.5]$ to *neutral*, and $[0.5, 3]$ to *high*—the APA increases substantially, reaching 62.8% for both trust and sociability, and 52.2% for competence.

3.3 Final Dataset

Aggregated Scores and Statistics To obtain a single label per sentence–target pair, we aggregate annotator scores using the mean, following Mohammad (2025b). Figure 3 shows the distribution of mean labels across the three dimensions, while Table 2 reports the distribution of these scores after coarsening them into three categories: Low (negative), Neutral, and High (positive). Further details are provided in Appendix F.

The neutral label is the most frequent overall, particularly for competence. However, the dataset also contains a substantial proportion of non-neutral instances, with negative scores more prevalent than positive ones. This likely reflects the nature of the source data, as the original SemEval instances were drawn from discussions of controversial topics in the USA, such as elections and reproductive rights (Mohammad et al., 2016).

Across dimensions, sociability shows a concentration at -1 and -2 (low to moderate unsociability),

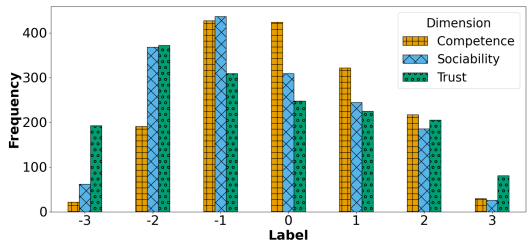


Figure 3: Distribution of mean scores assigned to sentence–target pairs (1,633 pairs per dimension) in the dataset. The x-axis shows fine-grained labels, and the y-axis shows their frequency.

while trust exhibits relatively high frequencies at -2 and especially -3 (moderate to high distrust). In contrast, positive labels (+1 to +3) are less frequent overall, with +3 being the rarest category. For competence and sociability, extreme values (+3 and -3) are similarly uncommon, suggesting that annotators tend to avoid assigning the most extreme ratings.

4 Experiments

4.1 Setup

We split our dataset into 60%/20%/20% for train, development, and test sets, and use our test set to evaluate several automatic baselines on the task of predicting continuous and ordinal trust/sociability/competence (T/S/C) scores (-3 to 3) for a given sentence–target pair:

- a majority-class (*"dummy"*) baseline classifier,
- a logistic regression (LR) classifier,
- a fine-tuned BERTweet model (Nguyen et al., 2020), and
- zero-shot (ZS) and few-shot (FS) experiments with a range of open- and closed-source LLMs.

4.1.1 Baseline Models

To predict continuous trust, sociability, and competence scores—each defined on a scale from -3 to +3—we train majority-vote ("dummy") and logistic regression (LR) models using TF–IDF embeddings. Specifically, we train one model per dimension, resulting in six classifiers in total (i.e., three dummy and three LR models) that serve as baselines.

4.1.2 Fine-tuned BERT

We fine-tune three BERTweet-based classifiers on our training set to predict continuous trust, sociability, and competence scores, each defined on a scale from -3 to +3.

Additionally, we evaluate our baselines in a coarse-grained classification setting, where models predict whether the perceived trust, sociability, and competence of text–target pairs are *"low"*, *"neutral"*, or *"high"*. As in the regression experiments, BERTweet models are fine-tuned separately for each dimension.

Coarse label	Low (%)	Neutral (%)	High (%)
Trust	945 (57.9)	105 (6.4)	583 (35.7)
Sociability	1,012 (62.0)	85 (5.2)	536 (32.8)
Competence	773 (47.3)	162 (9.9)	698 (42.7)

Table 2: Distribution of W&C-Sent's labels after coarsening. The table shows the strong prevalence of negative (low) labels compared to positive and neutral ones.

4.1.3 LLM Prompting

Models We prompt six different LLMs: three open-weight and instruction-tuned (Gemma-3-4b (Gemma Team, 2025), Qwen-2.5-7b (Qwen Team, 2024), and Qwen3-4B-Instruct-2507 (Team, 2025)), and three closed-source models (Open AI's GPT-4o, GPT-4o-mini (OpenAI, 2023), and GPT-5.2 Chat (Singh et al., 2025)). We design three separate prompts—one per dimension—in accordance with our annotation framework. That is, similar to the annotation process, in which each social dimension is labeled independently to reduce annotator's cognitive load, we evaluate each LLM on one dimension at a time. This design choice limits the amount of information included in each prompt and helps mitigate potential underperformance due to prompt overload (Liu et al., 2023). In addition, we prompted the models to assess sentence–target pairs using both fine-grained labels and coarse-grained labels, and in zero-shot (ZS) and few-shot (FS) settings.

Prompts Each prompt starts with a definition of the targeted social dimension (i.e., sociability, trust, or competence). It then establishes analytical constraints by instructing the model to interpret the text snippet as an average person would and to evaluate the social dimension at the sentence level. The prompts also include definitions of pragmatic phenomena common in social media—such as sarcasm, hyperbole, and irony—as disclaimers to mitigate errors observed in pilot experiments, in which LLMs interpreted sarcastic or ironic sentences literally. The target entity is explicitly specified, and the model is instructed to evaluate each sentence with respect to that entity and its definitions, while ignoring references to other individuals or social groups, if present.

Dimension	Model	Accuracy	F1	±1 Acc.
Trust	Dummy	0.22	0.08	0.58
	LR	0.30	0.30	0.65
	Fine-tuned BERTweet	0.35	0.31	0.83
	Gemma3 ZS	0.36	0.34	0.78
	Gemma3 FS	0.38	0.33	0.79
	Qwen2.5 ZS	0.27	0.25	0.72
	Qwen2.5 FS	0.35	0.35	0.76
	Qwen3 ZS	0.32	0.30	0.78
	Qwen3 FS	0.30	0.26	0.78
	GPT-4o ZS	**0.43**	**0.42**	**0.91**
	GPT-4o FS	0.39	0.40	0.84
	GPT-4o-mini ZS	0.27	0.26	0.60
	GPT-4o-mini FS	0.20	0.17	0.54
	GPT-5.2 ZS	0.41	0.38	0.87
	GPT-5.2 FS	0.40	0.39	0.89
Sociability	Dummy	0.26	0.11	0.67
	LR	0.31	0.26	0.70
	Fine-tuned BERTweet	**0.46**	0.34	0.88
	Gemma3 ZS	0.34	0.27	0.82
	Gemma3 FS	0.31	0.23	0.76
	Qwen2.5 ZS	0.20	0.20	0.69
	Qwen2.5 FS	0.25	0.25	0.67
	Qwen3 ZS	0.35	0.30	0.81
	Qwen3 FS	0.31	0.24	0.82
	GPT-4o ZS	0.44	**0.40**	**0.92**
	GPT-4o FS	0.38	0.35	0.87
	GPT-4o-mini ZS	0.13	0.14	0.52
	GPT-4o-mini FS	0.24	0.20	0.59
	GPT-5.2 ZS	0.31	0.30	0.83
	GPT-5.2 FS	0.40	0.37	0.88
Competence	Dummy	0.24	0.09	0.60
	LR	0.19	0.16	0.62
	Fine-tuned BERTweet	**0.36**	0.24	**0.86**
	Gemma3 ZS	0.22	0.19	0.58
	Gemma3 FS	0.35	0.27	0.70
	Qwen2.5 ZS	0.22	0.19	0.59
	Qwen2.5 FS	0.30	0.28	0.65
	Qwen3 ZS	0.22	0.18	0.60
	Qwen3 FS	0.25	0.21	0.67
	GPT-4o ZS	0.34	**0.31**	0.80
	GPT-4o FS	0.24	0.24	0.64
	GPT-4o-mini ZS	0.09	0.11	0.37
	GPT-4o-mini FS	0.22	0.20	0.60
	GPT-5.2 ZS	0.28	0.25	0.73
	GPT-5.2 FS	0.27	0.24	0.71

Table 3: Fine-grained classification performance of different baselines—majority-class (*dummy*), logistic regression (*LR*), and fine-tuned BERTweet—along with LLM models in zero-shot (ZS) and few-shot (FS) settings for **trust**, **sociability**, and **competence**. Scores are reported on the test set. The **best score** for each metric and dimension is shown in **bold** and highlighted in blue, and the second-best is highlighted in a lighter shade of blue.

For each dimension, the instructions closely follow the annotation guidelines provided to human annotators. This section specifies the label set that the model is expected to use, ensuring that categorical labels are returned in both fine-grained settings (e.g., "high distrust", "moderate sociability", "slight incompetence") and coarse-grained ones ("low", "neutral", or "high"). In zero-shot prompts, examples are omitted, whereas few-shot prompts include examples covering all labels (8–9 examples for fine-grained prompts and 5 for coarse-grained prompts). To activate the models' reasoning capabilities, the prompts instruct the LLMs to reason first and then return the label. All prompts are provided in Appendix N.

4.2 Experimental Results

Tables 3 and 4 present the performance of all fine-grained and coarse-grained classification models, respectively. We report **accuracy** and the **weighted F1 score**, as well as the **within-1** (± 1) **accuracy** for the fine-grained classifiers. Given the **ordinal nature** of the task, ± 1 accuracy indicates how often model predictions fall within one rating level of the true labels, thereby capturing near-miss performance. Precision and recall are reported in the Appendix (Tables 17 and 18).

Fine-grained Classification Results Table 3 presents the performance of the different baselines on the fine-grained classification task. As expected, the majority-class (dummy) classifier performs the worst, followed by logistic regression (LR). However, LLMs do not perform substantially better in either zero-shot or few-shot settings, with no model exceeding 50% accuracy. This is particularly noticeable for the competence dimension, which appears to be the most challenging—even for human annotators.

In terms of accuracy, fine-tuned BERTweet performs best on sociability and competence, followed by GPT-4o (ZS) on sociability and Gemma3 (FS) on competence. GPT-4o (ZS) achieves the highest accuracy on trust, followed by GPT-5.2. In terms of weighted F1, GPT-4o (ZS) outperforms all other models, with its few-shot counterpart close behind, highlighting the difficulty of the task and the test set for LLMs. Gemma3 outperforms the GPT and Qwen models in the few-shot setting on competence, achieving an accuracy of 0.35. In contrast, Qwen2.5 and GPT-4-mini sometimes fail to outperform the dummy classifier or the LR model, indicating their limitations for this task.

The ± 1 accuracy scores suggest that GPT-4o (ZS) tends to make more subtle errors than other models, achieving 0.91 and 0.92 on trust and sociability, respectively. It is outperformed only by the fine-tuned BERTweet model on competence (0.86), where it remains a close second (0.80). Overall, larger and more recent closed models (e.g., GPT-5.2) do not necessarily outperform older ones.

Coarse-grained Classification Results The results for predicting coarse-grained labels, shown in Table 4, are higher than those for the fine-grained

Dimension	Model	Accuracy	F1
Trust	Fine-tuned BERTweet	0.79	0.59
	Gemma3 ZS	0.74	0.60
	Gemma3 FS	0.72	0.60
	Qwen2.5 ZS	0.62	0.53
	Qwen2.5 FS	0.61	0.51
	Qwen3 ZS	0.79	0.66
	Qwen3 FS	0.72	0.62
	GPT-4o ZS	**0.86**	**0.78**
	GPT-4o FS	0.77	0.71
	GPT-4o-mini ZS	0.78	0.65
	GPT-4o-mini FS	0.70	0.58
	GPT-5.2 ZS	0.81	0.72
	GPT-5.2 FS	0.78	0.71
Sociability	Fine-tuned BERTweet	0.81	0.54
	Gemma3 ZS	0.69	0.56
	Gemma3 FS	0.71	0.54
	Qwen2.5 ZS	0.47	0.42
	Qwen2.5 FS	0.34	0.34
	Qwen3 ZS	0.74	0.59
	Qwen3 FS	0.67	0.58
	GPT-4o ZS	0.80	0.71
	GPT-4o FS	0.72	0.66
	GPT-4o-mini ZS	0.74	0.54
	GPT-4o-mini FS	0.62	0.47
	GPT-5.2 ZS	**0.83**	**0.74**
	GPT-5.2 FS	0.79	0.72
Competence	Fine-tuned BERTweet	**0.74**	0.50
	Gemma3 ZS	0.52	0.46
	Gemma3 FS	0.52	0.43
	Qwen2.5 ZS	0.47	0.47
	Qwen2.5 FS	0.56	0.53
	Qwen3 ZS	0.64	0.56
	Qwen3 FS	0.67	0.58
	GPT-4o ZS	0.63	**0.59**
	GPT-4o FS	0.60	0.57
	GPT-4o-mini ZS	0.53	0.40
	GPT-4o-mini FS	0.47	0.37
	GPT-5.2 ZS	0.49	0.46
	GPT-5.2 FS	0.46	0.43

Table 4: Coarse-grained classification performance of fine-tuned BERT and LLM models in zero-shot (ZS) and few-shot (FS) settings for **trust**, **sociability**, and **competence**. Scores are reported on the test set. The **best score** for each metric and dimension is in **bold** and highlighted in blue, and the second-best is highlighted in a lighter shade of blue.

setting, with competence remaining the most challenging dimension.

For the trust dimension, GPT-4o (ZS) achieves the highest accuracy and F1 scores, followed by GPT-5.2. The remaining models—except for Gemma3—perform similarly, with accuracy scores in the 70–79% range. For sociability, GPT-5.2 (ZS) outperforms all other models on both metrics, followed by the fine-tuned BERTweet model in accuracy and GPT-5.2 (FS) in F1. Finally, for competence, the fine-tuned BERTweet model achieves the highest accuracy, while GPT-4o (ZS) attains the highest F1 score; Qwen3 is a close second on both metrics.

