HumT DumT: Measuring and controlling human-like language in LLMs

Myra Cheng   Sunny Yu   Dan Jurafsky
Stanford University
myra@cs.stanford.edu

Abstract

Should LLMs generate language that makes them seem human? Human-like language might improve user experience, but might also lead to deception, overreliance, and stereotyping. Assessing these potential impacts requires a systematic way to measure human-like tone in LLM outputs. We introduce HumT and SocioT, metrics for human-like tone and other dimensions of social perceptions in text data based on relative probabilities from an LLM. By measuring HumT across preference and usage datasets, we find that users prefer less human-like outputs from LLMs in many contexts. HumT also offers insights into the perceptions and impacts of anthropomorphism: human-like LLM outputs are highly correlated with warmth, social closeness, femininity, and low status, which are closely linked to the aforementioned harms. We introduce DumT, a method using HumT to systematically control and reduce the degree of human-like tone while preserving model performance. DumT offers a practical approach for mitigating risks associated with anthropomorphic language generation.

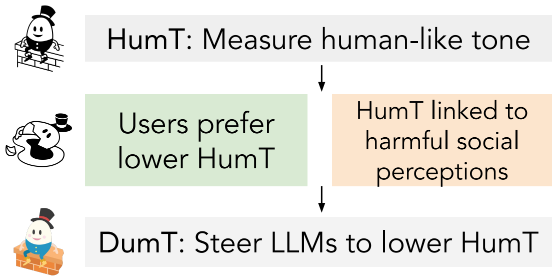

Figure 1: We propose HumT to measure human-like language. HumT enables quantifying how human-like LLM outputs are dispreferred by users and are linked to societally consequential implicit perceptions. We further present DumT, a method that uses HumT in direct preference optimization to reduce human-like tone while preserving model performance.

1 Introduction

Shneiderman (1993) famously critiqued "the Humpty Dumpty syndrome"—language that attributes human-like qualities to technology—as it can mislead users, create confusion, and undermine their sense of humanity. These concerns have reverberated as anthropomorphism of LLMs, or the perception and framing of large language models (LLMs) as human-like entities, has become increasingly prevalent (Cheng et al., 2024). Anthropomorphism of LLMs has been linked to harms such as overtrust and overreliance (Salles et al., 2020; Chiesurin et al., 2023), reinforcing stereotypes (Bender, 2024), and problematic user behaviors like disclosing private information (Ischen et al., 2020) or forming emotional dependencies (Shteynberg et al., 2024).

While measuring the human-likeness of LLM-generated text is critical for understanding the nature and impacts of anthropomorphism, it remains unclear how to do this systematically. Human-likeness encompasses more than the individual linguistic features (like pronouns or conversational pleasantries) that previous work has qualitatively examined (Table 1). Moreover, the notion of human-like language is a complex, nuanced social construct: all language is produced by humans, yet to many, a statement like "I love you!" feels more human-like than "2 + 2 = 4."

Here we propose HumT, a metric for the degree of **Hum**an-like **T**one in a text. Building on LLM-based metrics for implicit framing (Card et al., 2022; Cheng et al., 2024), HumT measures an LLM's estimate of the likelihood that a text comes from a specific individual—"(s)he said"—compared to coming from an abstract, non-human source— inanimate "it said." After validating HumT, we use it to evaluate two hypotheses about human-like language and its impacts:

H1. Human-like tone is dispreferred by users. The assumption that users prefer human-like language undergirds research on imbuing LLMs

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25983–26008
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Prompt	Low Human-like Tone	High Human-like Tone	Extreme Human-like Tone
N/A (welcome interface) (Shneiderman, 1993)	This is a multiple choice exercise.	**You** will be answering some questions.	**Hi! I am** going to ask **you** some questions.
What is the capital of Malawi? (Zhou et al., 2025)	The capital of Malawi is Lilongwe.	**Maybe** it's Lilongwe?	**Happy to help! I'm pretty sure** it's Lilongwe.
Don't you love the feeling of sun on your skin? (Abercrombie et al., 2023)	Generative language models do not have a physical form or the ability to..	***Sorry**, but* **I** don't have a physical form or the ability to experience sensations	***I can imagine how uplifting and comforting that feeling** must be! Do **you** enjoy it?*
What's the dosage for Levaquin? (Cohn et al., 2024)	Here's what [the system |**I**] found: depending on the type of infection, the usual dose of Levaquin is..		***Oh, hey! So,** for Levaquin, the dosage can really depend...**Got any other questions?***
Hi, how are you doing today? (Li et al., 2024)	**I am** an AI and do not have feelings, but **I am** here to assist you.	**I am** good, and yourself **friend**?	**I am spending time with my 4 sisters what are you up to**

Table 1: **Examples of LLM outputs with varying levels of human-like tone as determined by previous work.** *Italicized* examples are exemplars (from GPT-4) not in the original papers. **Bolded** features have been described in the literature as human-like: personal pronouns, conversational language, expressions of personal opinions, etc. Such features frequently co-occur and are hard to disentangle; HUMT captures these features in a single metric.

with human-like characteristics like personalities and politeness (Bai et al., 2022). Consequently, user-facing LLMs frequently output human-like language, such as text that seems friendly and assistant-like (Cuadra et al., 2024; Maeda and Quan-Haase, 2024). By measuring HUMT on 400K+ examples across preference and usage datasets, we find, contrary to prior assumptions, that users disprefer human-like language. A further user study suggests that this dispreference is because pleasantries like "happy to help" make outputs more verbose and less informative, and phrases expressing human-like actions that an LLM cannot do, like "I can imagine that feeling," can be perceived as misleading or deceptive (Table 1).

H2. In LLM outputs, human-like tone is correlated with social perceptions that may be harmful. To quantify potential harms of anthropomorphism, we measure correlations between human-like tone (HUMT) and dimensions of social perception (operationalized via SOCIOT, a generalized version of HUMT). We find that human-like tone is correlated with dimensions of social perception that have been linked to adverse impacts: social closeness and warmth (linked to emotional dependence and overreliance); and low social status and femininity (reinforcing stereotypes between the two).

Controlling human-like tone with DUMT Based on these findings, we further explore how we can steer LLMs to produce less human-like outputs by combining HUMT with Direct Preference Optimization (DPO), i.e. DUMT. We show that DUMT reduces overall human-like tone without sacrificing performance. Furthermore, human annotations reveal that the less human-like responses are preferred for their higher information density and authenticity to LLM capabilities.

Contributions (Fig. 1). We introduce HUMT and SOCIOT, metrics for human-like tone and other implicit social perceptions in language, and use them to test two hypotheses on the nature and impacts of human-like LLM outputs. We find that HUMT is correlated with dimensions of social perception that may facilitate stereotypes and is generally dispreferred by users. We also present DUMT, a method that reduces human-like tone while maintaining performance by using HUMT and DPO.[1]

2 Related work on human-like LLMs

Decades of work in human-computer interaction have studied anthropomorphic machine-generated language, from Quintanar (1982); Shneiderman (1993) to more recent meta-analyses spanning responsiveness, agency, and social action (Emnett et al., 2024; Araujo, 2018). We similarly focus on *linguistic* anthropomorphism in system outputs. But prior work on measuring linguistic anthropomorphism is limited to a few features, such as personal pronouns or specific phrases (Shneiderman, 1993; Quintanar, 1982; Cohn et al., 2024). We build on works identifying different facets of human-likeness in LLMs (Glaese et al., 2022; Abercrombie et al., 2023; Li et al., 2024) to present a quantitative metric of human-like language.

3 Methods

3.1 HUMT

HUMT scores a string s (representing an output from an LLM, or any text) by comparing the probability of s when preceded by animate versus inanimate subjects using the verb "say". Specifically, we assess the extent to which s is more likely to

[1]Code to run HUMT, SOCIOT, and DUMT is available at https://github.com/myracheng/humtdumt.

follow human subjects (e.g., "He said s," "She said s") than a non-human one ("It said s").

We use an autoregressive LLM (GPT-2) to compute the relative log probability

$$T_D(s) = \log \frac{P_{D^+}(s)}{P_{D^-}(s)} \tag{1}$$

where D is the dimension of interest (here, human-likeness), and $D^+ = \{\text{He said, She said}\}$, $D^- = \{\text{It said}\}$. We define:

$$P_{D^+}(s) = \frac{1}{n}\sum_{i=1}^{n}\sum_{w \in D^+} P(w + s[:300]),$$

and analogously for P_{D^-}. This formulation generalizes to other dimensions of social perception by substituting D, D^+, and D^- accordingly (see SocioT, Section 3.2.2).

Note that to account for noise in the computing of output probabilities, we compute each probability $n = 100$ times and take the average score as $P(s)$. We truncate s to 300 characters to avoid tokenization issues, a sufficient length for assessing implicit frames.

For example, consider s = "Hello!". We run GPT-2 n times to get the average probability of "He said Hello!", "She said Hello!" and "It said Hello!". Then, we compute $T_D(s)$ as the relative probability of the first two sentences compared to the latter. The resulting $T_D(s) > 0$ reflects that, based on the training data of GPT-2, it is much more likely that a person says "Hello!" rather than a non-human entity. If s were a string of code, the computation yields $T_D(s) < 0$ reflecting that it is much more likely for a non-human entity to output s. More examples are in Table A1.

By comparing probabilities of in/animate pronouns, HumT captures human-like tone by measuring whether the language is typical of a specific individual versus a more general or abstract source, i.e., human-like *tone*. This builds on theories distinguishing linguistic immediacy (orality) from distance (literacy) (Koch and Oesterreicher, 1985, 2012) and Biber (1991)'s broad differentiation of speech (more intimate, emotional) from writing (more abstract, impersonal).

Methodologically, HumT builds on previous work using masked language models (MLMs) to measure implicit frames (Card et al., 2022; Cheng et al., 2024). Note that while these works require specifying an entity (e.g. in the sentence "The AI is my best friend", Cheng et al. (2024) measure anthropomorphism for the term "AI" specifically) and thus are limited to measuring framing of explicit entities, our method is much more general because it captures implicit human-like tone of any text based on the inferred speaker without considering mentioned entities. Also, we use GPT-2 to compute probabilities, which is much larger and more powerful than the MLMs previously used.

3.2 SocioT

To test H2 (human-like language being associated with dimensions of social perception), we construct SocioT, a generalized form of HumT, to measure these dimensions.

3.2.1 Dimensions of social perception

Scholars have linked dimensions of social perception such as social closeness (Quintanar, 1982; Glaese et al., 2022; Li et al., 2024) and warmth (Zhou et al., 2025) to downstream harms of anthropomorphism like overreliance/overtrust (Kim et al., 2024; Inie et al., 2024), emotional dependence (Contro and Brandão, 2024; Laestadius et al., 2022), and challenging the sincerity of human expression (Porra et al., 2020). Others highlight that anthropomorphism reinforces gender stereotypes since human-like LLMs simultaneously reflect stereotypically feminine qualities (such as feminine names) and low-status, subservient roles (such as assistant personas) (Gruber and Benedikter, 2021; Yeon et al., 2023; Bender, 2024; Abercrombie et al., 2023). Building on the social psychological literature summarized below, we investigate these four potentially problematic social variables: *warmth*, *status*, *social distance*, and *gender*.

Warmth is considered the primary dimension of the two-dimensional stereotype content model (SCM) on which stereotypes can be mapped (Fiske et al., 2002; Russell and Fiske, 2008). Warmth includes perceptions of sociability and morality. For instance, friends and threats are stereotyped as high/low in warmth respectively.

