The Moral Foundations Reddit Corpus

Jackson Trager*, Alireza S. Ziabari*, Elnaz Rahmati, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Georgios Chochlakis, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani

University of Southern California
Corresponding author: jptrager@usc.edu

Abstract

Moral framing and sentiment can affect a variety of online and offline behaviors, including donation, environmental action, political engagement, and protest. Various computational methods in Natural Language Processing (NLP) have been used to detect moral sentiment from textual data, but achieving strong performance in such subjective tasks requires large, hand-annotated datasets. Previous corpora annotated for moral sentiment have proven valuable, and have generated new insights both within NLP and across the social sciences, but have been limited to Twitter. To facilitate improving our understanding of the role of moral rhetoric, we present the Moral Foundations Reddit Corpus, a collection of 16,123 English Reddit comments that have been curated from 12 distinct subreddits, hand-annotated by at least three trained annotators for 8 categories of moral sentiment (i.e., Care, Proportionality, Equality, Purity, Authority, Loyalty, Thin Morality, Implicit/Explicit Morality) based on the updated Moral Foundations Theory (MFT) framework. We evaluate baselines using large language models (Llama3-8B, Ministral-8B) in zero-shot, few-shot, and PEFT (Parameter-Efficient Fine-Tuning) settings, comparing their performance to fine-tuned encoder-only models like BERT (Bidirectional Encoder Representations from Transformers). The results show that LLMs continue to lag behind fine-tuned encoders on this subjective task, underscoring the ongoing need for human-annotated moral corpora for AI alignment evaluation.

Keywords: moral sentiment annotation, moral values, moral foundations theory, multi-label text classification, large language models, benchmark dataset, evaluation and alignment resource

1. Introduction

Moral rhetoric and framing play a role in increasing polarization and divisions in our societies (Marietta, 2008; Dehghani et al., 2016; Brady et al., 2020), but also in a wide range of pro-social behaviors that can potentially bring people together (Voelkel et al., 2022; Wolsko, 2017; Kidwell et al., 2013; Moaz, 2020). In order to understand the relationship between hate, division, compassion, and unity in the digital age, we need to understand the dynamics of moral language online. In particular, capturing and investigating the moral sentiment of text can allow for the study of how individuals' and groups' expressed moral sentiment relate to various downstream online and offline behaviors.

Moral sentiment assessment and classification are subjective tasks, and when done automatically using Natural Language Processing (NLP) techniques, this subjectivity results in the need for large and diverse, both in terms of topics and coders, sets of annotations. The Moral Foundations Twitter Corpus (MFTC; Hoover et al., 2020), a collection of 35,108 tweets that have been curated from seven distinct domains of discourse and hand annotated for 10 categories of moral sentiment (care, harm, fairness, inequality, loyalty, betrayal, authority, subversion, purity, and degradation) based on the Moral Foundations Theory (Haidt, 2012; Graham et al., 2011), was released a few years ago. This corpus has been used to design novel methods for moral sentiment classification (Asprino et al., 2022; Lan and Paraboni, 2022; Burton, 2022; Araque et al., 2020; Wang and Liu, 2021; Wu and Huang, 2022), utilized in models to investigate the impacts of moral framing in other domains (e.g., misinformation, polarization, and hate; Mutlu et al., 2020; Abdurahman et al., 2025; Ruch et al., 2022; Kennedy et al., 2023), and has been applied to train models that produce morally salient text (e.g., arguments and jokes; Alshomary et al., 2022; Yamane et al., 2021). In the era of large language models (LLMs), MFTC has been employed in evaluating the moral reasoning of LLMs (Ramezani and Xu, 2024) and serves as a foundation for new moral-reasoning benchmarks (Trager et al., 2025), underscoring how morally annotated corpora support both evaluation and alignment work. However, as useful as MFTC is, its training dataset is limited to the social-media platform Twitter.