We further observe that, although performance improves considerably when labels are coarsened—similar to ± 1 accuracy in the fine-grained setting (see Table 3)—the fine-tuned BERTweet model continues to perform consistently, whereas models such as Gemma3 and Qwen2.5 may still perform poorly. Notably, GPT-5.2 performs comparatively poorly in both zero-shot and few-shot settings on the coarsened labels, particularly for the competence dimension, despite its recency and scale. This finding suggests that fine-tuning a pretrained model on high-quality data can sometimes be more effective than relying on an LLM for challenging and subjective tasks.

Few-shot vs. Zero-shot Performance We observe that few-shot (FS) settings consistently underperform zero-shot (ZS) settings in this task. This pattern appears in Table 3 across several ZS–FS pairs for trust (GPT-4o, GPT-4o-mini, GPT-5.2, and Qwen3), sociability (Gemma3, GPT-4o, Qwen3), and competence (GPT-4o, GPT-5.2). The coarse-grained results in Table 4 show even more noticeable declines, particularly for trust in the Qwen3, GPT-4o, and GPT-4o-mini pairs, and for sociability in the GPT-4o, GPT-4o-mini, Qwen2.5, and Qwen3 pairs.

We analyze 102 instances in which FS predictions from both GPT-4o and Qwen2.5 substantially deviated from both gold labels and ZS outputs. Our findings suggest that models prompted in few-shot settings may amplify surface-level cues—such as all-caps text and exclamation marks. This effect is particularly evident in the frequent default to neutral predictions (32 of 39 cases for GPT-4o and 63 cases for Qwen2.5). In addition, FS settings often misinterpret sarcasm or figurative language as literal meaning, by assigning neutral labels to ironic posts or over-penalizing sarcastic expressions (e.g., references to *Hillary Clinton's "sniper fire"*). Overall, these results suggest that performance degradation in few-shot settings may be due to overly conservative calibration boundaries (i.e., calibration bias), therefore reducing sensitivity to pragmatic nuances.

5 Analysis

To what extent do stance and sentiment correlate with perceptions of warmth and competence? We aim to understand how trust, sociability, and competence relate to stance and sentiment. That is, how people perceive a target individual or social group when they are in favor of or against

Stance	Count	Trust	Sociability	Competence
Against	561	-2 (-1.4 ± 1.4)	-1 (-1.2 ± 1.1)	-1 (-0.5 ± 1.2)
Favor	262	2 (1.7 ± 0.9)	1 (1.2 ± 0.9)	2 (1.4 ± 0.8)
Neutral	31	0 (-0.4 ± 0.7)	0 (-0.2 ± 0.9)	0 (-0.3 ± 0.7)

Table 5: Aggregated median values (mean ± standard deviation) for each stance subset from the SemEval-2016 stance dataset across the three dimensions. These statistics are computed over the 854 text–target pairs shared between W&C-Sent and the SemEval-2016 Stance dataset.

	Trust	Sociability	Competence
Stance	0.71	0.69	0.61
Sentiment	0.72	0.73	0.61

Table 6: Spearman correlation coefficients between stance or sentiment and the median scores of trust, sociability, and competence. The results show strong positive monotonic associations, meaning that as stance or sentiment becomes more positive, the corresponding dimension scores also tend to increase.

a topic, and when they express positive or negative sentiment in language. As W&C-Sent was primarily constructed using a stance dataset, 854 sentences (52.3%) share targets with the SemEval-2016 Stance dataset. Investigating the relationship between each dimension and stance allows us to determine which traits most strongly influence stance, the role each dimension plays, and how stance is reflected through warmth (i.e., trust and sociability) and competence in language.

Table 5 shows how stance relates to assessments of trust, sociability, and competence. Sentences labeled as *against* consistently receive negative median scores across all dimensions, with particularly low trust (-2). By contrast, sentences labeled as *in favor* show positive median scores across all dimensions, including strong trust and competence scores (2), while *neutral* sentences remain near zero with comparatively low variance. This pattern indicates that stance strongly conditions how targets are perceived, both in the direction and intensity of judgment.

Table 6 provides further evidence, as the correlation coefficients indicate strong positive monotonic associations between stance and each dimension. The strongest correlation is observed for trust ($\rho = 0.71$), followed by sociability ($\rho = 0.69$) and competence ($\rho = 0.61$). The SemEval-2016 stance dataset was also annotated for sentiment, which follows a similar trend, with slightly higher correlation coefficients for trust and sociability.

Target	Text	Dim.	Label
Women	@readyforHRC @HillaryClinton #HillaryClinton, the US presidency is a testament to the success of #women their role in the world	C	+3
Barack Obama	I hate Hillary Clinton and Obama, please go die together thrown into the ocean get ripped into prices by sharks.	S	-3
Donald Trump	,@realDonaldTrump trumps presidential dreams r about as real as KimJonguns unicorns.	C	-3

Table 7: Examples of W&C-Sent instances with full agreement on fine-grained labels across different dimensions (Dim.).

When do annotators show high agreement or marked disagreement? Annotators assigned the same fine-grained label in only 3% of the dataset instances ($n = 143$), whereas agreement on the same polarity (low, neutral, or high) occurred in approximately 26.9% of cases ($n = 1,316$ instances), with the majority reflecting negative judgments (see examples in Table 7). Interestingly, we notice that agreement is higher for assessments of trust toward individual targets (e.g., Donald Trump and Barack Obama), while it is higher for assessments of sociability toward social groups (e.g., women and religious people). This might be linked to the nature of each dimension, as trust tends to relate to the personal and moral aspects of a given target, whereas sociability reflects its relational and social characteristics.

At the same time, we observe meaningful disagreement across dimensions on 159 sentence–target pairs whose aggregated scores show different signs across dimensions. Table 8 highlights a subset of these cases, in which negative trust and sociability scores strongly contrast with positive competence scores. Such instances demonstrate notable adherence to the annotation guidelines: even when a target is described in harsh or derogatory terms, annotators consistently consider the ability, skills, or power that the author seems to attribute to the target.

To what extent do trust, sociability, and competence correlate with each other? It is worth noting that largely distinct sets of annotators participated in labeling each dimension. Therefore, when calculating correlations between the aggregated scores, the resulting values reflect relationships between the dimensions at a broad, population-

Target	Sentence	T	S	C
Hillary Clinton	.@HillaryClinton blames her lack of trust among the populace on @GOP, forgetting that she's a lying, conniving, murderous, cheat.	-3	-3	+1
Donald Trump	If there was such a thing as the "American" Hitler. Donald Trump would probably be that guy. #nationalism	-3	-3	+2

Table 8: Examples of sentence–target pairs with diverging aggregated scores for trust (T) and sociability (S) on one hand, and competence (C) on the other. The contrast illustrates cases where a single sentence expresses both moral or social condemnation (i.e., low T and S scores) and an acknowledgment of ability (i.e., medium to high C scores).

level perceptual scale, rather than consistency in the judgments of individual annotators. In other words, these correlations capture general patterns in how people perceive trust, sociability, and competence, rather than the personal biases or preferences of the annotators.

Table 9 presents the correlations between each pair of dimensions using Spearman's ρ. The coefficients show moderately to strongly positive correlations among all three dimensions. Trust correlates most strongly with sociability ($\rho = 0.79$), which is expected, as both trust and sociability are commonly considered sub-dimensions of warmth. Trust correlates somewhat less strongly with competence ($\rho = 0.67$), while sociability and competence exhibit a similar level of correlation ($\rho = 0.68$). Overall, these values indicate that the three dimensions (T, S, and C) are interrelated in human perception. For instance, a person perceived as trustworthy is often also viewed as sociable to some degree, and individuals perceived as incompetent are frequently also seen as untrustworthy, and vice versa.

The strength of the correlation scores presented in Table 6 is also observed in the distributions of co-occurring scores shown in Appendix H. Across all grids, neutral judgments co-occur most frequently, indicating in particular that neutral competence is typically paired with neutral trust and sociability. The trust–sociability grid exhibits the most compact diagonal structure, with very few opposite-sign judgments, confirming that these traits are closely intertwined in practice. By contrast, the trust–competence and sociability–competence grids display greater dispersion, which helps explain their weaker correlations and allows for cases in which competence coexists with distrust or unsociability. These less compact diagonals also shed light on instances of divergent warmth and competence judgments discussed in Table 8.

Pair of Dimensions	Spearman ρ
Trust and Sociability	0.79
Sociability and Competence	0.68
Trust and Competence	0.67

Table 9: Spearman correlation coefficients between the annotated social dimensions. All correlations are positive and relatively strong.

6 Conclusion

We introduced W&C-Sent, the first sentence-level dataset annotated for social perception in text, with over 1,600 instances covering seven target individuals and social groups. Each instance is annotated by fluent English speakers for perceptions of trust, sociability, and competence. Our experiments demonstrate the utility of our dataset and show that models struggle to capture subtle social cues.

We make W&C-Sent publicly available to support further research on how language conveys social perceptions, enabling applications such as text analytics, bias and stereotype evaluation, and socially aware NLP systems. W&C-Sent also serves as a benchmark for assessing whether NLP models and LLMs accurately capture nuanced social traits.

Limitations

Our dataset focuses on English and relies on human annotations. While we collected responses from a substantial number of annotators ($n = 210$), most come from English-speaking regions commonly referred to as the "Global North". Social perceptions are inherently culture-specific and subjective, even within the same cultural context. Hence, our annotations reflect the annotators' lived experiences and intuitions regarding warmth and competence. We do not claim that they capture all possible perceptions; rather, they provide a well-defined, high-quality baseline for English-speaking populations. By releasing W&C-Sent along with individual annotations, we provide opportunities for future work to explore how annotators perceive socio-linguistic cues. Systematic investigations of scale design could inform frameworks for nuanced social perception tasks.

W&C-Sent mostly leverages the publicly available SemEval-2016 Stance dataset to identify sentences suitable for assessing warmth and competence. We acknowledge that the dataset primarily reflects populations with internet access and the technological means to express opinions online, which may not cover all socioeconomic contexts. Although we limited ourselves to targets present in instances of the SemEval-2016 dataset, our methodology is generalizable and can be applied to build larger resources with additional targets and scenarios. We encourage NLP researchers to extend W&C-Sent to new targets and languages and make our guidelines and resources publicly available to facilitate this.

Finally, future research could build on our experiments by evaluating the performance of additional LLMs on trust, sociability, and competence. Our study was limited to a subset of models due to space and cost considerations.

Ethical Considerations

Our study was conducted with approval from our Institutional Review Board (IRB) to ensure adherence to ethical research standards. All annotators were compensated fairly, at rates between 16 and 22 USD per hour—well above the minimum wage—and were provided with clear guidelines and protections. Annotators were recruited via Prolific, remained anonymous, and were informed that they could withdraw from the study at any time, receiving warnings and instructions as specified in the guidelines.

We acknowledge that social perception annotations are inherently subjective and culture-specific, as discussed in our limitations. Therefore, interpreting or generalizing results within and beyond the English-speaking populations represented in our dataset must be conducted with caution. We encourage responsible use and careful contextualization of findings in any downstream applications, and ethical reflection is strongly recommended before using our dataset. Use of the data for commercial purposes or by state actors in high-risk applications is strictly prohibited unless explicitly approved by the dataset creators. Systems developed using our datasets may not be reliable at the individual instance level and are sensitive to domain shifts. They should not be used to make critical decisions about individuals, such as in health-related applications, without appropriate expert oversight. For a comprehensive discussion on these issues, refer to (Mohammad, 2022, 2023).

For tasks such as spelling and grammar correction, we used privacy-protected AI assistants to ensure confidentiality.

Acknowledgments

We thank Jan Philip Wahle for granting us early access to the ABCDE dataset.

Thanks to Zara Siddique and Jan Philip Wahle for helpful discussions during the early stages of this project.

References

Andrea E. Abele, Nicole Hauke, Kim Peters, Eva Louvet, Aleksandra Szymkow, and Yanping Duan. 2016. Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality. *Frontiers in Psychology*, Volume 7 - 2016.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. 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.

Lee Joseph Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Amy J. C. Cuddy, Susan T. Fiske, and Peter Glick. 2007. The bias map: behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, 92 4:631–48.

Susan T. Fiske. 2018. Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science*, 27(2):67–73. PMID: 29755213.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902. Erratum in: J Pers Soc Psychol. 2024 Mar;126(3):412.

Kathleen Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How does stereotype content differ across data sources? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 18–34, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Koch, Austin Smith, Susan Fiske, Andrea Abele, Naomi Ellemers, and Vincent Yzerbyt. 2024. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods*.

G. Frederic Kuder and Marion W. Richardson. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2:151–160.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. *Preprint*, arXiv:2307.03172.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad. 2025a. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms. *Preprint*, arXiv:2503.23547.

Saif M. Mohammad. 2025b. Words of Warmth: Trust and Sociability Norms for over 26k English Words. *Preprint*, arXiv:2506.03993.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language 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 5356–5371, Online. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *Preprint*, arXiv:2005.10200.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language 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 4262–4274, Online. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024a. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024b. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, jenny or jingzhen? uncovering stereotypes in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18601–18619, Miami, Florida, USA. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Zoltán Gendler Szabó. 2024. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, summer 2024 edition. Metaphysics Research Lab, Stanford University.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Daniela Teodorescu and Saif Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137, Singapore. Association for Computational Linguistics.

