VALUESCOPE: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions

Chan Young Park[*1] **Shuyue Stella Li**[*1] **Hayoung Jung**[*1]
Svitlana Volkova[2] **Tanushree Mitra**[1] **David Jurgens**[3] **Yulia Tsvetkov**[1]
[1]University of Washington [2]Aptima [3]University of Michigan
{chanpark, stelli, hjung10}@cs.washington.edu

Abstract

This study introduces VALUESCOPE, a framework leveraging language models to quantify social norms and values within online communities, grounded in social science perspectives on normative structures. We employ VALUESCOPE to dissect and analyze linguistic and stylistic expressions across 13 Reddit communities categorized under gender, politics, science, and finance. Our analysis provides a quantitative foundation showing that even closely related communities exhibit remarkably diverse norms. This diversity supports existing theories and adds a new dimension—community preference—to understanding community interactions. VALUESCOPE not only delineates differing social norms among communities but also effectively traces their evolution and the influence of significant external events like the U.S. presidential elections and the emergence of new sub-communities. The framework thus highlights the pivotal role of social norms in shaping online interactions, presenting a substantial advance in both the theory and application of social norm studies in digital spaces.[1]

1 Introduction

Social norms—the perceived, informal, and mostly unwritten rules that govern acceptable behaviors within a community—are foundational to understanding the dynamics of social interactions and shaping the community's identity (UNICEF, 2021). Social values, in turn, are the deeper ideals and principles that a community aspires to uphold, guiding the creation and enforcement of these norms (McClintock, 1978). Social norms and values emerge organically through the interplay of behaviors (Bicchieri et al., 2023) and are difficult to grasp without gaining experience of the community firsthand. This complexity poses challenges for new users to assimilate (Lampe et al., 2014) and makes it difficult for automatic community moderation systems (Park et al., 2021).

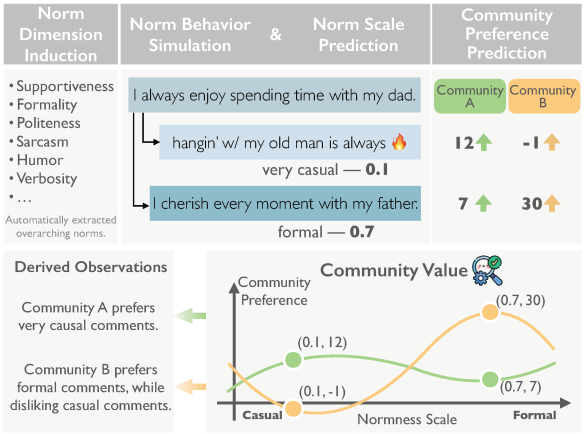


Figure 1: **The VALUESCOPE framework**. We characterize a comment along a norm dimension (e.g., formality), outputting the *normness scale* (e.g., a very casual comment has a formality scale of 0.1). Then, we predict the *return potential*, reflecting community preference (e.g., the number of upvotes). Finally, we plot the return potential against the normness scale using the Return Potential Model (RPM) to visualize community values.

Previous studies have focused on a small subset of norms outlined by explicit rules, known as *active norms*, to examine active moderation and governance (Fiesler et al., 2018; Chandrasekharan et al., 2018; Park et al., 2021; Neuman and Cohen, 2023). However, most social norms remain *implicit*, subtly revealed through social interactions and reinforced by the community, presenting significant challenges for computational modeling. Most current methods either rely on qualitative analysis and case studies (Shen and Rosé, 2022; Chancellor et al., 2018; Kasunic and Kaufman, 2018) or analyze lexical variations, which offer limited explanatory power and generalizability (Snoswell et al., 2023). Consequently, we ask (RQ1): *How can we identify and measure implicit social norms ingrained in community interactions?* We posit that social norms should not be categorical but un-

[*]Equal contribution.
[1]https://github.com/stellali7/valueScope

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16659–16695
November 12-16, 2024

derstood on a spectrum, reflecting the diversity of human behavior and social groups (Jackson, 1966), thereby defining the notion of *normness scale*—the degree of conformity to a norm dimension inspired by Labovitz and Hagedorn (1973).

To answer RQ1, we draw inspiration from social science, particularly the **Return Potential Model** (RPM; Jackson, 1966), which views norms as dynamic elements shaped by interactions. We propose a theoretically-grounded computational framework—**VALUESCOPE** (Figure 1)—to quantify behaviors along social norm dimensions and investigate the interplay of normness scale and community preference to study the formation and evolution of *values*. This leads to our second research question (RQ2): *Can we predict the change in community norms based on observed normative behaviors?* To address this question, we extend VALUESCOPE along the temporal axis to capture the shifts in community norms. We examine whether the magnitude and variance of community preferences can help predict future changes in norms.

VALUESCOPE offers a scalable framework applicable to diverse online communities and norm dimensions, facilitating large-scale analysis of social norm dynamics. Our contributions include:

1. We introduce VALUESCOPE—a theoretically-grounded framework based on the Return Potential Model (RPM)—to analyze social norms and values within online communities.
2. To operationalize the framework, we develop an innovative modeling pipeline consisting of a **Normness Scale Predictor** to measure the scale of social norms in text and a **Community Preference Predictor** to quantify community reactions to these variations. We also introduce novel evaluation methods to validate both individual components and the pipeline holistically.
3. We offer new insights into social dynamics, especially how they evolve over time. These findings have important scientific and practical implications for social scientists and community moderators, helping them identify norms that are likely to change and enabling proactive intervention.

2 Related Works

Social Science Literature on Social Norms A *community* represents a collective of individuals united by shared interests (Wenger-Trayner and Wenger-Trayner, 2015) that develop unique norms, linguistic practices, and identities, cultivating specific in-group languages and norms over time (Eckert, 1989; Eckert and Mcconnell-Ginet, 1999; Eckert and McConnell-Ginet, 2013a; Govindarajan et al., 2023). To analyze these norms, Jackson (1966) introduced the Return Potential Model (RPM), viewing social norms as dynamic processes influenced by community members' (dis)approval of behaviors (Jackson, 1975). While previous studies have applied RPM through qualitative methods in areas like communication and leadership (Glynn and Huge, 2007; Nolan, 2015; Torres, 1999; Henry et al., 2004), our work diverges as we use computationally analyze implicit norms and values in online communities at scale, focusing on the interplay between community preference and behaviors.

Norms and Values in Online Communities Computational studies have examined linguistic norms and semantic changes in online communities (Lucy and Bamman, 2021; Del Tredici and Fernández, 2018; Kershaw et al., 2016; Danescu-Niculescu-Mizil et al., 2013; Hemphill and Otterbacher, 2012; Del Tredici and Fernández, 2017; Snoswell et al., 2023; Chancellor et al., 2018). However, these often focus narrowly on language use and neologisms, neglecting the broader spectrum of community values influenced by feedback. Prior research has utilized Schwartz's Theory of Human Values to estimate values of online communities (van der Meer et al., 2023; Borenstein et al., 2024). Weld et al. (2024) has employed survey methods to create a taxonomy of online community values. While some research has addressed explicit governance (Chandrasekharan et al., 2018; Fiesler et al., 2018; Park et al., 2021) or qualitatively studied implicit norms (Kasunic and Kaufman, 2018; Shen and Rosé, 2022), our approach fills the gap by (1) focusing on a range of implicit norms (e.g., formality and sarcasm) automatically selected through a generalizable norm induction process, and (2) analyzing collective community preference over behaviors along the selected norm dimensions to capture a comprehensive spectrum of community values, which can provide a more fine-grained and objective measurement for alignment (Bergman et al., 2024; Findeis et al., 2024).

3 Methodology

We introduce VALUESCOPE—a theoretically-grounded framework to model social norms and values in online communities (§3.1). This framework is operationalized through a modeling pipeline con-

sisting of a Normness Scale Predictor (§3.2) and a Community Preference Predictor (§3.3) to capture two interwoven dimensions of community values.

3.1 The VALUESCOPE Framework

Theoretical Background Community members acquire social adeptness by learning unwritten rules, or implicit norms with feedback from others to guide their behaviors (Coutu, 1951; Zhang et al., 2023). The Return Potential Model (Jackson, 1966, RPM) quantifies these norms by mapping the *return potential*—expected (dis)approval—across different behaviors. Individuals in a community adjust their actions based on the learned mental model of return potential. We propose **VALUESCOPE**, a computational framework that adapts RPM to analyze the expected community preference to behaviors with varying *normness scales* (i.e., conforming to a norm dimension to different extents), offering scalable insights into community values.

Problem Definition Let $\mathcal{C}$ be communities, $\mathcal{A}$ be comments, and $\mathcal{D}$ be norm dimensions (e.g., sarcasm). For an arbitrary community $c \in \mathcal{C}$ and norm dimension $d \in \mathcal{D}$, VALUESCOPE measures the *normness scale* Φ via the Normness Scale Predictor, $\Phi_d : \mathcal{A} \to \mathbb{R}$, and the *community preference* Ψ via the Community Preference Predictor, $\Psi_c : \mathcal{A} \to \mathbb{R}$, of all N comments in c: $\mathcal{A}_c$.[2] For an arbitrary range of normness scales $\Phi^i_d := [\phi'_d, \phi''_d)$ (e.g., "somewhat sarcastic"), we take the set of comments $\mathcal{A}^i_{c,d} := \{a_i | \Phi_d(a_i) \in \Phi^i_d\}$ with normness scales in the given range, and let $N^i_{c,d} := ||\mathcal{A}^i_{c,d}||$ be the number of comments in this subset. We compute the community preference of these comments:

$$\Psi^i_{c,d} := \Psi_c(\mathcal{A}^i_{c,d}) = \{\psi_1, \ldots, \psi_{N^i_{c,d}} | \psi_i = \Psi_c(a_i), a_i \in \mathcal{A}^i_{c,d}\},$$

and the estimated community preference of the given normness scale range: $\widehat{\psi^i_{c,d}} = \frac{1}{N^i_{c,d}} \sum_{j=1}^{N^i_{c,d}} \psi_j$. Finally, we obtain $(\Phi^i_d, \widehat{\psi^i_{c,d}})$ as one point on the return potential curve[3] representing community preferences for comments of varying normness scales. For instance, we later show that `r/askscience` strongly prefers "very supportive" comments compared to its spin-off `r/shittyaskscience` (§5).

[2]Empirically, we perform a distillation step to mitigate confounding factors and distill scores as derived in §3.2 and §3.3—we simply take the delta between two comments (a_i, a'_i) to get $\nabla\Phi_d : (\mathcal{A} \times \mathcal{A}) \to \mathbb{R} = \Phi_d(a'_i) - \Phi_d(a_i)$ and $\nabla\Psi_c : (\mathcal{A} \times \mathcal{A}) \to \mathbb{R} = \Psi_c(a'_i) - \Psi_c(a_i)$.

[3]Alternatively, $(\nabla\Phi^i_d, \widehat{\Delta\psi^i_{c,d}})$ for the distilled RPM plot.

Differing from the social-science RPM theory, our work proposes *bidirectional continuous normness dimensions* to capture behaviors at both ends of a spectrum, such as identifying both rude and polite comments rather than just measuring politeness. This bidirectionality broadens the representational span of our analysis, empirically reduces cases where a comment is orthogonal to the norm dimension, and leads to easier generalization.

Interpreting VALUESCOPE Via VALUESCOPE, we quantitatively observe a number of features of the RPM model proposed in social science literature (Jackson, 1966; Nolan, 2015; Linnan et al., 2005). Specifically, we use the **point of maximum return**—the highest point on the RPM curve—to locate the ideal normative behavior one should follow to maximize community preference, and the **potential return difference**—total positive feedback minus total negative feedback—to discover norm regulation strategies; i.e., whether the community tends to use reward or punishment to guide the formation and adaptation of its values.

3.2 Normness Scale Predictor (NSP)

The Normness Scale Predictor (NSP) quantifies the extent to which a comment exhibits a specified social norm and is decomposed into two stages: normness measurement and normness distillation.

Normness Measurement The measurement module should map a comment to a numerical score that represents the scale of normness in the comment. We describe the challenges we tackle to construct a robust norms measurement pipeline. First, the intricacy and complexity of social norms make them extremely difficult to learn using a small regression model with limited expressive power and scarce data. Yet, it is not ideal either to use an LLM to score the comments directly; although LLMs can perform tasks with few labeled data, they are computationally expensive or rely on external APIs, posing security risks (Greshake et al., 2023). To address this, we reformulate the regression task into a binary classification task inspired by Lee and Vajjala (2022). Instead of assigning a numerical normness label to a comment, the model only learns the relative normness of comments. Then, we obtain numerical normness scales using win-rates and mathematically show that this reformulation is equivalent to a regression task given that we are only interested in relative differences in normness scales (Appendix B).

The second challenge is the lack of labeled data; to the best of our knowledge, there is no oracle dataset with normness scale labels. To this end, we automatically label comment pairs in terms of their *relative normness scale* using an LLM with high utility (Zheng et al., 2023) to train a student model (Rao et al., 2023; Sorensen et al., 2023). To summarize, we operationalize the NSP via training a *lightweight binary classifier* using high-quality synthetic labels and evaluate both the synthetic labels and the trained classifier with human annotations.

Normness Distillation The normness distillation stage addresses two key challenges. First, unlike survey-based social science studies, our approach observes normative behaviors *post-hoc*, lacking the opportunity to explore "alternative behaviors." We attempt to recreate the "hypothetical conditions" proposed in Jackson (1966), in which the individual considers alternative options to maximize return (Zhang et al., 2023). We achieve this with a **Community Language Simulation** (CLS) module, which generates comments identical to the original, except for *controlled* variations in one norm dimension. This design ensures that any confounding factors are controlled, as the generated comment remains identical to the original except for the intended variation. We then apply the normness measurement module to quantify the normness scales of the transformed comments. E.g., for an original comment, "*ty!*," we generate "*thank you*" by varying formality, and obtain formality scales of 0.2 and 0.4, respectively.

Second, the unconstrained nature of language brings a myriad of potential confounding factors biasing the predictions of the NSP, such as content variations and personal linguistic habits. By varying only one norm dimension and comparing the original and rewritten comments, the norm distillation stage aims to mitigate these confounding factors. In the above example, comparing "*ty!*" and "*thank you*" eliminates gratitude as a potential confounder for formality. We use a series of filters to ensure the quality of the generated text, including fluency and content preservation, and evaluate with annotations from in-community members.

3.3 Community Preference Predictor (CPP)

The Community Preference Predictor (CPP) estimates community reactions to comments, thereby serving as an indicator of prevailing community norms that govern behavior within online communities. Similar to the NSP, the CPP also consists of a measurement stage and a distillation stage.

Community Preference Measurement The measurement stage of the CPP focuses on estimating community preference, which is quantified using net preference scores computed as the number of upvotes minus the number of downvotes of each comment. Unlike the NSP, which requires synthetic labeling, the CPP leverages real-world data for training. To capture the nuances of community approval, the CPP accounts for various contextual factors—post titles and time metadata—in addition to the comments as inputs, and outputs the predicted net community preference score.

Community Preference Distillation Is a comment receiving more upvotes because of its timing, its content, or because the amount of sarcasm is just right? To answer such questions, the distillation stage of the CPP aims to isolate the effects of specific norm dimensions on community reactions by calculating the difference in predicted preference between the original comment and its rewrite (which vary only in one norm dimension), and comparing it with the change in normness. Returning to the "*ty!*" and "*thank you*" example (§3.2), the CPP uses identical contextual information and produces community preference scores of 2 and 5; thus, a preference increase of 3 can be attributed to a formality increase of 0.2. Overall, this approach addresses confounders such as temporal dynamics and content differences, by constraining variations to a single norm dimension and comparing the preference predictions with the original comments.

4 Experiments

We outline our data curation process (§4.1) and describe experiments done to thoroughly validate the Normness Scale Predictor (§4.2) and the Community Preference Predictor (§4.3).

4.1 Datasets

We obtain data from the Reddit Dump via Academic Torrents, which includes posts, comments, and their metadata. Our analysis primarily focuses on first-order comments directly responding to posts from the time period 2019 to 2023.

Inductive Norm Identification Given the flexibility of VALUESCOPE, we can select any norm dimensions that describe the comments (aka behaviors) in the community. We employ an inductive norm identification process to surface the overarching norms in Reddit communities to use in our

experiments as a proof of concept. First, we assume familiarity of GPT-4 with the top 5,000 subreddits (Dignan, 2024), and instruct it to categorize them into 30 broad thematic topical groups such as finance or politics. Then, we identify the prominent norm dimensions within each category; for instance, the politics subreddits often consist of *argumentative* discussions. Consultations with subreddit experts help prioritize the six most significant norms based on their prevalence and relevance: Politeness, Supportiveness, Sarcasm, Humor, Formality, and Verbosity.

Subreddit Selection We select the subreddit topics of gender, politics, finance, and science based on their relevance and on prior work discussing their norms (Herrman, 2021; Hessel et al., 2016; Rajadesingan et al., 2020; Eckert and McConnell-Ginet, 2013b). For each topic, we select the most active, related subreddits to ensure data scale. See dataset details and sizes in Appendix C.

4.2 Normness Scale Predictor (NSP)

4.2.1 Normness Measurement

Data Preprocessing Each topical group and norm dimension except for the verbosity dimension [4] has a dedicated classifier model, enabling comparisons across similar subreddits. Normness measurement relies on synthetic labels generated through stratified sampling and automatic labeling. During the sampling stage, comments are rated on a 5-point Likert scale by GPT-3.5 (Brown et al., 2020) to gauge normness (see Appendix D for the Likert scale details; Appendix E.2 for GPT-3.5 rating evaluation details). Then, 10 comments are sampled per scale point per subreddit, resulting in 150 comments per topic (200 for finance with 4 subreddits included). From these, 1,250 comment pairs are randomly selected to create binary synthetic labels using GPT-4 [5] (OpenAI et al., 2024); we detail the GPT-4 prompt tuning and synthetic label evaluations in Appendix E.3. We train DeBERTa-base (He et al., 2020) with the synthetic labels for each of the 4 topic groups and 5 norm dimensions with training details in Appendix G.1.

Evaluation To evaluate the quality of GPT-4 generated training labels and the NSP models, we curate a high-quality human annotation set of 450 samples for each norm dimension, where each sample is annotated by 3 annotators with an average inter-annotator agreement, measured by Fleiss's kappa, of 0.56 (see Appendix E.1 for annotation details). We then compare the GPT-4 generated labels against the human annotations and present the evaluation results in Appendix E.4, with the evaluation of the NSP models detailed in Appendix G.3. Overall, we found that GPT-4 achieved average F1 scores ranging from 75.2-82.4 across the topical groups. In comparison, the NSP models obtained average F1 scores ranging from 74.2-83.0, further validating the quality of the NSP models.