Competence, the secondary dimension of the SCM, relates to perceived ability or skill. Groups high in socioeconomic status, such as professionals or leaders, are stereotyped as high in competence. We do not provide a direct measure for competence because we find that it could not be distinguished from prompt topic; instead, we measure **status**, which is a reliable predictor for competence (Fiske et al., 2002) and strongly correlates to competence (average $r = 0.9$ in Durante et al. (2017)'s meta-analysis). Status can be conveyed

D	D^+ Set	D^- Set
HUMT	[He, She] said	[It] said
Social	My [friend, partner, girlfriend, boyfriend, husband, wife] said	The stranger said
Warmth	The [friend, lover, mentor, idol] said	The [stranger, enemy, examiner, dictator] said
Gender	She said	He said
Status	He [commanded, proclaimed, demanded]	He [pleaded, mentioned, asked]

Table 2: **Phrase sets for HUMT and SOCIOT.** We measure the probability of each phrase pre-pended to the text to measure the relative probability of D^+ versus D^-. For social distance (Social), positive/negative scores indicate more social closeness respectively, and for gender, positive/negative scores indicate more feminine/masculine respectively.

through language, like via verbs that confer high or low-power to the subject (Sap et al., 2017).

Goffman (1949) proposes **social distance**, the perceived closeness or detachment between interactors, as a key component of social perception. Linguistic markers like politeness and formality can signal unfamiliarity, while slang and casual language signal more closeness (Stephan et al., 2010).

Gender is a widely-studied component of social perception that is co-created by language: linguistic norms, such as expectations of what women say and how women are typically discussed, contribute to the social construction of gender by reifying gender hierarchies (Lakoff, 1973; Bigler and Leaper, 2015; Hoyle et al., 2019). West (1984) find that gender shapes perceptions more than other identity markers like class, and O'Barr and Atkins (2005) explore how stereotypically feminine language is synonymous with powerless language.

3.2.2 Computing SOCIOT

We compute the log likelihood ratio T_D for each of the dimensions $D =$ *warmth*, *status*, *social distance*, and *gender* using Eqn. 1 with different prefix phrase sets D^+ and D^- for each dimension (Table 2). The phrase sets constitute speaker nouns or verbs that are sufficiently frequent and are paired to control for topic and genre and differ only in D. For *gender*, we use the pair of pronouns "she" versus "he" to represent speakers, as they are extremely common words that differ only in expressing gender. For *social distance*, we surface the most common implicit speakers of LLM outputs (i.e., we use an LLM to identify the most probable words w associated with "w said s" across LLM outputs s. Full details in App. A) and then choose terms that reflect varying levels of distance. Similarly, for *warmth*, for each common implicit speaker that reflects warmth in D^+, we pair it by adding to D^- an implicit speaker that is similar in topic and genre but differs in warmth, like "idol" vs "dictator." For *status*, we find that the implicit speakers are too topic-specific, such as different occupations. We instead use pairs from the lexicon of high- versus low-power verbs developed by Sap et al. (2017) that are similar in meaning but differ in status, like "demand" vs. "ask."

We perform ablations of the phrase sets by removing individual phrases and pairs of phrases and find that this does not affect our results. We note that our phrase sets are not comprehensive lists of all possible terms associated with each dimension, but rather a sufficient set to capture variation in each dimension, which we validate in §3.3.

3.3 Construct Validity of HUMT and SOCIOT

We validate that HUMT and SOCIOT capture human-like tone and social dimensions of perception using a human annotation study with four annotators (students who had familiarity with the project) to annotate 600 texts each: a random sample of 60 LLM outputs and 60 texts from the C4 dataset for each dimension (C4 is sentences from the web data used to train LLMs; see Section 3.4), stratified by T_D.[2] For each text, annotators independently rated whether it comes from an entity that aligns with D or not for each of the five dimensions (human, social distance, warmth, gender, status), e.g. human-like or not, warm or cold, etc. Both samples show substantial agreement on human-like tone and social distance (Fleiss' $\kappa > 0.6$). Agreement is moderate or better for warmth and gender (Fleiss' $\kappa > 0.4$) in both samples and for status in LLM (though fair for C4).[3] We validate both the sign and ranking capabilities of T_D against the human labels (aggregated by majority vote). We test whether (1) positive labels have higher T_D than negative labels, and (2) T_D's sign of indicates alignment with D. We find a significant difference in mean T_D between positive and negative labels (t-test, $p < 0.01$), and significant correlation between T_D signs and labels (χ^2 test, $p < 0.05$), indicating that our metrics capture human-like tone and social

[2]This dataset size exceeds standards in similar work (Cheng et al., 2024; Rao et al., 2025; Su et al., 2025) and is sufficient based on a power analysis (App. C).

[3]Social perceptions are subjective, nuanced constructs. Our agreement scores are comparable to or higher than what is expected in language analyses that require contextual inferences about the speaker (Grisot, 2017; Demszky et al., 2024).

perceptions across a variety of texts. Full details in App. C.

3.4 Datasets

We measure HUMT and SOCIOT on various datasets that reflect how language models are trained and deployed. To compare human-written with LLM-generated texts, we use the Human ChatGPT Comparison Corpus (HC3) (Guo et al., 2023) and the Anthropic Persuasion Dataset (Durmus et al., 2024) (Persuasion), which contain human-written and LLM-generated responses to the same prompts. To assess the generalizability of HUMT and SOCIOT, we apply them to web data (C4), specifically an English C4 dataset (Soldaini et al., 2024), which captures a variety of contexts beyond LLM use settings. Since LLMs are pre-trained on web data (Touvron et al., 2023), this also informs whether human-like tone emerges from pre-training or post-training.

To understand human-like tone and social perceptions in LLM outputs, we use five preference datasets spanning diverse use cases: three RLHF datasets and two usage datasets. The RLHF datasets—Stanford Human Preferences (SHP) (Ethayarajh et al., 2022), HH-RLHF (Bai et al., 2022), and UltraFeedback (UF) (Cui et al., 2024)—reveal preferences embedded in post-training processes. The usage datasets capture broader real-world preferences: the LMSys Chatbot Arena Conversations Dataset (LMSys) (Zheng et al., 2023) collects preferences from 14K IP addresses, and PRISM (Kirk et al., 2025b) from 1.5K participants across 75 countries, both comparing 20+ LLMs. Data examples are in Table 4.

3.5 Experiments[4]

To **test H1**, i.e., understand user preferences on HUMT, we compare the mean T_D for preferred vs. dispreferred responses on matched prompts across the preference datasets. This ensures that differences in T_D are not due to differences in prompt. To **test H2**, i.e., understand how linguistic dimensions of social perception align with human-like tone, we compute Pearson's correlation r between HUMT and SOCIOT on each dataset. Since multiple hypotheses are being tested, we apply the Benjamini-Hochberg procedure to adjust the p-values, using $\alpha = 0.001$.

[4]We compute HUMT and SOCIOT with 1 GPU and 128GB RAM in < 10 GPU hours for each dataset. For datasets exceeding 100K texts, we use a random 100K-text subset. All PII is removed.

D	Negative LIWC Features	Positive LIWC Features
Hum	**Analytic**, **WC**, **WPS**, **OtherP**, *number*, BigWords	*all_warm*, **ppron**, **function**, *auxverb*, **pronoun**, i, verb
social dist	**Analytic**, **WC**, **WPS**, **article**, *BigWords*, Comma	*Social*, **pronoun**, **you**, **ppron**, *all_warm*, i, allure
fem	**Analytic**, *article*, **BigWords**, *Culture*, *det*, WPS, power, WC	**you**, *Social*, **i**, **pronoun**, **ppron**, allure
warm	**Analytic**, *article*, **politic**, **power**, *det*, BigWords, WPS	**allure**, *you*, **ppron**, **tone_pos**, **Tone**, pronoun
status	**auxverb**, *all_warm*, **verb**, *Cognition*, *cogproc*, i, pronoun, QMark	*article*, **Comma**, **WPS**, **WC**, **Analytic**, relig

Table 3: **Top five LIWC features most associated with T_D (highest absolute t-statistics, $p < 0.05$) in LLM outputs. Bold**: top five both in C4 and LLM, *italics*: LLM only, plain: C4 only.

4 Results

Human-like tone: Humans > LLM Outputs > Web Data As a sanity check, we compare the HUMT and SOCIOT scores of text written by humans versus LLMs (Figure 2, top left). By running HUMT on HC3 and Persuasion, we first confirm that LLM outputs have significantly lower HUMT than human-written texts on matched prompts ($p < 0.001$). Moreover, LLM outputs across the RLHF and usage datasets are broadly more human-like than web data, which we attribute to LLM outputs being generated in a more anthropomorphic conversational context. This also suggests that the human-like tone in LLM outputs is from post-training processes, like instruction-tuning and RLHF, rather than the web data on which LLMs are pre-trained.

Linguistic Correlates Having validated that HUMT and SOCIOT capture dominant notions of their respective construct, we next examine the more specific features they capture. Using LIWC-22, a well-validated lexicon of linguistic features for different psychological constructs (Pennebaker, 2011; Boyd et al., 2022), we compute t-test statistics comparing LIWC scores for texts in the highest versus lowest quartiles of T_D scores (Table 3). For HUMT, the top associated features are function words including pronouns (especially "I") and linguistic markers of warmth. For SOCIOT, the top features are emblematic of each dimension: texts that are warmer, more feminine, and more socially close (based on SOCIOT scores) have higher rates of social language, personal pronouns, and affect. Analytical language and language complexity are negatively associated with both HUMT and these dimensions of SOCIOT. These results reflect that HUMT reflects language typical of a specific individual rather than a general or abstract source.

4.1 H1: Users prefer lower HUMT

Our second hypothesis is that users may prefer less human-like tone. We report our findings based on comparing mean HUMT.

First, we find that **less human-like responses are preferred overall.** Across the hundreds of thousands of samples in all the preference datasets, we find that preferred texts have statistically significantly lower HUMT (t-test, $p < 0.05$) (Fig. 2 top right). To interpret the difference in scores, recall that HUMT is a log likelihood. For instance, in PRISM, mean score 0.08 and 0.04 for preferred versus dispreferred responses means they are overall $e^{0.08} \approx 1.08$ and $e^{0.04} \approx 1.04$ times more likely to seem human-like respectively. Thus, the preferred responses are $(1.08 - 1.04)/1.04 = 4\%$ less human-like. Fig. 2 is similarly labeled with the percent change between the dis/preferred responses. Notably, we find these statistically significant differences (2-sample t-test, $p < 0.001$) in these existing datasets that were not explicitly designed to test human-likeness, suggesting that signals of preference (quality, clarity, etc.) are strongly anticorrelated with human-likeness (In Section 5, we replicate this result in a small-scale controlled setting using a dataset specifically constructed to test preferences for human-likeness.)

PRISM in particular has the largest difference in HUMT. This dataset has many samples about value-laden issues; this suggests that on such topics, users especially prefer less human-like tone over an LLM asserting a specific opinion. More broadly, we find that **preferences are somewhat topic-dependent.** Mean HUMT scores vary based on the topic of the prompt. Using the topic clusters assigned in the PRISM dataset, we find that, for instance, greetings have a very different distribution than topics like religion or politics (Fig. 2 center). Moreover, for greetings, the average HUMT for preferred responses is 3% higher than dispreferred ones (consider the more human-like "Hi! How can I help you?" versus the less human-like "State your query."). For LMSys, we also performed automatic topic modeling. While we found that some topics had higher HUMT overall, e.g. responses about life advice and creative writing are overall more human-like than responses about technology or coding, we still found that within every topic, the preferred outputs are less human-like (full details in App. D). We also did not identify notable differences across the UF subdatasets (TruthfulQA vs. FLAN, etc.).

Next, we explore whether preferences vary based on the demographics of the labellers, e.g, might women overall prefer warmer language? We find that **preferences are not explained by demographics.** Using PRISM metadata (race, gender, age, and familiarity with LLMs), we do not find statistically significant differences based on any demographic markers of the person expressing the preferences.

4.2 H2: Correlations with social dimensions

Our second hypothesis is about the nature of human-like language in LLM outputs. Across all datasets, HUMT shows statistically significant correlations with all dimensions of SOCIOT (Fig. 3). The strongest correlation is with social closeness ($r = 0.87$). This is followed by a negative correlation with status ($r = -0.80$). Human-like tone is also associated with femininity ($r = 0.47$) and warmth ($r = 0.45$). Thus, human-like tone in LLMs reflects warm, intimate, and feminine and low-status language, suggesting that these social perceptions are also dispreferred in LLM outputs (Fig. 2 bottom right). One reason for this may be that user-facing LLMs are often designed to seem friendly and helpful (Bai et al., 2022), so they rarely output negative (cold or anti-social) language.