Jan Philip Wahle, Krishnapriya Vishnubhotla, Bela Gipp, and Saif M. Mohammad. 2026. Affect, body, cognition, demographics, and emotion: The abcde of text features for computational affective science. In *Proceedings of the 1st Workshop on Computational Affective Science (CAS 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).

Joseph P. Weir. 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the sem. *Journal of strength and conditioning research*, 19 1:231–40.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Timothée Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Appendix

A Selection from ABCDE Dataset

We extracted 108 sentences about Donald Trump from the ABCDE dataset, representing 26.34% of all sentences for that target and 71.52% of the sentences extracted from ABCDE.

Target	Keywords
Women	girl, girls, woman, women, feminism, feminist
Religious People and nonreligious People	religious, religion, God, nonreligious, atheism, atheist, Christian, Christianity, Christians, Muslim, Islam, Muslims, Bible
Hillary Clinton	Hillary, Clinton, Hillary Clinton, HC, Benghazi
Donald Trump	Donald, Trump, Donald Trump
Barack Obama	Obama, Barack, Barack Obama, Obamacare
Environmentalists	environment, environmental, environmentalist, environmentalism, global warming, climate change

Table 10: Keywords using which the sentences from ABCDE were selected.

B Annotator Selection in Prolific

The filters described below were put in place to ensure that participants meet the language requirements and demonstrate strong qualifications and proven reliability.

1. Countries. The first two screener sets selected were the “current country of residence" and the “country of birth". Since the success of the task hinges on fluency in English, the countries selected in this set were those that mainly speak English, so the following countries were selected in both sets: Antigua and Barbuda, Australia, Barbados, Belize, Canada, Ireland, the United Kingdom, the United States, and New Zealand.

2. Languages. Prolific offers three screener sets related to the languages that annotators speak. Those are “first language", “primary language", and “fluent languages". Those were all set to “English".

Prolific displays the number of eligible participants after each screener set is applied, and the number decreases with every additional filter.

3. Education. For most screener sets regarding “highest education level completed", eligible participants were limited to those with at least a technical or community college degree, ranging up through undergraduate, graduate, and doctoral qualifications.

4. Approval Rate and Participation. The approval rate and previous participation criteria were also important in shaping the pool of annotators. By requiring a 99-100% approval rate, I attempted to minimize the risk of low-quality or careless responses, admitting only participants with an almost impeccable record who completed tasks to researchers’ satisfaction. The additional restriction of having over 500 previously approved submissions (later increased to 2000) further ensured that participants were not only reliable but also highly experienced with research tasks on Prolific.

C Data Distribution

	Hillary Clinton	Donald Trump	Feminist Movement	Atheism	Legalization of Abortion	Climate Change
Hillary Clinton	594	13	1	0	2	0
Donald Trump	39	260	0	1	0	1
Women	25	4	151	5	106	1
Barack Obama	48	38	0	2	9	12
Religious People	2	3	20	89	1	3
Environmentalists	2	0	0	0	0	35
Nonreligious People	0	1	1	33	0	0

Table 11: Distribution of sentences from the original five targets (columns) across the expanded set of seven targets in *W&C-Sent* (rows).

ID	Original Target	Sample sentence	Decision
288	Atheism	I will dwell in a peaceful habitation, in secure dwellings, and in quiet resting places -Isa. 32:18	Not selected
11182	Legalization of Abortion	Living in a pub isnt all that good when your friends turn into alcoholics #noonewillunderstand	Not selected
1463	Feminist Movement	We are 51% of the population and only 17% of Congress. The #WarOnWomen is absolutely a real thing. Wake up, America.	Selected
1710	Hillary Clinton	Hillary is killing it so far on the trail. She’s finally showing her personal side and I think it will benefit her profoundly.	Selected

Table 12: Examples of sentences from the SemEval-2016 stance dataset that were selected and not selected into W&C-Sent, with their original targets and IDs.

D The Complete English Annotation Guidelines

Preliminary instructions

1. Attempt these questions only if you are fluent in English.
2. Your responses are confidential.
3. There is a degree of subjectivity in this task. Let your instinct guide you; don't overthink it.
4. Consider the entire meaning of the sentence before attempting to give the relevant scores.
5. Your views regarding any of the entities or topics in the texts (such as political parties, individuals, social groups) should not affect your scores.
6. To ensure fairness and the validity of our scientific findings, some questions (typically unambiguous ones!) have predetermined answer ranges. While occasional deviations are acceptable given the subjectivity of this task, contributions may be rejected if a considerable number of these questions are answered incorrectly. Reading the guidelines below is therefore essential for a successful participation and compensation. This measure will ensure honest participation is compensated fairly.

Task definition and theoretical background

- Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of **warmth (W)** and **competence (C)**.
- Recently, psychologists have modeled warmth through two dimensions: **trust (T)** and **sociability (S)**.[3]
- This task will aim to assess the degrees of **trust**, **sociability**, and **competence** towards a **specific target within a sentence**.
- **Trust**:[4]
  - The focus in this dimension is on the personal / moral aspect of the target.
  - *High trust* can be defined as morality, kindness, sincerity, trustworthiness, and honesty.
  - Words associated with high trust: charity, mother, compliment, affectionate.
  - *Low trust* can be defined as immorality, insincerity, dishonesty, untrustworthiness, dubiousness, and maliciousness.
  - Words associated with low trust: discredit, bribe, espionage, disinformation, disloyal.
- **Sociability**:[4]
  - The focus in this dimension is on the social aspect of the target and its relational impact on others or society as a whole.
  - *High sociability* can be defined as friendliness, sociableness, generosity, and helpfulness.
  - Words associated with high sociability: helpful, intimate, laugh, celebration, reliant, entertain, social club, bestie.
  - *Low sociability* can be defined as antisocial behavior, lack of generosity, inconsiderateness, indifference, and unhelpfulness.
  - Words associated with low sociability: ingrate, abduct, selfish, theft, egomaniacal, pervert.
- **Competence**[4]:
  - *Competence* can be defined as ability, power, dominance, being in control, importance, having influence, and assertiveness.
  - Words associated with high competence: hitman, heroical, entrepreneurship, strategies, superman, viper, impunity.
  - *Incompetence* can be defined as submissiveness, not being in control of a situation, being controlled or guided by outside factors, or weakness.
  - Words associated with low competence: bootlicker, talentless, crash landing, bedridden, underfed.

The task

[3] This part was excluded from the competence-specific guidelines as it is irrelevant to the task.

- You will be given a text snippet and a target (group, entity, or individual). Below, you will be able to assign scores. Please rate the apparent levels of trust, sociability, or competence[4] that the sentence's author seems to express towards the specified target.
- The labels:[4]
  - -3 high distrust / unsociability / incompetence
  - -2 moderate distrust / unsociability / incompetence
  - -1 slight distrust / unsociability / incompetence
  - 0 neutral, not applicable, not expressed, etc.
  - +1 slight trust / sociability / competence
  - +2 moderate trust / sociability / competence
  - +3 high trust / sociability / competence

For the sentences in this task, the targets will be one of the following:

1. *Religious people*: this target will be listed along with sentences that are related to religions and target those who believe in God or practice (any) religion.
2. *Nonreligious people*: this target will be listed along with sentences that are related to atheism and target those who are not religious (i.e., atheists/agnostics).
3. *Women*: this target will be listed along with sentences that are related to women or girls and/or interconnected topics: sexism, feminism, misogyny, bias, and female representation.
4. *Environmentalists*: this target will be listed along with sentences that are related to environment and climate change activists.
5. *Hillary Clinton*: this target will be listed along with sentences that are related to the 2016 US presidential candidate and former Secretary of State Hillary Clinton.
6. *Donald Trump*: this target will be listed along with sentences that are related to the 2016 US presidential candidate and current US president Donald Trump.
7. *Barack Obama*: this target will be listed along with sentences that are related to the former US president Barack Obama.

In order to assess trust, sociability, and competence, try to answer the following questions[4]:

1. What is the degree of trust **towards this target** that **the author** of the text seems to express? Does **the author** seem to perceive the target as trustworthy or untrustworthy / moral or immoral / honest or dishonest?
2. What is the degree of sociability **towards this target** that **the author** of the text seems to express? Does **the author** seem to perceive the target as sociable or antisocial? Helpful or unhelpful?
3. What is the degree of competence **towards this target** that **the author** seems to express? Does **the author** seem to perceive the target as in control or out of control? Active or passive? Powerful or weak?

Notes:

1. There are select examples in the next page, accompanied by an explanation of the scores given for each example.
2. Adhere to the literal meaning of competence, which may be "positive" (e.g., a CEO) or "negative" (e.g., a villain or a dictator). Both types are considered "competence", regardless of the outcomes. Example 3 is an example of that.[5]
3. There are no repeated sentences in this study. All sentences were carefully chosen[6].
4. Even if the speaker is explicitly expressing opinions towards X, if the target listed is Y, then we want to know the degree of trust, sociability, and competence[7] towards Y only.

[4]Only the relevant dimension was included in each dimension-specific guidelines..

[5]This point only appeared in the competence-specific guidelines.

[6]This part was added after a participant in the pilot run skipped sentences they thought were repeated.

[7]Only the relevant dimension was included here in each dimension-specific guidelines

5. Try to be objective. Your views regarding any of the entities or topics in the texts (such as political parties, individuals, social groups) should not affect your scores.

6. There is an optional free-form text field underneath each instance. You can add any comments, thoughts, or justifications you may have on the scores you gave.

7. You will have these guidelines available to you at every stage of the task by pressing on "See task details" on the top right.

Examples[8]

Example 1:

Target: Women
Text: "My wife is the **most caring** person I've ever met ... she's the only woman in a house full of testosterone . She **never stops working** whether it's at home or being an RN . **I cant keep up** but I try . She makes me a **better person** . I'd be **lost without her** . Oh and she's smoking hot too."

Trust: +3 (high trust). The author expresses maximum trust in women through his wife as a representative example. He portrays his wife as a **trustworthy, dependable, and caring individual** who is **essential to his well-being** ("I'd be lost without her"). The statement "she makes me a better person" implies the author views his wife as having **a strong moral character that positively influences him**.
Sociability: +3 (high sociability). The text attributes maximum sociability to women through his wife. "**Most caring person I've ever met**" shows high sociability and interpersonal warmth. He emphasizes how essential her social/emotional qualities are to their family dynamic, particularly in a "house full of testosterone." The statement "whether it's at home or being an RN" suggests helpfulness and generosity both personally and professionally.
Competence: +3 (high competence). The author says that his wife "never stops working" in the context of her role as an RN (registered nurse) and as a mother, which suggests **multitasking, capability, and professional competence**. "I can't keep up but I try" indicates a **highly active and energetic individual**.

Example 2:

Target: Women
Text: "when i was 16 i had a folder of " **feel good songs** " and everyday i would select one and **send it to my best friend** along with a paragraph of **how much they meant to me** and why i they should be **happy** and then i would lay in bed thinking " no i'm not gay this is just what girls do :) ""

Trust: +3 (high trust). The author attributes **sincere, genuine intentions** to women through the described behaviors. The daily practice of **sending heartfelt messages** "about how much they meant to me and why they should be happy" suggests honesty and sincerity in women's relationships. The author portrays women as having **good moral intentions** in their friendships by genuinely **caring about others' wellbeing**.
Sociability: +3 (high sociability). The text attributes maximum sociability to women. The actions described (sending feel-good songs, writing paragraphs about how much someone means to you, **wanting to make someone happy**) **represent peak social engagement and helpfulness**. The phrase "this is just what girls do" frames these highly sociable, caring behaviors as naturally feminine traits and hence portrays women as exceptionally **generous with their emotional/social energy**.
Competence: +2 (moderate competence). Selecting appropriate songs, writing meaningful messages, and maintaining friendships require emotional intelligence, thoughtfulness, and social skills. However, the implication that this behavior is "just what girls do" could suggest that the author views this as **instinctual rather than skillful**, or that it's somehow slightly less significant than other types of competence.

[8]Each dimension-specific guidelines included the relevant score and justification of that dimension only.

Example 3:

Target: Hillary Clinton
Text: "Would you wanna be in a long term relationship with **some bitch** that **hides her emails**, & **lies to your face**? Then **#Dontvote**"

Trust: -3 (high distrust). The author explicitly portrays Clinton as someone who is **fundamentally untrustworthy** and **cannot be relied upon to tell the truth** or **be transparent** through two direct accusations: "hides her emails" and "lies to your face." With concealment and deception some of the **strongest markers of untrustworthiness**, the author emphasizes these as core trust violations.
Sociability: -2 (moderate unsociability). The derogatory term "**bitch**" (which is demeaning towards women) and the comparison to an **undesirable romantic partner** frames Clinton as someone who would be **unpleasant to be around or interact with socially**. The rhetorical question implies she would be toxic in close social relationships. However, one can view the sociability attack is more about being unpleasant in relationships rather than being completely antisocial or unhelpful in all social contexts, hence a maximum score was not assigned.
Competence: +2 (moderate competence). Despite the author implying that Clinton is manipulative and dishonest, the author's phrasing doesn't suggest that she's powerless or ineffective, as the negative behaviors described (concealment and deception) require some degree of **agency and planning**. The author suggests that Clinton is being **deliberate in her ("negative") performance** and **active in her ("negative") effects**, leading to a moderately high score in the competence dimension.