Different online social media platforms have different linguistic and social structural environments that may result in variations in moral language and behavior (Curiskis et al., 2020). Different platforms have varying character limits (e.g. 280 characters on Twitter compared to 10k-40k on Reddit) which

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6383–6407
11–16 May 2026.

alters the language usage of users (Boot et al., 2019) and therefore may contribute to differences in the use and effectiveness of moral rhetoric (Candia et al., 2022). Additionally, different platforms have different policies with respect to the levels of user anonymity and sensitive content moderation which may influence the domains of morality discussed given the potential judgements from others. Research has shown that higher levels of anonymity reduces the feeling of responsibility and alters moral behavior online (Simfors and Rudling, 2020). Lastly, modern NLP methods are known to require massive training data for producing sufficiently accurate, generalizable, and robust models. It has empirically been shown that diverse sets of training data, from different platforms and on different topics, can help improve the classification results by allowing the models to obtain generalized domain knowledge (see Kennedy et al., 2022a), rather than surface knowledge restricted to a particular platform and a small set of topics.

As mentioned previously, the MFTC relied on the Moral Foundation Theory's framework which is a pluralistic perspective of moral cognition and identifies multiple dimensions of moral values that have evolved to facilitate individual well-being, coalitional unity, and cooperation with strangers (Haidt, 2012; Graham et al., 2013). The original version of the theory identified five separate but interrelated categories of moral concerns: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Purity/Degradation. Recently, a revision to the Moral Foundations Theory (Atari et al., 2023) split Fairness into the distinct and new foundations of Equality and Proportionality, while retaining the other four existing foundations. This split aims to capture the distinct moral concerns of fairness in procedure (proportionality) and equality of outcome (equality). In order to better understand the different nuances that will result from this theoretical change, we need to have an updated annotated corpus that complies with the latest theoretical revisions.

Together, the above reasons call for a corpus from a different platform, focused on a diverse set of topics, and annotated by a diverse group of annotators. Here we address this need by introducing the Moral Foundations Reddit Corpus (MFRC), a collection of 16,123 `Reddit.com` comments annotated for 8 categories of moral sentiment and curated from 12 morally-relevant subreddits.

Reddit is a public social media platform with over 52 million daily users who post content in over 138,000 active "subreddits" (user-created and user-moderated communities about different subjects) (Upvoted Staff, 2021). Shared content and comments in subreddits are "voted" on by users which is used to decide the visibility of the post. Activity on Reddit has been the center of numerous prominent cultural moments including coordinated attempts to challenge short-sellers of GameStop stock (Roose, 2021) or attempts to identify the Boston bombing terrorists (Starbird et al., 2014).

We focused our corpus compilation effort on Reddit for a number of reasons. First, in comparison to Twitter, Reddit shares many of the same research friendly features (e.g., responsiveness to current events, public posts, and available APIs) (Proferes et al., 2021), but is organized into what are called subreddits. Different subreddits have distinct topics and consistent communities (Datta and Adar, 2019; Soliman et al., 2019) with varying cultures and norms (Chandrasekharan et al., 2018; Fiesler, 2019). In relation to morality, these distinct communities and norms have led researchers to use Reddit to investigate moral conflicts across groups (e.g., Kumar et al., 2018), a phenomena that is harder to investigate on Twitter (which does not have organized groups) or Facebook (where many groups are private). Second, Reddit provides more anonymity than many other social media platforms, potentially enabling users to more freely speak their minds and express their opinions (e.g., Triggs et al., 2021; De Choudhury and De, 2014; Simfors and Rudling, 2020). Third, in addition to general differences in language usage (Boot et al., 2019), the lack of restriction on the length of posts on Reddit may be particularly beneficial for training models. Fourth, we believe that Reddit has played a distinct role in contemporary politics. For example, the r/TheDonald and r/incels subreddits have been linked to political extremism (Gaudette et al., 2021) and mass shootings (Helm et al., 2022).