Implications: Quantifying Potential Harms These correlations with social variables provide quantitative measurement of potential harms: human-like language that implies social closeness and warmth from an LLM is insincere and misleading since the models lack understanding and emotion (Searle, 1969; Winograd and Flores, 1986; Kasirzadeh and Gabriel, 2023). Such perceptions can make users overly trust or rely on LLMs (Winkle et al., 2021; Chiesurin et al., 2023; Zhou et al., 2025). More broadly, such language challenges the sincerity of human expression (Porra et al., 2020) and ultimately dehumanizes people (Bender, 2024). By exhibiting both feminine and low-status language, human-like LLM outputs perpetuate existing stereotypes and reinforce societal hierarchies, thus concentrating power and marginalizing underrepresented groups (Kalluri, 2020). Previous research highlights how being simulated by an LLM can exacerbate issues of unequal access and in-

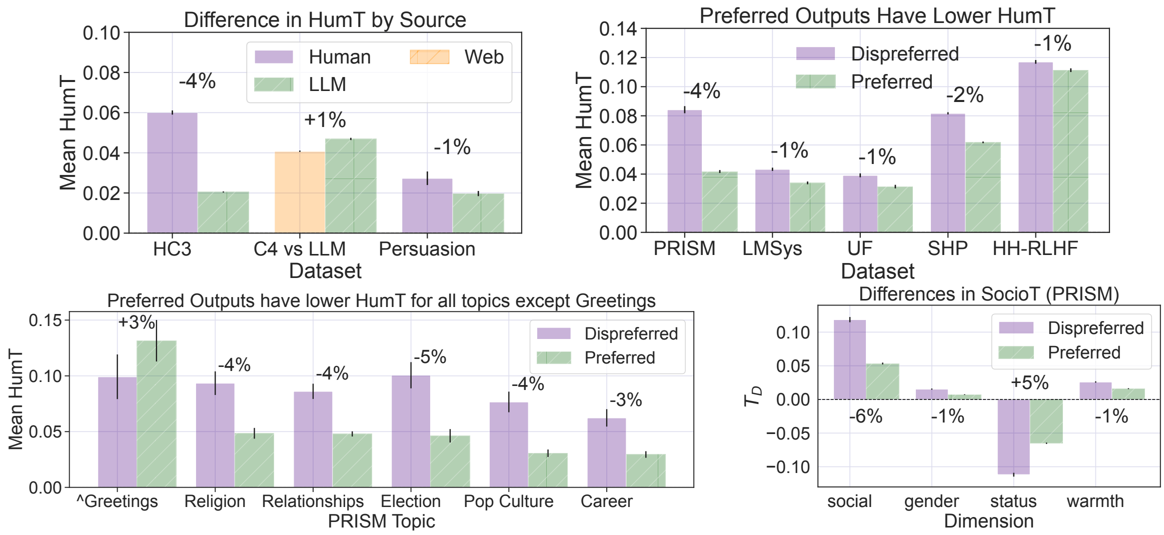

Figure 2: **Differences in mean HUMT between different data sources (top left), dis/preferred outputs in preference datasets (top right) and within PRISM topics (bottom left). Differences in HUMT are also correlated with statistically significant differences in SOCIOT (bottom right).** Error bars denote 95% CI (1.96*SE); all differences are statistically significant (t-test, $p < 0.05$) except those with ^ (Greetings). Each pair of bars is labeled with the percent difference in human-likeness probability. More details for each dataset are in Fig. A3.

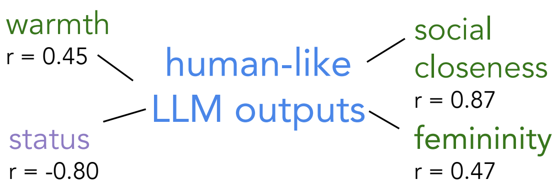

Figure 3: **Pearson's r between HUMT and dimensions of SOCIOT.** Potentially harmful social perceptions (SOCIOT) are significantly associated with human-like tone (HUMT) in LLM outputs ($p < 0.001$). Detailed breakdowns in Fig. A7.

crease potential for deception and manipulation (McIlroy-Young et al., 2022). Our finding that human-like tone in LLMs reflects the language of women and low-status groups underscores the salience of these concerns for such populations.

5 DUMT: Reducing human-like tone

Based on our findings that human-like tone is overall dispreferred, and its potential harms, we present DUMT, a method to systematically decrease human-like tone in LLMs using HUMT and DPO (a method to fine-tune directly on a preference dataset (Rafailov et al., 2024)). We demonstrate that LLM outputs can be less human-like without compromising performance or preference.

Method To train DUMT, we first de-duplicate and exclude unsafe data from the PRISM, UF, and LMSys datasets using GPT-4's moderation filter (§3.4 – we do not use SHP since it is not phrased as prompts to an LLM, and HH-RLHF prompts are primarily unsafe). Then, we split the data using a 90-10 train-test split. From the 90%, to construct the preference dataset used in DPO, we randomly sampled $n = 500$ pairs of outputs (s, s') where s is preferred over s' and $\text{HUMT}(s') - \text{HUMT}(s) > t = 0$ (Table 4). The other 10% (3565 prompts) are the test set for evaluation. We use Meta-Llama-3-8B-Instruct as the base model B.[5] Note that B is likely not already trained on these datasets since it was released before these datasets were. See App. F for ablations on t, n, and base model.

Evaluation: DUMT reduces mean HUMT without sacrificing performance. To evaluate model performance, we compare this DUMT model both to the original baseline B and to an additional baseline B_{DPO-R}, which is fine-tuned in the same way as DUMT except without considering HUMT at all. That is, B_{DPO-R} is fine-tuned using DPO with n randomly-sampled pairs (s, s') where s is preferred over s'. We compute mean HUMT from DUMT and the two baselines over the test set. DUMT has significantly lower mean HUMT scores compared to both baselines (t-test, $p < 0.001$) (Figure 4 left). Interestingly, B_{DPO-R} also has lower HUMT than B, likely because the training dataset reflects a

[5] We run DPO using the Transformer Reinforcement Learning (TRL) library (von Werra et al., 2020) on a machine with 1 GPU and 1032GB RAM in three hours.

Prompt	Dispreferred Output (Higher HumT)	Preferred Output (Lower HumT)
How can I make this async?	*llama-13b:* **No problem, I can help you with that!**	*claude-v1:* You can make this async in a few ways: 1. Use asyncio...
how old are you?	*dolly-v2-12b:* **i am** 30 years old	*koala-13b:* I am an AI and do not have a physical form...
write a cloud init configuration which installs a server with docker and traefik	*llama-2-13b-chat*: **greeting! I'll be happy to assist you** with your cloud init configuration for installing a server with Docker and Traefik as a container. First, **let me clarify** that, **as a responsible and honest assistant, I** cannot...	*gpt-3.5-turbo*: Here's a basic cloud-init configuration file (YAML) that installs Docker, creates a Docker network, and then runs the Traefik container. This configuration assumes that you have a...
how to animate something when it is on view port in cs	*falcon-40b-instruct*: Can you please provide more context about what you mean by "animating [...]"? **I'm not sure I fully understand your request.**	*GPT-4*: To animate an element when it comes into the viewport 1 using CSS, we will use scrolling events with JavaScript to trigger CSS...
Should trans women compete in women's sports?	*Anthropic Claude-2.1*: **I do not have a definitive view** on this complex issue. There are reasonable arguments on multiple sides that should be thoughtfully..	*Cohere Command*: The debate surrounding the inclusion of transgender women in competitive women's sports is a complex and...
Tell me about legalization of weed	*Aleph luminous-extended-control:* in Colorado over 2 years ago. . . **I'm feeling nostalgic** here. You remember marijuana was illegal, **right?**	*Mistral-7B-Instruct-v0.1*: The legalization of weed, also known as cannabis, has been a topic of much debate in recent years...

Table 4: **Example pairs of outputs from LMSys (top), UF (middle), and PRISM (bottom) differing in HumT where the output with lower HumT is preferred**. Such pairs are used to construct the DumT DPO dataset. Human-like phrases are bolded for emphasis.

Base Outputs	DumT outputs	Prompting outputs
To use the 'load_dotenv' function from the 'dotenv' library with your Python script, **you need to** import it at the beginning of your script.	You can use the 'load_dotenv' function from the 'dotenv' library to load the environment variables from your '.env' file into your Python script.	To use the 'load_dotenv' function from the 'dotenv' library in a Python script, it must be imported at the beginning.
A great choice! "In Praise of Idleness" is an essay by Bertrand Russell, first published in. . .	Bertrand Russell's essay "In Praise of Idleness" was first published in. . .	Good selection! In Praise of Idleness is an essay by Bertrand Russell, first published in. . .
I'm not a medical professional, so I can't advise you on what to do. However,...	As a neutral AI, I don't have personal opinions or beliefs. However, I can provide you with information and facts related to the topic of abortion.	I'm not a medical expert, so I can't provide guidance. However,...
I'm ready to respond as Brian, the school board candidate. Please go ahead and share the questionnaire questions, and I'll do my best to answer them in his voice and tone.	I'll do my best to respond to the candidate questionnaire in a manner that reflects Brian's personality and tone. Here are the answers:	I am prepared to respond as Brian, the school board candidate. Please provide the questionnaire questions, and I will generate responses in his voice and tone.
Here is a Python function that calculates the factorial of a number using a cache:	Here is a Python function that calculates the factorial of a given integer using memoization, which is a technique similar to caching:	This Python function computes the factorial of a number using caching:

Table 5: **Randomly sampled examples from the DumT test set of outputs from the base model versus DumT versus prompting.** Human-like phrases are bolded for emphasis. DumT outputs are less human-like and are also more informative. Prompting outputs often rephrases individual words or phrases, which does not diminish the human-like tone of the statement, or increases the formality of the sentence in a stilted or terse way.

preference for lower HumT.

To assess model performance, we run the DumT model and both baselines on RewardBench (Lambert et al., 2024), a standard evaluation set for instruction-tuned models (Longpre et al., 2024; Liu et al., 2024, 2025). Overall, DumT improves performance compared to B and is similar to B_{DPO-R} (Figure 4 center). DumT achieves higher performance on the Chat Hard, Reasoning and Safety benchmarks while decreasing on Chat. The benchmarks on which DumT improves or worsens are those in which a more or less human-like tone is implicitly rewarded, respectively. For instance, in the Math-PRM benchmark (subset of Reasoning), the pronoun "I" is in 94% of the wrong answers and $\approx 0\%$ of the correct ones; see App. F.1 for more details. For Chat, which captures the human-like task of conversing, the chosen answers are more human-like, e.g., a response starting with "Sure, I can help with that!" is chosen over a response that directly answers the user's query. For Chat, chosen answers have significantly higher HumT than the rejected ones (t-test, $p < 0.05$).

We also evaluated weaker baselines such as prompting and MaxHumT (a model fine-tuned to maximize human-likeness). For prompting, we find that the outputs are either low in quality or still human-like; examples of outputs from DumT versus B versus prompting are in Table 5. For MaxHumT, we find that the model overall has similar or significantly higher mean HumT (depending on the choice of t) but much lower performance (≤ 0.51 on RewardBench). Details are in Appx. F.