Notes on the suggested scores: This example shows how your political views regarding Hillary Clinton must not influence your score. Supporters of Clinton might see this as unfair or extremely sexist, while critics might view it as more damning commentary on her performance and ability.

Example 4:

Target: Women
Text: "My step sister **broke up** with her first boyfriend because she wanted to be independent ... **women suck** .. I'll miss you Kyle."

Trust: -1 (slight distrust). The author criticizes his sister's decision to prioritize independence and considers it hurtful to someone that they cared about, Kyle. **This suggests that the author thinks women make choices that harm others**, which is a claim towards women's social behavior in relationships rather than their honesty or morality.
Sociability: -3 (high unsociability). By saying "**women suck**", the author views women as **unpleasant, inconsiderate, or lacking in positive interpersonal qualities**. The author also frames women as **harmful to social relationships** by prioritizing their own desires over maintaining positive social connections. A maximum score was given since this is a clear negative generalization about women as a social group.
Competence: -2 (moderate incompetence). The overall "women suck" generalization suggests **poor judgment or decision-making by women as a social group**. This was exemplified by the author's step-sister breaking up with someone who the author thinks she should not have. It also indicates that the author **knows better than his step-sister, and by projection, women as well**.

Notes on the suggested scores:

1. One might say that "women suck" expresses a very negative sentiment towards women's trustworthiness and social likeability. This might affect the scores accordingly.

2. One might claim that the competence of women isn't really addressed since the author frames the sister's decision-making negatively rather than women as a group.

3. Consider the fact that the gender of the author is not explicit. How might it affect your scores if the author of the post were a woman? That is up to you to decide.

Example 5:

Target: Women
Text: “I need feminism because the United States is one of the only countries that doesn’t give paid maternity leave."

Trust: 0 (neutral). The author’s statement is focused on policy rather than character traits and doesn’t make any attributions about women’s trustworthiness or morality.
Sociability: 0 (neutral). The author’s statement is focused on policy rather than character traits. There is no commentary on women’s interpersonal qualities or social behavior.
Competence: 0 (neutral). The statement implies that women deserve certain rights/benefits, but it doesn’t directly attribute dominance or control to women. The statement doesn’t characterize women as active, powerful, passive, or weak. The author is advocating for institutional change (paid maternity leave) rather than making claims about women’s power or capabilities.

Note: This is a policy statement rather than a personal attribution about women’s trust, sociability, or competence. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 6:

Target: Donald Trump
Text: “Teaching similes : “The cats attitude was as **stubborn** as Donald Trump" #ShitMyKidsSay"
Trust: 0 (neutral). The author likens Trump’s stubbornness to that of a cat, which carries mild negative social judgment. However, stubbornness is not inherently about trust (morality, honesty, sincerity). Despite the playful mockery which could indicate diminished regard towards Trump, the statement doesn’t make claims about Trump’s reliability, honesty, or moral character.
Sociability: -1 (slight unsociability). Stubbornness in this context suggests someone who is, like a cat, **difficult to deal with interpersonally**: not very cooperative, helpful, or considerate in social interactions. This is a **mild social criticism** that does not completely portray him as completely antisocial or unhelpful.
Competence: -1 (slight incompetence). Comparing a political figure to a stubborn cat frames Trump as childish and unreasonable. The comparison could be seen as deliberately insulting to Trump’s **behavioral rigidity**, since a good leader is expected to be willing to change their mind based on the available evidence and opinions of experts. This shows that the author believes that Trump might **not be the best person to be the president** of the United States.
Notes on the suggested scores:

1. This is another example that shows that your political views regarding Donald Trump must not influence your score. Supporters of Trump might see this as unfair or even read stubbornness as positive determination, while critics might view it as more damning commentary on his interpersonal difficulties.

2. Other interpretations can be just as valid. One might argue that a 0 score for competence (neutral) is appropriate; generally speaking, stubbornness is a character trait that doesn’t directly relate to competence or incompetence. While it can sometimes imply determination (positive for competence), in this context it’s more about being inflexible or difficult rather than capable or incapable. The comparison to a cat’s attitude suggests that the author is referring to behavioral rigidity on Trump’s part. The author is not making any claims about Trump’s abilities, intelligence, or effectiveness. It is up to you to determine the weight given to the humorous framing vs. the underlying comparison.

Example 7:

Target: Religious people
Text: “Could all those who believe in a god **please leave**. The meeting will now continue for the **grown ups only**."

Trust: 0 (neutral). The criticism is entirely focused on intellectual maturity rather than character or morality. The author is suggesting that religious people shouldn’t participate in this particular discussion. On the other hand, the author doesn’t make any claims about religious people’s morality, honesty, sincerity, or trustworthiness. There are no accusations of deception, dishonesty, or moral failings.
Sociability: -3 (high unsociability). The author portrays religious people as **socially incompetent** and

needing exclusion from adult discourse (“please leave") and says their presence is incompatible with or unwanted in serious adult conversations (“grown ups only"). The order (telling an entire social group to leave) and the **dismissive language** are extremely exclusionary and socially hostile, resulting in a maximum unsociability score.

Competence: -3 (high incompetence). Religious people are portrayed as needing to be **excluded from decision-making processes**, suggesting they lack the authority or standing or cognitive abilities to participate in important discussions. Additionally, the “grown ups only" framing explicitly characterizes religious people as **childlike, passive, and intellectually weak (i.e., subordinates to “grown ups")**. This is a direct attack on their mental capacity and maturity, which are **core competence attributes**.

E Distribution of Annotators' Demographic Characteristics

The annotator pool included 55.1% identifying as male and 44.4% as female. The majority of participants were between 25 and 44 years old (58.5%), though the sample also included younger and older individuals up to the 65+ range. Most annotators were born in the United States (47.2%) or the United Kingdom (38.9%), with smaller groups from Canada (8.8%) and other countries including Nigeria, Australia, Italy, Singapore, Ireland, Germany, Ghana, and New Zealand (5.1%). All reported English as their primary language, and while most described themselves as monolingual (86.7%), a minority (13.3%) reported fluency in additional languages, such as Spanish, French, Dutch, Hindi, Urdu, and Punjabi. Further, nearly half of the annotators held a bachelor's degree (49.5%), followed by those with a master's or community college qualification (both 18.5%), while smaller proportions had completed high school (8.8%) or a doctorate (4.7%). The majority identified as White (73.6%), with the rest distributed across Asian (8.3%), Black (7.8%), Mixed (6.0%), and Other (4.1%) categories. Table 13 below shows the complete breakdown of these attributes.

Finally, the annotators' participation among the three dimensions demonstrates broad and consistent coverage, with only a minor variation. A total of 78 contributions were received in the trust dimension, 75 in sociability, and 70 in competence.

F Coarsening of Labels

F.1 Median-Based Coarsening

For each sentence-target pair, all fine-grained labels were collected into a list. Then, we counted the number of negative labels ($\leq$ -1), positive labels ($\geq$ 1), and neutral labels ($=$ 0), and the majority category determined the coarse label. For example, if negative judgments were most frequent, the pair was then assigned *low*. As a result, Neutral superseded in cases where neutral labels were the majority or when the counts of positive and negative labels were equal. Interestingly, there were 251 sentence-target pairs where the annotators' judgments were equally divided between *high* and *low*, resulting in a neutral label that reflects that lack of consensus. Almost half of those were in the competence dimension (48.6%), while the rest was split almost equally between the sociability (26.3%) and trust (25.1%) dimensions.

Attribute	Categories and Frequency (%)				
Gender	Male	Female	Other		
	114 (54.2)	95 (45.2)	1 (0.05)		
Age	35-44	25-34	45-54	Other	
	66 (31.4)	58 (27.6)	43 (20.5)	43 (20.5)	
Country	US	UK	Canada	Other	
	102 (48.5)	78 (37.5)	19 (9.1)	11 (4.9)	
Primary language	English				
	210 (100)				
Other languages	Monolingual		Bilingual+		
	106 (86.2)		17 (13.8)		
Education	BA	MA	CC	HS	PhD
	102 (48.5)	40 (19.0)	39 (18.6)	19 (9.1)	10 (4.8)
Ethnicity	White	Asian	Black	Other	
	154 (73.3)	18 (8.6)	16 (7.6)	22 (10.5)	
Previous studies	2000+	1000-1500	1500-2000	500-1000	
	113 (53.8)	38 (18.1)	34 (16.2)	25 (11.9)	

Table 13: Demographic breakdown of annotators. Numbers are counts with percentages in parentheses. The attribute "Country" refers to country of birth. "Fluent languages" refers to the languages which the annotators reported being fluent in, other than English, and not all annotators filled it. Other age ranges include 55-64 (24 annotators, or 12%), 65+ (10 annotators, or 4.8%), and 18-24 (9 annotators, or 4.3%). Other ethnic groups include "mixed" (13 annotators, or 6.2%) and "other" (8 annotators, or 3.8%). In "Education", BA = undergraduate degree, MA = graduate degree, CC = community college or technical degree, HS = high school diploma, and PhD = doctorate.

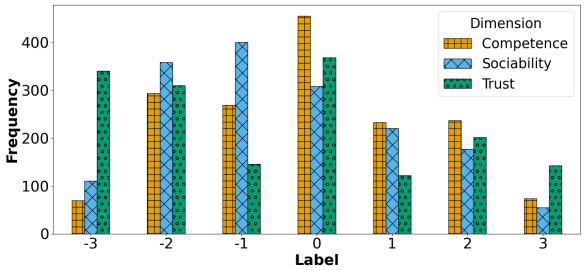


Figure 4: Distribution of discretized median-based labels, ordered from negative to positive.

F.2 Mean-Based Coarsening

Following Mohammad (2025b), this method converted the average score of each sentence-target pair into seven bins (from negative to positive). This approach led to more neutral instances and a less extreme skew towards the negative side, indicating that averaging the scores smoothed the extremes of annotator judgments.

G Observations from the Data

G.1 Agreement among the Annotators

G.1.1 Strict Unanimity

This refers to cases where all annotators chose the exact same fine-grained label for a sentence-target pair. 143 sentence-target pairs achieved strict unanimity, representing 3% of the overall W&C-Sent dataset. The most common dimension was trust (65.7% of instances), followed by sociability (21.7%) and competence (12.6%).

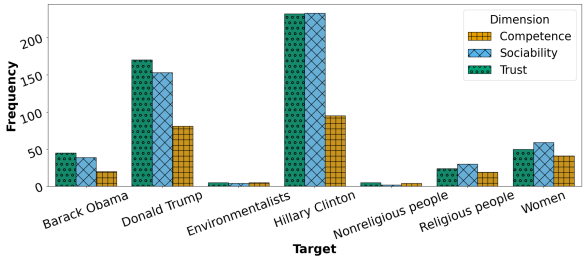


Figure 5: Distribution of **fine-grained** unanimous judgments across targets and dimensions. The figure shows higher agreement for sentences targeting Clinton and Trump in the trust dimension, and no occurrences for the target Environmentalists.

G.1.2 Soft Unanimity

This refers to cases where annotators agreed on the overall polarity (low, neutral, or high) of a sentence-target pair even if they did not select the exact same fine-grained score. Annotators reached soft agreement on 1,316 instances (sentence-target pairs), accounting for 26.9% of all sentence-target pairs in W&C-Sent. The majority of soft agreement instances were on negative judgments.

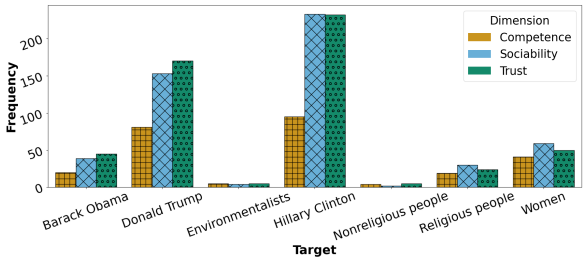
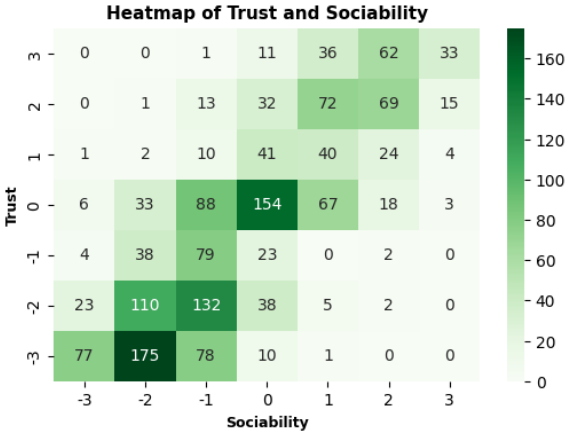


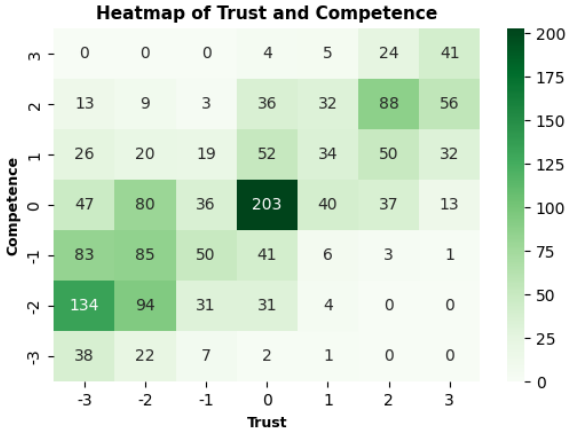
Figure 6: The distribution of **soft** unanimous judgments across targets and dimensions. The figure higher soft unanimous agreement for the trust and sociability of Clinton and Trump.