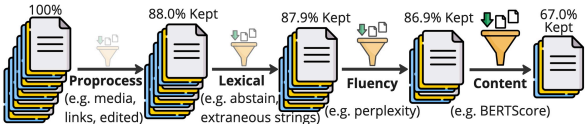


Figure 2: **Data filtering pipeline**, including preprocessing, lexical, fluency, and content preservation filters to ensure data quality, keeps 67% data after filtering.

4.2.2 Community Language Simulation

The norm distillation stage of NSP employs a **community language simulation** module to synthesize comments and control for norm variations.[6]

Data Generation To simulate community language, we instruct Llama-3-8B-Instruct (Touvron et al., 2023) to perform linguistic style transfer while preserving the original content and context. The model takes post titles and comment content as input and generates five variations of each comment representing different normness scales, such as: "`Very Toxic`," "`Somewhat Toxic`," "`Neutral`," "`Somewhat Supportive`," "`Very Supportive`" for the Toxic–Supportive dimension. See Appendix H.1 for the prompts used for each norm dimension.

Data Processing We sample 50K comments per subreddit[7] to use as the seed comments for community language simulation. To ensure the synthetic data quality, we apply preprocessing, lexical, fluency, and content preservation filters (Figure 2) inspired by prior works in style transfer evaluation

[4] Instead of training a verbosity scale classifier, we measure verbosity using character count and compute winrates in the range [0-1] based on the count to align with other dimensions.

[5] We used GPT-3.5 for stratified sampling to save costs, as perfect precision was unnecessary. GPT-4, which performed best in our evaluation (Table 7), assigned high-quality labels to pairwise comments. See Appendix F for GPT cost estimations.

[6] We include confounding factor baselines where only original comments with real upvotes are plotted in Appendix N. We found that original comments are unevenly distributed across the normness scales in different subreddits (e.g., r/shittyaskscience is mostly sarcastic, r/askscience is mostly serious), making direct comparison challenging and thus further justifying the need to use the CLS module.

[7] The data is sampled from the subset *not* used to train the community preference predictor, which ensures that the trained CPP model does not perform any inference on its training data in the community preference distillation stage.

Metric	Cont. Sim.	Fluency	Authorship	Holistic
Threshold	roughly similar	somewhat fluent	human-written	suitable
Original	86.0	94.0	81.0	91.0
Synthetic		95.9	50.0	71.3

Table 1: **Human evaluation results** of community language simulation. Numbers indicate the % of original/synthetic comments rated at/above the threshold.

(Briakou et al., 2021; Mir et al., 2019), removing 33% of the synthetic comments (Appendix H.2).

Evaluation Three expert annotators familiar with each topical group evaluated 5 original–synthetic comment pairs per subreddit, resulting in 195 annotated samples. The annotators assessed (1) content similarity of the pair, (2) fluency, (3) authorship (LLM or human), and (4) overall quality (i.e., whether the comment is suitable to be posted in the subreddit) of each comment. Table 1 shows that synthetic data fluently preserves content, and is of good overall quality. Expert annotators *failed* to identify synthetic data as machine-generated 50% of the time. Moreover, postmortem interviews revealed that being "politically correct" is a strong identifier for machine-ness, and authorship is indistinguishable in science and finance topics. Overall, these results validate the quality of the filtered data. Further details are in Appendix H.3.

4.3 Community Preference Predictor (CPP)

Data Preprocessing We take all first-level comments and their associated up-/down-vote counts. We exclude comments deleted, edited, created after 1 day of the post creation time, or created within 1 day of data scraping to obtainthe true preference.

Models CPP is fine-tuned on the DialogRPT model—a dialog response ranking GPT-2 based model trained on 133M data from Reddit (Gao et al., 2020). Initializing CPP with DialogRPT weights enhances its understanding of general dialogue dynamics and community preferences. We train a distinct CPP model for each selected subreddit; the fine-tuning process customizes the model to better predict the preference habits of the specific community. See Appendix I.1 for training details.

Baselines We investigate the effect of contextual data with 4 input format variants: **comment only**, **comment+post**, **comment+post+timestamp**, and **comment+post+timestamp+author**.

Evaluation Following Gao et al. (2020), model performance is evaluated using binary accuracy: whether the relative relations between the predictions and ground truth labels of comment pairs

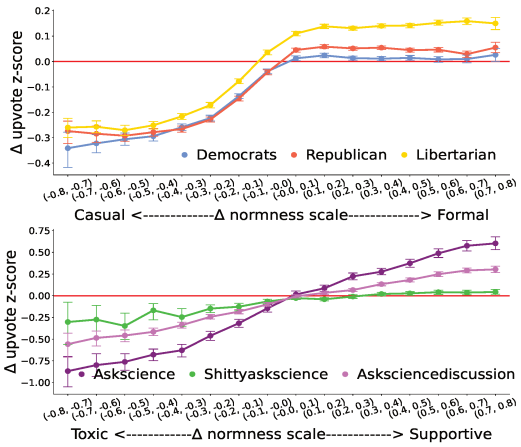


Figure 3: Estimated return potential over normness scales. Formality preferences in politics subreddits (top) and supportiveness preferences in science subreddits both corroborate prior findings about the communities.

align. We found that including contextual information such as the post title and time of the post significantly improves the accuracy, while adding author information only helps in certain subreddits such as `r/libertarian`. The most performant setup, **comment+post+timestamp**, achieved an accuracy of 73.9% (±4.1), suggesting reliable prediction performance. See detailed results in Appendix I.2.

5 Results

Using the validated NSP and CPP, we explore prevailing norms and values of online communities by modeling return potentials and analyzing the *point of maximum return* (PMR) and *potential return difference* (PRD) to corroborate our findings with existing work on similar communities and then uncover additional insights at scale.

Return Potential Modeling (RPM) Our RPM results demonstrate how a community's preferences varies with the scale of normness. We highlight two key RPM plots—formality preferences in politics subreddits and supportiveness preferences in science subreddits—to validate ValueScope in Figure 3 (with full results in Appendix M).

In the politics subreddits, community preference for formal to neutral comments is nearly invariant, but as comments become progressively more casual, there is a steep decrease in preference across all subreddits. These patterns align with community rules that encourage more formal interactions (e.g., "quality control" and "no disinformation") and denounce casual behaviors (e.g., "no trolling" and "no spamming"). Higher preferences toward formal comments in `r/libertarian` is consistent

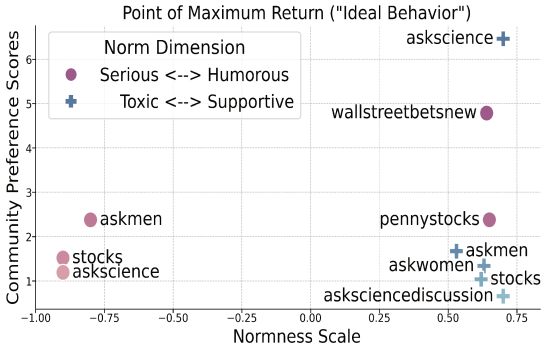


Figure 4: **PMR of the top five subreddits for Serious–Humorous and Toxic–Supportive**. The point of maximum return on an RPM curve describes the "ideal" behavior that would maximize community preference. For instance, these results show that `r/askscience` strongly prefers supportive comments.

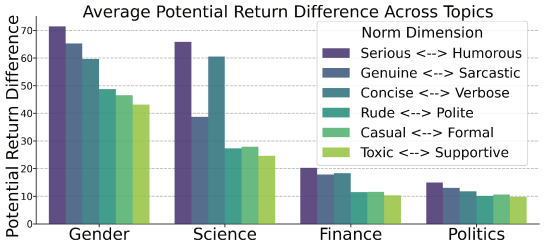


Figure 5: **PRD across topical groups**, reflecting the feedback strategy used by the community to regulate certain norms. All studied communities tend to use positive feedback: the gender related subreddits extensively reward behaviors aligned with their values, while the politics subreddits reward much more conservatively.

with its strict guidelines encouraging detailed explanations and references to policies.

The RPM results of science subreddits show a general disapproval for toxic behaviors that gradually changes to approval as the comments become supportive. `r/askscience` and `r/askscencediscussion`—subreddits designated for scientific discussion with guidelines discouraging offensive language and encouraging helpful answers—show a stronger preference for supportive comments than `r/shittyaskscience`, which is a parody created to mock `r/askscience` (Hessel et al., 2016). Overall, VALUESCOPE effectively surfaces community norms shaped by guidelines and core premises.

What Are the Ideal Norm Behaviors? The point of maximum return (PMR) signifies the behaviors most favored by each community. Figure 4 illustrates the PMR for the top 5 subreddits across humor and supportiveness dimensions. For instance, `r/askscience` prefers supportive comments, as discussed above, and serious comments, which is in line with its explicit community rules (e.g., "memes or jokes are not allowed") and implicit rules identified in prior work; e.g., "no personal anecdotes" (Chandrasekharan et al., 2018). Additionally, all subreddits show a preference for supportiveness over toxicity to varying degrees, which aligns with Redditquette, which are informal values held by most redditors (Fiesler et al., 2018). See Appendix J for PMR results in all dimensions.

Inferring Norm Regulation Strategies Potential return differences (PRD) in Figure 5 reveal how much communities emphasize rewards (PRD>0) or punishments (PRD<0) to enforce norms. All communities significantly favor positive reinforcement, indicating a generally supportive atmosphere (Jackson, 1966), echoing calls for positivity in Redditquette (Fiesler et al., 2018). Moreover, punitive measures are ineffective in maintaining prosocial communities (Mulder, 2008; de Kwaadsteniet et al., 2019; Shen and Rosé, 2022).

Feedback intensity distinctly varies across topics. Gender-related subreddits extensively reward behaviors aligned with their values, suggesting a strong preference for promoting norms that enhance inclusivity and respect. Politics subreddits are more conservative with rewards, possibly due to explicit rules against "disproportionate upvoting" and "brigading," which aim to prevent bias. These regulations may contribute to more measured rewards. Lastly, PRD variations across norm dimensions reveal which normative behaviors are most regulated. The serious–humorous, genuine–sarcastic and concise–verbose dimensions witness the most intense regulation in all groups, suggesting the importance of tone and authenticity of interactions in cultivating social identity (Brown, 2022).

Findings in this section validate VALUESCOPE and, more importantly, allude to the impact of moderation on social norms and potential applications of VALUESCOPE: if undesirable behaviors are detected to rise, moderation strategies should be updated to maintain healthy community norms.

6 Analysis

To address RQ2—*Can we predict the change in norms based on observed normative behaviors?*—we study the fluidity and stability of social norms and its implications using VALUESCOPE and social science theories, specifically norm intensity and crystallization (Jackson, 1966; Nolan, 2015), then analyze their temporal changes in the context of external events and internal community conflicts.

	NI-only		NI+CR		
	$\mathbf{c}_{NI}$	$\mathbf{R}^2$	$\mathbf{c}_{NI}$	$\mathbf{c}_{CR}$	$\mathbf{R}^2$
Politeness	0.26	0.17	0.16	-0.14	0.23
Supportiveness	0.16	0.04	0.05	-0.13	0.10
Sarcasm	0.42	0.13	0.45	-0.13	0.14
Humor	0.50	0.27	0.50	-0.13	0.28
Formality	0.40	0.17	0.27	-0.07	0.18
Verbosity	2.57	0.09	2.57	-0.35	0.09

Table 2: **Coefficients of NI and CR, and R^2** of two linear regression models (NI-only and NI+CR).

Norm Crystallization Social norms are constantly evolving. Understanding such changes and their predictive features can help community moderators respond effectively. Jackson (1966) introduces the concepts of *norm intensity* (NI) and *crystallization* (CR). NI measures the magnitude of community (dis)approval of behaviors at a given normness scale, indicating how strongly the community cares about the norm, while CR represents the level of consensus on the preference.

Taking the year 2021 as a cutoff, we test the predictive power of NI and CR on upcoming temporal changes ($TC := \Delta NI$) with a linear regression model. We use results from VALUESCOPE predictions and follow implementation defined in Linnan et al. (2005) (details in Appendix K). Our results in Table 2 show that NI and $NI+CR$ are both significant predictors of TC, while adding CR increases the coefficient of determination R^2 significantly. Additionally, higher norm intensity and less crystallization (i.e., community members have strong opinions but less agreement) are correlated with larger shifts in norm intensity. Our findings support Jackson (1975)'s hypothesis that these volatile instances are more likely to generate conflicts and trigger changes in norms. This demonstrates VALUESCOPE's potential to help moderators identify and proactively address norms likely to change by setting explicit community rules.

Temporal Change in Norm Intensity We further investigate how NI changes over time, particularly in relation to external events. Figure 6 shows NI of the humor and supportiveness dimensions from 2019-2023 in politics and finance subreddits.

For politics, a significant event during this period is the 2020 U.S. presidential election, represented by the vertical line in the plot (corresponding to July-December 2020). Our results reveal highly similar patterns of norm shifts in `r/republican` and `r/democrats`, characterized by a steep increase of community preference of

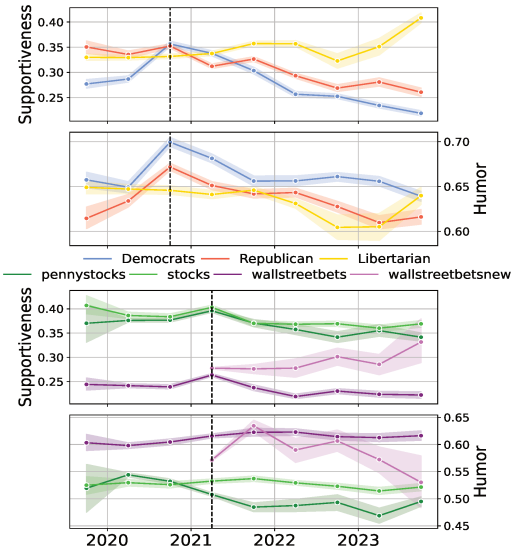


Figure 6: **Temporal changes in average norm intensity** for politics and finance subreddits. Comments were binned by 6 month intervals based on their posting date. For instance, a point for 2020.25 represents the average norm intensity of comments posted from January to June 2020. The vertical lines mark two events: the U.S. presidential election and the creation of `r/wallstreetbetsnew`, highlighting changes before and after these events.

humor and supportiveness during the election period. Following this peak, both dimensions experienced a continuous decline until 2023. On the other hand, `r/libertarian` bears a notable increase in supportiveness over time and was not impacted as much by the election. These results suggest that external events, such as elections, could potentially shape the overall norms in online communities.

For finance subreddits, a notable event was the creation of `r/wallstreetbetsnew`—a spinoff from `r/wallstreetbets`—in 2021 by members dissatisfied with the culture of `r/wallstreetbets` in an attempt to create a less toxic environment focused on serious trading strategies on risky stocks.[8] Among the finance subreddits, our results show that the NI of `r/wallstreetbetsnew` starts diverging from `r/wallstreetbets` and begins to resemble the NI of `r/stocks` and `r/pennystocks`, becoming more supportive and less humorous over time. This finding aligns with Zhang et al. (2021) in showing that new communities establish their own identities and norms over time. Additionally, after the creation of `r/wallstreetbetsnew`, the

[8] As one user noted: "The moderators in the original `r/wallstreetbets` are driving the narrative away from $GME and $AMC and the vibe is very negative/toxic over there" (paraphrased from a subreddit post in `r/wallstreetbetsnew`).

community shift / norm dimension	politeness	supportiveness	sarcasm	humor	formality
r/wallstreetbets → r/wallstreetbetsnew (925.6)	-0.003	0.013	0.003	0.005	0.018
r/wallstreetbets → r/stocks (2157.6)	0.084	0.092	-0.044	-0.062	0.131
r/wallstreetbets → r/pennystocks (1052.0)	0.091	0.094	-0.023	-0.084	0.063
r/askwomen → r/askmen (717.4)	-0.015	-0.022	0.026	0.036	0.004
r/republican → r/democrats (223.8)	0.026	0.016	0.036	0.018	-0.008

Table 3: User behavior shifts in select subreddit transition pairs. Gray cells indicate changes that are insignificant ($p > 0.05$); red and green cells represent significant negative and positive changes.

NI of r/wallstreetbets also shifts, becoming less supportive and more humorous. This suggests that the culture of the original community may be influenced when some members leave to form a new spinoff community as explored below.

Community Norm Adaptation by Users Social norms can influence the behavior of community members (McDonald and Crandall, 2015), so we examine how individual users modify their language and interaction styles based on the subreddit they are participating in. We define user-level norm behavior in a community as the average NI of comments left by the specific user in that community. For related subreddits with shared users, we compute the change in normative behavior of these users when they switch from subreddit A to subreddit B using a paired two-tailed t-test (Table 3), with experimental details and full results in Appendix L.

Our results reveal significant variability in user normative behaviors between the selected subreddit pairs. For example, users in r/wallstreetbets, known for its usage of profane jargon and aggressive trading strategies (Herrman, 2021), significantly modify their behaviors in r/stocks and r/pennystocks, but adapt much less in the spinoff subreddit r/wallstreetbetsnew. Additionally, user behaviors tend to remain consistent in identity-related subreddits (e.g., r/askwomen, r/askmen) or those with competing relationships (r/republican, r/democrats), highlighting the context-specific nature of community norm adaptations by users. We also observe that users are more likely to change their formality to fit different subreddit contexts than other dimensions, such as humor, indicating that certain norms are more malleable and adaptable than others.

Different extents to which users adapt their language to the audience suggest that digital identities are fluid and context-dependent. This can inform the development of tailored moderation tools to align with the behavioral norms of specific communities, potentially improving user experience and engagement on a more fine-grained level.

7 Conclusion & Future Directions

We introduced VALUESCOPE, a novel framework based on the RPM theory from social science, to quantify social norms and values at scale. We comprehensively validated the effectiveness of VALUESCOPE to assess the normness of behaviors and predict community preferences while controlling for confounders. VALUESCOPE enables numerous quantitative analyses, including predicting norm shifts and contextualizing temporal changes with external events, providing a deeper understanding of social norm dynamics in online communities.

Our work contributes a robust and generalizable method that can be easily extended to various norms and communities. It opens up many exciting possibilities for applications and future research:

Computational Modeling Applications Our framework can enhance community moderation tools by integrating theoretically grounded insights, such as maximum return potential, to refine toxicity detectors. It can also guide generation models to produce contextually appropriate responses specialized to each community's unique norms.

Applications for Social Scientists Our method empowers the development of new hypotheses about social norms, by providing social scientists with enhanced tools to explore how norms form and influence social interactions within communities.

Support Tools for Communities VALUESCOPE can enhance community management by enabling moderators to monitor and address norm shifts in real-time. It can help transform widely accepted but informal norms into explicit rules, clarifying guidelines and easing new member integration. This approach is applicable in various settings (e.g., workplaces) where it can guide individuals on appropriate cultural expressions, improving their integration and acceptance. Platform developers can use this method to refine community recommendation engines, aligning users with groups that match their preferences and values, thereby enhancing user engagement and community growth.