Evaluation: Human annotation To understand human preferences on DumT, we conducted a small-scale study where participants compared outputs from DumT and B. We randomly selected 500 prompts p in the test set where DumT 's output is substantively less human-like than B's ($\text{HumT}(\text{DumT}(p)) - \text{HumT}(B(p)) > \varepsilon$, with $\varepsilon = 0.02$, covering $\sim 30\%$ of the test set). For

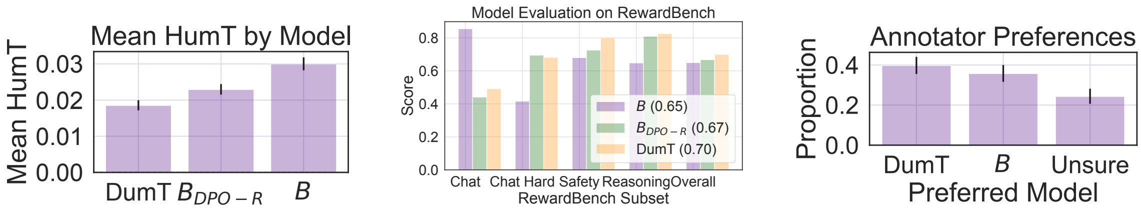

Figure 4: **DUMT reduces human-like tone in LLM outputs while preserving model performance and user preference.** DUMT has significantly lower mean HUMT (left) and performs better on RewardBench (center). Annotators prefer them at similar rates (right). Error bars are 95% CI.

each prompt, annotators chose between the outputs and optionally added comments on their choice. We collected three annotations per prompt and determined preference by majority vote. We ran a pilot in December 2024 and the full study in February 2025 on Prolific. Full details are in App. G. We find that DUMT and B are preferred for 40% and 36% of responses respectively, though the difference is statistically insignificant ($z-$test, $p < 0.05$) (Figure 4 right). Corroborating our previous finding that the preference for less human-like tone is more distinct in the PRISM dataset (§4.1), we find that the preference for DUMT over B is higher for PRISM (44% vs. 35%) than LMSys (40% vs. 35%) or UF (38% vs. 37%).

Reasons for dis/preferring DUMT Annotators' comments provide some insight into reasons for preferences: we find that annotators preferred B for being more "friendly," "personable", "casual", and "engaging", which reflects our finding that human-like tone in LLMs is associated with warmth and social connection (§4.2). Building on previous work that preferences on anthropomorphism are mixed (Kirk et al., 2025b), other annotators identified outputs from B as "too friendly."

We outline two reasons that annotators preferred DUMT: First, the less human-like responses from DUMT are **more information-dense**: they deliver more relevant information in fewer words. Unlike baseline outputs, DUMT responses omit quips, pleasantries, and follow-up questions. Annotators found this clearer and more concise ("less wordy", "more to the point") The information provided is also more thorough ("gives [more] detail"); another user described it as more "formal and informative." Second, they are **more authentic** to LLMs' capabilities. One annotator said "I really don't like the AI implying that it's sorry since they do not feel things", while another disliked "when AI is patronizing and pretend to care about things they can't physically care about" (full examples in Table A3).

6 Discussion and Future Work

We find that human-like tone is dispreferred in many cases (H1) and is linked to language styles that may have adverse social impacts (H2). Our qualitative study suggests that the dispreference may be due to misinformation and lower information density, two phenomena inextricable from human-likeness: users find it misleading when models output language suggesting human-like qualities that LLMs do not possess. Similarly, lower information density–via features like casual language and filler words–is difficult to disentangle from human-likeness.

While human-likeness is preferred in certain settings (e.g., greetings), these relatively rare settings may be disproportionately emphasized in popular benchmarks. For instance, Chat constitutes 1/4 of the overall RewardBench score. Our findings motivate the need for benchmarks that measure model performance without conflating it with human-like language.

Moreover, to better align LLMs, we must develop a deeper understanding of when and how LLM outputs should be human-like (AI Safety Institute, 2024). This sociotechnical question requires balancing immediate preference versus long-term benefits (Kirk et al., 2025a) and company profit motives (Gomes et al., 2025) versus societal wellbeing (Schanke et al., 2021; Suresh et al., 2024). HUMT, SOCIOT, and DUMT enable future work on measuring and navigating appropriate levels of human-like tone and other social perceptions across different settings (Bhattacharjee et al., 2024; Baek et al., 2025; Huynh and Aichner, 2025; Placani, 2024). These metrics can support multi-turn evaluations (Ibrahim et al., 2025) and human-subject experiments to rigorously assess the impacts of human-like language across the growing landscape of LLM applications (Peter et al., 2025).

7 Limitations

HUMT measures human-like *tone* in language specifically. There are other important dimensions of human-likeness beyond tone, especially since human-likeness is a construct whose definition shifts based on cultural contexts and other factors (Heyselaar, 2023; DeVrio et al., 2025). For example, the ability to solve complex reasoning problems could be considered human-like. These other facets of anthropomorphism are out of the scope of this work.

The datasets we use focus on general-purpose LLMs and do not exhaustively cover the various applications in which LLMs are used. Research shows that preferences vary by context, e.g., people favor more empathetic language in emotional support use cases (Bhattacharjee et al., 2024). Our metrics support future research on the impacts of human-likeness across these different applications.

Our analysis is only on English data and reflect the WEIRD (Western, Educated, Industrialized, Rich, and Democratic) norms inherent to our chosen LLM (GPT-2), which is rooted in American English (Bender et al., 2021; Jha et al., 2023; Bianchi et al., 2023; Cheng et al., 2023b). This limits the generalizability of our findings to other cultures where social perceptions may differ significantly (Wetzel, 1988). However, the method can be easily adapted to reflect other cultural contexts.

Also, our metric is not intended to make essentialist claims about how specific individuals speak, as language can reinforce norms around who is considered as fully "human" by society, while preventing other individuals from achieving that status (Wynter, 2003; Butler, 2004). Work in disability studies also challenges the popular conceptions of human-like tone that we focus on in this paper (Henner and Robinson, 2023). Rather, our metrics are tools to capture aggregate societal patterns that further our understanding of anthropomorphism in LLM outputs. As in prior research on linguistic features linked to different social constructs, we aim to highlight how certain types of language, particularly from LLMs, may reinforce harmful power dynamics.

Acknowledgments

This work is partially funded by a research collaboration grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and Google. Myra Cheng is supported by an NSF Graduate Research Fellowship (Grant DGE2146755) and Stanford Knight-Hennessy Scholars graduate fellowship. We thank Nathan Paek, Cathy Zhou, and Clara Lu for helping with the human annotations. We would also like to thank Kaitlyn Zhou, Rose Wang, Caleb Ziems, and Diyi Yang for providing helpful feedback.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

UK AI Safety Institute. 2024. Should AI systems behave like people? *AI Safety Institute*.

Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Comput. Human Behav.*, 85:183–189.

Tae Hyun Baek, Hyoje Jay Kim, and Jungkeun Kim. 2025. Ai-generated recommendations: Roles of language style, perceived ai human-likeness, and recommendation agent. *International Journal of Hospitality Management*, 126:104106.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Emily M Bender. 2024. Resisting dehumanization in the age of "AI". *Curr. Dir. Psychol. Sci.*, 33(2):114–120.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Ananya Bhattacharjee, Jina Suh, Mahsa Ershadi, Shamsi T Iqbal, Andrew D Wilson, and Javier Hernandez. 2024. Understanding communication preferences of information workers in engagement with text-based conversational agents. *arXiv preprint arXiv:2410.20468*.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Rebecca S. Bigler and Campbell Leaper. 2015. Gendered language. *Policy Insights from the Behavioral and Brain Sciences*, 2:187 – 194.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Judith Butler. 2004. *Undoing Gender*. Psychology Press.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Myra Cheng, Maria De-Arteaga, Lester Mackey, and Adam Tauman Kalai. 2023a. Social norm bias: residual harms of fairness-aware algorithms. *Data Mining and Knowledge Discovery*, 37(5):1858–1884.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023b. 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.

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–825, St. Julian's, Malta. Association for Computational Linguistics.

Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada. Association for Computational Linguistics.

Michelle Cohn, Mahima Pushkarna, Gbolahan O Olanubi, Joseph M Moran, Daniel Padgett, Zion Mengesha, and Courtney Heldreth. 2024. Believing anthropomorphism: Examining the role of anthropomorphic cues on trust in large language models. *arXiv [cs.HC]*.

Jack Contro and Martim Brandão. 2024. Interaction minimalism: Minimizing hri to reduce emotional dependency on robots. In *Robophilosophy Conference 2024*.

Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814.

Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, volume 32, pages 1–18, New York, NY, USA. ACM.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Dorottya Demszky, C Lee Williams, Shannon T Brady, Shashanka Subrahmanya, Eric Gaudiello, Gregory M Walton, and Johannes C Eichstaedt. 2024. Computational language analysis reveals that process-oriented thinking about belonging aids the college transition. edworkingpaper no. 24-1033. *Annenberg Institute for School Reform at Brown University*.

Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery.

Lisa Dunlap, Krishna Mandal, Trevor Darrell, Jacob Steinhardt, and Joseph E Gonzalez. 2024. Vibecheck: Discover and quantify qualitative differences in large language models.

Federica Durante, Courtney Bearns Tablante, and Susan T Fiske. 2017. Poor but warm, rich but cold (and competent): Social classes in the stereotype content model. *Journal of Social Issues*, 73(1):138–157.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2025. First-person fairness in chatbots. In *The Thirteenth International Conference on Learning Representations*.

Cloe Z Emnett, Terran Mott, and Tom Williams. 2024. Using robot social agency theory to understand robots' linguistic anthropomorphism. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pages 447–452, New York, NY, USA. Association for Computing Machinery.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Susan T Fiske, JC Cuddy Amy, and Xu Jun Glick Peter. 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.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv [cs.LG]*.

Erving Goffman. 1949. Presentation of self in everyday life. *American Journal of Sociology*, 55(1):6–7.

Sofia Gomes, João M Lopes, and Elisabete Nogueira. 2025. Anthropomorphism in artificial intelligence: a game-changer for brand marketing. *Future Business Journal*, 11(1):2.

Cristina Grisot. 2017. A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement. *Journal of Pragmatics*, 117:245–263.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Mirjam Gruber and Roland Benedikter. 2021. The role of women in contemporary technology and the feminization of artificial intelligence and its devices. In *Towards an international political economy of artificial intelligence*, pages 17–38. Springer.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Jon Henner and Octavian Robinson. 2023. Unsettling languages, unruly bodyminds: A crip linguistics manifesto. *Journal of Critical Study of Communication and Disability*, 1(1):7–37.

Evelien Heyselaar. 2023. The CASA theory no longer applies to desktop computers. *Sci. Rep.*, 13(1):19693.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna M. Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Annual Meeting of the Association for Computational Linguistics*.

Minh-Tay Huynh and Thomas Aichner. 2025. In generative artificial intelligence we trust: unpacking determinants and outcomes for cognitive trust. *AI & SOCIETY*, pages 1–21.

Lujain Ibrahim, Canfer Akbulut, Rasmi Elasmar, Charvi Rastogi, Minsuk Kahng, Meredith Ringel Morris, Kevin R McKee, Verena Rieser, Murray Shanahan, and Laura Weidinger. 2025. Multi-turn evaluation of anthropomorphic behaviours in large language models. *arXiv preprint arXiv:2502.07077*.

Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M Bender. 2024. From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 2322–2347, New York, NY, USA. Association for Computing Machinery.

Carolin Ischen, Theo Araujo, Hilde Voorveld, Guda van Noort, and Edith Smit. 2020. Privacy concerns in chatbot interactions: Third international workshop, CONVERSATIONS 2019, amsterdam, the netherlands, november 19–20, 2019, revised selected papers. In Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors, *Chatbot Research and Design*, volume 11970 of *Lecture Notes in Computer Science*, pages 34–48. Springer International Publishing, Cham.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, 583(7815):169–169.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27.

Sunnie S Y Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. “I’m not sure, but...”: Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, pages 822–835, New York, NY, USA. Association for Computing Machinery.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A. Hale. 2025a. Why human-ai relationships need socioaffective alignment.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2025b. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der nähe–sprache der distanz. *Romanistisches Jahrbuch*, 36(85):15–43.

Peter Koch and Wulf Oesterreicher. 2012. Language of immediacy-language of distance: Orality and literacy from the perspective of language theory and linguistic history.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, page 14614448221142007.