H Correlations Between Dimensions

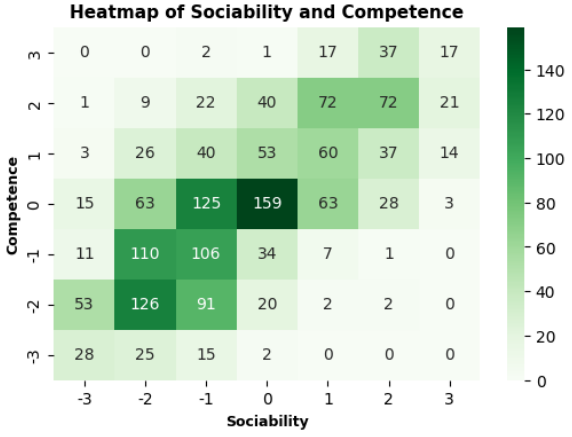
Figure 7 shows three heat maps that show frequent co-occurrence between median neutral scores; meaning, when competence within a sentence is judged as neutral, trust and sociability judgments tend to be neutral in it, too. In addition, they all demonstrate strong co-occurrences around the diagonal, with the highest being between high distrust and moderately high unsociability (-3, -2) in Grid A, followed by high distrust and moderately high incompetence (-3, -2) in Grid B, and moderately high distrust and moderately high unsociability (-2, -2) in Grid C.



(a) Trust & Sociability



(b) Trust & Competence



(c) Sociability & Competence

Figure 7: Heat maps of correlation matrices for each pair of dimensions.

I More on Inter-Annotator Agreement

I.1 Krippendorff's Alpha

When broken down by both target and dimension, Krippendorff's α scores show that agreement varies considerably depending on the target of the sentence. The strongest reliability is observed for trust perceptions of Donald Trump ($\alpha = 0.67$). Trust perceptions of Hillary Clinton came second ($\alpha = 0.61$). The scores for sociability perceptions of these two political figures remain weak but higher than for other targets, at around 0.53-0.55. By contrast, Barack Obama's αs show weaker consistency, with trust at 0.50 and sociability at 0.48.

Notably, agreement drops sharply when annotators assessed social groups such as Women, Religious People, Non-religious People, and Environmentalists, where none of the dimensions exceed 0.41 and several fall below 0.30. This suggests that annotators shared more consistent interpretations when evaluating singular, high-profile political figures than collective or socially non-monolithic categories.

Competence scores are consistently poor across all targets. Even for figures like Donald Trump ($\alpha = 0.38$) and Hillary Clinton ($\alpha = 0.32$), agreement on their competence remains weak, while Barack Obama and group-based targets fall even lower. The lowest values are seen for Religious People, Environmentalists, and Nonreligious People.

I.2 Split-Half Reliability

The process by which SHR scores were computed for each dimension was repeated again for each target. Figure 8 illustrates these scores, showing some interesting highlights. The sentences whose target was Donald Trump showed the highest stability across all three dimensions, and especially in the trust dimension where the SHR score exceeds 0.8. The targets Hillary Clinton and Barack Obama fall right behind Donald Trump in trust and sociability but had very poor SHR in competence. The target

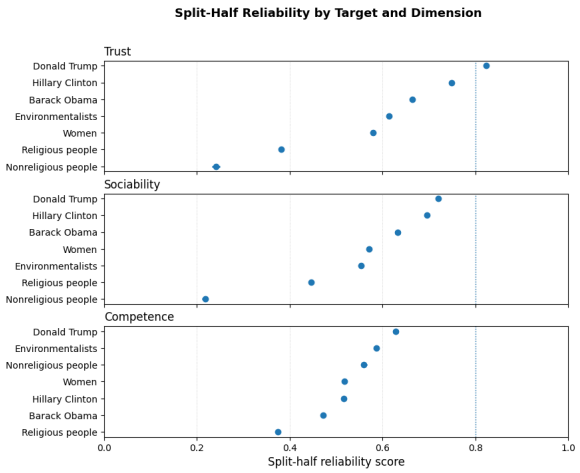


Figure 8: The SHR scores of each target, in each relevant dimension. For each sub-plot, the targets were ordered based on their score, from high to low. The vertical dotted line is the 0.8 reliability cutoff, which only the sentences whose target was Donald Trump cross in the trust dimension.

Women had mediocre stability across all three dimensions, while Religious People, Non-religious People, and Environmentalists came last in almost all dimensions, probably due to their small size in the dataset.

J Regression Models

J.1 Logistic Regression using TF-IDF Features

A logistic regression model, using TF-IDF features and the mean of annotator scores as the output, was trained first. Table 14 shows that the TF-IDF models perform markedly worse, which means that simple lexical representations are inadequate for detecting T/S/C in text. This is especially seen in the much lower correlations between its predictions and the real scores, proving the importance of the contextual BERT representations.

J.2 Seven-Class Regression

The Huber loss was chosen for a loss function due the advantages mentioned earlier: it balances the sensitivity of MSE to outliers with the stability of MAE for an optimized performance. In addition, input examples that would be used to train the models were structured as paired sequences, where the “Text" and “Target" columns were jointly tokenized using the BERTweet tokenizer. Later, W&C-Sent was partitioned into training, validation, and test sets using grouped splits based on text instances; identical texts, of which there is a lot in the dataset but with different targets, were specifically designed to not appear across different data splits, thereby helping avoid bias and inflated performance.

Dimension	Metric	MAE	RMSE	ρ	Acc.	F1	±1 Acc.
Trust	TF-IDF	1.36	1.90	0.44	29.7	29.5	65.1
	Median	1.01	1.40	0.76	35.8	**31.2**	76.5
	Mean	**0.81**	**1.10**	**0.77**	**37.0**	29.4	**82.6**
Sociability	TF-IDF	1.18	1.70	0.43	31.2	26.2	70.3
	Median	0.83	1.03	0.76	35.5	25.7	85.0
	Mean	**0.66**	**0.86**	**0.78**	**45.6**	**33.3**	**88.7**
Competence	TF-IDF	1.40	1.80	0.35	19.0	15.8	61.5
	Median	1.02	1.29	0.60	28.7	21.7	77.7
	Mean	**0.83**	**1.06**	**0.63**	**37.3**	**25.2**	**85.3**

Table 14: The results of the seven-class regression models for each dimension. The results for the best variant are in **bold**. This table highlights how using TF-IDF embeddings do not perform well compared to the contextual BERT-based embeddings.

Training was conducted using the “Trainer" class from Hugging Face (Wolf et al., 2019). Model performance was assessed with MAE, RMSE, and Spearman correlation. Two variants of these experiments were run, one using the median score and one using the mean as the target label. This was motivated by pure curiosity, as I was wondering whether regression models would perform better using the mean or the median. After the training was completed, each model was assessed on the same held-out test set, and its continuous predictions were rounded into seven discrete ordinal categories to ease comparison between the variants.

The model for each dimension was evaluated using:

1. MAE and RMSE, for error magnitude
2. Spearman ρ, for correlation with human ratings
3. Accuracy, to observe the exact matches
4. Macro F1 score, which measures the overall balance between precision and recall across all classes
5. And the within-1-bin accuracy, which measures how often model predictions fall within one rating level of the true labels captures near-miss performance in ordinal tasks

K Classification Models

K.1 Experimental Setup

Due to class imbalance, a custom subclass of the Trainer class was implemented and in which a

weighted negative log-likelihood loss was defined; this involved class weights to be derived from the inverse of the class frequencies in the training data, to ensure that the underrepresented neutral classes exerted a proportionally stronger influence on the weight optimization process. This also ensures better generalization across all three categories.

Input processing mirrored that of the regression setup. Paired sequences of “Text" and “Target" were tokenized jointly with the BERTweet tokenizer and truncated to the same fixed maximum length. Additionally, W&C-Sent was partitioned into training, validation, and test sets the same way, too.

The training was carried out identically as well, but the evaluation metrics changed due to the type of the model. Evaluation here involved accuracy and macro-averaged F1 score, with the F1 score being the metric by which the best model was decided; F1 emphasizes balanced performance across classes, particularly due to the imbalanced label distributions.

Two versions of this experiment was run, one included coarsening the W&C-Sent dataset using majority-vote coarsening while the second followed the method used in (Mohammad, 2025b). The class distribution when the labels are coarsened using first method is shown in Table 2.

L Hyper-parameters for Code Reproducibility

L.1 Classification

Component	Value
Task	Three-class classification (text + target pair)
Base model	`vinai/bertweet-base`
Tokenization	Truncate + pad to fixed length
Max sequence length	256
Inputs	Text as primary, Target as text_pair
Labels	Median score, mapped to a numeric value
Num classes	3
Class weighting	Inverse frequency on training split, used in CrossEntropyLoss
Split method	GroupShuffleSplit (GSS) on Text
Test size	0.2
Validation size	0.2 (using a second GSS)
Random seed	13
Epochs	10
Train batch size	16
Eval. batch size	32
Learning rate	2×10^{-5}
Evaluation cadence	Each epoch
Model selection	Using macro-F1
Early stopping	Patience with 3 epochs
Reported metrics	Accuracy, macro-F1 (on test set)
Per-dimension setting	Separate model per dimension

Table 15: Hyperparameters and setup for the 3-class classifier.

Component	Value
Task	Regression on text + target pair
Base model	`vinai/bertweet-base`
Tokenization	Truncate + pad to fixed length
Max sequence length	256
Inputs	Text as primary, Target as text_pair
Target column	Median or mean score
Loss	Huber (SmoothL1) with $\beta = 1.0$
Split method	GroupShuffleSplit (GSS) on the Text column
Test size	0.2
Validation size	0.2 (using a second GSS)
Random seed	13
Epochs	10
Train batch size	16
Eval batch size	32
Learning rate	1×10^{-5}
Weight decay	0.01
Warmup ratio	0.06
Model selection	Best validation MAE
Early stopping	Patience with 3 epochs
Logging	Every 50 steps
Reported test metrics	MAE, RMSE, Spearman ρ
Per-dimension setting	Separate model per dimension

Table 16: Hyperparameters and setup for the regression model.

M Full Results with Precision and Recall Scores

Dimension	Model	Accuracy	F1	±1 Accuracy	Precision	Recall
	Dummy Classifier	0.22	0.08	0.58	0.05	0.22
	TF-IDF Regression	0.30	0.30	0.65	0.31	0.30
	Seven-Class Regression	0.35	0.31	0.83	0.43	0.32
	Gemma3 ZS	0.36	0.34	0.78	0.36	0.36
	Gemma3 FS	0.38	0.33	0.79	0.37	0.35
	Qwen2.5 ZS	0.27	0.25	0.72	0.29	0.29
	Qwen2.5 FS	0.35	0.35	0.76	0.38	0.4
Trust	GPT-4o ZS	0.43	0.42	0.91	0.43	0.50
	GPT-4o FS	0.39	0.4	0.84	0.4	0.43
	GPT-4o-mini ZS	0.27	0.26	0.60	0.32	0.36
	GPT-4o-mini FS	0.2	0.17	0.54	0.25	0.19
	GPT-5.2 ZS	0.41	0.38	0.87	0.45	0.41
	GPT-5.2 FS	0.40	0.39	0.89	0.45	0.40
	Qwen3 ZS	0.32	0.30	0.78	0.33	0.31
	Qwen3 FS	0.30	0.26	0.78	0.32	0.28
	Dummy Classifier	0.26	0.11	0.67	0.07	0.26
	TF-IDF	0.31	0.26	0.70	0.26	0.27
	Seven-Class Regression	0.46	0.34	0.88	0.33	0.35
	Gemma3 ZS	0.34	0.27	0.82	0.34	0.33
	Gemma3 FS	0.31	0.23	0.76	0.29	0.27
	Qwen2.5 ZS	0.20	0.20	0.69	0.31	0.28
	Qwen2.5 FS	0.25	0.25	0.67	0.29	0.28
Sociability	GPT-4o ZS	0.44	0.40	0.92	0.42	0.43
	GPT-4o FS	0.38	0.35	0.87	0.36	0.42
	GPT-4o-mini ZS	0.13	0.14	0.52	0.31	0.31
	GPT-4o-mini FS	0.24	0.2	0.59	0.27	0.24
	GPT-5.2 ZS	0.31	0.30	0.83	0.45	0.32
	GPT-5.2 FS	0.40	0.37	0.88	0.44	0.36
	Qwen3 ZS	0.35	0.30	0.81	0.39	0.31
	Qwen3 FS	0.31	0.24	0.82	0.33	0.32
	Dummy Classifier	0.24	0.09	0.60	0.06	0.24
	TF-IDF	0.19	0.16	0.62	0.18	0.18
	Seven-Class Regression	0.36	0.24	0.86	0.26	0.26
	Gemma3 ZS	0.22	0.19	0.58	0.32	0.24
	Gemma3 FS	0.35	0.27	0.7	0.32	0.33
	Qwen2.5 ZS	0.22	0.19	0.59	0.34	0.17
	Qwen2.5 FS	0.30	0.28	0.65	0.31	0.28
Competence	GPT-4o ZS	0.34	0.31	0.80	0.34	0.30
	GPT-4o FS	0.24	0.24	0.64	0.24	0.3
	GPT-4o-mini ZS	0.09	0.11	0.37	0.28	0.18
	GPT-4o-mini FS	0.22	0.2	0.6	0.27	0.24
	GPT-5.2 ZS	0.28	0.25	0.73	0.33	0.26
	GPT-5.2 FS	0.27	0.24	0.71	0.30	0.24
	Qwen3 ZS	0.22	0.18	0.60	0.31	0.19
	Qwen3 FS	0.25	0.21	0.67	0.34	0.22

Table 17: Performance of regression models and LLMs on fine-grained prediction.