To examine the effectiveness of LLMs for moral concern detection, we evaluate Ministral (Jiang et al., 2024) and Llama (Dubey et al., 2024) under both zero-shot and few-shot prompting settings for single-label and multi-label text classification, following standard prompting methodologies (Brown et al., 2020). Beyond prompting, we also perform LoRA parameter-efficient fine-tuning (PEFT; Hu et al., 2022) of Llama models for single-label and multi-label classification tasks (Tsoumakas and Katakis, 2007). For comparison, we include fully fine-tuned BERT models (Devlin et al., 2019), trained in both single-label and multi-label configurations. These baselines allow us to understand the validity and relative performance of different text classification methods for identifying the moral concerns manifested in the MFRC.

Finally, in order to facilitate research into annotator response patterns and bias, as recommended by Prabhakaran et al. (2021), we provide psychological and demographic metadata of our annotators. The background and biases of human annotators have been shown to impact their annotations (Hovy and Prabhumoye, 2021; Davani et al., 2022, 2023;

Bolukbasi et al., 2016) with particularly damaging effects that amplify pre-existing biases (Mujtaba and Mahapatra, 2019; Zhao et al., 2017). Annotators' biases may be particularly relevant in domains characterized by high subjectivity, such as moral values (Garten et al., 2019a,b; Hoover et al., 2020). While, for example, an annotator's political ideology might not have a substantial influence on how they annotate "positive" and "negative" sentiment in a corpus of hotel reviews, it seems likely that their ideology could substantially influence how they annotate expressions of justice and respect in a politically relevant corpora.

The contributions of this paper are threefold: (1) We introduce MFRC, a collection of 16,123 comments curated from 12 morally relevant subreddits and annotated for 8 moral sentiment categories based on the revised MFT. (2) We provide demographic and psychological metadata of annotators to facilitate future research on subjectivity and bias in moral sentiment annotation. (3) We evaluate a range of baseline models, including fine-tuned encoder-only models and recent LLMs in zero-shot, few-shot, and PEFT settings, establishing the MFRC as a benchmark for both moral sentiment classification and AI alignment evaluation. These contributions collectively extend prior work such as the MFTC (Hoover et al., 2020) to a new platform and theoretical framework, enabling richer cross-platform and cross-cultural analyses of moral language online.

2. Corpus Overview

As noted above, the MFRC consists of 16,123 Reddit comments drawn from 12 different subreddits. These subreddits were chosen based on the following criteria: First, we focused on subreddits that we expected to contain a wide range of moral concerns. Second, the chosen subreddits had to be active and have sufficient data. Third, we aimed to have a non-US based political subreddit with focus on current events that could be of use for different research communities. Accordingly, our corpus consists of 12 subreddits organized into three buckets; US politics, French politics, and Everyday moral life. The US politics bucket contains comments from 3 subreddits from the dates 1/1/2020 - 1/31/2021; *r/politics* which captures political moral language generally, *r/conservative* which covers moral rhetoric of the right, and *r/antiwork* which covers different, but still political moral language from the left. The everyday moral life bucket is a collection of topics related to various aspects of everyday life, collected for their non-political moral judgement and moral emotions which includes comments from the 4 subreddits of *r/nostalgia*, *r/AmItheAsshole*, *r/confession*, and *r/relationshipadvice* between the dates of 1/1/2020 - 1/31/2021. The third bucket on French politics and contains comments from the subreddits of *r/europe*, *r/worldnews*, *r/neoliberal*, *r/geopolitics*, and *r/Conservative* that had the relevant keywords related to the presidential race including 'Macron', 'Le Pen', 'France', 'French', and 'Hollande' (see below for the full set of keywords used) from the dates of 01/01/2017 to 06/30/2017 and had at least 10 likes/comments in order to control for sufficient engagement.

The MFRC is available for download as a HuggingFace dataset [1].