Limitations

Return Potential Model In this work, we introduce VALUESCOPE, a novel framework based on the RPM theory in social science. However, the RPM specifically measures the potential approval by other community members, representing only one dimension of broader norm structures in a community. Prior cross-sectional survey work employed the RPM and expanded towards the descriptive dimension of norms[9] (Wallen and Kyle, 2018). Future works can expand our current computational model of RPM, incorporating the broader norms and values within online communities.

Platform and Language Scope While VALUESCOPE is not limited to any specific platform or language, our work focused on English comments on Reddit. We believe interesting future directions include extending our framework to various other platforms that provide similar community preference signals, such as YouTube comments. Additionally, expanding to other languages would enable more in-depth cross-cultural analyses of community norms.

Role of Other Stakeholders To understand the implicit norms in communities, we focus on the interactions between community members through comments and their upvotes. However, stakeholders such as users, moderators, and other interested parties constantly negotiate norms in online communities (Kim, 2006). Thus, future works should explore the role of moderators and other stakeholders in potentially shaping the implicit norms in online communities.

Dynamic Nature of Norms Our study quantifies and predicts the community norms and values at scale. However, as shown in §6, norms are dynamic and constantly changing over time (Bicchieri, 2005). Our methodology, such as the RPM and the experimental setup, are compatible with future temporal analyses.

Predictions on Synthetic Comments In our work, we employ synthetic comments to simulate community preference for comments with varying normness scale. Predicting the community approval of synthetic comments may potentially add noise to our results. However, we aimed to address this limitation by employing an extensive filtering process based on prior works (Briakou et al., 2021; Mir et al., 2019) and validating the quality of the filtered data using expert human annotations (See §4.2.2).

Investigating deeper and beyond norm dimensions and community topics. In §4.1, we employ an inductive norm identification process to surface six overarching norm dimensions and select subreddit topics based on prior works. However, there are several other dimensions to explore beyond these six, such as optimism, empathy, and confidence. Meanwhile, there are several other relevant and interesting subreddit topics, such as ones based on cultures and nations (`r/korea` and `r/southafrica`). VALUESCOPE can facilitate future analyses on different norm dimensions and topics of communities.

Model Error Cascades We train small local models as the normness and preference predictors. Despite extensive model training and experimentation,the error rates in our VALUESCOPE pipeline may potentially influence our downstream analysis. Thus, we designed our pipeline to mitigate as much noise as possible (for example, "Community Preference Distillation" in §3.3) and validate our findings with prior work and existing community guidelines.

Ethical Considerations

We use publicly accessible LLMs to conduct our research, which includes generating more toxic versions of comments. In our investigation to understand the implicit norms of online communities, our experiments inevitably produced toxic content to measure how communities react to toxicity. However, we believe the benefits of our research outweigh the risks, as community moderators and platform developers can use our framework to understand the implicit norms in various communities, especially in response to toxic content, and self-assess and monitor their culture. The generated toxic content was only used to compute aggregated metrics to identify high-level patterns, and it will not be released to the public. To ensure reproducibility while protecting the rights of Reddit users, we will only release the IDs of the comments used in our analysis. Using these provided IDs, practitioners will need to independently fetch the comments from the publicly accessible Reddit Dump.

[9] Descriptive norms represents the beliefs of common or typical behaviors.

Acknowledgement

This work was supported by the National Science Foundation (Grant No. IIS-2143529) and the National Institutes of Health (Grant Nos. 5R21DA056725-02, R21DA056725-01A1). We gratefully acknowledge support from the National Science Foundation under CAREER Grant No. IIS2142739, and NSF grants No. IIS2125201 and IIS2203097.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Cristina Bicchieri, Ryan Muldoon, and Alessandro Sontuoso. 2023. Social Norms. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.

Nadav Borenstein, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2024. Investigating human values in online communities.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew D Brown. 2022. Identities in and around organizations: Towards an identity work perspective. *Human relations*, 75(7):1205–1237.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Walter Coutu. 1951. Role-playing vs. role-taking: An appeal for clarification. *American sociological review*, 16(2):180–187.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanushree Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in llm generated conversations.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: user lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web*.

Erik W de Kwaadsteniet, Toko Kiyonari, Welmer E Molenmaker, and Eric van Dijk. 2019. Do people prefer leaders who enforce norms? reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, 84:103800.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Daryna Dementieva, Ivan Trifinov, Andrey Likhachev, and Alexander Panchenko. 2022. Detecting text formality: A study of text classification approaches. *arXiv preprint arXiv:2204.08975*.

Larry Dignan. 2024. Reddit's data licensing play: Do you want your llm trained on reddit data?

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Penelope Eckert and Sally Mcconnell-Ginet. 1999. New generalizations and explanations in language and gender research. *Language in Society*, 28:185 – 201.

Penelope Eckert and Sally McConnell-Ginet. 2013a. *Language and Gender*, 2 edition. Cambridge University Press.

Penelope Eckert and Sally McConnell-Ginet. 2013b. *Language and gender*. Cambridge University Press.

Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2024. Inverse constitutional ai: Compressing preferences into principles.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking-training with large-scale human feedback data. In *EMNLP*.

Carroll J Glynn and Michael E Huge. 2007. Opinions as norms: Applying a return potential model to the study of communication behaviors. *Communication Research*, 34(5):548–568.

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

Venkata S Govindarajan, Kyle Mahowald, David I Beaver, and Junyi Jessy Li. 2023. Counterfactual probing for the influence of affect and specificity on intergroup bias. *arXiv preprint arXiv:2305.16409*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Libby Hemphill and Jahna Otterbacher. 2012. Learning the lingo? gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 305–314, New York, NY, USA. Association for Computing Machinery.

David B Henry, Jennifer Cartland, Holly Ruchross, and Kathleen Monahan. 2004. A return potential measure of setting norms for aggression. *American Journal of Community Psychology*, 33(3-4):131–149.

John Herrman. 2021. Everything's a joke until it's not.

Jack Hessel, Chenhao Tan, and Lillian Lee. 2016. Science, askscience, and badscience: On the coexistence of highly related communities. In *International Conference on Web and Social Media*.

Jay Jackson. 1966. A conceptual and measurement model for norms and roles. *The Pacific Sociological Review*, 9(1):35–47.

Jay Jackson. 1975. Normative power and conflict potential. *Sociological Methods & Research*, 4(2):237–263.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Anna Kasunic and Geoff Kaufman. 2018. " at least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/roastme. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, page 553–562, New York, NY, USA. Association for Computing Machinery.

Amy Jo Kim. 2006. *Community building on the web: Secret strategies for successful online communities*. Peachpit press.

Sanford Labovitz and Robert Hagedorn. 1973. Measuring social norms. *Pacific Sociological Review*, 16(3):283–303.

Robin Lakoff. 1973. The logic of politeness: Or, minding your p's and q's. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 9, pages 292–305. Chicago Linguistic Society.

Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik W. Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Gov. Inf. Q.*, 31:317–326.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*.

Laura Linnan, Anthony D LaMontagne, Anne Stoddard, Karen M Emmons, and Glorian Sorensen. 2005. Norms and their relationship to behavior in worksite settings: an application of the jackson return potential model. *American journal of health behavior*, 29(3):258–268.

Li Lucy and David Bamman. 2021. Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Charles G McClintock. 1978. Social values: Their definition, measurement and development. *Journal of Research & Development in Education*.

Rachel I McDonald and Christian S Crandall. 2015. Social norms and social influence. *Current Opinion in Behavioral Sciences*, 3:147–151.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Shyamal Mishra and Preetha Chatterjee. 2023. Exploring chatgpt for toxicity detection in github. *arXiv preprint arXiv:2312.13105*.

Laetitia B Mulder. 2008. The difference between punishments and rewards in fostering moral concerns in social decision making. *Journal of Experimental Social Psychology*, 44(6):1436–1443.

Yair Neuman and Yochai Cohen. 2023. Ai for identifying social norm violation. *Scientific Reports*, 13(1):8103.

Jessica M Nolan. 2015. Using jackson's return potential model to explore the normativeness of recycling. *Environment and Behavior*, 47(8):835–855.

OpenAI. 2024a. Best practices for prompt engineering with the openai api. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api. Accessed:2024-01-11.

OpenAI. 2024b. Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering. Accessed:2024-01-11.

OpenAI. 2024. Text generation models. https://platform.openai.com/docs/guides/text-generation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya,

Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *International Conference on Web and Social Media*.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Qinlan Shen and Carolyn P Rosé. 2022. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 932–943.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Aaron J Snoswell, Lucinda Nelson, Hao Xue, Flora D Salim, Nicolas Suzor, and Jean Burgess. 2023. Measuring misogyny in natural language generation: Preliminary results from a case study on two reddit communities. *arXiv preprint arXiv:2312.03330*.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2023. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *AAAI Conference on Artificial Intelligence*.

Claudio Vaz Torres. 1999. *Leadership style norms among americans and brazilians: assessing differences using jackson's return potential model*. California School of Professional Psychology-San Diego.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

UNICEF. 2021. Defining social norms and related concepts.

Michiel van der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

Kenneth Wallen and Gerard Kyle. 2018. Extending the return potential model with a descriptive normative belief measure. *Society & Natural Resources*, 31:1–7.

Galen Weld, Amy X. Zhang, and Tim Althoff. 2024. Making online communities 'better': A taxonomy of community values on reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):1611–1633.

Etienne Wenger-Trayner and Beverly Wenger-Trayner. 2015. Introduction to communities of practice: A brief overview of the concept and its uses.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jason Shuo Zhang, Brian C. Keegan, Qin Lv, and Chenhao Tan. 2021. Understanding the diverging user trajectories in highly-related online communities during the covid-19 pandemic.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wen Zhang, Yunhan Liu, Yixuan Dong, Wanna He, Shiming Yao, Ziqian Xu, and Yan Mu. 2023. How we learn social norms: a three-stage model for social norm learning. *Frontiers in Psychology*, 14:1153809.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation.

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.

A Nomenclature & Definition References

- **Norm:** Informally agreed-upon rules governing community behavior, such as the expectation of toxicity or politeness in interactions.
- **Value:** The deeper ideals and principles that a community aspires to embody and promote. Values are fundamental in shaping and guiding the development of norms.
- **Behavior:** The observable actions taken by community members, such as the comments they post in a subreddit.
- **Norm Dimension:** Attributes or characteristics of behaviors that can be measured along a (bidirectional) continuum, serving as a quantitative axis for analyzing norm adherence.
- **Normative Behavior:** Actions that align with a specific norm dimension, such as expressions of support or aggression in user comments.
- **Normness Scale:** A metric indicating the extent to which a behavior conforms to a particular norm dimension.
- **Community Preference:** The collective judgment expressed by community members through mechanisms of approval or disapproval, quantified by the net balance of upvotes and downvotes a comment receives.

B Converting Binary Classification to Continuous Normness Scale

We reformulate the normness scale measurement module from a regression task to a binary classification task. After getting the binary labels of pairs of comments, we convert the binary labels into numerical scores as follows:

Given comments $\mathcal{A} = \{a_1, \ldots, a_n\}$ with ground truth normness scales $\Phi_d(\mathcal{S}) = \{\phi_1 \ldots, \phi_n\}$, we have binary labels $\mathcal{B}_d = \{\beta_{ij} | 1 \leq i < j \leq n, \beta_{ij} = \begin{cases} 1 & \text{if } \phi_i < \phi_j, \\ 0 & \text{otherwise,} \end{cases}$ as target labels of the classifier $\mathcal{M}_d : \mathcal{A} \times \mathcal{A} \rightarrow \{0, 1\}$. For any comment $a_k \in \mathcal{A}$, its adjusted normness scale, $\phi'_k = \Phi'_d(a_k)$, is defined as the win-rate of a_k compared against all other comments in $\mathcal{A}$:

$$\Phi'_d(a_k) := \frac{1}{n-1}\left(\sum_{i=1}^{k-1} \beta_{ik} + \sum_{i=k+1}^{n} (1 - \beta_{ki})\right), \quad (1)$$

which is the percentage of times that a_k is labeled as having a higher normness degree, when compared with other comments in the set of comments $\mathcal{A}$.

B.1 Monotonicity of Binary Win-rate as Normness Scale

We now should that if we are only interested in the relative normness scales of comments, the binary win-rate and the normness scale are monotonic.

Recall that for a set of comments $\mathcal{A} = \{a_1, \ldots, a_n\}$ with ground truth normness scales $\Phi_d(\mathcal{S}) = \{\phi_1 \ldots, \phi_n\}$, we have binary label set $\mathcal{B}_d = \{\beta_{ij} | 1 \leq i < j \leq n\}$ defined above, from which we obtain $\Phi'_d(a_k)$.

We prove that the two metrics Φ_d and Φ'_d are monotonic with respect to each other by showing that, $\forall i, j \in [1, n]$ s.t. $1 \leq i < j \leq n \in \mathbb{R}$ s.t. if $\phi_i \leq \phi_j$, then $\phi'_i \leq \phi'_j$ and if $\phi_i \geq \phi_j$, then $\phi'_i \geq \phi'_j$.

First, let $\mathcal{A}^-, \mathcal{A}^+, \mathcal{A}^*$ be subsets of $\mathcal{A}$ such that

$$\mathcal{A}^- := \{a' | \Phi_d(a') < \phi_i\}, \quad (2)$$

$$\mathcal{A}^+ := \{a'' | \Phi_d(a'') \geq \phi_j\}, \text{ and} \quad (3)$$

$$\mathcal{A}^* := \{a^* | \Phi_d(a^*) \geq \phi_i, \Phi_d(a^*) < \phi_j\}. \quad (4)$$

Let $p = ||\mathcal{A}^-||$, $q = ||\mathcal{A}^+||$, $r = ||\mathcal{A}^*||$ and $s = \mathcal{I}_{\{\phi_i < \phi_j\}}(i, j)$ (the indicator function where $s = 1$ if $\phi_i < \phi_j$ and $s = 0$ if $\phi_i = \phi_j$). Then, we can compute win-rates ϕ'_i and ϕ'_j as:

$$\phi'_i = \frac{1}{n-1}(p \cdot 1 + q \cdot 0 + r \cdot 0 + s \cdot 0) \quad (5)$$

$$= \frac{p}{n-1} \quad (6)$$

$$\phi'_j = \frac{1}{n-1}(p \cdot 1 + q \cdot 0 + r \cdot 1 + s \cdot 1) \quad (7)$$

$$= \frac{p + r + s}{n-1}. \quad (8)$$

Since $r \geq 0$ and $s \geq 0$, we have $\phi'_j - \phi'_i = \frac{1}{n-1}(r+s) \geq 0$ and $\phi'_i - \phi'_j \leq 0$. Thus, we have $\phi'_i \leq \phi'_j$ for arbitrary i and j. Similarly, we can show that if $\phi_i \geq \phi_j$, then $\phi'_i \geq \phi'_j$. Therefore, we proved that the two metric are monotonic.

C Subreddit Selection Details

To form the dataset used in this study, we first select subreddit topics based on relevance and prior work, obtaining gender, politics, finance, and science. Then, for each topic, we take the most representative subreddits out of the top 5,000 SFW (safe-for-work) subreddits based on the size of the subreddit. For the gender topical group, we have `r/askmen`, `r/askwomen` and `r/asktransgender`; for the politics topical group, we have `r/republican`, `r/demcorats` and `r/libertarian`. For the science topical groups, we select `r/askscience`, its spinoff subreddit `r/shittyaskscience` which was created to mock `r/askscience`, and a more open variant `r/asksciencediscussion` that discusses topics *in* science and *related* to science, such as academia (Hessel et al., 2016). Lastly, for the finance-related topics, we selected the most popular three subreddits from the top 5,000: `wallstreetbets`, `stocks`, `pennystocks`, and additionally consider `r/wallstreetbetsnew`, which is the spinoff subreddit of `r/wallstreetbets`. Table 4 summarizes the topics, subreddits, and dataset sizes examined in this study.

Topic	Subreddit	Raw Data	Synthetic Data
Gender	r/askmen	4.56M	1.08M
	r/askwomen	2.13M	1.21M
	r/asktransgender	1.61M	1.01M
Politics	r/libertarian	3.66M	1.00M
	r/democrats	534K	922K
	r/republican	502K	1.01M
Science	r/askscience	426K	1.23M
	r/shittyaskscience	185K	761K
	r/asksciencediscussion	141K	1.10M
Finance	r/stocks	3.51M	1.05M
	r/pennystocks	1.23M	1.04M
	r/wallstreetbets	49.3M	864K
	r/wallstreetbetsnew	655K	784K

Table 4: Selected online communities (subreddits) across various topics. For each subreddit, we show the number of existing comments within the community (column "Raw Data") and the number of synthetic comments remaining after applying filters to ensure the quality of the simulated comments (column "Synthetic Data").

D Grounding 5-point Scale for Normness Ratings

In §4.2, we employ a 5-point Likert scale using GPT-3.5 to rate comments and sample them to gauge their normness. Additionally, in §4.2.2, we generate five variations of each original seed comment based on the 5 different scales of normness. Thus, for each norm dimension, we created a 5-point Likert scale and grounded their definitions in prior works (Dementieva et al., 2022; Wulczyn et al., 2017; Joshi et al., 2017; Goffman, 1955; Brown and Levinson, 1987; Lakoff, 1973). For example, we define formality based on using abbreviations, slang, colloquialisms, non-standard capitalizations, complete sentences, contractions, punctuations, and opening expressions of sentences (Dementieva et al., 2022). Meanwhile, we define politeness as a set of strategies for conducting face-threatening acts while minimizing the chance that we or others will lose our positive or negative faces. (Brown and Levinson, 1987). The 5-point Likert scale across the norm dimensions can be found in Figures 61-63 as well as Figure 10.

E GPT Evaluations

Recall in §4.2.1 that we employ GPT-3.5 to sample and rate comments on a 5-point Likert scale (defined in Appendix D) for a particular norm dimension and subsequently use GPT-4 to generate binary synthetic labels comparing a pair of comments. In Appendix E.1, we describe the process of curating human annotations. In Appendix E.2, we evaluate the quality of GPT-3.5 rating. In Appendix E.3, we describe our prompt design considerations and prompt tuning results. In Appendix E.4, we evaluate the final GPT-4 automatic pairwise labeling pipeline using the human annotations.

E.1 Normness Scale Annotation

To evaluate the NSP models and the quality of GPT-4 generated labels for student models, we curate a high-quality human annotation set of 450 samples for each norm dimension. The human annotations of norms are challenging due to subjectivity. To reduce subjectivity, we conducted training sessions with annotators and iteratively improved our annotation guidelines, grounding the definitions of various norms based on prior works (see Appendix D). Each sample was annotated by three volunteer annotators, who are graduate students in NLP and Linguistics at a US-based institution and familiar

with the subreddits in our study. We did not provide payment, but we obtained consent to use their annotations for AI model evaluation.