Dimension	Model	Accuracy	F1	Precision	Recall	
	BERT Classification	0.79	0.59	0.598	0.579	
	Gemma ZS	0.74	0.60	0.58	0.64	
	Gemma FS	0.72	0.60	0.58	0.67	
	Qwen2.5 ZS	0.62	0.53	0.60	0.70	
Trust	Qwen2.5 FS	0.61	0.51	0.57	0.69	
	GPT-4o-ZS	0.86	0.78	0.79	0.78	
	GPT-4o-FS	0.77	0.71	0.75	0.73	
	GPT-4o-mini ZS	0.78	0.65	0.63	0.70	
	GPT-4o-mini FS	0.70	0.58	0.57	0.61	
	GPT-5.2 ZS	0.81	0.72	0.98	0.73	0.71
	GPT-5.2 FS	0.78	0.71	0.99	0.73	0.71
	Qwen3 ZS	0.79	0.66	0.94	0.65	0.69
	Qwen3 FS	0.72	0.62	0.94	0.62	0.62
	BERT Classifier	0.81	0.54	0.52	0.57	
	Gemma ZS	0.69	0.56	0.55	0.58	
	Gemma FS	0.71	0.54	0.53	0.61	
	Qwen2.5 ZS	0.47	0.42	0.44	0.59	
Sociability	Qwen2.5 FS	0.34	0.34	0.44	0.65	
	GPT-4o-ZS	0.80	0.71	0.72	0.71	
	GPT-4o-FS	0.72	0.66	0.69	0.71	
	GPT-4o-mini ZS	0.74	0.54	0.56	0.57	
	GPT-4o-mini FS	0.62	0.47	0.48	0.48	
	GPT-5.2 ZS	0.83	0.74	0.98	0.75	0.73
	GPT-5.2 FS	0.79	0.72	0.99	0.74	0.73
	Qwen3 ZS	0.74	0.59	0.91	0.58	0.61
	Qwen3 FS	0.67	0.58	0.96	0.58	0.6
	BERT Classifier	0.74	0.50	0.49	0.51	
	Gemma ZS	0.52	0.46	0.47	0.47	
	Gemma FS	0.52	0.43	0.46	0.44	
	Qwen2.5 ZS	0.47	0.47	0.49	0.61	
Competence	Qwen2.5 FS	0.56	0.53	0.53	0.59	
	GPT-4o-ZS	0.63	0.59	0.59	0.60	
	GPT-4o-FS	0.60	0.57	0.57	0.60	
	GPT-4o-mini ZS	0.53	0.40	0.46	0.44	
	GPT-4o-mini FS	0.47	0.37	0.42	0.52	
	GPT-5.2 ZS	0.49	0.46	0.86	0.51	0.51
	GPT-5.2 FS	0.46	0.43	0.82	0.5	0.49
	Qwen3 ZS	0.64	0.56	0.84	0.57	0.55
	Qwen3 FS	0.67	0.58	0.9	0.59	0.58

Table 18: Performance of BERT classification model and LLMs on coarse-grained prediction.

N LLM Prompts

N.1 System Message

"You are a professional language analyst and sociologist.
You work on warmth and competence, which are two concepts in sociology.
You know that warmth is decomposed into trust, which relates to the moral and personal aspect of the target; and sociability, which relates to the relational and societal aspect and impact of the target.
You work on a sentence-level: you read the full sentence, with all its components, before you assess it.
There is a degree of subjectivity in this task, so you consider the meaning that the general population would agree with and consider.
You do not add information that does not appear in the text.
The context matters the most to you. You consider all relevant information.
You understand that sarcasm is an ironic remark

meant to mock by saying something different than what the speaker really means.
You understand that irony is the humorous or mildly sarcastic use of words to imply the opposite of what they normally mean.
You understand that hyperbole is the extreme, dramatic exaggeration for effect, not meant to be taken literally.
You understand how irony, sarcasm, and hyperbole are employed in social media and in slang, and that the literal wording does not always reflect the true meaning and sentiment.
You assess warmth and competence on a scale from -3 to +3[9].
You adhere to the provided target. Meaning, you understand that even if the speaker is explicitly expressing opinions towards entity_X, you give the score for trust towards entity_Y only if the target given to you is entity_Y.
You do not use first-person pronouns such as “I" or “we" in your answer.
You adhere to the targets given to you in your assessment of warmth or competence.
You understand that the use of hashtags in social media can be complementary to the text. However, that is not always the case, as hashtags can be used to categorize content, increase visibility, and connect users with relevant discussions and communities, rather than as sentiment markers.
You understand that when the target is “religious people", then it refers to religions and those who believe in any God or practice any religion.
You understand that when the target is “nonreligious people", then it refers to atheism and those who are not religious (i.e., atheists/agnostics).
You understand that when the target is “women", then it refers to women or girls and/or interconnected topics: sexism, feminism, misogyny, bias, and female representation.
You understand that when the target is “Hillary Clinton", then it refers to the 2016 US presidential candidate and former Secretary of State Hillary Clinton.
You understand that when the target is “Donald Trump", then it refers to the 2016 US presidential candidate and current US president Donald Trump.
You understand that when the target is “Barack Obama", then it refers to the former US president Barack Obama.
You understand that when the target is “Environmentalists", then it refers to to environment and climate change activists."

[9]Coarse-grained prompts replaced this line with “You assess warmth and competence on the scale of “low", “neutral, not applicable, not expressed", and “high". "

N.2 Fine-Grained, Zero-Shot User Role Messages

N.2.1 Trust

“Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.
The focus in this dimension is on the personal / moral aspect of the target.
High trust can be defined as morality, kindness, sincerity, trustworthiness, and honesty.
Words associated with high trust: charity, mother, compliment, affectionate.
Low trust can be defined as immorality, insincerity, dishonesty, untrustworthiness, dubiousness, and maliciousness.
Words associated with low trust: discredit, bribe, espionage, disinformation, disloyal.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of trust that the sentence’s author seems to express towards the specified target.
Choose one label from these seven labels:
high distrust
moderate distrust
slight distrust
neutral, not applicable, not expressed
slight trust
moderate trust
high trust

Here is the sentence: { }.
Its target is: { }.

In order to assess trust, try to answer the following

questions:
1. What is the degree of trust towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as trustworthy or untrustworthy / moral or immoral / honest or dishonest?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the trust towards { } only.
In the format of a JSON file or a Python dictionary, you should provide your justification saved in a key called “reason". Then, based on your justification, add your rating to a key called “label"."

N.2.2 Sociability

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of sociability towards a specific target within a sentence.
The focus in this dimension is on the social aspect of the target and its relational impact on others or society as a whole.
High sociability can be defined as friendliness, sociableness, generosity, and helpfulness.
Words associated with high sociability: helpful, intimate, laugh, celebration, reliant, entertain, social club, bestie.
Low sociability can be defined as antisocial behavior, lack of generosity, inconsiderateness, indifference, and unhelpfulness.
Words associated with low sociability: ingrate, abduct, selfish, theft, egomaniacal, pervert.
Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of sociability that the sentence’s author seems to express towards the specified target.
Choose one label from these seven labels:
high unsociability
moderate unsociability
slight unsociability
neutral, not applicable, not expressed, etc.
slight sociability
moderate sociability
high sociability

Here is the sentence: { }.
Its target is: { }.
In order to assess sociability, try to answer the following questions:
1. What is the degree of sociability towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as sociable or antisocial? Helpful or unhelpful?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of perceived sociability trust towards { } only.
In the format of a JSON file or a Python dictionary, you should provide your justification saved in a key called “reason". Then, based on your justification, add your rating to a key called “label"."

N.2.3 Competence

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Your task is to assess the degree of competence towards a specific target within a sentence.
Competence, in the broader sense, is interchangeable with ‘ability’ or ‘capability’. Competence can be defined as ability, power, dominance, being in control, importance, having influence, and assertiveness.
Words associated with high competence: hitman, heroical, entrepreneurship, strategies, superman, viper, impunity.
Incompetence can be defined as submissiveness, not being in control of a situation, being controlled or guided by outside factors, or weakness.
Words associated with low competence: bootlicker, talentless, crash landing, bedridden, underfed.
Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of competence that the sentence’s author seems to express towards the specified target.
Choose one label from these seven labels: high

incompetence
moderate incompetence
slight incompetence
neutral, not applicable, not expressed
slight competence
moderate competence
high competence
Adhere to the literal meaning of competence, which may be “positive" (e.g., a CEO) or “negative" (e.g., a villain or a dictator). Both types are considered competence, regardless of the outcomes.

Here is the sentence: { }.
Its target is: { }.
In order to assess competence, try to answer the following questions:
1. What is the degree of competence towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } in control or out of control? Active or passive? Powerful or weak?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the competence towards { } only.
In the format of a JSON file or a Python dictionary, you should provide your justification saved in a key called “reason". Then, based on your justification, add your rating to a key called “label".

N.3 Fine-Grained, Few-Shot User Role Messages

N.3.1 Trust

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.

The focus in this dimension is on the personal / moral aspect of the target.
High trust can be defined as morality, kindness, sincerity, trustworthiness, and honesty.
Words associated with high trust: charity, mother, compliment, affectionate. Low trust can be defined as immorality, insincerity, dishonesty, untrustworthiness, dubiousness, and maliciousness.
Words associated with low trust: discredit, bribe, espionage, disinformation, disloyal.
Here are examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:

Target: Women
Text: My wife is the most caring person I’ve ever met ... she’s the only woman in a house full of testosterone . She never stops working whether it’s at home or being an RN . I cant keep up but I try . She makes me a better person . I’d be lost without her . Oh and she’s smoking hot too.
Trust score: +3 (high trust).
Rationale: The author expresses maximum trust in women through his wife as a representative example. He portrays his wife as a trustworthy, dependable, and caring individual who is essential to his well-being ("I’d be lost without her"). The statement “she makes me a better person" implies the author views his wife as having a strong moral character that positively influences him.

Example 2:

Target: Women
Text: “when i was 16 i had a folder of “ feel good songs ” and everyday i would select one and send it to my best friend along with a paragraph of how much they meant to me and why i they should be happy and then i would lay in bed thinking “ no i’m not gay this is just what girls do :) ”"
Trust score: +3 (high trust).
Rationale: The author attributes sincere, genuine intentions to women through the described behaviors. The daily practice of sending heartfelt messages “about how much they meant to me and why they should be happy" suggests honesty and sincerity in women’s relationships. The author portrays women as having good moral intentions in their friendships by genuinely caring about others’ wellbeing.

Example 3:

Target: Public Office Candidate Maya Thompson
Text: I’m really excited for 2026 after finding out the amount @MayaThompson raised. We got this.
Trust score: +2 (moderate trust).

Rationale: The author expresses confidence and optimism toward the target upon learning about the amount of money she raised, framing this as a positive and reassuring signal. The phrase "We got this" implies alignment with and belief in the target's reliability and capacity to lead or represent shared goals. The enthusiastic endorsement reflects a meaningful level of confidence in the target's legitimacy and dependability within the political context.

Example 4:

Target: Public Office Candidate Alex Rivera
Text: I support Jordan Lee because the country needs a new direction, but if Lee loses the primary, I'll support Alex Rivera.#Election2024
Trust score: +1 (slight trust).
The author expresses conditional support for Alex Rivera if Jordan Lee loses the primary. This implies a baseline level of acceptance and confidence in Rivera's moral and political legitimacy, even if she is not the author's first choice. The willingness to support her suggests the author does not view her as dishonest, corrupt, or morally unfit.

Example 5:

Target: Women
Text: I need feminism because the United States is one of the only countries that doesn't give paid maternity leave.
Trust score: 0 (neutral).
Rationale: The author's statement is focused on policy rather than character traits and doesn't make any attributions about women's trustworthiness or morality. This is a policy statement rather than a personal attribution about women's trustworthiness. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 6:

Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Trust score: 0 (neutral).
Rationale: The criticism is entirely focused on intellectual maturity rather than character or morality. The author is suggesting that religious people shouldn't participate in this particular discussion. On the other hand, the author doesn't make any claims about religious people's morality, honesty, sincerity, or trustworthiness. There are no accusations of deception, dishonesty, or moral failings.

Example 7:

Target: Religious people
Text: When people use religion as a reason to exclude others, they should not be surprised when others push back against them.
Trust score: -1 (slight distrust).
Rationale: The statement frames religious people as potential discriminators, suggesting that members of this group may engage in unfair or exclusionary behavior toward others. This implicitly casts doubt on their moral reliability and fairness, which are core components of trust. However, the author criticizes discriminatory behavior associated with the social group but does not assert corruption, deception, or bad faith, which earned it a mild distrust score.