2.1. General Sampling Procedure

In assembling the MFRC, we sampled Reddit posts from a larger set of each subreddit. Our initial filtering criteria selected posts of sufficient length (at least 10 tokens), and de-selected any posts that were automatically marked by reddit as a bot, with the text "I am a bot" appended to the end of the post. The French politics bucket was also filtered for comments that mentioned at least one of the French presidential candidates ('macron', 'le pen', 'hollande', 'dupont-aignan', 'hamon', 'arthaud', 'poutou', 'cheminade', 'lassalle', 'melenchon', 'asselineau', 'fillon') in which most hits were for frontrunners Macron, Le Pen, Hollande, and Fillon. For the US Politics and Everyday buckets, comments were selected that had a comment score of at least 10.

In this filtered set, the proportion of comments that contained moral sentiment proved too low to conduct fully randomized sampling in a way that would result in a sufficient amount of moral examples. While these subreddits were chosen specifically for their potential moral salience, research has shown that use of moral language in some domains are rare (Atari et al., 2022). To address this issue, a semi-supervised method was used to up-sample from moralized posts (Kennedy et al., 2022b). Specifically, we first used word embeddings and a list of moral foundations seed words to compute a moral loading score (Distributed Dictionary Representations; Garten et al., 2018) for every comment. Next, for each moral concern, we computed the 95% percentile scores to mark the highly moral comments and set bin size. Finally, in order to have a diverse range of moral posts, we compiled comments in a manner that each subreddit bucket consisted of 1/2 comments with high moral loading ($> 95\%$), and 1/2 comments with less high moral loading ($\leq 95\%$).

This filtering and sampling procedure yielded approximately 6,000 comments for US Politics, 6,000 for Everyday Politics, and 8,000 comments for French politics. However, since vice, virtue, and

[1] https://huggingface.co/datasets/USC-MOLA-Lab/MFRC

multiple foundations regularly co-occur in an individual comment, duplicates occurred and were subsequently removed, resulting in a smaller final sample size for each bucket (US Politics: 4,821; French Politics: 6,489; Everyday Morality: 4,813; Total: 16,123).

3. Annotation

Every post in the MFRC has been labeled by at least three trained annotators from a set of five (see Table 1 for the distribution of annotators for the corpus) for 8 categories of moral sentiment as outlined in the new version of our annotation manual (See Appendix A).

3.1. Moral sentiment

Moral sentiment labels are drawn from the recently revised typology of Moral Foundations Theory (Atari et al., 2023), which proposes a six-factor taxonomy of morality. In this model, each factor includes both virtues, or prescriptive moral concerns, and vices, prohibitive moral concerns. The proposed moral foundations are (Atari et al., 2023):

Care/Harm: Intuitions about avoiding emotional and physical damage or harm to others. It underlies virtues of kindness, gentleness, and nurturing, and vices of meanness, violence, and abuse.

Equality/Inequality: Intuitions about egalitarian treatment and equal outcome for all individuals and groups. It underlies virtues of social justice and equality, and vices of discrimination and prejudice.

Proportionality/Disproportionately: Intuitions about individuals getting rewarded in proportion to their merit (i.e., effort, talent, or input). It underlies virtues of meritocracy, productiveness, and deservingness, and vices of corruption and nepotism.

Loyalty/Betrayal: Intuitions about in-group cooperation and out-group competition. It underlies virtues of patriotism and self-sacrifice for the group, and vices of abandonment, cheating, and treason.

Authority/Subversion: Intuitions about deference toward legitimate authorities and high-status individuals. It underlies virtues of leadership and respect for tradition, and vices of disorderliness and resenting hierarchy.

Purity/Degradation: Intuitions about avoiding bodily and spiritual contamination and degradation. It underlies virtues of sanctity, nobility, and cleanliness and vices of grossness, impurity, and sin.

Notably, unlike the MFTC, in the MFRC we did not code for Fairness. Rather, following the latest theoretical developments in the field (Atari et al., 2023), we coded for separate foundations of Proportionality and Equality.