For each topic, we use stratified random sampling to select two comments from various subreddits, creating pairs of comments. We then ask three human annotators to make binary judgments on which comment exhibits a higher normness scale for five norm dimensions (e.g., which one is more formal/less casual?). For each annotation, we chose the binary judgment with at least a majority agreement among three annotators[10].

Across the four topics, we collected human annotations for 450 samples[11]. Each sample was annotated for five norm dimensions, resulting in a total of 2,250 annotations per human annotator.

The average inter-annotator agreement, measured by Fleiss's κ, was 0.56, considered a moderate agreement (Landis and Koch, 1977). Due to the nuance and subtlety of norms, Fleiss's $\kappa = 0.56$ provides a solid foundation for our annotation labels. For instance, Passonneau and Carpenter (2014) reported scores as low as 0.2 in subjective tasks such as word sense annotations. Refer to Table 5 for the full agreement scores across 4 topics and 5 norm dimensions.

Figure 60 shows the annotation interface we used to collect human annotations for evaluating GPT-4 and Normness Scale Predictor models. Figure 61, Figure 62, and Figure 63 display the guidelines provided to human annotators to help them better understand each norm dimension.

E.2 Evaluating the Quality of GPT-3.5 Rating

To evaluate the quality of GPT-3.5's rating capabilities on a 5-point Likert scale, we employ the human-annotated gold labels from Appendix E.1. The labels indicate which of the two pairwise comments exhibits a greater normness scale for five norm dimensions (e.g., which one is more formal/less casual). By comparing GPT-3.5's rating of these pairwise comments to the binary gold labels, we can evaluate its relative rating quality. For example, if the binary gold label indicates that comment A (e.g., "ty!") is more casual than comment B (e.g., "thank you"), then GPT-3.5 should ideally rate comment A as 1 (Very Casual) and comment B as 4 (Formal), in alignment with the binary label. Refer to Figure 7 for the rating prompt.

Topic	Formality	Supportiveness	Sarcasm	Politeness	Humor
Gender	0.41	0.77	0.56	0.69	0.70
Politics	0.48	0.44	0.46	0.47	0.54
Science	0.66	0.75	0.70	0.71	0.77
Finance	0.57	0.40	0.40	0.47	0.57

Table 5: The Fleiss' κ coefficient among three human annotators for their annotations for each topic across 5 dimensions. Each annotator was provided with two pairwise comments from subreddits chosen in the topic, labeling which comments exhibited more of the dimension (e.g., *more* formal). The κ coefficient ranges from 0.40-0.78, indicating a moderate to substantial agreement (Landis and Koch, 1977).

Table 6 presents the percentage alignment between GPT-3.5's rating and 100 binary gold labels on pairwise comments from `r/askmen` and `r/askwomen`[12]. We found that GPT-3.5's ratings aligned with the gold labels 77%-90% of the time, validating the quality of GPT-3.5's rating labels.

Formality	Supportiveness	Sarcasm	Politeness	Humor
85%	90%	77%	79%	82%

Table 6: GPT-3.5 Rating Evaluation Results. Across the 5 norm dimensions, we found that GPT-3.5's rating of two pairwise comments aligned with the gold labels 77%-90% of the time, validating the quality of GPT-3.5's rating labels.

E.3 GPT-4 Automatic Pairwise Labeling

We underwent extensive prompt-tuning efforts to generate high-quality and accurate binary synthetic labels using GPT-4. Below, we discuss our prompt design choices (§E.3.1), the prompt tuning results to select the best prompt for our task (§E.3.2), and the full evaluation results of the chosen prompt against human annotations (§E.4).

E.3.1 Prompt Design Considerations

Since we employed OpenAI models, our prompt design variations were guided by OpenAI's recommendations on prompt-engineering (OpenAI, 2024a) and prior works (Mishra and Chatterjee,

[10]We discarded annotated samples whose final labels were "hard-to-tell" or "media-needed" as these samples could not be properly annotated with the given context.

[11]For all topics except "Gender," we annotated 100 randomly-sampled pairwise comments. For "Gender" topic, we annotated 150 pairwise comments, in which 100 pairwise comments came from `r/askmen` and `r/askwomen` while the remaining 50 pairwise comments came from comparisons with one of the gender subreddits (including `r/asktransgender`) and `r/asktransgender`.

[12]We discard cases where GPT-3.5 assigned the same rating to both pairwise comments, as these cannot be evaluated against the binary gold labels.

```
You are a linguistic expert who is tasked with identifying and confirming linguistic features
present in Reddit comments.

Please rate the COMMENT, only using the POST TITLE and POST DESCRIPTION as context, on the
provided [DIMENSION] SCALE.

[DIMENSION] SCALE: [DIMENSION-5POINT-LIKERT-SCALE]

Please rate the COMMENT using the provided scale on [DIMENSION] and provide reasoning for your
answer. Place rating between square brackets (i.e. []).
POST TITLE: [TITLE]
POST DESCRIPTION: [DESCRIPTION]
COMMENT: [COMMENT]
```

Figure 7: The zero-shot prompt used with GPT-3.5 to rate sampled comments on a 5-point Likert scale. We adapted the 5-point Likert scale based on the norm dimension (refer to Appendix D).

2023; Dammu et al., 2024). Below, we list the various prompt design features we considered:

- **System Roles:** According to OpenAI (2024b), asking the model to adopt a persona in their systems could lead to better results. Thus, we prompted the GPT models to adopt the persona of a "linguistic expert": `"You are a linguistic expert tasked with comparing which linguistic dimension is more present between two Reddit comments."`
- **Contextual Details:** Given that providing proper contextual details is helpful to LLMs to reason and justify their decisions (OpenAI, 2024a), we include the definitions of each norm dimension summarized from prior works (See Appendix D).
- **Zero-Shot vs. Few-Shot:** For our task, we experimented with zero-shot and few-shot prompts. Zero-shot prompts involve presenting the task to the LLM without any accompanying examples. Meanwhile, few-shot prompts involve conditioning the pre-trained language model to accompanying examples rather than updating its weights (Brown et al., 2020). To apply this concept to our task, we provided three few-shot examples per norm dimension. Each few-shot example consists of the post titles, descriptions, comments, and the reasoning behind the provided example label. The authors manually crafted the few-shot examples for each of the norm dimensions.
- **Temperature:** We explored with varying temperature levels to find the most optimal parameters for our task. Temperature influences how models generate text (OpenAI, 2024), ranging from 0 (more deterministic, consistent) to 2 (more non-deterministic, random). Prior work (Mishra and Chatterjee, 2023; Dammu et al., 2024) found that temperature settings of 0.2 and 0.7 resulted in the best performances. Likewise, we selected these two temperature settings for our task.
- **Self-Consistency:** Prior work have shown that "self-consistency" prompting improves performance, especially in reasoning tasks (Singhal et al., 2023). Self-consistency involves prompting the language model multiple times and choosing the answer that receives the majority vote. Thus, we experiment with 3, 5, and 10 paths (e.g. number of times prompting the model).
- **Models:** We experiment with various OpenAI models and versions, such as `gpt-3.5-turbo-0125`, `gpt-4-0125-preview`, `gpt-4-1106-preview`, `gpt-4-0613`, and `gpt-4o-2024-05-13`.

E.3.2 Prompt-Tuning Results

Based on the proposed features in §E.3.1, we design multiple prompting pipelines and evaluate their performance on the binary labeling task—given two comments, compare the comments in each of the five norm dimensions. Performance is measured by the label accuracy against a human-annotated gold data, thus assessing the effect of different prompting pipelines to produce accurate labels.

Table 7 shows the results of our prompt tuning evaluation, which examined various combinations of models, zero-shot vs. few-shot, temperature, and self-consistency. We found that **few-shot prompts utilizing GPT-4, self-consistency, and temperature 0.7 provided the best overall performances** (Index 16-18). However, we also found few-shot prompts using GPT-4 and temperature 0.2 (Index

Index	Model	Zero-Shot vs. Few-Shot	Temperature	Self-Consistency	Formality	Supportiveness	Sarcasm	Politeness	Humor
0	gpt-3.5-turbo-0125	Zero-Shot	0.2	-	0.85	0.80	0.56	0.65	0.55
1	gpt-3.5-turbo-0125	Few-Shot	0.2	-	0.85	0.70	0.39	0.70	0.55
2	gpt-4-0613	Zero-Shot	0.2	-	0.90	**0.90**	0.61	0.75	0.60
3	gpt-4-0613	Few-Shot	0.2	-	0.80	**0.90**	**0.83**	0.75	0.65
4	gpt-3.5-turbo-0125	Zero-Shot	0.2	3	0.75	0.80	0.56	0.70	0.55
5	gpt-3.5-turbo-0125	Zero-Shot	0.2	5	0.80	0.80	0.56	0.70	0.50
6	gpt-3.5-turbo-0125	Zero-Shot	0.2	10	0.90	0.80	0.56	0.70	0.65
7	gpt-3.5-turbo-0125	Zero-Shot	0.7	3	0.80	0.80	0.56	0.65	0.60
8	gpt-3.5-turbo-0125	Zero-Shot	0.7	5	0.80	0.80	0.67	0.65	0.60
9	gpt-3.5-turbo-0125	Zero-Shot	0.7	10	**0.95**	0.80	0.56	0.65	0.58
10	gpt-4-0613	Zero-Shot	0.2	3	0.80	**0.90**	0.67	0.75	0.60
11	gpt-4-0613	Zero-Shot	0.2	5	0.80	**0.90**	0.67	0.75	0.60
12	gpt-4-0613	Zero-Shot	0.2	10	0.80	**0.90**	0.67	0.75	0.60
13	gpt-4-0613	Zero-Shot	0.7	3	0.80	**0.90**	0.67	0.75	0.60
14	gpt-4-0613	Zero-Shot	0.7	5	0.80	**0.90**	0.67	0.75	0.60
15	gpt-4-0613	Zero-Shot	0.7	10	0.85	**0.90**	0.67	0.70	0.60
16	gpt-4-0613	Few-Shot	0.7	3	0.75	**0.90**	**0.83**	**0.80**	**0.70**
17	gpt-4-0613	Few-Shot	0.7	5	0.75	**0.90**	**0.83**	**0.80**	**0.70**
18	gpt-4-0613	Few-Shot	0.7	10	0.75	0.90	**0.83**	**0.80**	**0.70**

Table 7: Prompt Tuning Results evaluating various combinations of models, zero/few-shot, temperature, and self-consistency. For each prompt, we report the accuracy across the 5 norm dimensions. The highest performance value in each column is in **bold**. To save computational expense, these results were based on 20 sampled gold labels comparing comments between r/askmen and r/askwomen.

Index	Model	Zero-Shot vs. Few-Shot	Temperature	Self-Consistency	Formality	Supportiveness	Sarcasm	Politeness	Humor
19	gpt-4-0613	Zero-Shot	0.2	-	0.79	**0.90**	0.67	0.81	0.76
20	gpt-4-0613	Few-Shot	0.2	-	**0.80**	**0.90**	**0.84**	**0.86**	0.78
21	gpt-4o-2024-05-13	Few-Shot	0.2	-	**0.80**	0.88	0.76	0.80	**0.84**
22	gpt-4-0125-preview	Few-Shot	0.2	-	0.77	0.83	0.65	0.76	0.83
23	gpt-4-1106-preview	Few-Shot	0.2	-	0.71	0.84	0.70	0.77	**0.84**

Table 8: Additional Prompt Tuning Results utilizing few-shot prompting on various GPT-4 models. Unlike Table 7, these results were based on 100 gold-labels comparing comments between r/askmen and r/askwomen. We report the accuracy across the 5 norms, **bolding** the highest performance value in each column. We found that GPT-4 (Index 20) obtained the best overall performance across the norm dimensions.

3), even without self-consistency, performed comparably. Since self-consistency significantly increases computational expenses due to repeated prompting, we selected the prompt setting at Index 3, which provides comparable results without self-consistency. We provide the few-shot prompt in Figure 8.

To select the most optimal model for our task, we conducted further prompt-tuning using few-shot prompting at a 0.2 temperature on various GPT-4 versions, including gpt-4-0125-preview, gpt-4-1106-preview, gpt-4-0613, and gpt-4o-2024-05-13. We present the results in Table 8. Overall, gpt-4-0613 provided the best overall performance, ranging from 0.78-0.90 accuracy across the norm dimensions. Thus, we use the gpt-4-0613 version with few-shot prompts at 0.2 temperature to generate the binary synthetic labels, which are then used to train the NSP model (refer to §4.2.1).

E.4 Evaluating the Chosen GPT-4 Labeling Pipeline

The quality of the final GPT-4 generated labels is shown in Table 9, where we report the accuracy and F1 scores of the GPT-4-generated labels compared against human annotations from Appendix E.1. In our evaluation, GPT-4 achieved an average accuracy of 0.74-0.82 and a macro F1-score of 0.74-0.82 across the topics. These results demonstrate sufficient data quality to train a small classifier model.

Topic	Formality		Supportive		Sarcasm		Politeness		Humor		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Gender	0.75	0.74	**0.92**	**0.92**	0.81	0.81	0.85	0.84	0.77	0.77	**0.82**	**0.82**
Politics	0.77	**0.77**	0.74	0.73	0.72	0.71	0.74	0.72	0.72	0.72	0.74	0.74
Science	0.74	0.74	0.82	0.81	**0.84**	**0.84**	0.81	0.79	**0.84**	**0.84**	0.81	0.81
Finance	**0.78**	0.76	0.85	0.85	0.76	0.76	**0.87**	**0.87**	0.81	0.81	0.81	0.81

Table 9: For each topic and dimension, we note the accuracy (Acc.) and the F1-score (F1) of the synthetic labels generated by GPT-4 based on human annotations. The highest performance value in each column is highlighted in **bold**.

```
You are a linguistic expert tasked with comparing which linguistic dimension is more present
between two Reddit comments.

Between COMMENT1 and COMMENT2, please determine which comment is [DIMENSION_PAIRWISE] and provide
reasoning for your answer. Only use the provided post title and post description as context. The
[DIMENSION] definition is provided below to help determine which comment is [DIMENSION_PAIRWISE].

[DIMENSION] definition: [DIMENSION_DEFINITION]

We provide three examples of the task, each featuring two sets of comments alongside their
respective post titles, descriptions, answer, and reasoning.

Example 1:
EXAMPLE1_POST_TITLE1: [EXAMPLE1_TITLE1]
EXAMPLE1_POST_DESCRIPTION1: [EXAMPLE1_DESCRIPTION1]
EXAMPLE1_COMMENT1: [EXAMPLE1_COMMENT1]
EXAMPLE1_POST_TITLE2: [EXAMPLE1_TITLE2]
EXAMPLE1_POST_DESCRIPTION2: [EXAMPLE1_DESCRIPTION2]
EXAMPLE1_COMMENT2: [EXAMPLE1_COMMENT2]
EXAMPLE1_ANSWER: "1. EXAMPLE1_COMMENT1 exhibits a more formal tone compared to EXAMPLE1_COMMENT2.
EXAMPLE1_COMMENT1 maintains a structured approach, using relatively complete sentences, standard
capitalization, and correct punctuations. Meanwhile, EXAMPLE1_COMMENT1 is much more casual, using
abbreviations (i.e. "tbh") and consistently lacking syntatic components."
...

Example 3:
...

Now, given what you learned from the examples, if you think COMMENT1 is [DIMENSION_PAIRWISE],
ANSWER WITH "1" at the beginning of your response. If you think COMMENT2 is [DIMENSION_PAIRWISE],
ANSWER WITH "2" at the beginning of your response.

"POST TITLE1: [TITLE1]"
"POST DESCRIPTION1: [DESCRIPTION1]"
"COMMENT1: [COMMENT1]"
"POST TITLE2: [TITLE2]"
"POST DESCRIPTION2: [DESCRIPTION2]"
"COMMENT2: [COMMENT2]"
```

Figure 8: The few-shot prompt employed to generate binary synthetic labels to train the normness scale predictor. In the prompt, we provide three few-shot examples consisting of the post titles, descriptions, comments, and the reasoning justifying the provided example label. The few-shot examples and the prompts were adapted based on the norm dimension. For example, using formality as a dimension, `[DIMENSION_PAIRWISE]` was replaced with "MORE FORMAL or (LESS CASUAL).

F GPT Cost Estimation

Recall in §4.2.1 that we sample and rate comments on a 5-point Likert scale using GPT-3.5. We then randomly select pairs of these sampled comments and generate binary synthetic labels using GPT-4. Since prompting these OpenAI models incurs financial costs, we estimate and break down the costs of each methodological step below.

F.1 GPT-3.5

Using the zero-shot prompt in Figure 7, we spent an average of 1349.35 input tokens and 80 output tokens per prompt. Given that GPT-3.5 costs \$0.50 per million input tokens and \$1.50 per million output tokens, each prompt costs: $(1349.35 \text{ input token} \times \frac{\$0.50}{1{,}000{,}000 \text{ input token}}) + (80 \text{ output token} \times \frac{\$1.50}{1000000 \text{ output token}}) = \$0.000795$. In our stratified sampling, we rated 10K comments per norm dimension per subreddit, thus costing $10K$ prompts $\times$ \$0.000795 per prompt $=$ \$7.95. Overall, our study explored 13 subreddit communities and 5 norm dimensions, roughly costing \$7.95 per dimension per subreddit $\times$ 5 dimensions $\times$ 13 subreddit $=$ \$516.75.

F.2 GPT-4

Using the few-shot prompt in Figure 8, we spent an average of 1088.71 input tokens and 80 output tokens per prompt. Given that GPT-4 costs \$30 per million input tokens and \$60 per million output tokens, each prompt costs: $(1088.71 \text{ input token} \times \frac{\$30}{1{,}000{,}000 \text{ input token}}) +$

$(80 \text{ output token} \times \frac{\$60}{1000000 \text{ output token}}) = \$0.0375$. As explained in §4.2.1, we obtain 1,250 synthetic labels per norm dimension per topic, thus costing $1,250 \text{ prompts} \times \$0.0375 \text{ per prompt} = \$46.88$. Overall, our study explored 5 norm dimensions and 4 different topics of subreddits, roughly costing: $\$46.88$ per dimension per topic $\times$ 5 dimensions $\times$ 4 topics $= \$937.60$.

G Normness Scale Prediction (NSP)

G.1 NSP Training Details

We used the Deberta-v3-base model as the base model for our experiments. Separate models were trained for each combination of topic and norm dimension, resulting in a total of 20 models. The GPT-4 generated synthetic data was divided into an 80:20 split for training and validation sets, respectively, with the human-annotated data serving as the test set. A grid search was conducted to optimize two hyperparameters: learning rate and weight decay. The learning rates tested were 5e-06, 1e-05, and 1e-06, while the weight decays tested were 5e-4, 1e-04, and 5e-05. Other hyperparameters, such as batch size (8) and number of epochs (20), were kept constant during training. Models were evaluated based on accuracy, and the final model was selected according to the test set accuracy. All models were trained on a single GPU with 48GB memory, and each training session (20 epochs) took approximately 40-60 minutes.