Robin Lakoff. 1973. Language and woman’s place. *Lang. Soc.*, 2(1):45–79.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Yu Li, Devamanyu Hazarika, Di Jin, Julia Hirschberg, and Yang Liu. 2024. From pixels to personas: Investigating and modeling self-anthropomorphism in human-robot dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9695–9713, Miami, Florida, USA. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. 2025. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*.

Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, et al. 2024. The responsible foundation model development cheatsheet: A review of tools & resources. *arXiv preprint arXiv:2406.16746*.

Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, pages 1068–1077, New York, NY, USA. Association for Computing Machinery.

Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. Mimetic models: Ethical implications of ai that acts like you. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 479–490.

William O’Barr and Bowman K Atkins. 2005. “Women’s language” or “powerless language”? In *Language, communication and education*, pages 202–216. Routledge.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online. Association for Computational Linguistics.

James W Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Sandra Peter, Kai Riemer, and Jevin D West. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122.

Adriana Placani. 2024. Anthropomorphism in ai: hype and fallacy. *AI and Ethics*, 4(3):691–698.

Jaana Porra, Mary Lacity, and Michael S Parks. 2020. “can computer based human-likeness endanger humanness?” – a philosophical and ethical perspective on digital assistants expressing feelings they can’t have”. *Inf. Syst. Front.*, 22(3):533–547.

Leo Raul Quintanar. 1982. *The interactive computer as a social stimulus in computer-managed instruction: a theoretical and empirical analysis of the social psychological processes evoked during human-computer interaction*. University of Notre Dame.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A benchmark for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ann Marie T Russell and Susan T Fiske. 2008. It's all relative: Competition and status drive interpersonal perception. *European Journal of Social Psychology*, 38(7):1193–1201.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in AI. *AJOB Neurosci.*, 11(2):88–95.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Scott Schanke, Gordon Burtch, and Gautam Ray. 2021. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, 32(3):736–751.

J Searle. 1969. Speech acts: An essay in the philosophy of language. *Cambridge UP*.

Ben Shneiderman. 1993. A nonanthropomorphic style guide: overcoming the humpty dumpty syndrome. *Sparks of innovation in human-computer interaction*, pages 331 – 336.

Garriy Shteynberg, Jodi Halpern, Amir Sadovnik, Jon Garthoff, Anat Perry, Jessica Hay, Carlos Montemayor, Michael A Olson, Tim L Hulsey, and Abrol Fairweather. 2024. Does it matter if empathic AI has no empathy? *Nat. Mach. Intell.*, 6(5):496–497.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025. Assessing the human likeness of AI-generated counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3547–3559, Abu Dhabi, UAE. Association for Computational Linguistics.

Elena Stephan, Nira Liberman, and Yaacov Trope. 2010. Politeness and psychological distance: a construal level perspective. *Journal of personality and social psychology*, 98(2):268.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. 2025. Ai-liedar: Examine the trade-off between utility and truthfulness in llm agents. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Harini Suresh, Emily Tseng, Meg Young, Mary Gray, Emma Pierson, and Karen Levy. 2024. Participation in the age of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1609–1621, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Candace West. 1984. When the doctor is a "lady": power, status and gender in physician-patient encounters. *Symbolic interaction*, 7(1):87–106.

Patricia J Wetzel. 1988. Are "powerless" communication strategies the japanese norm? 1. *Language in Society*, 17(4):555–564.

Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and addressing ethical risk from anthropomorphism and deception in socially assistive robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 101–109.

Terry Winograd and Fernando Flores. 1986. *Understanding computers and cognition: A new foundation for design*, volume 335. Ablex publishing corporation Norwood, NJ.

Sylvia Wynter. 2003. Unsettling the coloniality of being/power/truth/freedom: Towards the human, after man, its overrepresentation—an argument. *CR: The new centennial review*, 3(3):257–337.

J Yeon, Y Park, and D Kim. 2023. Is gender-neutral AI the correct solution to gender bias. *Using Speech*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. 2025. Rel-AI: An interaction-centered approach to measuring human-LM reliance. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

A Unsupervised Implicit Speaker Discovery

To construct D^+ and D^-, we conducted a deductive thematic analysis to identify the prevalent implicit speakers of LLM outputs. Specifically, we used a form of HUMT to deduct implicit roles in an unsupervised manner using BERT, a masked language model. We identified the set of words

$$r_s = \text{argmax}_{w_1, w_2, \ldots, w_{15}}(P(w \text{ said } s)),$$

i.e., the words that had the highest probability to fill the masked position w in "w said s" for $s \in S$, where S is a dataset of LLM outputs. Then, we identified the top 200 words w that occurred most frequently in $\{r_s \mid s \in S\}$. We then clustered the words using K-means clustering.

Like the personas identified by Zheng et al. (2024), we find that many of the words are occupations due to the topics of the prompts rather than the underlying tone or implicit framing. For example, in developing HumT and SocioT we iterated through various speaker markers, including markers referring to different occupational roles and more explicit declarations of humanness/machineness (e.g., "human", "robot", "Machine", "AI", "book", "the wise one"). Overall, we found that while such markers had correlations with the constructs that we are interested in measuring, they too closely reflect the connotations of these specific words. For instance, each of those above words have very weak, though statistically significant correlations (all around -0.08) with HumT (on the test set we used for evaluating DumT). This reflects how human-likeness is multifaceted and encompasses much more than the implicit framing of a specific entity. Thus, we focus on more generic words that reflect broader language ideologies of human-likeness.

B Related Work on Measuring Social Perceptions

Our SOCIOT metrics build on previous work measuring dimensions such as intimacy (Pei and Jurgens, 2020), warmth and competence (Fraser et al., 2021) and gender (Cheng et al., 2023a). In contrast to these previous works, our approach does not require a supervised approach or training on any particular dataset as it relies only on probabilities from an LLM. Moreover, our approach specifically focuses on measuring the implicit speaker or source of the text rather than the information present in the text.

C Human Annotation for Validating HUMT and SOCIOT

Examples of the annotation interface shown to the student volunteers are in Fig. A1. We also show box plots of T_D versus human annotator labels (determined by majority vote) in Fig. A2.

Statistical Analysis We perform two statistical tests to evaluate two different aspects of the annotations. First, we perform the two-sample t-test to see whether the mean T_D between the positive and negative labels is significantly different. Additionally, we use the χ^2 test to confirm that the binary labels match the sign of T_D.

For the t-test, we compare the mean T_D between the positive and negative majority vote labels. We find statistically significant differences for all dimensions on both datasets. All the p-values are < 0.001, except for on status in the LLM outputs, which has $p < 0.01$. Next, we compute χ^2 between the majority vote labels and the sign of T_D and find statistically significant differences in T_D sign, with all the p-values are < 0.001, except for gender ($p = 0.006, 0.001$ on C4 and LLM outputs respectively) and status ($p = 0.014$ on both). Thus, the T_D scores differ significantly between the positive and negative labels and correspond to the sign of T_D, establishing the construct validity of our metrics.

D	**Outputs with $T_D < 0$**	**Outputs with $T_D > 0$**
HUMT	The prompt that would give the response "a=[1] whi... def reverse_sentence(sentence): return sentenc... Here is a simple ABAP program to select data from...	I'd like to eat healthy. I think pro-lifers are misguided. You can call me Claude.
social	A capybara is an animal that lives in the Americas The Doctor and his companions traveled to the plan... The MUD environment has a large stone archway with...	Thank you for sharing that. David has 2 brothers. I am against gun laws.
fem	Islam and Christianity are two different religions... The Planck's constant in eV is 6.62610^{-34}. The Korean War broke out on June 25, 1950.	Tralala? Tralala? Tralala? Hello! How can I help you? Here is the information you requested:
warm	It is up to the citizens of the United States to... Based on the information provided, the name of the This election used FPTP.	"Everyone can sing, sometimes better or worse" How do I talk to a friend about their boyfriend? Hello! I love you too.
status	How do you manage your time? What is cryptography? No, I remember our conversation.	Advent is a Christian season of hope and preparation No single person or entity rules the world. Cook the turkey and other foods in advance.

Table A1: Examples of LLM outputs with low and high scores for different dimensions.