In addition to these six foundations, annotators were trained to look for an additional construct: *Thin Morality* – a moral judgment or concern which is voiced without clearly referring to one of the six moral domains (Atari et al., 2022). This brings the total categories of moral sentiment to 7. Annotators also had a formal category for *Implicit/Explicit Morality* – whether the moral sentiment in the comment was expressed explicitly or implicitly. Lastly, the annotators were asked to report their overall level of confidence in their annotation as not confident, somewhat confident, or very confident.

Table 1: Number of Reddit Posts Annotated by N Annotators for Each Subreddit

	N Annotator		
Subreddit	3	4	5
r/AmItheAsshole	1009	330	-
r/Conservative(French politics)	75	69	-
r/Conservative(US politics)	870	898	8
r/antiwork	885	882	4
r/confession	993	338	-
r/europe	1338	1302	7
r/geopolitics	53	59	1
r/neoliberal	846	815	12
r/nostalgia	994	348	-
r/politics	864	894	10
r/relationship_advice	1021	331	1
r/worldnews	1267	1288	9

3.2. Annotators

We started with a larger pool of 27 annotators, all undergraduate research assistants who completed two months of training sessions to develop expert-level familiarity with MFT. Training consisted of lectures, discussions, readings, and practice annotations with inter-annotation agreement analysis. In early annotation stages, annotator disagreement was addressed through discussion and, if necessary, certain labels were modified. However, given the subjective nature of the task, in many cases, it is difficult to make the determination of whether or not a document expresses moral sentiment or which category of moral sentiment it expresses. While it is necessary to have consistent annotator training, a focus on maximizing annotator agreement risks artificially inflating agreement at the cost of suppressing the natural variability of moral sentiment (Hoover et al., 2020). Accordingly, our annotators were trained to both strive for consistency, while also encouraged to avoid stereotypes that may increase agreement with other annotators but would lead them to ignore their own beliefs. Out of these original set of annotators, we selected the top five performing annotators, based on both inter-coder reliability assessments and the commitment of the annotators to the project, to become primary annotators and complete the rest of the annotations in

our corpus.

Our trained annotators were independently assigned to label each comment from a subset of comments sampled from a corpus associated with one of the 12 subreddits (see Table 2). The annotations were performed on Prodigy.[2] Each post was assigned a label indicating the absence or presence of the six foundations, thin morality, explicit/implicit expression, and the confidence level of the annotator (MFT Coding Manual in Appendix A).

3.3. Annotator Metadata

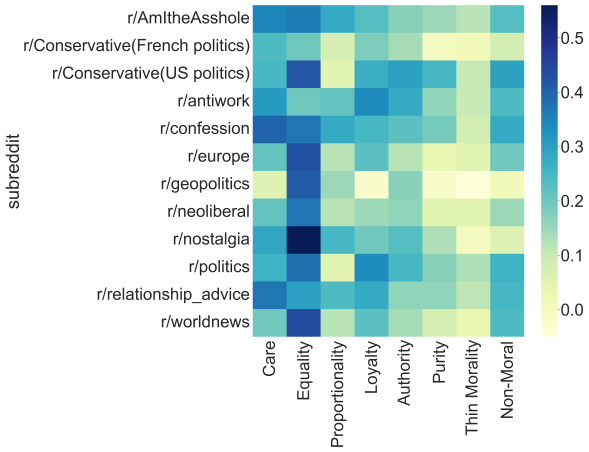
We have also collected responses to a range of psychological and demographic measures from our annotators. While keeping their identities anonymous, we provide measures of each annotator's gender, sexual orientation, age, household income, first language, political ideology along a liberal-conservative scale, religious affiliation, and moral values via the Moral Foundation Questionnaire-2 (Atari et al., 2023). Basic analyses demonstrate that our annotators' political ideology and morality skews liberal while family income skews wealthier than the average American. Based on the recommendations of Prabhakaran et al. (2021), in order to increase utility and transparency of this corpus, these measures are provided by request (for privacy concerns) and encourage research into their potential impact on annotations, and the subsequent biases in the machine learning models.