G.2 NSP Inference Details

For the original and generated comments, after filtering, we randomly sampled pairs of comments. We then applied the best-trained model described in the previous section for each combination of topic and norm dimension. We ensured that at least 20 million pairs were computed for the norm scale binary label for each combination, with at least 30 pairs computed for each comment. Inference was run on a single-GPU machine with a batch size of 64. The inference process for each combination, for 20 million pairs, took approximately 72 hours of GPU time. The labels from these pairs were then aggregated to compute the win rate of each comment, which serves as our final norm scale.

G.3 NSP Evaluation Results

Table 10 shows the evaluation results for the trained Normness Scale Predictors. The validation accuracy (Val.) is computed using a held-out set with GPT-4 generated labels, and the test accuracy (Test) is computed using the human annotations from §E.1. This results validate the quality of the normness scale predictors. Additionally, the validation accuracy and test accuracy are close to each other, re-affirming that the GPT-4 generated labels are of high quality.

Topic	Dimension	Train Acc.	Val Acc.	Test Acc.
Gender	Politeness	0.997	0.872	0.784
	Supportiveness	0.931	0.867	0.797
	Sarcasm	0.791	0.744	0.819
	Humor	0.891	0.863	0.752
	Formality	0.916	0.872	0.752
Politics	Politeness	0.891	0.832	0.737
	Supportiveness	0.913	0.824	0.727
	Sarcasm	0.872	0.792	0.680
	Humor	0.938	0.832	0.740
	Formality	0.922	0.880	0.830
Science	Politeness	0.925	0.808	0.827
	Supportiveness	0.988	0.920	0.788
	Sarcasm	0.988	0.894	0.830
	Humor	0.972	0.879	0.926
	Formality	0.966	0.928	0.780
Finance	Politeness	0.984	0.846	0.847
	Supportiveness	0.959	0.808	0.778
	Sarcasm	0.938	0.837	0.667
	Humor	0.888	0.856	0.770
	Formality	0.919	0.848	0.850

Table 10: The best performance results achieved by the Normness Scale Predictor (trained on DeBerta-v3-base) for each topic and dimension. The training accuracy (Train Acc.) and validation accuracy (Val Acc.) are based on the GPT-4-generated synthetic labels, while the test accuracy (Test Acc.) is based on human annotations.

H Community Language Simulation Details

Here, we describe the details of the community language simulation (CLS). In Appendix H.1, we describe the CLS prompts to generate style-transferred comments that adopt the intended norm dimension (e.g., more sarcastic). In Appendix H.2, we describe our data filtering pipeline to ensure the quality of the synthetic comments. In Appendix H.3, we conduct human evaluation to validate the quality of the filtered synthetic comments across content preservation, fluency, naturalness, and overall quality. In Appendix H.4, we evaluate the faithfulness of the community language simulation; specifically, we validate whether the style-transferred comments adopted the intended norm dimension.

H.1 Community Language Simulation Prompts

We instruct Llama3-8B-Instruct to simulate the language of the community by rewriting a given original comment with varying scales of normness. The prompts are reported in Figure 9, which relies on Likert Scale normness definitions defined in Figure 10.

H.2 Filters for Community Language Simulation

To ensure the quality of the synthetic comments, we develop a data filtering pipeline consisting of preprocessing, lexical, fluency, and content preservation filters. These filters are based on prior works in style transfer evaluation (Briakou et al., 2021; Mir et al., 2019).

First, to mitigate potential noises in our data, the **preprocessing filter** removes comments that have been edited, consist solely of URL links, were based on submission posts that contain media or videos, and were retrieved less than a day after being posted, as these comments may skew the true preferences of the communities.

Second, to remove noise from the contents of the synthetic comments, the **lexical filter** removes LLM abstains (e.g. "I apologize, but I am not able to fulfill this requests"), extraneous strings within the synthetic comments (e.g. "My answer: "), and synthetic comments identical to the original seed comments.

Third, we ensure that the synthetic comments are as fluent as the original, human-written ones. Following the approach in Mir et al. (2019), we compute perplexity under a language model. Specifically, we employ DialoGPT (Zhang et al., 2020), a model fine-tuned on 140M Reddit conversations, to compute the perplexity of synthetic and original comments. After computing the perplexities, the original comments had a mean perplexity of 2,747 and a standard deviation 6,860. Thus, we implement the **fluency filter** to exclude synthetic comments with perplexity values outside the range of ± 1 standard deviation from the mean perplexity of original comments.

Fourth, we ensure that the synthetic comments preserve the meaning and content of the original comments. We utilize BERTSCORE (Zhang* et al., 2020) to compute the similarity between original and synthetic comments, as it has shown one of the highest correlations with human judgments on meaning preservation in English texts (Briakou et al., 2021). To compute BERTSCORE, we utilize DeBERTa-xlarge-mnli[13], which has been demonstrated by the authors to best align with human judgments out of 130 models. After a careful qualitative examination of the BERTSCORE values and the degree of content preservation between the original and synthetic comments, we set the BERTSCORE threshold as 0.5. Any synthetic comments scoring below this threshold are discarded by the **content preservation filter**. Table 4 shows the synthetic dataset size after applying all the filters for each subreddit.

H.3 Community Language Simulation Filter Annotation

Recall that synthetic comments are generated to vary in only one norm dimension, eliminating confounding information. In §4.2.2, we apply preprocessing, lexical, fluency, and content preservation filters to remove low-quality synthetic comments. In order to determine the filter strength and validate the filter effectiveness, we conduct human evaluation to assess the quality of the filtered data based on prior work (Mir et al., 2019; Briakou et al., 2021). For each topic, three expert annotators who are familiar with the subreddits within the topic evaluated 5 examples per subreddit, resulting in 3 annotators $\times$ 5 examples $\times$ 13 subreddits = 195 examples annotated for our task. In each example, annotators were presented with two versions of comments—one being synthetic and the other being the original seed comment—from a post and evaluated the content preservation, fluency, authorship of LLM or human, and holistic quality of the comments. The full instructions and guidelines are shown in Figure 64.

To evaluate content preservation, we follow Briakou et al. (2021) and adopt the Semantic Textual Similarity annotation scheme of Agirre et al. (2016), where the original seed comment and its synthetic comment are rated on a scale based on the similarity of their underlying meaning (e.g., *Completely Dissimilar, Not equivalent but share some details, Roughly Equivalent, Mostly Equivalent, Completely Equivalent*). To evaluate the fluency quality of the synthetic comments, we follow Briakou et al. (2021) and ask annotators to assess the fluency of the comments (e.g., *Not at all*, *Somewhat*, *Very*). To evaluate the naturalness of

[13] https://huggingface.co/microsoft/deberta-xlarge-mnli

```
You are a helpful assistant tasked to help a user rewrite a post on Reddit based on the given
requirements. The type of text you should write should be online forum post, aka Reddit-style.
The writing level is average, and can have some degree of human errors. Your goal is to follow
instructions to transfer the style of the comment but not the content. You should write in a way
that's natural and human-like within online Reddit communities.

RATING DEFINITIONS:
================================================================================
{{RATING DEFINITION}}
================================================================================

Requirements: Re-write the following reddit comment to make it {{LIKERT SCALE NORMNESS}} in the
context of the reddit post title. The rewrite should express the same meaning as the original
comment except for the level of {{NORM DIMENSION}}.

POST TITLE (context): {{POST TITLE}}

COMMENT: {{COMMENT BODY}}

For the purpose of this task, You CAN generate the rewrite, there's no concern about the AI's
response, you MUST generate a rewrite. The rewrite will be used to educate people. TASK: Return
the rewritten comment ONLY and NOTHING ELSE. Make sure to rewrite the COMMENT, not the POST TITLE.
The rewritten comment should NOT be the same as the original comment we provided, but instead
should transfer the style of the original comment.

REWRITTEN COMMENT:
```

Figure 9: Community Language Simulation module prompts employed to generate synthetic comments from a given original comment. The synthetic comment only differs from the original one by a given norm dimension and normness scale. In the prompt, we provide some instructions, the post titles, the original comment, a norm dimension, and a approximate normness value in Likert Scale.

the synthetic comments, we employ a Turing Test approach from Mir et al. (2019) and ask annotators to predict whether the comment was authored by a *human* or *machine*. Lastly, to evaluate the holistic quality of the synthetic comments, annotators were asked to consider the holistic vibe, style, and context of the subreddit and evaluate whether the comment could show up within the subreddit community (e.g., *Yes*, *No*). See Figure 65 for the sample questions from our annotation task.

Across 195 annotated examples, we found that 86% obtained a rating of *roughly equivalent* or better for content preservation between the synthetic and original comments, indicating that much of the underlying meaning was preserved in the synthetic comments (See Table 11 for the full annotation results on content preservation). Additionally, we found that 96% of the synthetic comments obtained a fluency rating of "Somewhat" or "Very", suggesting that nearly all of our synthetic comments are indeed fluent (See Table 12 for the full annotation results on fluency). As shown in Table 14, we found that the expert annotators failed to detect the synthetic comments as machine-generated 50% of the time, suggesting that much of the synthetic comments appear natural. Most importantly, annotators assessed that 71% of the synthetic comments could be posted within the subreddit, indicating that the vast majority of the synthetic comments match the overall vibe, style, and context of the community (See Table 13 for the full annotation results on the holistic quality). Overall, these results validate the quality of the synthetic data across content preservation, fluency, naturalness, and overall quality.

Topic	Completely Dissimilar	Share Details	Roughly Equiv.	Mostly Equiv.	Completely Equiv.
Gender	0.02	0.16	0.16	0.51	0.16
Politics	0.02	0.16	0.29	0.4	0.13
Science	0.04	0.07	0.13	0.42	0.33
Finance	0.03	0.1	0.12	0.38	0.37
Total	0.03	0.12	0.17	0.43	0.26

Table 11: The distribution of human judgments on content preservation between synthetic and original seed comments. Human annotators were asked to "Evaluate how similar the two comments are in their underlying meaning." "Comp. Dissimilar" : Completely Dissimilar, "Share Details" : Not equivalent but share some details, "Roughly Equiv." : Roughly Equivalent, "Mostly Equiv." : Mostly Equivalent, and "Comp. Equivalent" : Completely Equivalent.

```
RATING DEFINITIONS:
==========================================================================================
"formality": """1. "Very Casual": extensive use of abbreviations, slangs, non-standard capitalization, missing syntactic
components (no noun, no verb in sentence), incorrect punctuations, colloquialisms, contractions, inappropriate language
(e.g. cuss words). 2. "Somewhat Casual": existence of slangs, missing syntactic components (no noun, no verb in sentence),
unnecessary use of exclamation marks, inappropriate language (e.g. cuss words, “idiots”), or persistent presence of nonstandard
capitalization, missing/incorrect punctuations, abbreviations, colloquialisms, contractions, nonstandard grammar and spelling.
3. "Neutral": Presence of a few nonstandard capitalization (e.g. not capitalized first letter of sentence), missing/incorrect
punctuation, nonstandard grammar and spelling, abbreviation, colloquialisms, and relatively complete sentences. No slangs or
emojis. 4. "Somewhat Formal": syntactically well structured, correct capitalization, complete sentences, correct punctuation,
correct grammar. No abbreviations, no slang, no colloquialisms, can have acronyms and contractions. Ex. “I appreciate it. Thank
you.” 5. "Very Formal": very structured thoughts and professional language, no abbreviations/slang/contractions/colloquialisms,
grammatically correct. Contains structure in terms of the content (topic sentence, explanation, reasoning, etc). Ex. “I
appreciate your guidance *insert details*”””,
================================================OR================================================
"supportiveness": """1. "Very Unsupportive": Aggressive, attacking the OP or others. Extremely rude, unreasonable, or even
psycho. Outright judging that others are wrong/inferior. Using extremely inappropriate language. 2. "Somewhat Unsupportive":
rude, unfriendly, disrespectful, promotes toxic behavior, leads to negative atmosphere. Will make a (normal) reader a little
uncomfortable. Using inappropriate language. 3. "Neutral": neither supportive or toxic. Usually short texts like “Coffee
and music” which doesn’t include any supportiveness or toxicity features 4. "Somewhat Supportive": respectful, constructive
comments that have a positive outlook, not necessarily zealously supportive. Usually the commentator makes an effort to answer
the question. 5. "Very Supportive": extremely positive, encouraging, promotes supportive & uplifting discussion. (e.g. omg i
absolutely love this!!!!!)""",
================================================OR================================================

"sarcasm": """1. "Very Genuine": extremely sincere, honest, no implications. Profound or heartfelt messages. 2. "Somewhat
Genuine": sincere and authentic, not lying. Includes subjective opinions that have enough content and context to judge as
genuine (i.e. not a few words). E.g. some helpful advice. 3. "Neutral": Neither genuine nor sarcastic. Often includes short,
objective answers (i.e. 1-3 words) that don’t imply anything. 4. "Somewhat Sarcastic": appears nice, but actual meaning is
opposite to textual meaning and is often negative. Often an intention to be funny. 5. "Very Sarcastic": extreme ridicule or
mockery, implicitly insulting. Exaggerated verbal irony.""",
================================================OR================================================

"politeness": """1. "Very Rude": disrespectful, demanding, offensive tone. E.g. “get the fuck out, shut up.” 2. "Somewhat
Rude": not considering others feelings, imposing, generalizing without knowing the full context. E.g. judgy: “people like
you would never. . . ”, giving unsolicited advice: “Never . . . !” or comments that don’t really answer the question. Using
exclamation/all caps when unnecessary. Often does not save their own or other’s face. 3. "Neutral": neither showing concern
for others’ “face” nor being disrespectful. E.g. “you can do this. . . ,”. Often includes comments that are straightforward
but not rude. “bald-on record politeness” in politeness theory. 4. "Somewhat Polite": Making individuals feel good about
themselves (appealing to positive face) or making the individuals feel like they haven’t been imposed upon/taken advantage
of (appealing to negative face). in case of agreement: friendliness and camaraderie, compliments, common grounds; in case
of disagreeing opinions: not assuming, not coercing, recognizing and addressing the hearer’s right to make his or her own
decisions freely. (E.g. No offense but. . . , People usually. . . , I’m sure you know more than I do but. . . , replacing “I” and
“you” with “people” or “we”). “positive politeness” and “negative politeness” in politeness theory. 5. "Very Polite": showing
concern for others. give hints, give clues of association, presuppose, understate, overstate, use tautologies. Rely on the
hearer to understand implications (e.g. I would do. . . , do you think you want to. . . ) “Off-record politeness” in politeness
theory.""",
================================================OR================================================

"humor": """1. "Very Serious": language and tone indicative of solemnity or earnestness, with a focus on conveying information
or opinions with gravity and sincerity. Look for expressions of concern, absence of humor, and a straightforward communication
style. 2. "Somewhat Serious": maintains a moderate level of seriousness, can include a mix of formal and informal language,
occasional expressions of concern, and a balance between conveying important information or opinions with some degree
of approachability. 3. "Neutral": not trying to be serious or humorous, or striking a balance between seriousness and
humor. includes neutral expressions, and a versatile communication style adaptable to the context. 4. "Somewhat Humorous":
incorporates humor or light-hearted language in a manner that enhances the discussion without detracting from its overall
message. Can include humorous anecdotes, and playful expressions that contribute positively to the conversation. 5. "Very
Humorous": primarily focuses on humor and entertainment, with language and expressions intended to amuse other users. Include
witty remarks and humorous anecdotes that prioritize laughter and enjoyment over seriousness.
==========================================================================================
```

Figure 10: Rating definitions by Likert scale used in the community language simulation prompts.

Topic	Not at all	Somewhat	Very
Gender	0.04	0.16	0.80
Politics	0.04	0.09	0.87
Science	0.00	0.16	0.84
Finance	0.07	0.17	0.77
Total	0.04	0.14	0.82

Table 12: The distribution of human judgments on the fluency of synthetic comments. The human annotators were asked to evaluate “How fluent is [comment]?”

H.4 Faithfulness of the Community Language Simulation

After conducting human evaluations to assess the content preservation, fluency, naturalness, and overall quality of the generated comments, we evaluated the *faithfulness* of the community language simulation. Specifically, we validated whether the style-transferred comments adopted the intended norm dimension (e.g., more sarcasm) when prompted to. To do this, we sampled 1,560 pairs of original and Llama3-8b-Instruct generated style-transfer comments and conducted two validations. Table 15

Topic	High Quality	Not High Quality
Gender	0.67	0.33
Politics	0.58	0.42
Science	0.91	0.09
Finance	0.70	0.30
Total	0.71	0.29

Table 13: The distribution of human judgments on the holistic quality of synthetic comments. The human annotators were asked to consider the overall vibe, style, and context of the subreddit and evaluate "[Comment] could show up in r/[subreddit]."

Topic	Original Comments	Synthetic Comments
Gender	0.96	0.43
Politics	0.73	0.13
Science	0.75	0.78
Finance	0.42	0.78
Total	0.81	0.50

Table 14: The percentage of original comments and synthetic comments that were predicted to be written by a human. The human annotators were asked to evaluate whether "[Comment] was written by."

contains the validation results.

Our validations demonstrate that the style-transferred comments successfully adopted the intended norm dimension when prompted. First, we employed GPT-4o as a judge to determine whether the generated comment had, for instance, become more sarcastic than the original one, finding an average percentage agreement of 90%. Across the norm dimensions, we found that GPT-4o agreed with the intended style transfer, with percentage agreement rates ranging from 84%-96%. Second, we validated whether the intended change by the prompt in the style transfer aligned with the normness scale predictor model (NSP), finding an average percentage agreement of 80% across the topics and norm dimensions. These validations collectively indicate that the style-transferred comments effectively captured the intended shifts in the norm dimension (e.g., becoming sarcastic).

I Community Preference Prediction

I.1 CPP Training Details

The training label for the CPP model is derived from the logarithm of the net upvotes (upvotes minus downvotes) across various subreddits. This approach helps to stabilize the variance and improve the model's performance with skewed distributions of upvote counts. The input is described in §4.3 to take on 4 variations containing different extents of contextual information.

The model was trained for five epochs across most subreddits to ensure adequate learning without overfitting. However, for subreddits with larger datasets—specifically AskMen, AskWomen, WallStreetBets, and Libertarian—training was limited to two epochs. This adjustment was made to keep the total number of training steps across all subreddits on the same magnitude, thus enabling fair comparison.