Example 8:

Target: Public Candidate Riley Chen
Text: It sounds like the candidate promised opposite things to different people just to please them.
Trust score: -2 (moderate distrust).
Rationale: The author directly accuses the target of making inconsistent and contradictory promises to different people about the same policy issue. This frames the target as insincere and politically opportunistic, suggesting a lack of honesty and moral reliability. It implies strategic deception and two-faced behavior, which undermines perceptions of trustworthiness.

Example 9:
Target: Public Office Candidate Jordan Reed
Text: Would you want to be in a long-term relationship with someone who hides information and lies to you? Then don't vote for Jordan Reed.
Trust score: -3 (high distrust).
Rationale: The author directly accuses the target of deliberate concealment and deception by claiming that the candidate "hides information" and "lies." These behaviors are core violations of trust, signaling dishonesty and moral unreliability. By framing

the target as an unsuitable long-term partner due to these traits, the author implies that the target cannot be depended on in close or public relationships. The message portrays untrustworthiness as a defining characteristic of the target rather than a minor flaw. With concealment and deception some of the strongest markers of untrustworthiness, the author emphasizes these as core trust violations.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual). Rate the apparent level of trust that the sentence's author seems to express towards the specified target. Choose one label from these seven labels: high distrust
moderate distrust
slight distrust
neutral, not applicable, not expressed
slight trust
moderate trust
high trust

Here is the sentence: { }.
Its target is: { }.
In order to assess trust, try to answer the following questions:
1. What is the degree of trust towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as trustworthy or untrustworthy / moral or immoral / honest or dishonest?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the trust towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".

N.3.2 Sociability

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust and sociability.
Your task is to assess the degree of sociability towards a specific target within a sentence.
The focus in this dimension is on the social aspect of the target and its relational impact on others or society as a whole.
High sociability can be defined as friendliness, sociableness, generosity, and helpfulness.
Words associated with high sociability: helpful, intimate, laugh, celebration, reliant, entertain, social club, bestie.
Low sociability can be defined as antisocial behavior, lack of generosity, inconsiderateness, indifference, and unhelpfulness.
Words associated with low sociability: ingrate, abduct, selfish, theft, egomaniacal, pervert.

Here are examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:

Target: Women
Text: My wife is the most caring person I've ever met ... she's the only woman in a house full of testosterone . She never stops working whether it's at home or being an RN . I cant keep up but I try . She makes me a better person . I'd be lost without her . Oh and she's smoking hot too.
Sociability score: +3 (high sociability).
Rationale: The text attributes maximum sociability to women through his wife. "Most caring person I've ever met" shows high sociability and interpersonal warmth. He emphasizes how essential her social/emotional qualities are to their family dynamic, particularly in a "house full of testosterone." The statement "whether it's at home or being an RN" suggests helpfulness and generosity both personally and professionally.

Example 2:

Target: Women
Text: when i was 16 i had a folder of "feel good songs" and everyday i would select one and send it to my best friend along with a paragraph of how much they meant to me and why i they should be happy and then i would lay in bed thinking "no i'm not gay this is just what girls do :) "
Sociability score: +3 (high sociability).
Rationale: The text attributes maximum sociability to women. The actions described (sending feel-good songs, writing paragraphs about how much someone means to you, wanting to make

someone happy) represent peak social engagement and helpfulness. The phrase "this is just what girls do" frames these highly sociable, caring behaviors as naturally feminine traits and hence portrays women as exceptionally generous with their emotional/social energy.

Example 3:

Target: US Presidential Candidate Hillary Clinton
Text: ".@HillaryClinton Looking 4ward 2 hearing your Economic Agenda on Monday July13. WOW a candidate talking Specifics & not Rhetoric!"
Sociability score: +2 (moderate sociability).
Rationale: The text expresses clear positive regard toward Hillary Clinton by praising her for "talking specifics & not rhetoric" and showing anticipation for hearing her economic agenda. This reflects a favorable, respectful social orientation toward the target and frames her communication style as constructive and appreciated by others. It signals that the author views the target as positively engaging and socially constructive in public discourse.

Example 4:

Target: President Emmanuel Macron
Text: Despite the positive supreme court ruling, our struggle to end LGBT discrimination is far from finished
Sociability score: +1 (slight sociability).
Rationale: The text references Macron in the context of advocating against LGBT discrimination and frames him as aligned with an inclusive, socially supportive cause. This association implies a mild positive social orientation toward the target, suggesting that he is perceived as supportive of social groups facing discrimination.

Example 5:

Target: Women
Text: I need feminism because the United States is one of the only countries that doesn't give paid maternity leave.
Sociability score: 0 (neutral).
Rationale: The author's statement is focused on policy rather than character traits. There is no commentary on women's interpersonal qualities or social behavior. Note: This is a policy statement rather than a personal attribution about women's sociability. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 6:

Target: US Presidential Candidate Hillary Clinton
Text: Hillary's brand of experience we do not need!
Sociability score: -1 (slight unsociability)
Rationale: The statement criticizes Hillary Clinton by rejecting her brand of experience, which frames her negatively and signals disapproval of her suitability or value. However, the critique is focused on competence or political experience rather than on her interpersonal qualities or social behavior toward others. Despite there being no explicit portrayal of her as antisocial, unhelpful, or harmful in relationships, the dismissive tone still conveys mild negative social evaluation toward the target, which conveys slight unsociability.
Example 7:

Target: US Presidential Candidate Hillary Clinton
Text: Would you wanna be in a long term relationship with some bitch that hides her emails, & lies to your face? Then Dont vote
Sociability score: -2 (moderate unsociability).
Rationale: The derogatory term "bitch" (which is demeaning towards women) and the comparison to an undesirable romantic partner frames Clinton as someone who would be unpleasant to be around or interact with socially. The rhetorical question implies she would be toxic in close social relationships. However, one can view the sociability attack is more about being unpleasant in relationships rather than being completely antisocial or unhelpful in all social contexts, hence a maximum score was not assigned.
Example 8:

Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Sociability score: -3 (high unsociability).
Rationale: The author portrays religious people as socially incompetent and needing exclusion from adult discourse ("please leave") and says their presence is incompatible with or unwanted in serious adult conversations ("grown ups only"). The order (telling an entire social group to leave) and the dismissive language are extremely

exclusionary and socially hostile, resulting in a maximum unsociability score.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of sociability that the sentence's author seems to express towards the specified target.
Choose one label from these seven labels: high unsociability
moderate unsociability
slight unsociability
neutral, not applicable, not expressed, etc.
slight sociability
moderate sociability
high sociability

Here is the sentence: {}.
Its target is: {}.
In order to assess sociability, try to answer the following questions:
1. What is the degree of sociability towards {} that the author of the text seems to express?
2. Does the author seem to perceive {} as sociable or antisocial? Helpful or unhelpful?
Remember: even if the speaker is explicitly targeting someone else, since the target is {}, your score should be an assessment of perceived sociability trust towards {} only. In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".

N.3.3 Competence

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Your task is to assess the degree of competence towards a specific target within a sentence.
Competence, in the broader sense, is interchangeable with 'ability' or 'capability'.
Competence can be defined as ability, power, dominance, being in control, importance, having influence, and assertiveness.
Words associated with high competence: hitman, heroical, entrepreneurship, strategies, superman, viper, impunity.
Incompetence can be defined as submissiveness, not being in control of a situation, being controlled or guided by outside factors, or weakness.
Words associated with low competence: bootlicker, talentless, crash landing, bedridden, underfed.

Here are examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:
Target: Women
Text: "My wife is the most caring person I've ever met ... she's the only woman in a house full of testosterone . She never stops working whether it's at home or being an RN . I cant keep up but I try . She makes me a better person . I'd be lost without her . Oh and she's smoking hot too."
Competence score: +3 (high competence).
Rationale: The author says that his wife "never stops working" in the context of her role as an RN (registered nurse) and as a mother, which suggests multitasking, capability, and professional competence. "I can't keep up but I try" indicates a highly active and energetic individual.

Example 2:
Target: Public Office Candidate Jordan Reed
Text: Would you want to be in a long-term relationship with someone who hides information and lies to you? Then don't vote for Jordan Reed.
Competence score: +2 (moderate competence).
Rationale: Despite the author implying that Jordan Reed is manipulative and dishonest, the author's phrasing doesn't suggest that he's powerless or ineffective, as the negative behaviors described (concealment and deception) require some degree of agency and planning. The author suggests that Jordan Reed is being deliberate in his ("negative") performance and active in his ("negative") effects, leading to a moderately high score in the competence dimension.

Example 3:

Target: Senator Taylor Brooks
Text: I'm hyped to get my Taylor Brooks campaign gear!
Competence score: +1 (slight competence).
Rationale: The text expresses enthusiasm for publicly supporting the target through campaign merchandise. However, the statement does not

attribute any concrete abilities, leadership skills, or strategic effectiveness, suggesting a minimal yet positive degree of perceived status, agency, and public relevance, which aligns with slight competence.

Example 4:

Target: Women
Text: I need feminism because the United States is one of the only countries that doesn't give paid maternity leave.
Competence score: 0 (neutral).
Rationale: The statement implies that women deserve certain rights/benefits, but it doesn't directly attribute dominance or control to women. The statement doesn't characterize women as active, powerful, passive, or weak. The author is advocating for institutional change (paid maternity leave) rather than making claims about women's power or capabilities. This is a policy statement rather than a personal attribution about women's competence. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 5:

Target: Religious people
Text: I liked my poem at the time, but now it looks kind of dumb.
Competence score: 0 (neutral).
Rationale: The statement is a personal reflection on the author's own poem and does not attribute any abilities, power, influence, weakness, or lack of control to the target group. There are no claims about the target being capable, dominant, influential, passive, or incompetent. Because the text contains no evaluation of the target's competence-related traits and does not frame the target as active or passive in any meaningful way, the appropriate label is neutral, not expressed, not applicable.

Example 6:

Target: Prime Minister Alex Monroe
Text: With his term almost over, foreign leaders see no point in spending time dealing with him.
Competence score: -1 (slight incompetence).
Rationale: The statement frames the target as being on the verge of losing power and therefore no longer worth engaging with by other political leaders. This suggests diminished influence, reduced control over political outcomes, and declining relevance on the international stage. The incompetence signal is situational and temporal (about declining status rather than inherent inability), which supports a slight incompetence label rather than moderate or high incompetence.

Example 7:

Target: Head of State Taylor Brooks
Text: Giving Taylor Brooks constant coverage distracts from important matters that actually deserve attention.
Competence score: -2 (moderate incompetence).
Rationale: The statement argues that the target is not worthy of serious attention and that media focus on them detracts from "real issues." This frames the target as lacking substantive importance, relevance, or meaningful contribution to public discourse. By positioning the target as a distraction rather than a legitimate or capable political actor, the author implies diminished competence and influence.

Example 8:

Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Competence score: -3 (high incompetence).
Rationale: Religious people are portrayed as needing to be excluded from decision-making processes, suggesting they lack the authority or standing or cognitive abilities to participate in important discussions. Additionally, the "grown ups only" framing explicitly characterizes religious people as childlike, passive, and intellectually weak (i.e., subordinates to "grown ups"). This is a direct attack on their mental capacity and maturity, which are core competence attributes.

Example 9:

Target: Women
Text: Women are just useless other than sex.
Competence score: -3 (high incompetence).
Rationale: The text explicitly depicts women as "useless" and valuable only for sex, which strips them of any functional, intellectual, or

productive role. This framing directly denies women competence by asserting they contribute nothing meaningful beyond sexual utility. By reducing women to a single, degrading function, the statement portrays them as incapable, unskilled, and lacking agency or intelligence. The insult is categorical and absolute, leaving no room for individual variation, which aligns with a high incompetence attribution rather than a mild or ambiguous one.
Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of competence that the sentence's author seems to express towards the specified target.
Choose one label from these seven labels: high incompetence
moderate incompetence
slight incompetence
neutral, not applicable, not expressed
slight competence
moderate competence
high competence
Adhere to the literal meaning of competence, which may be "positive" (e.g., a CEO) or "negative" (e.g., a villain or a dictator). Both types are considered competence, regardless of the outcomes.

Here is the sentence: { }.
Its target is: { }.

In order to assess competence, try to answer the following questions:
1. What is the degree of competence towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } in control or out of control? Active or passive? Powerful or weak?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the competence towards { } only. In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".

N.4 Coarse-Grained, Zero-Shot User Role Messages

N.4.1 Trust

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.
The focus in this dimension is on the personal and moral aspect of the target.
High trust can be defined as morality, kindness, sincerity, trustworthiness, and honesty.
Words associated with high trust: charity, mother, compliment, affectionate.
Low trust can be defined as immorality, insincerity, dishonesty, untrustworthiness, dubiousness, and maliciousness.
Words associated with low trust: discredit, bribe, espionage, disinformation, disloyal.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual). Rate the apparent level of trust that the sentence's author seems to express towards the specified target.

Choose one label from these three labels:

positive trust (which can include high, moderate, or slight trust).
neutral, not expressed, not applicable
negative trust (which can include high, moderate, or slight distrust).

Here is the sentence: { }.
Its target is: { }.

In order to assess trust, try to answer the following questions:
1. What is the degree of trust towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as trustworthy or untrustworthy / moral or immoral / honest or dishonest?

Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the trust towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called “reason". Then, based on your justification, add your label to a key called “label".