4. Annotation Results

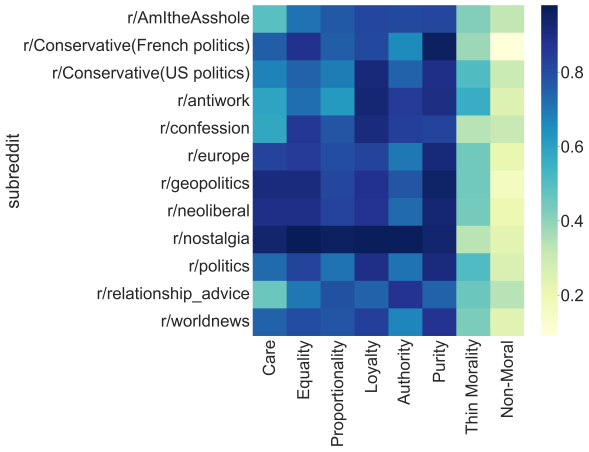
The annotation results can be seen in Table 2. Recall that each post was annotated for multiple labels by multiple annotators, and the frequencies reported in Table 2 are calculated based on annotators' majority vote (i.e. posts receiving at least 50% agreement for that label). For example, if a particular post is annotated as 'proportionality' by at least 50% of the annotators who coded that post, then the majority vote on 'proportionality' for that post is positive. We acknowledge the uneven distribution of labels across the various subreddits, in addition to the low base-rates of annotated posts for some the moral concerns, especially for Purity, Loyalty and Proportionality concerns.

The interannotator agreement results, using Fleiss's (Fleiss, 1971) kappa and prevalence-and bias-adjusted Fleiss's kappa (PABAK; Sim and Wright, 2005) for multiple annotators, are displayed in Figure 1. Fleiss's kappa is generally viewed as the gold standard measure for investigating agreement across many annotators, and it represents the degree of agreement beyond what is expected by chance. This measure though, is heavily influenced by the prevalence of positive cases. Given the subjective nature of our task, and the fact that positive cases are not prevalent given the often rarity of moral rhetoric (Atari et al., 2022), we use PABAK which adjusts kappa for prevalence and bias. As expected, given the low base rate of moral language across the various subreddits (i.e. low positive cases), most reported kappa's are low. However, once adjusted for the issue of prevalence, we see medium to high agreements across the subreddits.

Figure 1: The heatmaps show Interannotator Agreement (PABAK and Kappa) scores for all subreddits and foundations. Higher agreement corresponds with darker colors in both heatmaps.



(a) Fleiss' kappa



(b) PABAK

5. MFRC as a Community Resource

The preprint release of MFRC has already been widely adopted across disciplines, serving as a shared empirical resource for research on morality, values, and language. Beyond being cited as related work, many studies have directly reused MFRC as training data, an evaluation benchmark,

[2] https://prodi.gy/

Table 2: Frequency of Reddit posts per Foundation Calculated Based on Annotators' Majority Vote.

Subreddit	Care	Equality	Proportionality	Loyalty	Authority	Purity	Thin Morality	Non-Moral
r/AmItheAsshole	343	145	103	56	50	38	227	456
r/Conservative(French politics)	12	5	6	7	16	0	18	95
r/Conservative(US politics)	195	200	84	38	155	45	231	945
r/antiwork	304	132	202	46	86	29	186	930
r/confession	281	69	101	24	52	50	249	574
r/europe	105	180	117	107	174	18	324	1741
r/geopolitics	1	4	5	1	5	0	9	100
r/neoliberal	39	74	67	50	117	11	195	1210
r/nostalgia	20	11	15	5	8	9	207	1160
r/politics	148	119	72	61	161	34	256	1016
r/relationship_advice	418	137	84	98	31	56	191	450
r/worldnews	166	249	127	105	192	33	321	1543
All	2032	1325	983	598	1047	323	2414	10220

or a methodological reference point. These early applications provide evidence that MFRC fills a practical gap in the study of moral language and has begun to shape how researchers operationalize and evaluate moral constructs in computational settings.