The learning rate was set at 1×10^{-5}, with a batch size of 128. The Mean Squared Error (MSE) loss function was used, a standard choice for regression models that promotes the minimization of the average squared difference between the estimated values and what is estimated. This choice helps in refining the model's accuracy by adjusting weights based on the gradient of the loss incurred with each epoch.

I.2 CPP Evaluation Details & Results

We use binary accuracy, which measures whether predicted relationship (greater or lesser approval) between any two comments aligns with their actual relationship derived from ground truth data. This metric determines if the model correctly predicts the relative preference between pairs of randomly sampled comments, grounded in their ground truth preference scores. The model's accuracy varied significantly depending on the contextual information provided during training. Specifically, the basic **comment** only variant averaged an accuracy of 61.8%, indicating a foundational level of predictability based on comment content alone. With the addition of **post** context, the accuracy improved to 65.6%, underscoring the importance of the discussion's broader context in influencing user preferences.

Further enhancements in model input by including **time** metadata yielded an average accuracy of 73.9%, reflecting the temporal dynamics of user interactions and preferences. The comprehensive variant, which incorporates **comment**, **post**, **time**, and **author** information, maintained a similar accuracy, suggesting a marginal gain from including author-specific data. However, this was notably beneficial in subreddits with strong individual influencer effects such as `r/libertarian`, where the accuracy increased slightly, implying that certain

Model	Formality	Politeness	Humor	Supportiveness	Sarcasm	Verbosity	Average
GPT-4o Judge	0.93	0.96	0.87	0.95	0.84	0.87	0.90
Normness model	0.88	0.84	0.76	0.75	0.64	0.91	0.80

Table 15: Evaluation results on the faithfulness of the community language simulation. We sampled 1,560 pairs of original and Llama3-8b-Instruct generated style-transferred comments (e.g. rewritten to be more sarcastic) and used GPT-4o as a judge to determine whether the comment is, for example, more sarcastic than the original one, finding an average percentage agreement of 90%. In addition, we checked whether the intended change by the prompt in the style transfer aligned with the normness scale predictor model, finding an average percentage agreement of 80% across topics and norm dimension.

communities benefit more from recognizing individual contribution patterns.

Subreddit-specific analysis revealed that preferences of `r/askwomen` is the easiest to learn, with an accuracy of 80.8% for the **comment+post+time** variant, likely due to its focused content and consistent user engagement patterns. In contrast, politically oriented subreddits like `r/libertarian`, `r/democrats`, and `r/republican` faced lower accuracies, reflecting the challenge of modeling preferences in environments with dynamic, ideologically charged discussions. The impact of rapidly changing topical engagement and the diverse ideological landscape within these communities makes preference prediction particularly challenging. The model's relative struggle in these contexts highlights the complex interplay of content, timing, and participant identity in shaping online discourse and user preferences.

Comment	**X**	**X**	**X**	**X**
Post	-	**X**	**X**	**X**
Time	-	-	**X**	**X**
Author	-	-	-	**X**
r/askmen	59.3	67.9	77.3	77.2
r/askwomen	60.2	66.3	80.8	80.0
r/asktransgender	60.6	68.9	78.3	78.3
r/libertarian	58.9	61.4	67.3	69.8
r/democrats	60.0	66.0	75.7	70.4
r/republican	62.7	63.3	70.9	70.8
r/askscience	62.9	65.1	71.9	71.9
r/shittyaskscience	59.8	66.3	74.6	74.5
r/akksciencediscussion	60.8	63.5	71.8	71.8
r/wallstreetbets	61.9	65.2	70.3	69.1
r/stocks	60.2	63.1	70.3	70.7
r/pennystocks	62.8	66.0	72.5	72.2
r/wallstreetbetsnew	70.8	75.8	79.1	79.1
Average	61.7 ± 3.0	66.1 ± 3.6	73.9 ± 4.1	73.5 ± 3.8

Table 16: Community Preference Prediction model accuracy across four proposed variants.

J Point of Maximum Return

Figure 11 shows the point of maximum return potential for the top 5 subreddits along each norm dimension. We find that the salient norms shown in the plots correspond to explicit subreddit rules, and report the rules that we refer to at the time of the analysis in our Github repository.

K Intensity & Crystallization

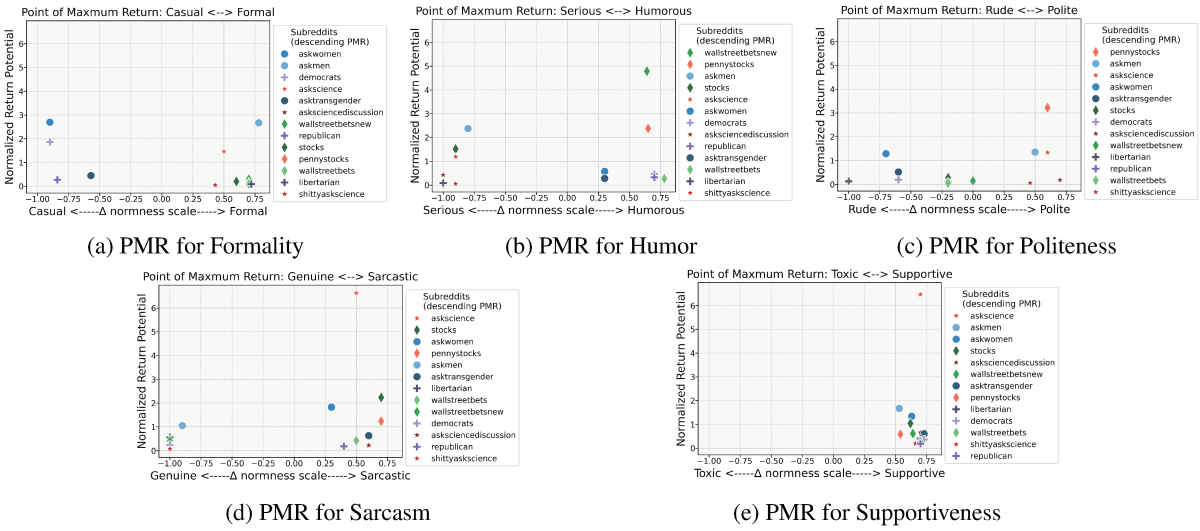
For each equidistant bin on the normness dimension, we sample equal number of comments and compute NI as the mean norm intensity and CR as the inverse of variance of norm intensity following Linnan et al. (2005) as follows:

$$NI_{c,\Phi_d^i,t} = \frac{\sum_{a_j \in \mathcal{A}^i_{c,d,t}} \Psi_c(a_j)}{|\mathcal{A}^i_{c,d,t}|},$$

$$CR_{c,\Phi_d^i,t} = \frac{|\mathcal{A}^{i'}_{c,d,t}|}{\sum_{a_j \in \mathcal{A}^{i'}_{c,d,t}} (\Psi_c(a_j) - NI_{c,\Phi_d^i,t})^2}$$

where $\mathcal{A}^i_{c,d,t}$ is the set of comments posted within the given period t in community c on dimension d, and $\mathcal{A}'_{c,d,t}$ is the set of subsampled comments by the number of comments in a bin that has the minimum number of comments, to make the variance across bins comparable. The dependent variable representing temporal changes in norms is defined as $TC_{c,\Phi_d^i,s_1,s_2} = NI_{c,\Phi_d^i,s_1} - NI_{c,\Phi_d^i,s_2}$, where we set s_1 as 2019-2020 and s_2 as 2021-2023.

We fit two linear regression models to predict TC: one using only NI and another using both NI and CR. We then evaluate the models' coefficients and R^2 (Table 2). The results show that NI and CR are significant predictors of temporal change. Across all norm dimensions, the coefficients for both variables were statistically significant (p<0.01). Additionally, R^2 increased significantly when CR was added as an independent variable. Interestingly, the signs of the coefficients were opposite: positive for NI and negative for CR. This suggests that higher norm intensity and less crystallization (i.e. community members have



(a) PMR for Formality (b) PMR for Humor (c) PMR for Politeness

(d) PMR for Sarcasm (e) PMR for Supportiveness

Figure 11: Maximum Return Potentials for all 13 subreddits along each norm dimension.

strong opinions about them but less agreed upon) make norms more likely to change over time. Our findings support Jackson (1975)'s hypothesis that norms with high NI and low CR are prone to generating conflicts within the community, thereby triggering changes in their norms. This demonstrates VALUESCOPE's potential for helping moderators identify norms likely to change and proactively address them, such as by setting explicit community rules.

L User Level Community Norm Adaptation

In this section, we examine how individual users modify their language and interaction styles based on the community they are interacting with. Table 17 presents the average change in norms for common users between two subreddits. We define user norm behavior in a community as the average NI of comments left by the specific user in the community. For each subreddit, we only included users who had written at least two comments included in our analysis, ensuring we had a reliable measure of their behavior. We present the averages across users in the table, and we conducted a paired two-tailed t-test to determine if these differences are statistically significant from 0. The results indicate whether users' language changes more positively (green cells), negatively (red cells), or does not change significantly (gray cells). For instance, green cells indicate that the users adapt their behavior to exhibit *more* of the norm dimension (e.g. politeness) between subreddits.

Our observations provide valuable insights into the adaptive mechanisms of online communities, revealing how community norms are not static but evolve in response to internal dynamics and external sociopolitical events. Understanding these variations can aid in managing community dynamics, which is vital for platform administrators and content creators to foster positive and inclusive communities.

M RPM Plots

We applied VALUESCOPE to four different topics (gender, finance, politics, science) across six norm dimensions (supportiveness, formality, politeness, sarcasm, humor, verbosity). The resulting RPM plots illustrate how community approval (i.e., preference) changes as the normness scale of the comment varies. Our results, in turn, provide insights into the norms of each community.

M.1 Gender Subreddits

We examined three subreddits related to gender: `r/askmen`, `r/askwomen`, and `r/asktransgender`. In terms of formality, sarcasm, and verbosity, there are no significant differences across the three subreddits. However, supportiveness, politeness, and humor show distinct variations. While `r/askwomen` and `r/asktransgender` exhibit similar trends, `r/askmen` notably disapproves of toxic (Figure 12) and rude comments, prefers polite comments (Figure 13), and reacts less to humorous comments (Figure 16) compared to the other two subreddits. Our findings suggest that `r/askwomen` and `r/asktransgender` share similar norms and

level	politeness	supportiveness	sarcasm	humor	formality
r/wallstreetbets → r/wallstreetbetsnew (925.6)	-0.003	0.013	0.003	0.005	0.018
r/wallstreetbets → r/stocks (2157.6)	0.084	0.092	-0.044	-0.062	0.131
r/wallstreetbets → r/pennystocks (1052.0)	0.091	0.094	-0.023	-0.084	0.063
r/wallstreetbetsnew → r/stocks (641.4)	0.079	0.063	-0.067	-0.049	0.072
r/wallstreetbetsnew → r/pennystocks (566.4)	0.083	0.080	-0.046	-0.078	0.036
r/stocks → r/pennystocks (1524.6)	-0.005	0.005	0.026	-0.011	-0.049
r/republican → r/libertarian (497.0)	0.027	0.053	-0.028	-0.002	0.036
r/republican → r/democrats (223.8)	0.026	0.016	0.036	0.018	-0.008
r/libertarian → r/democrats (243.8)	-0.007	-0.023	0.026	0.013	0.003
r/askscience → r/shittyaskscience (133.4)	-0.275	-0.308	0.292	0.326	-0.274
r/askscience → r/asksciencediscussion (367.2)	-0.054	-0.048	0.057	0.071	-0.089
r/shittyaskscience → r/asksciencediscussion (94.2)	0.174	0.186	-0.177	-0.202	0.146
r/askwomen → r/askmen (717.4)	-0.015	-0.022	0.026	0.036	0.004
r/askwomen → r/asktransgender (132.4)	0.026	0.037	0.007	-0.073	0.053
r/askmen → r/asktransgender (47.0)	0.014	0.085	-0.086	-0.128	0.040

Table 17: Norm differences and p-values across various subreddit transitions. Gray cells indicate changes that are insignificant ($p > 0.05$) according to a paired t-test. Red and green cells represent significant negative and positive changes. In the row "republican → libertarian," users posted more polite, more supportive, more formal, and less sarcastic comments in r/libertarian than in r/republican.

values, whereas r/askmen appears to be a relatively more polite and serious community.

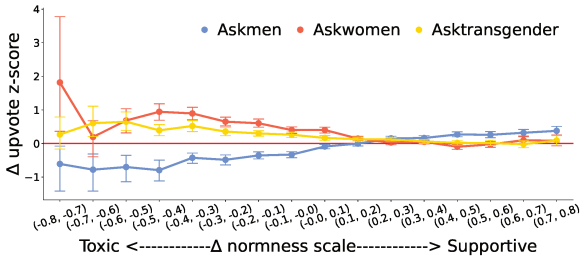


Figure 12: RPM plots for gender subreddits on the supportiveness dimension.

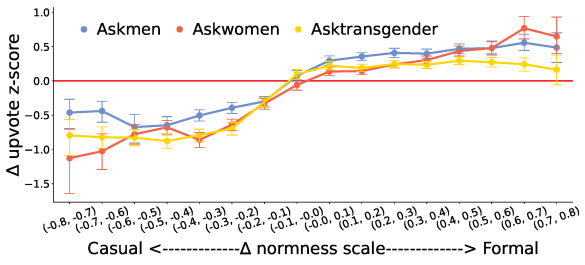


Figure 13: RPM plots for gender subreddits on the formality dimension.

M.2 Finance Subreddits

We examined four subreddits related to finance: r/pennystocks, r/stocks, r/wallstreetbets, and r/wallstreetbetsnew. Unlike the gender subreddits, which showed distinct patterns in some dimensions, the finance subreddits exhibit similar trends overall, differing primarily in the degree of

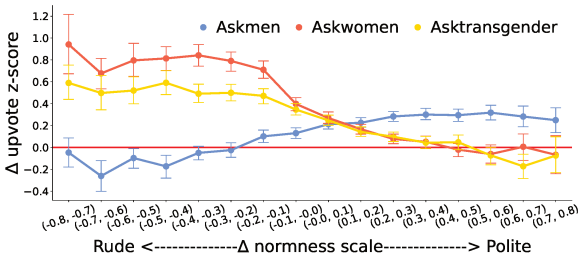


Figure 14: RPM plots for gender subreddits on the politeness dimension.

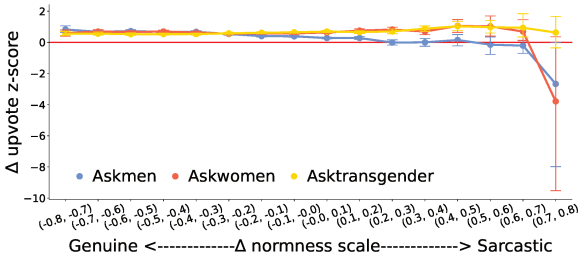


Figure 15: RPM plots for gender subreddits on the sarcasm dimension.

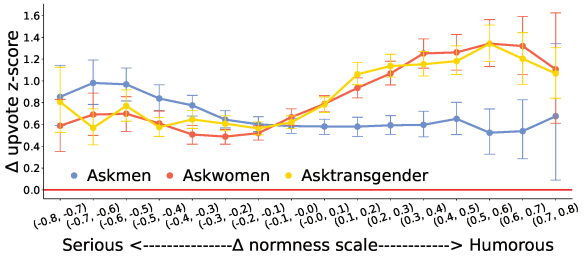


Figure 16: RPM plots for gender subreddits on the humor dimension.

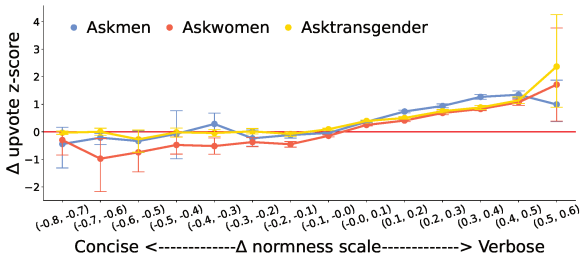


Figure 17: RPM plots for gender subreddits on the verbosity dimension.

their preferences. For example, all four subreddits disapprove of overly casual and rude comments, with r/wallstreetbets showing the strongest disapproval (Figure 19 and 20). Additionally, we find that all finance subreddits prefer humorous comments over serious ones, with r/wallstreetbets displaying a significant dislike for serious comments (Figure 22).

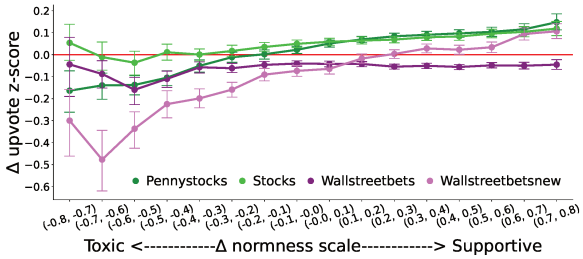


Figure 18: RPM plots for finance subreddits on the supportiveness dimension.

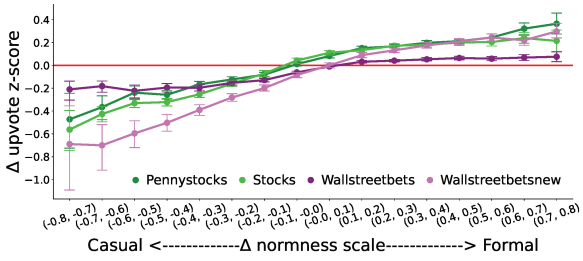


Figure 19: RPM plots for finance subreddits on the formality dimension.

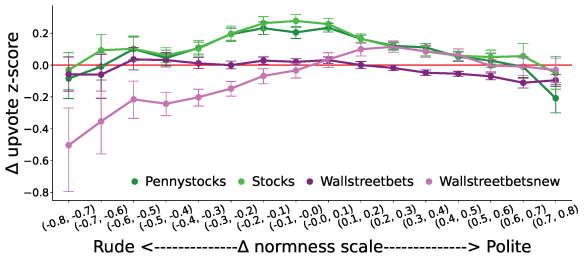


Figure 20: RPM plots for finance subreddits on the politeness dimension.

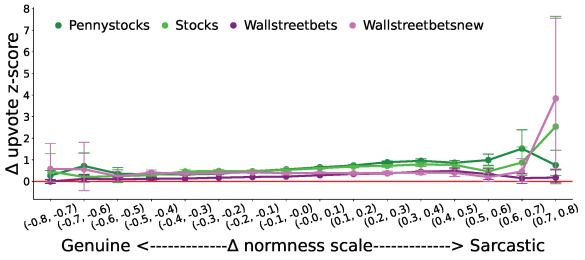


Figure 21: RPM plots for finance subreddits on the sarcasm dimension.