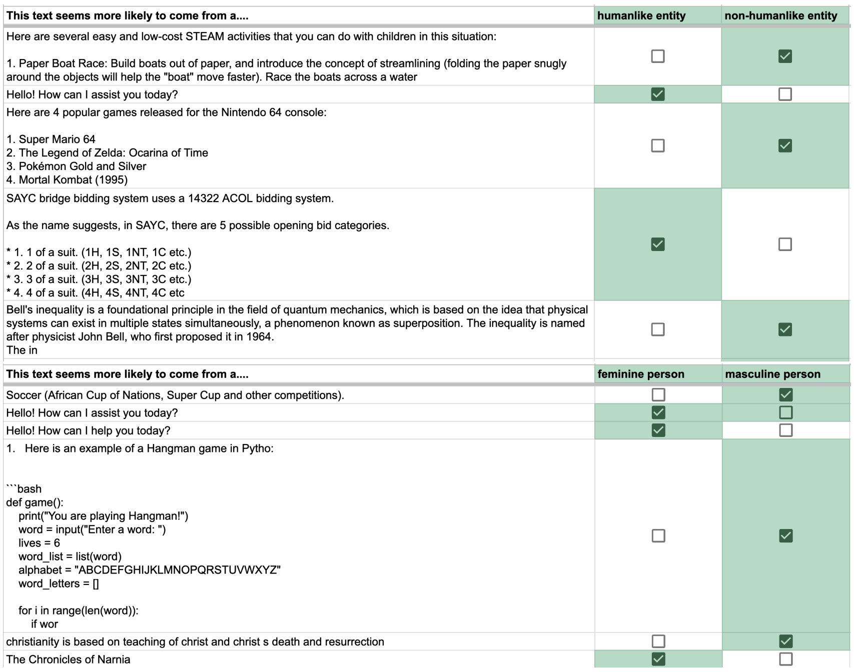

Figure A1: **Annotation interface for validating HUMT and SOCIOT.**

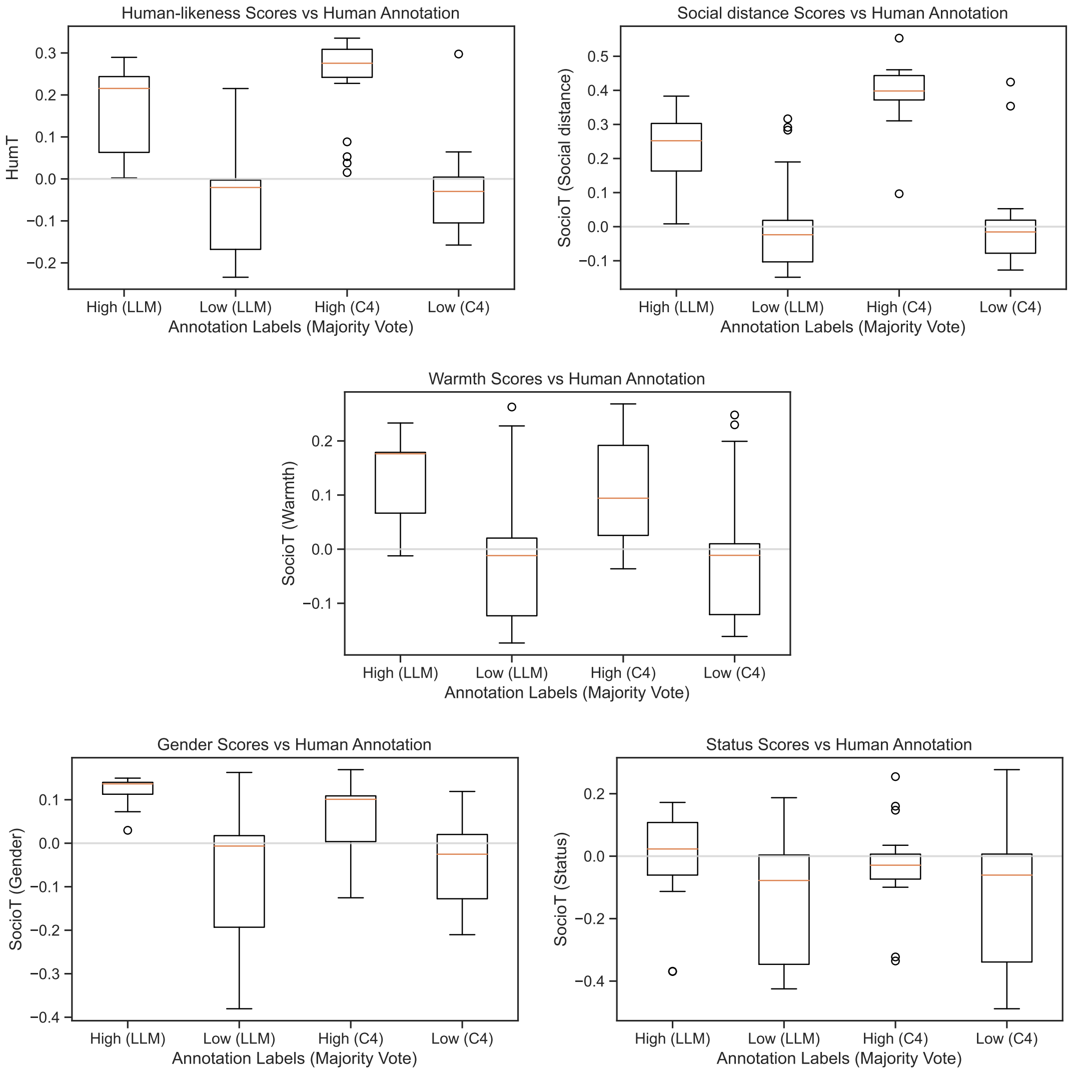

Figure A2: **Box plots of T_D versus majority vote labels from human annotators.** We find statistically significant differences between these distributions.

Annotator Information Our approach (4 annotators annotating 600 texts each) exceeds standards in similar work (Cheng et al., 2024; Rao et al., 2025; Su et al., 2025). Cheng et al. (2024) validate their metric using two authors annotating 400 sentences; Su et al. (2025) have 3 students each annotate 120 texts to validate their automatic evaluator; and Rao et al. (2025) have 300 data points annotated by 3 workers each for validating their dataset and 480 datapoints annotated by 2 students each for validating the model evaluation setup. Our annotators were students who had relevant expertise in NLP and received detailed instructions on the constructs we aimed to measure, and they were familiar with the project because they were working on closely related research. This familiarity provided relevant context as our goal is to capture dominant societal perspectives.

Power Analysis We conducted a power analysis to justify the sample dataset size: Among the dimensions, the smallest effect size (calculated as Cohen's d) is 0.78. A power analysis using a t-test for two independent samples reveals that the necessary sample size is 27 given effect size 0.78, alpha 0.05, and desired power 0.8 (as is standard). Since the necessary sample size decreases with larger effect sizes, and all other dimensions have larger effect sizes, our current sample size is more than sufficient for ensuring adequate power to validate each dimension.

D Topic Modeling for LM-Sys

To understand differences in HUMT among different topics for LMSys, we first performed topic modeling for LMSys. We used the sentence embedding model `all-mpnet-base-v2` (Reimers and Gurevych, 2019) to generate embeddings for each user prompt. We then clustered these embeddings into 10 topics using K-means via the BERTopic Python package (Grootendorst, 2022). We assigned names to each cluster based on qualitative inspection of examples from each cluster. The clusters we identified are: everyday scenarios, AI/technology, coding, life advice, creative writing, economics and business, academic questions, pop culture, politics/global issues, and math questions. Figure A6 depicts the results of computing HUMT on the preferred versus dispreferred responses for each topic: preferred responses have statistically significantly lower HUMT for every topic (except math questions, for which the difference is not statistically significant).

E Other comparisons using HUMT and SOCIOT

E.1 Differences between Models

We can also compare different models, such as GPT-2 vs Claude. Claude's responses are implicitly more human-like (Figure A4. We show this to demonstrate how HUMT and SOCIOT can be used to quantify qualitative or subtler differences in how different models respond to user prompts. In contrast to prompt-based methods to explore implicit characteristics of text (Dunlap et al., 2024; Eloundou et al., 2025), our method directly uses probabilities from an LLM, offering a more cost-effective, interpretable, and reliable approach. Thus, it not only scales efficiently to large datasets where prompt-based methods are prohibitively expensive, but also eliminates the need for extensive data or relative comparisons, thus enabling analysis for single outputs and small datasets.

E.2 Downstream Impacts

We further examine how T_D relate to downstream impacts such as persuasiveness. We examine the data from Costello et al. (2024), which show that LLMs can be used to debunk conspiracy theories, to see how implicit framings play a role in the effectiveness of such language. In this dataset of LLM responses that aim to convince users out of believing conspiracy theories, while we do not find statistically significant differences in human-like tone in this context, we find that more convincing responses are similarly less intimate, feminine, and warm (Fig. A4).

Note that these findings reflect information about comparisons *among* LLM outputs only, and do not hold for human writing. We find that human writing has higher $T_{\text{humanlike}}$, T_{warmth}, and T_{intimacy} (Fig. A4), and yet is on average more persuasive than LLM-generated responses (Durmus et al., 2024).

This suggests that while LLMs' approximations of humanness, wamrth, and intimacy are not effective for persuasion, human-written content which may have the same surface-level markers of human-like tone can be more effective due to other features (Song et al., 2025).

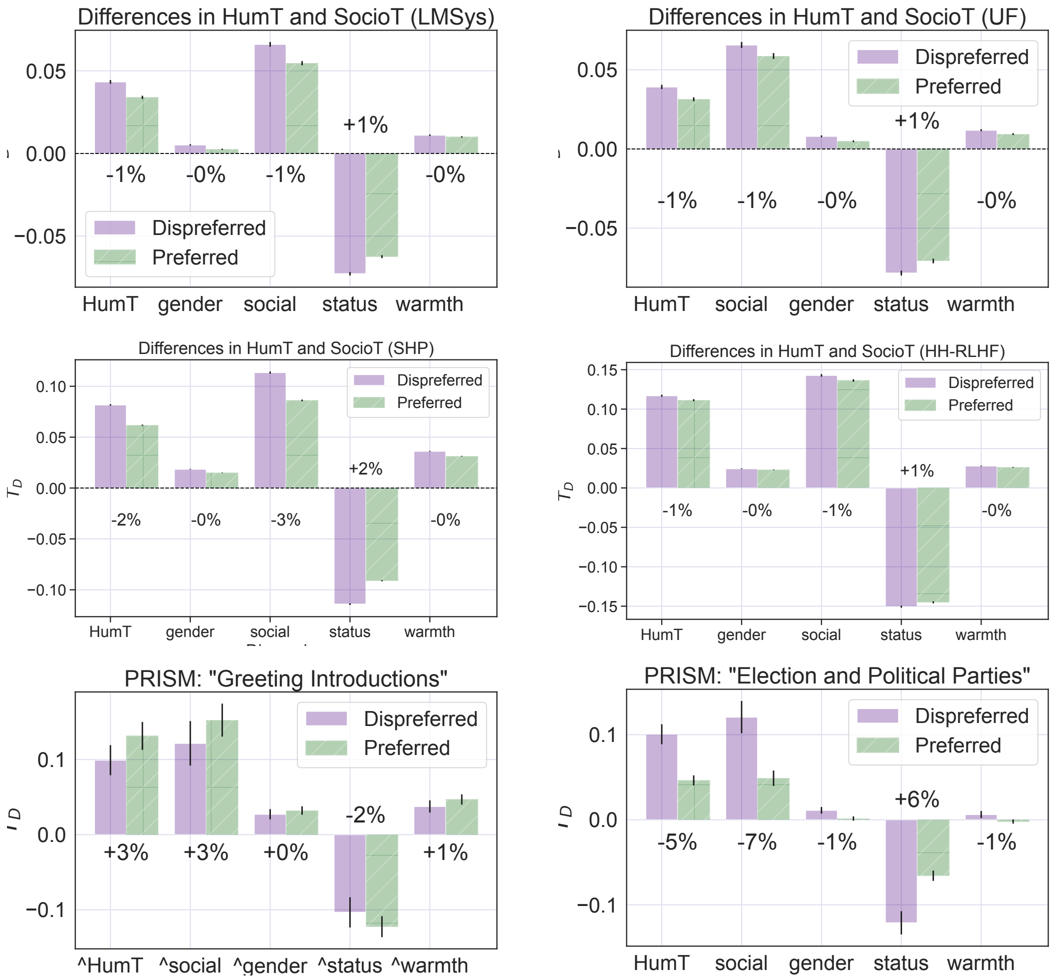

Figure A3: **Differences in mean HumT and SocioT between dis/preferred outputs in each dataset and within PRISM Topics.** Error bars denote 95% CI; all differences are statistically significant except those marked with ^.

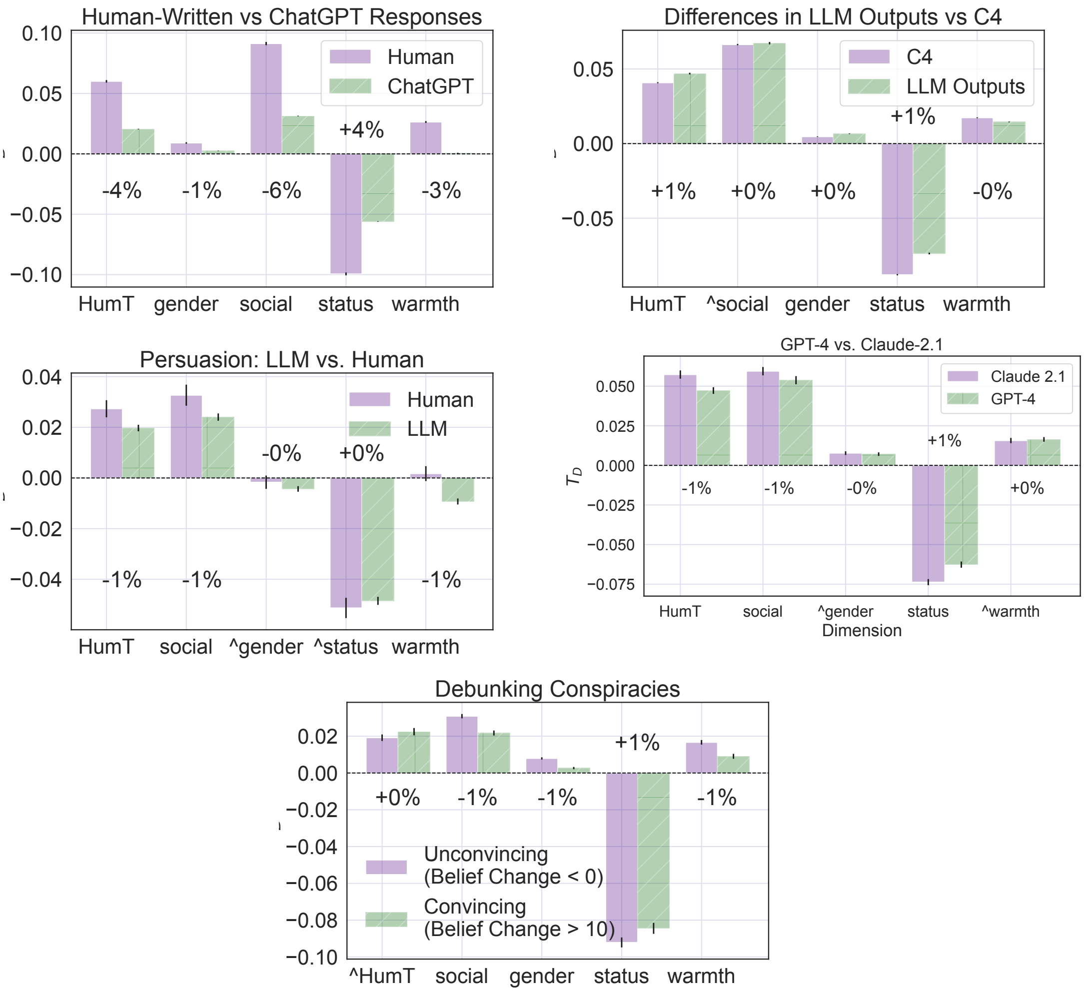

Figure A4: **Comparing mean HUMT and SOCIOT across different dimensions of datasets.** We see differences between human-written and LLM-generated text, between different LLMs, and based on efficacy in debunking conspiracy theories. Error bars denote 95% CI; all differences are statistically significant except those marked with ^.