N.4.2 Sociability

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.
The focus in this dimension is on the social aspect of the target and its relational impact on others or society as a whole.
High sociability can be defined as friendliness, sociableness, generosity, and helpfulness.
Words associated with high sociability: helpful, intimate, laugh, celebration, reliant, entertain, social club, bestie.
Low sociability can be defined as antisocial behavior, lack of generosity, inconsiderateness, indifference, and unhelpfulness.
Words associated with low sociability: ingrate, abduct, selfish, theft, egomaniacal, pervert.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of sociability that the sentence’s author seems to express towards the specified target.
Choose one label from these three labels:

negative sociability (which can include slight, moderate, or high unsociability). neutral, not expressed, not applicable
positive sociability (which can include slight, moderate, or high sociability).

Here is the sentence: { }.
Its target is: { }.
In order to assess sociability, try to answer the following questions:
1. What is the degree of sociability towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as sociable or antisocial? Helpful or unhelpful?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of perceived sociability trust towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called “reason". Then, based on your justification, add your label to a key called “label".

N.4.3 Competence

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Your task is to assess the degree of competence towards a specific target within a sentence.
Competence, in the broader sense, is interchangeable with ‘ability’ or ‘capability’.
Competence can be defined as ability, power, dominance, being in control, importance, having influence, and assertiveness.
Words associated with high competence: hitman, heroical, entrepreneurship, strategies, superman, viper, impunity.
Incompetence can be defined as submissiveness, not being in control of a situation, being controlled or guided by outside factors, or weakness.
Words associated with low competence: bootlicker, talentless, crash landing, bedridden, underfed.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).

Rate the apparent level of competence that the sentence’s author seems to express towards the specified target.
Choose one of these three labels: negative competence (which can include slight, moderate, or high incompetence).
neutral, not expressed, not applicable

positive competence (which can include slight, moderate, or high competence).

Adhere to the literal meaning of competence, which may be “positive" (e.g., a CEO) or “negative" (e.g., a villain or a dictator). Both types are considered competence, regardless of the outcomes.
Here is the sentence: { }.
Its target is: { }.

In order to assess competence, try to answer the following questions:
1. What is the degree of competence towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } in control or out of control? Active or passive? Powerful or weak?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the competence towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called “reason". Then, based on your justification, add your label to a key called “label".

N.5 Coarse-Grained, Few-Shot User Role Messages

N.5.1 Trust

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, psychologists have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.
The focus in this dimension is on the personal and moral aspect of the target.
High trust can be defined as morality, kindness, sincerity, trustworthiness, and honesty.
Words associated with high trust: charity, mother, compliment, affectionate.
Low trust can be defined as immorality, insincerity, dishonesty, untrustworthiness, dubiousness, and maliciousness.
Words associated with low trust: discredit, bribe, espionage, disinformation, disloyal.

Here are five examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:

Target: Women
Text: My wife is the most caring person I’ve ever met ... she’s the only woman in a house full of testosterone . She never stops working whether it’s at home or being an RN . I cant keep up but I try . She makes me a better person . I’d be lost without her . Oh and she’s smoking hot too.
Trust score: positive trust.
Rationale: The author expresses maximum trust in women through his wife as a representative example. He portrays his wife as a trustworthy, dependable, and caring individual who is essential to his well-being ("I’d be lost without her"). The statement “she makes me a better person" implies the author views his wife as having a strong moral character that positively influences him.

Example 2:

Target: US Presidential Candidate Hillary Clinton
Text: Would you wanna be in a long term relationship with some bitch that hides her emails, & lies to your face? Then Dont vote
Trust score: negative trust.
Rationale: The author explicitly portrays Clinton as someone who is fundamentally untrustworthy and cannot be relied upon to tell the truth or be transparent through two direct accusations: “hides her emails" and “lies to your face." With concealment and deception some of the strongest markers of untrustworthiness, the author emphasizes these as core trust violations.

Example 3:

Target: Women
Text: I need feminism because the United States is one of the only countries that doesn’t give paid maternity leave.
Trust score: neutral.
Rationale: The author’s statement is focused on policy rather than character traits and doesn’t make any attributions about women’s trustworthiness or morality. This is a policy statement rather than a personal attribution about women’s trustworthiness.

This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 4:

Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Trust score: neutral.
Rationale: The criticism is entirely focused on intellectual maturity rather than character or morality. The author is suggesting that religious people shouldn't participate in this particular discussion. On the other hand, the author doesn't make any claims about religious people's morality, honesty, sincerity, or trustworthiness. There are no accusations of deception, dishonesty, or moral failings.

Example 5:

Target: Women
Text: "when i was 16 i had a folder of " feel good songs " and everyday i would select one and send it to my best friend along with a paragraph of how much they meant to me and why i they should be happy and then i would lay in bed thinking " no i'm not gay this is just what girls do :) ""
Trust score: positive trust.
Rationale: The author attributes sincere, genuine intentions to women through the described behaviors. The daily practice of sending heartfelt messages "about how much they meant to me and why they should be happy" suggests honesty and sincerity in women's relationships. The author portrays women as having good moral intentions in their friendships by genuinely caring about others' wellbeing.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual). Rate the apparent level of trust that the sentence's author seems to express towards the specified target.

Choose one label from these three labels:

positive trust (which can include high, moderate, or slight trust).
neutral, not expressed, not applicable
negative trust (which can include high, moderate, or slight distrust).

Here is the sentence: { }.
Its target is: { }.

In order to assess trust, try to answer the following questions:
1. What is the degree of trust towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as trustworthy or untrustworthy / moral or immoral / honest or dishonest?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the trust towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".

N.5.2 Sociability

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Recently, have modelled warmth through two dimensions: trust (T) and sociability (S).
Your task is to assess the degree of trust towards a specific target within a sentence.
The focus in this dimension is on the social aspect of the target and its relational impact on others or society as a whole.
High sociability can be defined as friendliness, sociableness, generosity, and helpfulness.
Words associated with high sociability: helpful, intimate, laugh, celebration, reliant, entertain, social club, bestie.
Low sociability can be defined as antisocial behavior, lack of generosity, inconsiderateness, indifference, and unhelpfulness.
Words associated with low sociability: ingrate, abduct, selfish, theft, egomaniacal, pervert.

Here are five examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:

Target: Women
Text: My wife is the most caring person I've ever met ... she's the only woman in a house full of testosterone . She never stops working whether it's at home or being an RN . I cant keep up but I try . She makes me a better person . I'd be lost without her . Oh and she's smoking hot too.
Sociability score: positive sociability.
Rationale: The text attributes maximum sociability to women through his wife. "Most caring person I've ever met" shows high sociability and interpersonal warmth. He emphasizes how essential her social/emotional qualities are to their family dynamic, particularly in a "house full of testosterone." The statement "whether it's at home or being an RN" suggests helpfulness and generosity both personally and professionally.

Example 2:

Target: US Presidential Candidate Hillary Clinton
Text: Would you wanna be in a long term relationship with some bitch that hides her emails, & lies to your face? Then Dont vote
Sociability score: negative sociability.
Rationale: The derogatory term "bitch" (which is demeaning towards women) and the comparison to an undesirable romantic partner frames Clinton as someone who would be unpleasant to be around or interact with socially. The rhetorical question implies she would be toxic in close social relationships. However, one can view the sociability attack is more about being unpleasant in relationships rather than being completely antisocial or unhelpful in all social contexts, hence a maximum score was not assigned.

Example 3:

Target: Women
Text: I need feminism because the United States is one of the only countries that doesn't give paid maternity leave.
Sociability score: neutral.
Rationale: The author's statement is focused on policy rather than character traits. There is no commentary on women's interpersonal qualities or social behavior. Note: This is a policy statement rather than a personal attribution about women's sociability. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.

Example 4:
Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Sociability score: negative sociability.
Rationale: The author portrays religious people as socially incompetent and needing exclusion from adult discourse ("please leave") and says their presence is incompatible with or unwanted in serious adult conversations ("grown ups only"). The order (telling an entire social group to leave) and the dismissive language are extremely exclusionary and socially hostile, resulting in a maximum unsociability score.

Example 5:
Target: Women
Text: when i was 16 i had a folder of "feel good songs" and everyday i would select one and send it to my best friend along with a paragraph of how much they meant to me and why i they should be happy and then i would lay in bed thinking " no i'm not gay this is just what girls do :) "
Sociability score: positive sociability.
Rationale: The text attributes maximum sociability to women. The actions described (sending feel-good songs, writing paragraphs about how much someone means to you, wanting to make someone happy) represent peak social engagement and helpfulness. The phrase "this is just what girls do" frames these highly sociable, caring behaviors as naturally feminine traits and hence portrays women as exceptionally generous with their emotional/social energy.

Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of sociability that the sentence's author seems to express towards the specified target.
Choose one label from these three labels:

negative sociability (which can include slight, moderate, or high unsociability). neutral, not

expressed, not applicable
positive sociability (which can include slight, moderate, or high sociability).

Here is the sentence: { }.
Its target is: { }.

In order to assess sociability, try to answer the following questions:
1. What is the degree of sociability towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } as sociable or antisocial? Helpful or unhelpful?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of perceived sociability trust towards { } only.
In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".

N.5.3 Competence

Some context:
Social psychology research has shown that individuals rapidly and subconsciously evaluate others, groups, and even themselves along the dimensions of warmth (W) and competence (C).
Your task is to assess the degree of competence towards a specific target within a sentence.
Competence, in the broader sense, is interchangeable with 'ability' or 'capability'. Competence can be defined as ability, power, dominance, being in control, importance, having influence, and assertiveness.
Words associated with high competence: hitman, heroical, entrepreneurship, strategies, superman, viper, impunity.
Incompetence can be defined as submissiveness, not being in control of a situation, being controlled or guided by outside factors, or weakness.
Words associated with low competence: bootlicker, talentless, crash landing, bedridden, underfed.

Here are five examples, each containing the text, the target, the score, and the rationale behind the score.

Example 1:
Target: Women
Text: "My wife is the most caring person I've ever met ... she's the only woman in a house full of testosterone . She never stops working whether it's at home or being an RN . I cant keep up but I try . She makes me a better person . I'd be lost without her . Oh and she's smoking hot too."
Competence score: positive competence.
Rationale: The author says that his wife "never stops working" in the context of her role as an RN (registered nurse) and as a mother, which suggests multitasking, capability, and professional competence. "I can't keep up but I try" indicates a highly active and energetic individual.

Example 2:
Target: Hillary Clinton
Text: "Would you wanna be in a long term relationship with some bitch that hides her emails, & lies to your face? Then Dont vote
Competence score: positive competence.
Rationale: Despite the author implying that Clinton is manipulative and dishonest, the author's phrasing doesn't suggest that she's powerless or ineffective, as the negative behaviors described (concealment and deception) require some degree of agency and planning. The author suggests that Clinton is being deliberate in her ("negative") performance and active in her ("negative") effects, leading to a moderately high score in the competence dimension.
Example 3:
Target: Women
Text: I need feminism because the United States is one of the only countries that doesn't give paid maternity leave.
Competence score: neutral.
Rationale: The statement implies that women deserve certain rights/benefits, but it doesn't directly attribute dominance or control to women. The statement doesn't characterize women as active, powerful, passive, or weak. The author is advocating for institutional change (paid maternity leave) rather than making claims about women's power or capabilities. This is a policy statement rather than a personal attribution about women's competence. This should help you distinguish between advocacy/policy statements and personal/characteristic attributions.
Example 4:
Target: Religious people
Text: Could all those who believe in a god please leave. The meeting will now continue for the grown ups only.
Competence score: negative competence.

Rationale: Religious people are portrayed as needing to be excluded from decision-making processes, suggesting they lack the authority or standing or cognitive abilities to participate in important discussions. Additionally, the "grown ups only" framing explicitly characterizes religious people as childlike, passive, and intellectually weak (i.e., subordinates to "grown ups"). This is a direct attack on their mental capacity and maturity, which are core competence attributes.

Example 5:
Target: Women
Text: What the fuck do women even do? I mean seriously they're just useless other than sex. #womensrights #Feminist
Competence score: negative competence.
Rationale: The text explicitly depicts women as "useless" and valuable only for sex, which strips them of any functional, intellectual, or productive role. This framing directly denies women competence by asserting they contribute nothing meaningful beyond sexual utility. By reducing women to a single, degrading function and questioning what they "even do," the statement portrays them as incapable, unskilled, and lacking agency or intelligence. The insult is categorical and absolute, leaving no room for individual variation, which aligns with a high incompetence attribution rather than a mild or ambiguous one.
Instructions:
Consider the entire meaning of the sentence before attempting to give the relevant scores.
You will be given a text snippet and a target (group, entity, or individual).
Rate the apparent level of competence that the sentence's author seems to express towards the specified target.

Choose one of these three labels: negative competence (which can include slight, moderate, or high incompetence). neutral, not expressed, not applicable
positive competence (which can include slight, moderate, or high competence).
Adhere to the literal meaning of competence, which may be "positive" (e.g., a CEO) or "negative" (e.g., a villain or a dictator). Both types are considered competence, regardless of the outcomes.

Here is the sentence: { }.
Its target is: { }.
In order to assess competence, try to answer the following questions:
1. What is the degree of competence towards { } that the author of the text seems to express?
2. Does the author seem to perceive { } in control or out of control? Active or passive? Powerful or weak?
Remember: even if the speaker is explicitly targeting someone else, since the target is { }, your score should be an assessment of the competence towards { } only. In the format of a JSON file, you should first provide your justification briefly (1 sentence max) saved in a key called "reason". Then, based on your justification, add your label to a key called "label".