A prominent line of work uses MFRC to evaluate and align LLMs with human moral judgments. For example, Rathje et al. (2024) include MFRC among multiple psychological datasets to benchmark GPT-3.5 and GPT-4 on moral foundation detection, comparing model outputs directly against human annotations. Similarly, Abdurahman et al. (2024) analyze MFRC posts using both traditional NLP models and LLMs to demonstrate systematic divergences between machine-generated and human moral labels. More recent work explicitly leverages MFRC as a testbed for assessing the limits of LLM moral understanding, showing that even state-of-the-art models struggle with subtle or overlapping moral categories (Bulla et al., 2025; Skorski and Landowska, 2025). Collectively, these studies position MFRC as a grounding dataset for evaluating moral alignment and robustness in contemporary language models.

MFRC has also enabled methodological advances in moral NLP. Several studies use MFRC to train or benchmark new models for moral foundation classification, including domain-general classifiers such as MoralBERT (Preniqi et al., 2024), multi-domain and transfer-learning approaches (Guo et al., 2023; Chen et al., 2025), and architectures enriched with emotional or event-level features (Nguyen et al., 2024; Zangari et al., 2025a). Others exploit MFRC's multi-label and subjective structure to study annotation disagreement and evaluation itself, proposing improved metrics or auditing techniques for noisy moral labels (Mokhberian et al., 2022; Chochlakis et al., 2025b). Survey work further highlights MFRC as a central benchmark in the moral NLP literature, noting its consistent utility across platforms and domains (Zangari et al., 2025b). Together, these efforts underscore MFRC's role not only as a dataset, but as an infrastructure supporting model development, evaluation, and methodological reflection.

Finally, MFRC has served as a foundation for extending moral language analysis to new domains, languages, and modalities. Researchers have used MFRC-trained models to study moral disagreement in online discussions (Van Der Meer et al., 2023), to probe the limits of English-centric moral classifiers in multilingual political texts (Cheng and Hale, 2025), and to inspire domain-specific adaptations in areas such as news discourse, music lyrics, and multimodal content (Preniqi et al., 2024; Lei et al., 2024). Even when MFRC is not directly used, it frequently provides the moral taxonomy, baselines, or conceptual framing for new datasets and tasks. This breadth of uptake suggests that MFRC is already functioning as a general-purpose community resource, supporting cumulative research on how moral values are expressed, modeled, and evaluated in language.

6. Baseline Classification Models

In addition to compiling the corpus, we experiment with different models and provide baselines for predicting the annotators' majority vote of moral sentiment categories. Our goal is to simply establish baselines that future work can build on.

It should be noted that each post in MFRC was coded by multiple annotators for 8 different categories of moral sentiment discussed under the Section 3. This is a multi-label classification task; i.e., not only are the categories of moral sentiment not independent of one another, but understanding variance in one domain should theoretically inform about other related moral domains. Here though, we provide both single-label (treating labels as in-

dependent) and multi-label classification results.

Zero-shot and Few-shot Prompting To establish an initial baseline, we evaluate two widely used open-source LLMs, Meta-Llama-3-8B (Dubey et al., 2024) and Ministral-8B-Instruct-2410 (Jiang et al., 2024), using zero-shot prompting for both single-label and multi-label classification. We further explore few-shot prompting to assess whether including example instances for in-context learning (Brown et al., 2020) improves performance. For this setup, we select five stratified samples based on label distribution and include them in the prompt.

Classification Fine-tuning To examine the impact of training on LLM performance, we fine-tune the Llama model using parameter-efficient training (Hu et al., 2022) and evaluate its improvement on our dataset. Prior work has shown that for discriminative tasks, encoder-only models with full fine-tuning can outperform LLMs even after training (Qorib et al., 2024; Roccabruna et al., 2024). Therefore, we also include results from fine-tuning the BERT model (Devlin et al., 2019). Since this model is smaller, we apply full fine-tuning (FFT) instead of PEFT.