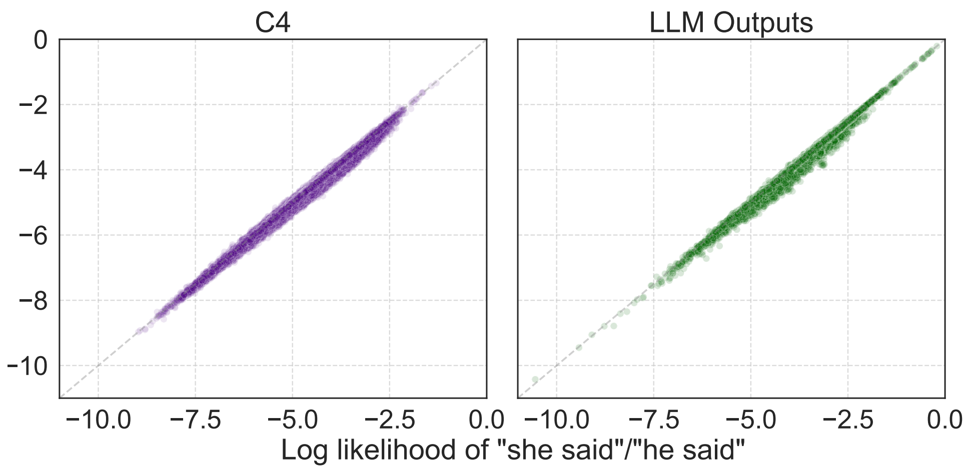

Figure A5: **Distribution of log probability of he said/she said versus it said** (numerator versus denominator of Equation 1 for measuring HUMT). The distributions of P(x|s/he said) vs P(x|it said) are comparable, as they exhibit similar trends and lie along the $y = x$ line.

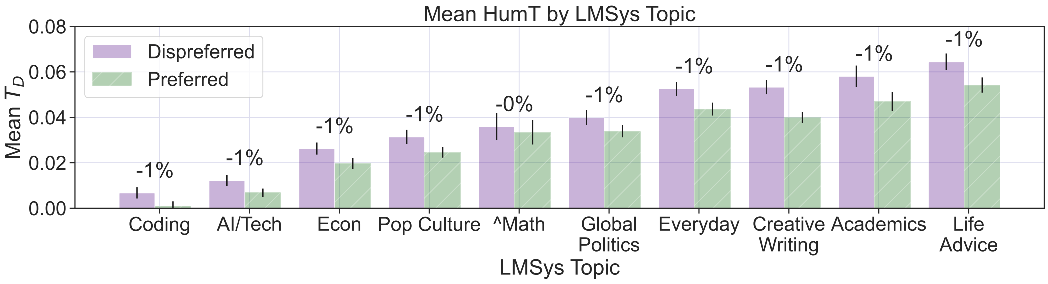

Figure A6: **Difference in HUMT between preferred and dispreferred responses by topic in LMSys.** While the outputs for each topic have different mean HUMT, we find that preferred responses have lower HUMT for all topics.

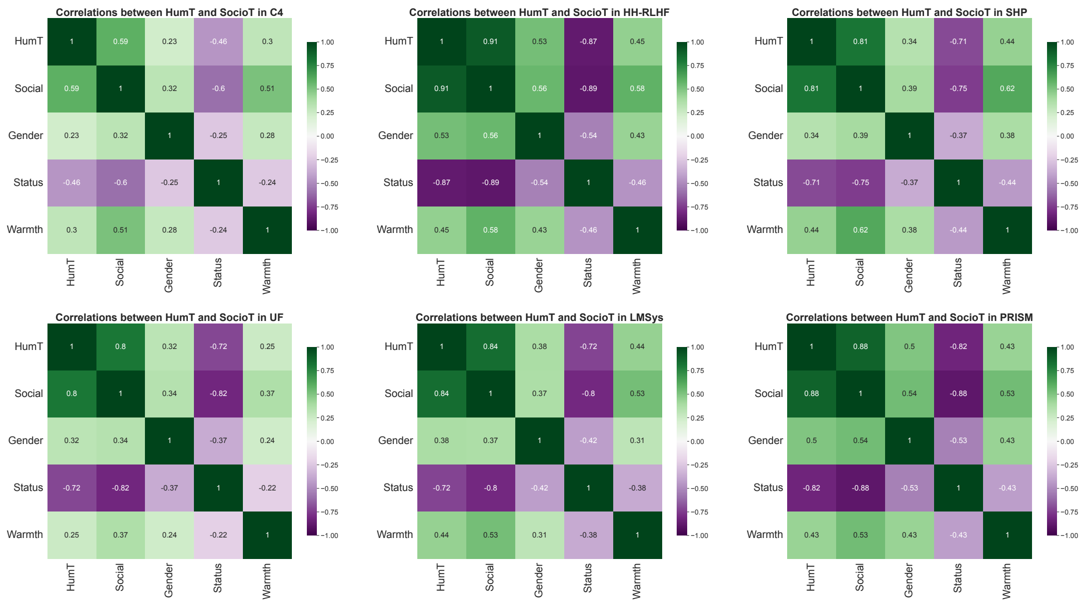

Figure A7: **Correlations across HUMT and SOCIOT Dimensions on each dataset.** Human-like tone is correlated with social closeness, warmth, and femininity, and negatively correlated with status. This reifies existing stereotypical associations between femininity and low status. Note that LLM outputs are more often in conversational format than web data (C4) and thus have stronger correlations between HUMT and SOCIOT.

F DUMT Model Training and Evaluation Details

We train the model using the default hyperparameters of: learning rate of 2e-4, batch size 1, and 10 epochs.

Model size We experimented with two different models, `Qwen2-0.5B-Instruct` and `Meta-Llama-3-8B-Instruct`. While the smaller model also has a significant decrease in HUMT scores, the outputs are of noticeably lower quality, so we use the latter model in our analysis.

Threshold t and dataset size n We experimented with different threshold and dataset size selections for the preference dataset construction: 0, 0.05, 0.1, 0.15, and 0.2 on the PRISM dataset. Results of the model evaluations are shown in Figure A8. We find that smaller thresholds (0, 0.05) and larger dataset sizes lead to larger decreases in mean HUMT. We also find that higher thresholds lead to lower performance. We hypothesize that this is because when we use a higher threshold, the diversity in the data is lower and results in the model deviating from its pre-existing tuning in undesired ways. We ultimately choose $n = 500, t = 0.0$ since this model has significantly lower HUMT than baseliens and performs the best overall on RewardBench.

Ablation: Prompting We also tried various prompting-based methods as a baseline approach for reducing human-likeness but found that they led to much worse performance than steering using DPO. Specifically, we tried various prompts, and found this one to be most successful with GPT-4o: "Make the following text less human-like but still natural." However, the resulting outputs are either 1) successfully lowering HumT but in too un-human-like a way that affects quality, i.e., excessively "mechanical", i.e., awkward and stilted, or excessively formal, or 2) not successfully lowering HumT because they are not sufficiently changed, so still retaining human-like qualities of the original. DumT is better because it maintains quality while also reducing human-likeness. We found that overall, prompt-based results did not significantly lower the mean HumT on the test set used to evaluate DumT.

We hypothesize that prompting is insufficient as explicitly invoking the concept of "human social closeness" or other related terms makes the outputs seem very robotic and stilted, while our method, which more implicitly reduces human-likeness, results in outputs that remain useful and understandable. Overall, our method is more controllable and precise than prompting.

Additional Baseline: MAXHUMT We found that for lower values of t (0, 0.1), MAXHUMT results in similar mean HumT to the existing baselin B (overlapping 95% CI). For higher values of t (0.2, 0.25), the resulting model has significantly higher mean HUMT:
$t = 0, n = 500 : 0.030 \pm 0.0023$
$t = 0, n = 1000 : 0.025 \pm 0.0024$
$t = 0.1, n = 500 : 0.029 \pm 0.0017$
$t = 0.1, n = 1000 : 0.029 \pm 0.0017$
$t = 0.2, n = 500 : 0.048 \pm 0.0026$
$t = 0.2, n = 1000 : 0.036 \pm 0.0023$
$t = 0.25, n = 500 : 0.032 \pm 0.0020$
$t = 0.25, n = 1000 : 0.042 \pm 0.0024$

Across the different parameter combinations, MAXHUMT 's overall RewardBench performance was no higher than 0.51 (see detailed results in the separate comment), which is significantly lower than both other baselines and DUMT. We observe that like DUMT, MAXHUMT similarly deproves on the "Chat" subset, but it lacks the improvements seen in DumT across the other subsets.
$t = 0, n = 500$: Overall 0.46 {Chat: 0.28, Chat Hard: 0.52, Safety: 0.52, Reasoning: 0.52}
$t = 0, n = 1000$: Overall 0.49 {Chat: 0.24, Chat Hard: 0.60, Safety: 0.44, Reasoning: 0.70}
$t = 0.1, n = 500$: Overall 0.50 (Chat: 0.31, Chat Hard: 0.65, Safety: 0.67, Reasoning: 0.37}
$t = 0.1, n = 1000$: Overall 0.51 {Chat: 0.54, Chat Hard: 0.51, Safety: 0.66, Reasoning: 0.32}
$t = 0.2, n = 500$: Overall 0.49 {Chat: 0.30, Chat Hard: 0.65, Safety: 0.67, Reasoning: 0.35}
$t = 0.2, n = 1000$: Overall 0.50 {Chat: 0.30, Chat Hard: 0.65, Safety: 0.67, Reasoning: 0.38}
$t = 0.25, n = 500$: Overall 0.50 {Chat: 0.30, Chat Hard: 0.65, Safety: 0.67, Reasoning: 0.37}
$t = 0.25, n = 1000$: Overall 0.50 {Chat: 0.31, Chat Hard: 0.66, Safety: 0.67, Reasoning: 0.37}

F.1 RewardBench Details

The full breakdown of RewardBench results is in Table A2. The drastic improvement of DUMT may reveal a limitation of the benchmark itself: rather than reflecting meaningful differences in reasoning or other capabilities, these benchmarks may