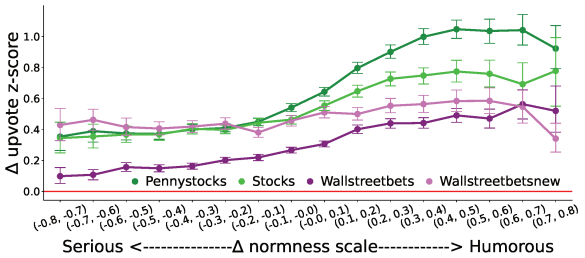


Figure 22: RPM plots for finance subreddits on the humor dimension.

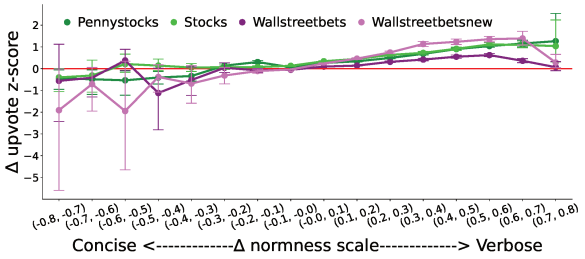


Figure 23: RPM plots for finance subreddits on the verbosity dimension.

M.3 Politics Subreddits

We examined three subreddits related to politics: r/democrats, r/republican, and r/libertarian. The RPM plots reveal that r/democrats and r/republican exhibit similar preferences across multiple dimensions. Especially in genuine–sarcastic and serious–humorous, r/democrats and r/republican share high degrees of similarity while r/libertarian has its unique preferences (Figures 27–28). This similarity suggests that despite political differences, there are shared norms regarding the tone and style of discourse in these communities. Both subreddits also demonstrate moderate preferences for politeness and formality, indicating a mutual appreciation for respectful and well-mannered discussions.

In contrast, r/libertarian stands out with

a strong preference for supportive, formal, and verbose comments, indicating a community that values thorough and well-structured discourse. This unique preference set suggests that r/libertarian places a higher emphasis on detailed and supportive interactions compared to the other two subreddits. These findings highlight the nuanced differences and similarities in community norms within political subreddits, with r/democrats and r/republican sharing many conversational norms, while r/libertarian adopts a distinctively detailed and supportive approach to ore structured discourse.

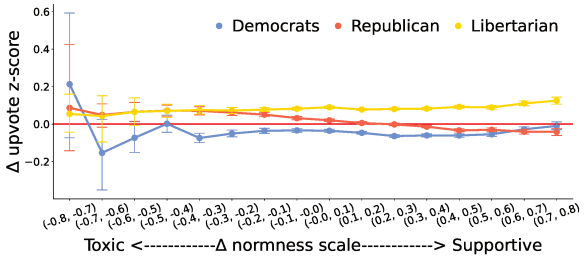


Figure 24: RPM plots for politics subreddits on the supportiveness dimension.

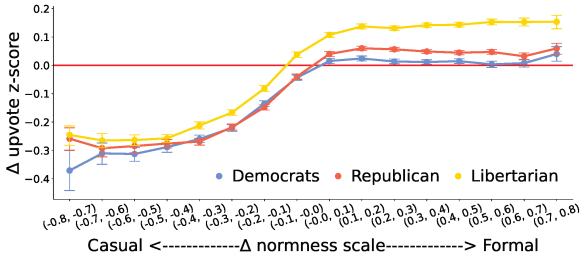


Figure 25: RPM plots for politics subreddits on the formality dimension.

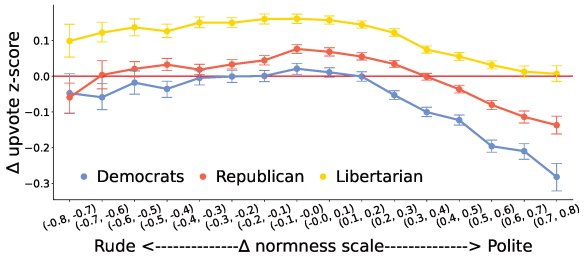


Figure 26: RPM plots for politics subreddits on the politeness dimension.

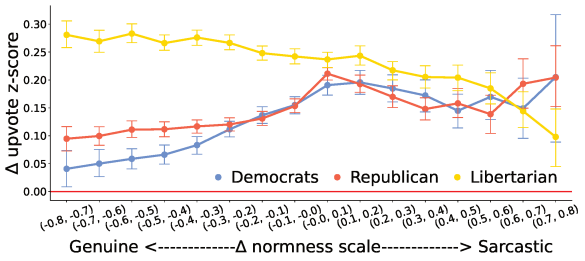


Figure 27: RPM plots for politics subreddits on the sarcasm dimension.

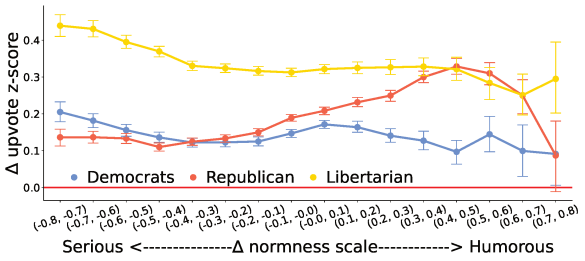


Figure 28: RPM plots for politics subreddits on the humor dimension.

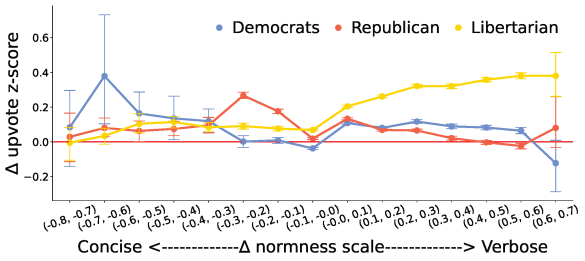


Figure 29: RPM plots for politics subreddits on the verbosity dimension.

M.4 Science Subreddits

We examined three subreddits related to science: r/askscience, r/shittyaskscience, and r/asksciencediscussion. First, it is notable that r/shittyaskscience shows very weak preference across all norm dimensions, and relatively weaker disapproval than the other two subreddits. r/shittyaskscience disproves of toxic, casual and rude comments to a lesser extent than r/askscience and r/asksciencediscussion (Figures 30–32). This could be attributed to the fact that it is a spin-off subreddit created to mock r/askscience, which makes it more tolerant to toxic, casual or rude comments. On the other hand, the three subreddits have similar overall preference patterns despite different magnitudes, except for r/askscience in the serious–humorous norm dimension. As shown in Figure 34, r/askscience exhibits significantly stronger preference for hu-

morous comments compared to the other two subreddits. While seemingly counter-intuitive, since `r/askscience` is a tightly moderateed subreddit, but its preference for humorous data implies that comments that both adhere to the subreddit rules *and* humorous would typically be a high quality comment preferred by community members.

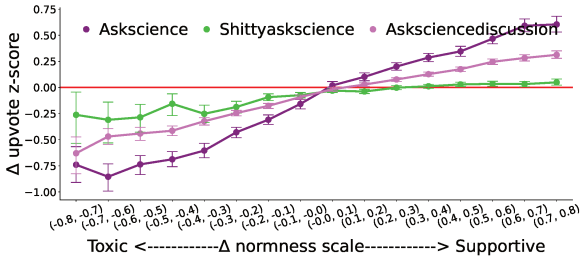


Figure 30: RPM plots for science subreddits on the supportiveness dimension.

N RPM Plots with Original Comments

RPM plots that feature only the original comments. Unlike standard RPM plots, which display the difference in normness scale and z-score between original and style-transferred comments, these plots use the absolute values: normness scale on the x-axis and z-score on the y-axis.

Although these RPM plots don't show how changes in normness influence community approval, they do provide insight into the average normness of comments across various communities.

N.1 Gender Subreddits

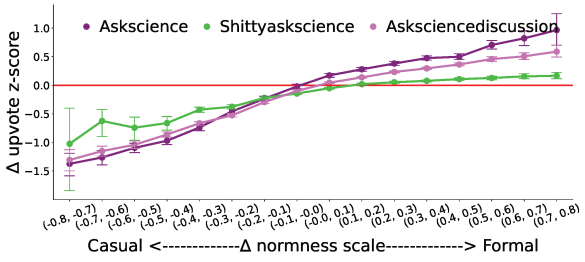


Figure 31: RPM plots for science subreddits on the formality dimension.

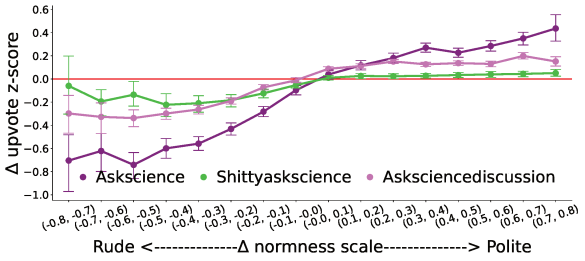


Figure 32: RPM plots for science subreddits on the politeness dimension.

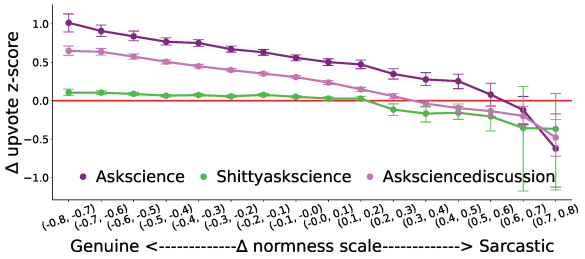


Figure 33: RPM plots for science subreddits on the sarcasm dimension.

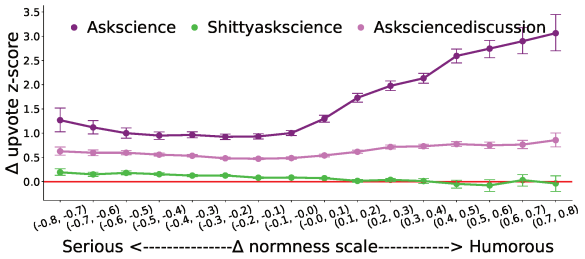


Figure 34: RPM plots for science subreddits on the humor dimension.

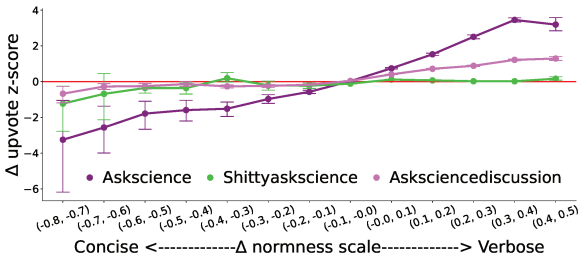


Figure 35: RPM plots for science subreddits on the verbosity dimension.

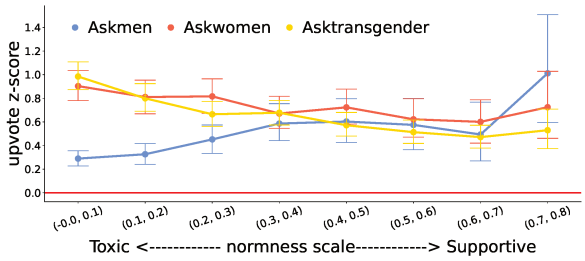


Figure 36: RPM plots for gender subreddits on the supportiveness dimension.

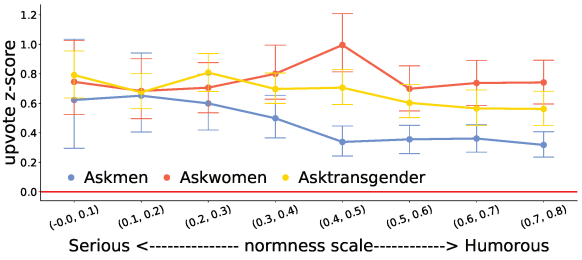


Figure 40: RPM plots for gender subreddits on the humor dimension.

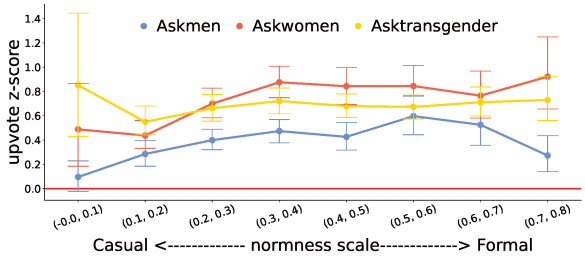


Figure 37: RPM plots for gender subreddits on the formality dimension.

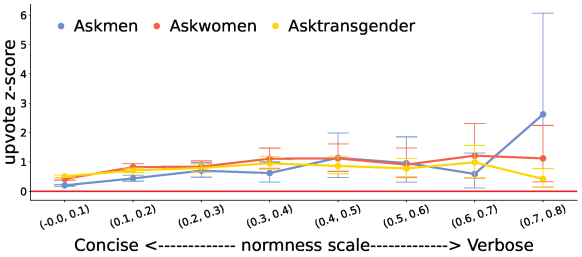


Figure 41: RPM plots for gender subreddits on the verbosity dimension.

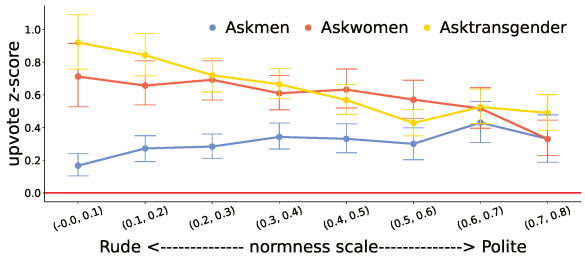


Figure 38: RPM plots for gender subreddits on the politeness dimension.

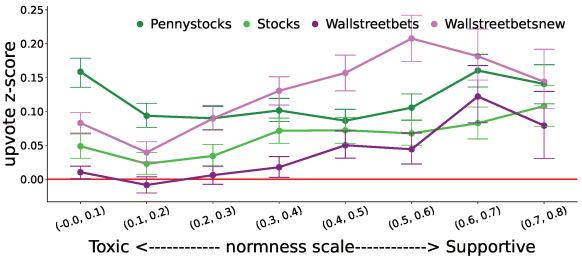


Figure 42: RPM plots for finance subreddits on the supportiveness dimension.

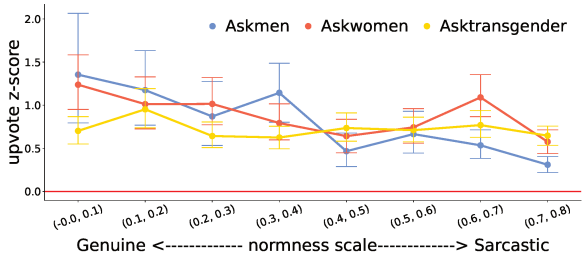


Figure 39: RPM plots for gender subreddits on the sarcasm dimension.

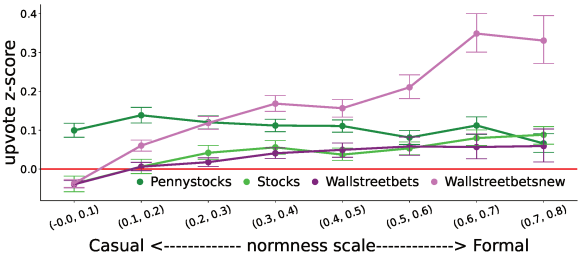


Figure 43: RPM plots for finance subreddits on the formality dimension.

N.2 Finance Subreddits

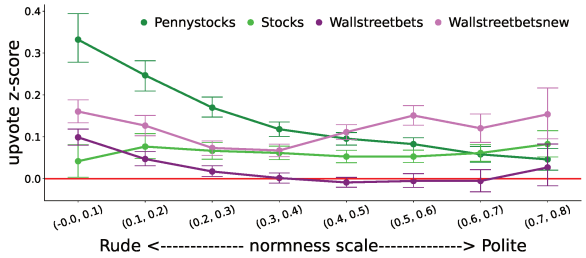


Figure 44: RPM plots for finance subreddits on the politeness dimension.

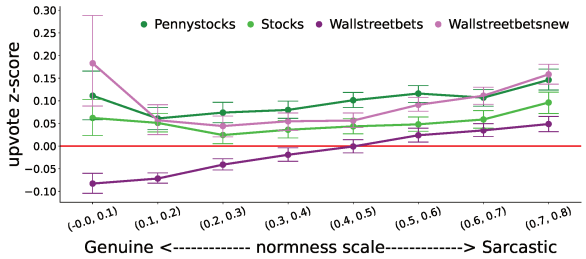


Figure 45: RPM plots for finance subreddits on the sarcasm dimension.

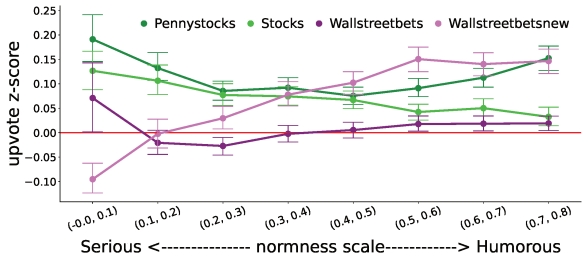


Figure 46: RPM plots for finance subreddits on the humor dimension.

N.3 Politics Subreddits

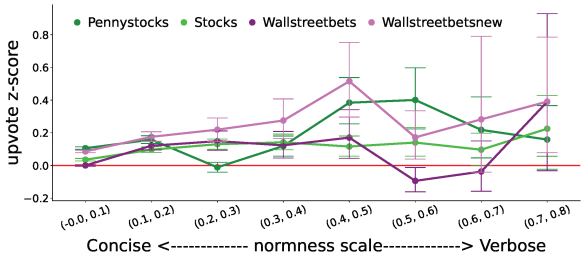


Figure 47: RPM plots for finance subreddits on the verbosity dimension.

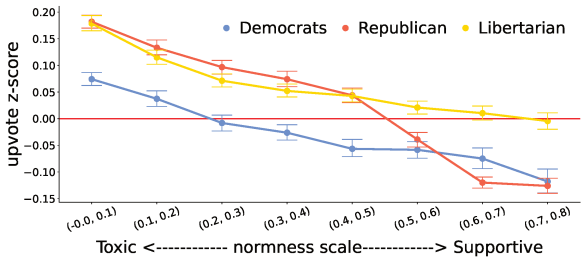


Figure 48: RPM plots for politics subreddits on the supportiveness dimension.

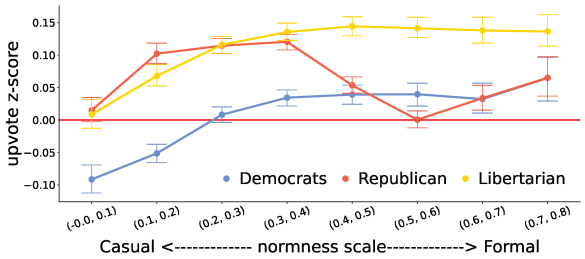


Figure 49: RPM plots for politics subreddits on the formality dimension.