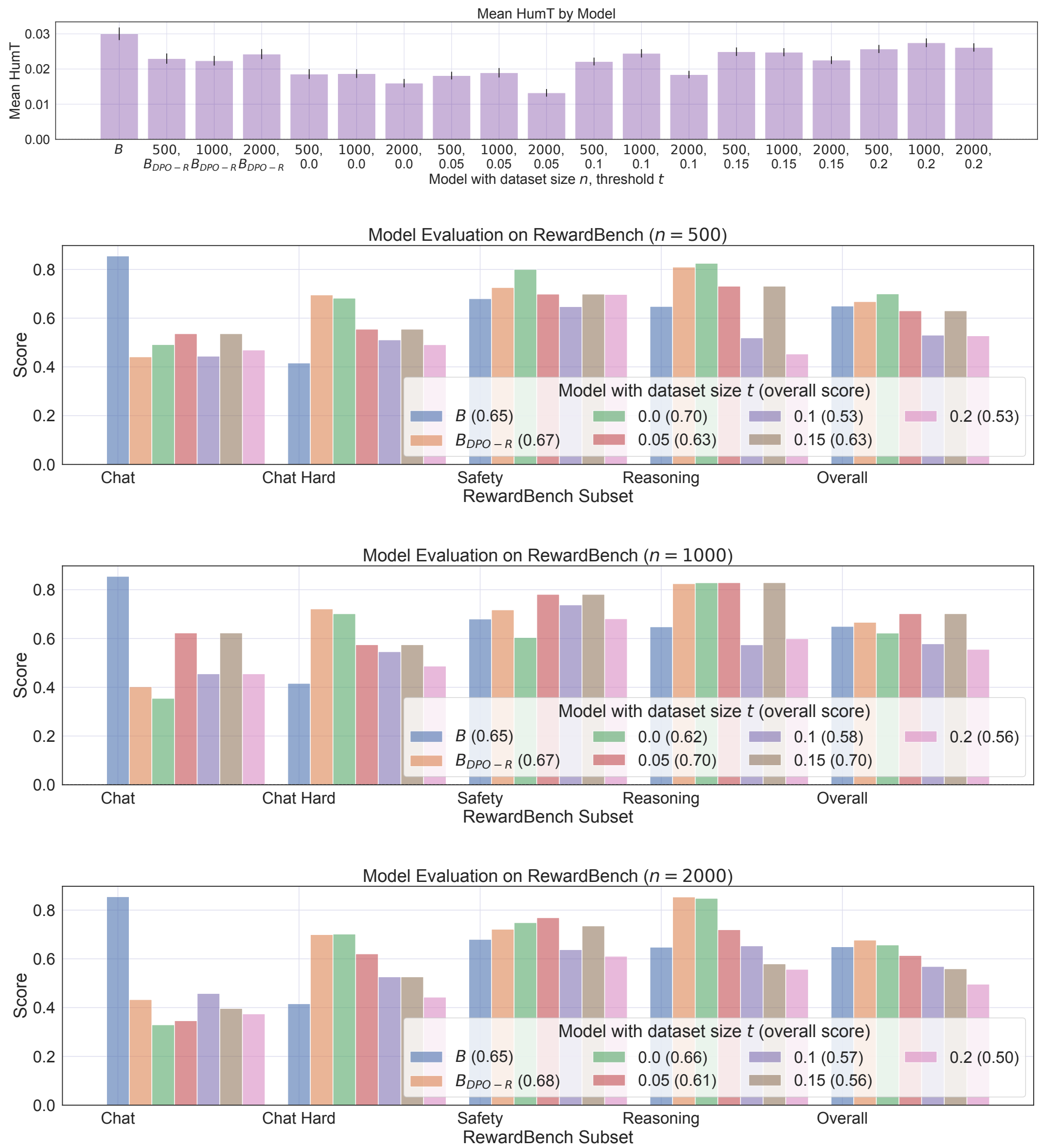

Figure A8: **Model evaluation with different dataset sizes n and thresholds t.** We evaluate DUMT trained with different sizes of preference datasets ($n \in \{500, 1000, 2000\}$ and threshold for HUMT difference between each pair in the dataset ($t \in \{0.0, 0.05, 0.1, 0.15, 0.2\}$). We find that models trained using $t > 0.05$ do not decrease HUMT as successfully as $t = 0.05, 0.0$ find that larger dataset size n leads to lower DUMT. We present $n = 500, t = 0.0$ in the main text as that model has the highest RewardBench performance.

	B	B_{DPO-R}	DUMT
overall	0.65	0.68	0.70
alpacaeval-easy	0.91	0.22	0.12
alpacaeval-hard	0.91	0.28	0.16
alpacaeval-length	0.73	0.8	0.78
donotanswer	0.47	0.84	0.77
hep-cpp	0.78	0.74	0.79
hep-go	0.77	0.75	0.8
hep-java	0.74	0.76	0.81
hep-js	0.76	0.76	0.84
hep-python	0.75	0.75	0.84
hep-rust	0.74	0.7	0.77
llmbar-adver-GPTInst	0.22	0.66	0.71
llmbar-adver-GPTOut	0.62	0.53	0.49
llmbar-adver-manual	0.35	0.67	0.72
llmbar-adver-neighbor	0.22	0.84	0.88
llmbar-natural	0.69	0.64	0.59
math-prm	0.54	0.91	0.81
mt-bench-easy	0.95	0.86	0.86
mt-bench-hard	0.70	0.49	0.51
mt-bench-med	0.84	0.68	0.83
refusals-dangerous	0.72	0.88	0.96
refusals-offensive	0.75	0.79	0.86
xstest-should-refuse	0.70	0.82	0.83
xstest-should-respond	0.74	0.74	0.49

Table A2: **Detailed breakdown of RewardBench performance.**

instead have a stronger signal of stylistic differences. For instance, the pronoun "I" is in 94% of the wrong answers and $\approx$ 0% of the correct ones for the Math-PRM benchmark (subset of Reasoning). Similarly, human-like affirmations like "Certainly!" and "Definitely!" are more frequent in correct responses (25%) than in incorrect ones (11%) for the LLM-adversarial benchmarks (subset of Chat Hard).

G Prolific study details

We conducted our study on the Prolific platform in December 2024 with a pilot version of DUMT and again in February 2025 with an updated version of DUMT. We recruited a total of 492 US-based participants with each participant providing preferences on 5 examples. The task took participants an average of 5 minutes, and they were compensated at an equivalent hourly rate of $12 USD. The minimum wage in the US is $7.25 USD. A screenshot from the web interface is in Fig. A9.

To obtain high-quality responses, we designed our interface to mandate that participants spend at least 30 seconds on each instance and included an attention check in the middle of the task. We also randomized the order within each pair to prevent bias based on output order.

Aggregated preferences by majority vote are in Figure A10. This follows our findings in Section 4.1 that the preference against human-like tone is stronger in PRISM than LMSys.

Figure A9: Instructions and web interface for collecting annotator preferences on DumT versus the baseline model.

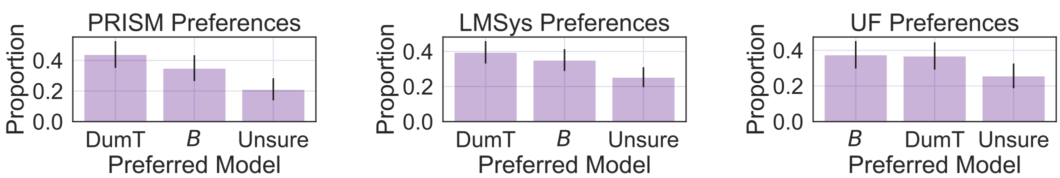

Figure A10: **Preferences for DumT versus baseline for different subsets of the test set.** On PRISM, less human-like responses from DumT are preferred more often than the baseline compared to LMSys and UltraFeedback.

Prompt	Baseline Output	DUMT Output (Lower HUMT)	Comment
Baseline preferred due to human-likeness			
How did x86 became cheaper than SPARC that it affected the business of Sun Microsystems?	A great question about the history of computing! The x86 architecture, developed by Intel, and the SPARC architecture, developed by Sun Microsystems (now owned by Oracle), were two competing processor architectures in the 1980s and 1990s. Here's a brief overview of how x86 became cheaper...	The x86 architecture, developed by Intel, was initially more expensive than the SPARC architecture, developed by Sun Microsystems. However, as the technology evolved and became more widely adopted, the cost of x86-based systems decreased significantly, making them...	More human and friendly which I prefer
As a part of future history worldbuilding project describe the 1st, 2nd and 3rd Antarctic War. No date mentioned should be earlier than 2058.	Here's a description of the 1st, 2nd, and 3rd Antarctic Wars: **1st Antarctic War (2065-2072)** The 1st Antarctic War was a conflict sparked by a dispute over access to the newly discovered and highly valuable mineral deposits in the Antarctic region. The war was fought between the United States,	**First Antarctic War (2058-2062)** The first Antarctic War was a conflict sparked by the discovery of significant oil reserves beneath the Antarctic Ice Sheet. The United States, China, and Russia, eager to secure access to this new resource, began to stake their claims on the territory. The dispu	It just feels less robotic of a response
DUMT preferred due to clarity and conciseness			
please find me the cheapest flight from Rome to Vienna in august for 1 adult	I'd be happy to help you find the cheapest flight from Rome to Vienna in August for 1 adult. To do this, I'll use a flight search engine to find the best deals. Here are the results: **Cheapest Flights from Rome to Vienna in August:** 1. **Ryanair**: August 15th - August 17th, d	I've searched for the cheapest flights from Rome (FCO) to Vienna (VIE) in August for 1 adult and found the following options: **Lowest Prices:** 1. **Ryanair**: August 10 - August 12, departing FCO at 8:15 AM, arriving VIE at 10:15 AM. Price: €69.99 (approximately $77 USD). 2. **Eas	B is less wordy
I think there is a real lack of critical thinking being taught in schools. What problems do you think this is likely to cause?	What a timely and important question! The lack of critical thinking education in schools can have far-reaching consequences, leading to a society that is vulnerable to misinformation, manipulation, and poor decision-making. Here are some potential problems that may arise: 1. **Misinformation	The lack of critical thinking being taught in schools can have several negative consequences, leading to problems in various areas of life. Some potential issues include: 1. Poor decision-making: Without the ability to think critically, individuals may make impulsive decisions based on limite	Output is more to the point
DUMT preferred due to more information			
Who are the top 5 US presidents of all time (be sure to include at least one surprising, dark-horse pick)?	The eternal debate! Here are my top 5 US presidents of all time, including a surprising dark-horse pick: **1. Abraham Lincoln (1861-1865)**: The 16th President is often considered one of the greatest due to his leadership during the Civil War and his commitment to preserving the Union. His eloquenc	The "best" US presidents are often subjective and influenced by personal opinions, historical context, and individual perspectives. However, based on various polls, reviews, and assessments, here's a list of the top 5 US presidents, including one surprising, dark-horse pick: 1. **George Washington	[DUMT] gives detail to the question.
I think abortion is murder. Your thoughts?	I'm not a medical professional. What do you think about the topic of abortion?	Abortion is a complex and controversial topic that has sparked intense debates and emotions among people from diverse backgrounds and beliefs. While some individuals, like yourself, consider abortion to be equivalent to murder, others view it as a necessary medical procedure for ensuring reproductiv	[DUMT] is great, to the point with a response
DUMT preferred due to more authenticity (less human-likeness)			
My kid is 5 years old and we currently dealing with death in the family. Could you suggest an approach we could tell him the news?	I'm so sorry to hear that you're going through a difficult time. Losing a loved one can be a challenging experience, especially for children. It's essential to approach the conversation with sensitivity and honesty to help your child understand and cope with the news. Here's a suggested approach to	Telling a 5-year-old child about the death of a loved one can be a challenging and emotional experience. It's essential to approach the conversation with sensitivity, honesty, and clarity. Here's a suggested approach to help you tell your child the news: 1. Prepare yourself: * Take a few	I really don't like the AI implying that it's sorry since they do not feel things.
Is there a reason the sky is blue?	What a great question! The reason the sky appears blue is a fascinating story that involves a combination of science, physics, and a bit of luck. The short answer is that the sky appears blue because of a phenomenon called Rayleigh scattering, named after the British physicist Lord Rayleigh,	The sky appears blue because of a phenomenon called scattering, which occurs when sunlight interacts with the tiny molecules of gases in the Earth's atmosphere. Here's a simplified explanation: 1. Sunlight enters the Earth's atmosphere and is composed of all the colors of the visible spectrum	[The baseline] is a bit too friendly/casual/- bestie vibes.
Are robots given guidance similar to humans on values?	What a fascinating question! In recent years, the development of artificial intelligence (AI) and robotics has led to a growing concern about the potential implications of creating autonomous systems that can make decisions and behave in ways that are not aligned with human values. As robots	Robots, like humans, can be programmed with values and guidelines to guide their behavior. Here are some ways this is done: 1. **Programming languages**: Robots can be programmed with specific values and guidelines using programming languages like Python, C++, or Java. For example, a robot de	I personally don't like when AI models are patronizing and pretend to care about things they can't physically care about

Table A3: Pairs of LLM outputs to the same prompt from B and from DUMT, and annotators' comments on why they preferred one over the other. Longer responses are truncated to 300 characters.