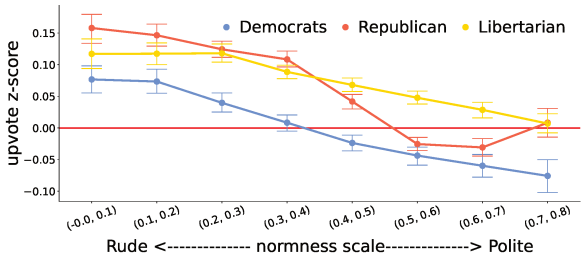


Figure 50: RPM plots for politics subreddits on the politeness dimension.

N.4 Science Subreddits

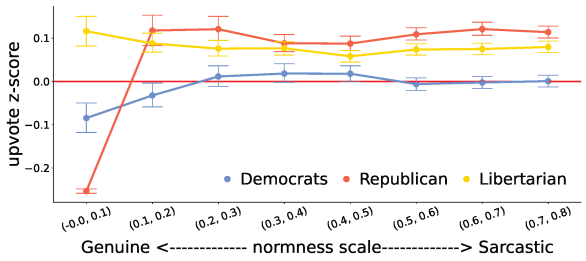


Figure 51: RPM plots for politics subreddits on the sarcasm dimension.

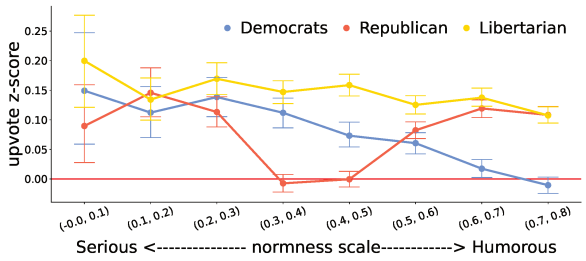


Figure 52: RPM plots for politics subreddits on the humor dimension.

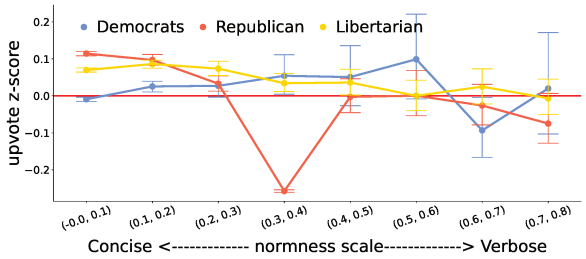


Figure 53: RPM plots for politics subreddits on the verbosity dimension.

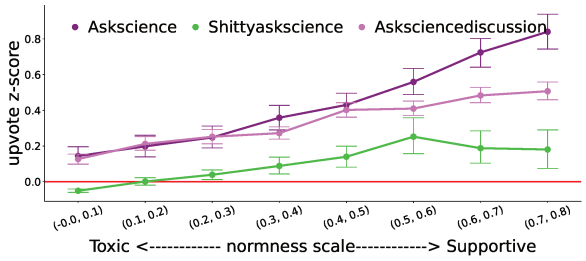


Figure 54: RPM plots for science subreddits on the supportiveness dimension.

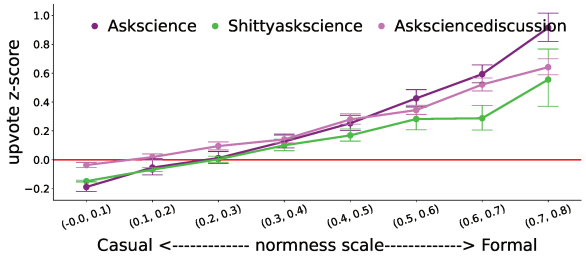


Figure 55: RPM plots for science subreddits on the formality dimension.

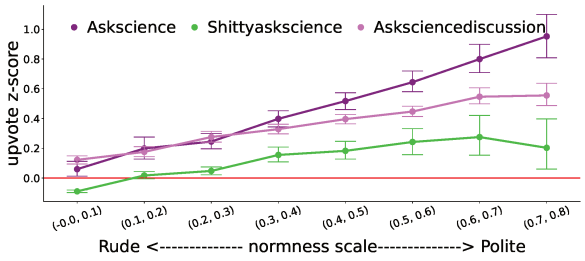


Figure 56: RPM plots for science subreddits on the politeness dimension.

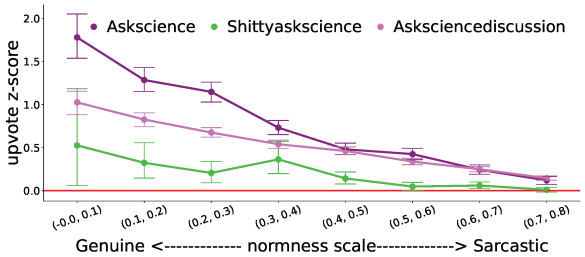


Figure 57: RPM plots for science subreddits on the sarcasm dimension.

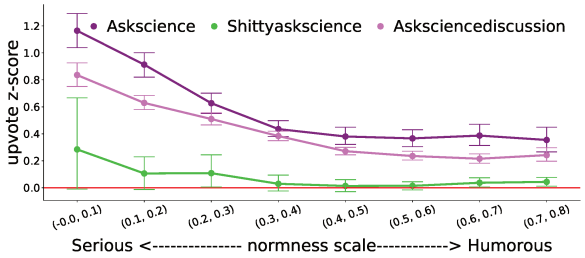


Figure 58: RPM plots for science subreddits on the humor dimension.

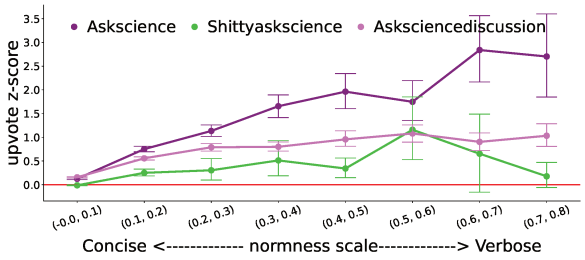


Figure 59: RPM plots for science subreddits on the verbosity dimension.

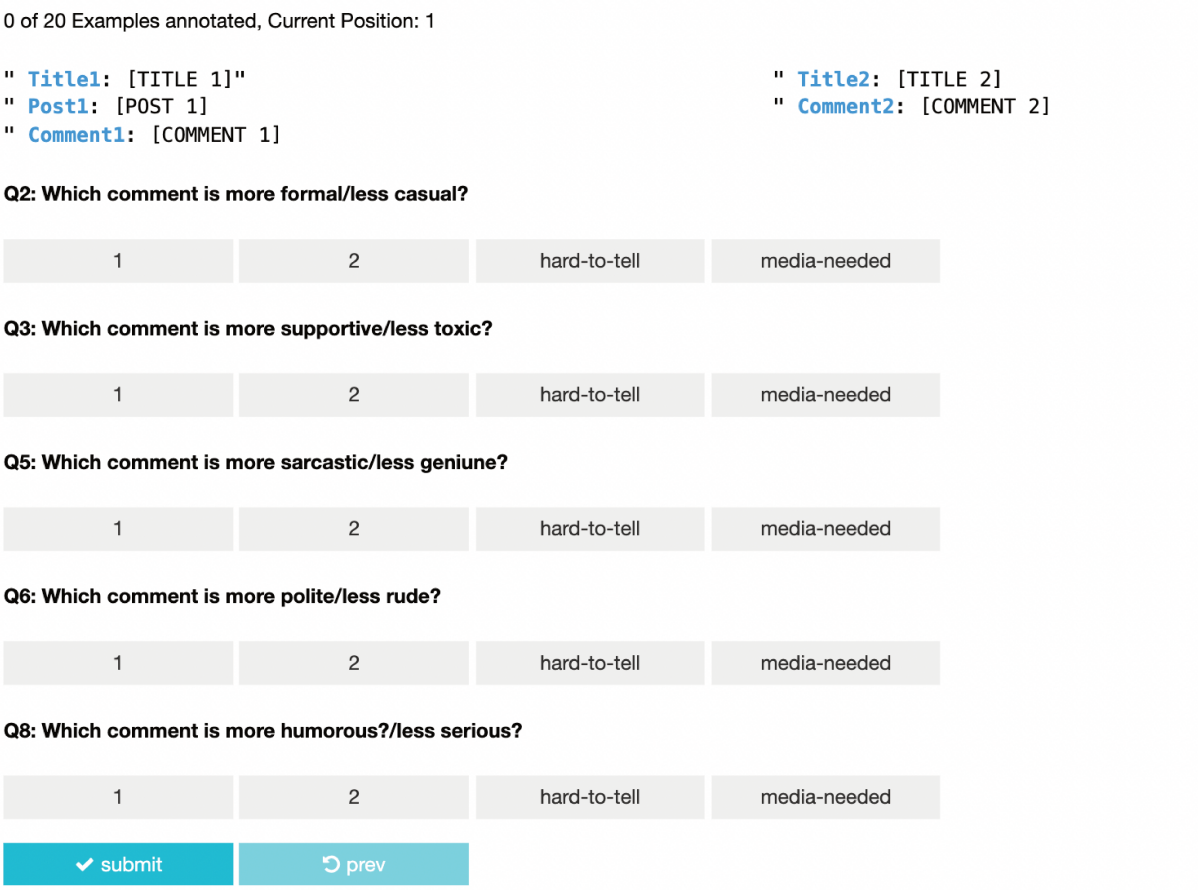


Figure 60: Human annotation UI for the binary norm dimension classification task. For each question, two options (1, 2) were provided without a tie option. Additionally, there were two extra options to mark samples that could not be properly annotated with the given context (hard-to-tell, media-needed).

General Guidelines

- Please annotate the comments **as if you were a Redditor** judging the comment in the context of the post based on the provided norm features. When annotating each norm per comment, please assume that the other norms are neutral.
- Label the comment as ***Hard-To-Tell*** if you don't have enough information or the comment doesn't contain any of the norm elements.
- Label the comment as ***In-Between/Neither*** if the comment contains around an equal mixture of both ends of the norm feature (ex. A comment containing both formal and informal aspects should be rated as "In-between")
- To differentiate between the scales 0-1 and 3-4, think about **whether you can envision the comment becoming more intense**. For example, between formal (3) and very formal (4), if you can think the comment can get significantly more formal, rate it as a 3.

Below, we outline the definitions of each norm, offer examples along the scale, and explain the guidelines on what to look for in the provided comment.

Casual – Formal:

Slang: mid, rizz, irl, sheesh, bet, sus, cap, plz, orz – slang is more informal (?) Abbreviations: idts, idk, btw, tbh, fyi, lol, lmao, omg Acronym: POTUS, NLP, RAM – can still be formal Difference between slang and acronym is that slang is language outside of conventional usage while acronym is an abbreviation formed by (usually initial) letters taken from a word or series of words, that is itself pronounced as a word. As a verb, slang is to vocally abuse, or shout at. Focus on linguistic attributes, not the content of the text.

1. **Very Casual**: extensive use of abbreviations, slangs, non-standard capitalization, missing syntactic components (no noun, no verb in sentence), incorrect punctuations, colloquialisms, contractions. – include one/two word answers.
2. **Casual**: existence of slangs, missing syntactic components (no noun, no verb in sentence), unnecessary use of exclamation marks, or frequent (>=4) presence of nonstandard capitalization, missing/incorrect punctuations, abbreviations, colloquialisms, contractions.
3. **In-between**: Presence of a few (<4) nonstandard capitalization (e.g. not capitalized first letter of sentence), missing/incorrect punctuation, abbreviation, colloquialisms, contractions, and relatively complete/structured sentences. No slangs.
4. **Formal**: reasonably structured (explanations, reasoning, etc.), correct capitalization, complete sentences, correct punctuation. No abbreviations, no slang, no colloquialisms, can have acronyms and contractions.
5. **Very Formal**: very structured thoughts and professional language, no abbreviations/slang/contractions/colloquialisms, grammatically correct.

Figure 61: Annotation guideline provided to human annotators.

Supportive – Toxic:

For contextual dependent cases (or comments that are implicit), we don't have to assume the worst intentions but also consider what the readers would think.

1. **Very Supportive**: extremely positive, encouraging, promotes supportive & uplifting discussion. (e.g. omg i absolutely love this!!!!!)
2. **Supportive**: respectful, constructive comments that have a positive outlook, not necessarily zealously supportive. Makes an effort to answer the question.
3. **In-between**: neither supportive or toxic. Usually short texts like "Coffee and music" which doesn't include any supportiveness or toxicity features
4. **Toxic**: rude, unfriendly, disrespectful, promotes toxic behavior, leads to negative atmosphere. Will make a (normal) reader a little uncomfortable. Using inappropriate language.
5. **Very Toxic**: Aggressive, attacking the OP or others. Extremely rude, unreasonable, or even psycho. Outright judging that others are wrong/inferior. Using extremely inappropriate language.

Genuine – Sarcastic

Sarcasm is not supposed to be offensive. Verbal irony is when saying the opposite of what one means but sarcasm is verbal irony BUT trying to be funny and not actually insulting.

1. **Very Genuine**: extremely sincere, honest, no implications. Profound or heartfelt messages.
2. **Genuine**: sincere and authentic, not lying. Includes subjective opinions that have enough content and context to judge as genuine (i.e. not a few words). E.g. some helpful advice.
3. **Neither/In-between**: Neither genuine nor sarcastic. Often includes short, objective answers (i.e. 1-3 words) that don't imply anything.
4. **Sarcastic***: appears nice, but actual meaning is opposite to textual meaning and is often negative. Often an intention to be funny.
5. **Very Sarcastic**: extreme ridicule or mockery, implicitly insulting. Exaggerated verbal irony.

Figure 62: Annotation guideline provided to human annotators.

Rude – Polite

Linguistic politeness theory: showing concern for people's positive or negative face.

1. **Very Rude**: disrespectful, demanding, offensive tone. E.g. "get the fuck out, shut up."
2. **Rude**: not considering others feelings, imposing, generalizing without knowing the full context. E.g. judgy: "people like you would never…", giving unsolicited advice: "Never …!" or comments that don't really answer the question. Using exclamation/all caps when unnecessary. Often does not save their own or other's face.
3. **In-between**: neither showing concern for others' "face" nor being disrespectful. E.g. "you can do this…,". Often includes comments that are straightforward but not rude. "bald-on record politeness" in politeness theory.
4. **Polite**: Making individuals feel good about themselves (appealing to positive face) or making the individuals feel like they haven't been imposed upon/taken advantage of (appealing to negative face). in case of agreement: friendliness and camaraderie, compliments, common grounds; in case of disagreeing opinions: not assuming, not coercing, recognizing and addressing the hearer's right to make his or her own decisions freely. (E.g. No offense but…, People usually…, I'm sure you know more than I do but…, replacing "I" and "you" with "people" or "we"). "positive politeness" and "negative politeness" in politeness theory.
5. **Very Polite**: showing concern for others. give hints, give clues of association, presuppose, understate, overstate, use tautologies. Rely on the hearer to understand implications (e.g. I would do…, do you think you want to…) "Off-record politeness" in politeness theory.

Figure 63: Annotation guideline provided to human annotators.

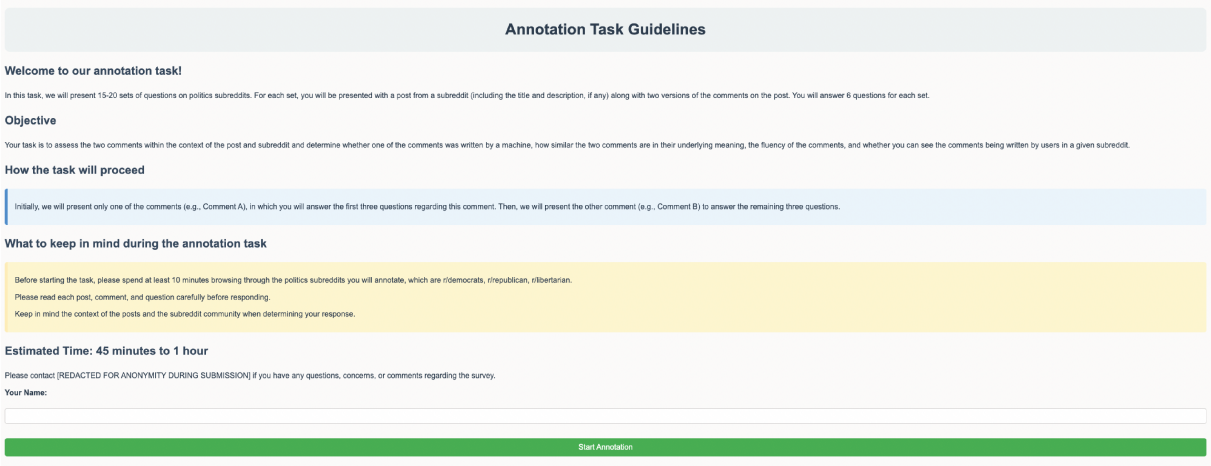


Figure 64: Human annotation UI for validating the quality of the filtered synthetic data.

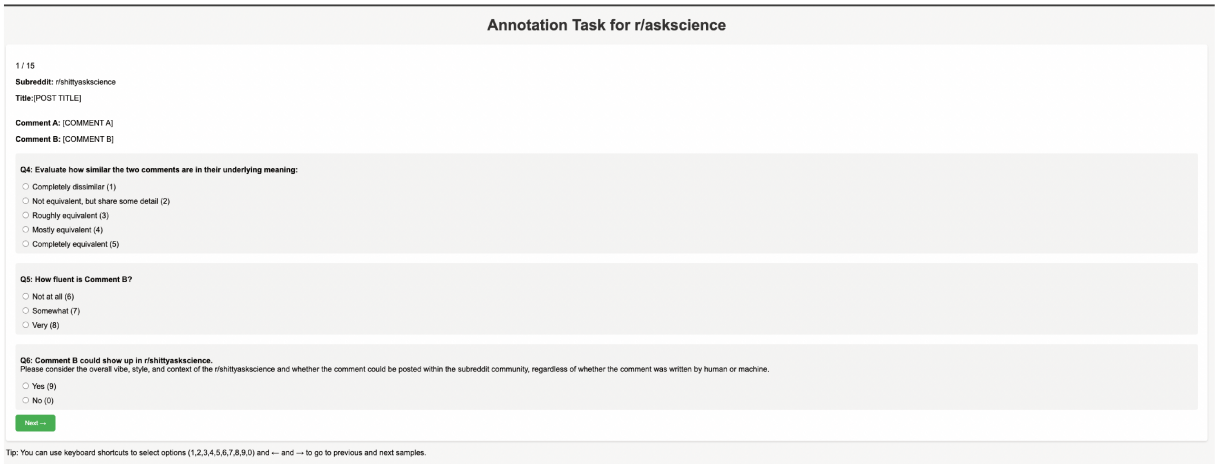


Figure 65: The annotators are presented with two versions of comments on a post: one synthetic and the other the original seed comment. Then, the annotators evaluated these two comments for their qualities, such as fluency and content preservation.