Due to the sparsity of our data, with far fewer moral posts than non-moral posts, we employ a weighted loss function. In this setup, the weight of sample i for label l is inversely proportional to the frequency of that label:

$$w_{i,l} \propto \frac{1}{\text{number of samples with label } l} \quad (1)$$

The baseline metrics we report are the F_1, precision, and recall, which are calculated across stratified 5-fold cross-validation. For the multi-label models, we stratified the data based on the moral sentiment label (the union of all the moral labels). We used 10% of the training data for validation, and we chose the best performing model based on binary F_1 for single-label classification and F_1 macro for multi-label BERT. We train LLMs with a learning rate of 1e-5, batch size of 8, and 4 epochs, while BERT models are trained with a learning rate of 2e-5, batch size of 8, and 5 epochs. For multi-label classification, we double the number of training epochs. Learning rates are determined empirically from a search over the range [1e-4, 1e-6].

7. Results

The results of the baseline models are provided in Table 3. We provide both the average F1 score and standard deviations for each moral category (All metrics including recall and precision are included in the appendix C). We ran each of the baseline models once for the entire corpus. Consistent with prior work (Qorib et al., 2024; Roccabruna et al., 2024), BERT models outperform the fine-tuned Llama models in terms of F_1. Interestingly, the multi-label BERT model performed worse than its single-label counterpart, whereas the multi-label trained Llama achieved comparable or better results than the single-label version.

We then evaluated the best performing BERT models on the three subreddit buckets separately (French Politics, Everyday, and U.S. Politics). Consistent with previous results, BERT (single-label) outperformed BERT (multi-label) across all moral categories in all three buckets. However, performance varied across subreddits: certain moral foundations were predicted more accurately within the Everyday bucket, while others showed higher scores in U.S. Politics. This pattern suggests that subreddit focus and community context influence model performance, reflecting how moral language varies across online domains.

While our results establish majority-vote baselines, moral judgments remain inherently subjective, with meaningful annotator heterogeneity. Motivated by concerns about aggregation in subjective tasks (Chochlakis et al., 2025a), we additionally examine annotator-level personalization. Personalization yields modest gains, particularly for annotators whose labeling patterns diverge more from the majority vote, but does not alter the overall pattern of findings, including the superiority of fine-tuned models. Details are provided in Section B of the Appendix.

8. Cross-Corpus Classification

In this section, we evaluate preliminary results on the transferability of models between MFTC and MFRC. Our hope is that future work can build on such cross-domain tasks in order to extract more generalized knowledge about moral rhetoric independent of the source and topic of the post.

In order to provide a fair comparison between MFRC and MFTC, we trained BERT models with the same hyperparameter discussed in the previous section. Additionally, we downsampled the MFTC dataset to have the same number of samples for each label as MFRC. The results are presented in Table 4. In general, models trained and tested on the MFTC have better classification performance than the results for the MFRC (See Table 3).

As mentioned before, the MFRC relies on the updated taxonomy of MFT in which the Fairness foundation is split into Proportionality and Equality. This makes the cross-corpus training for Fairness, Proportionality and Equality difficult. We hope the MFTC will be updated to use the more nuanced Proportionality and Equality labels in that corpus.

Table 3: F1 scores (mean ± standard deviation) by moral category on the full MFRC dataset, comparing models under different methods: zero-shot, few-shot prompting, parameter-efficient fine-tuning (PEFT), and full fine-tuning (FFT).

Method	Model	Care	Equality	Prop	Loyalty	Authority	Purity	Thin
0-shot	$\text{Llama}_{\text{single}}$	$0.40_{0.004}$	$0.29_{0.004}$	$0.21_{0.004}$	$0.17_{0.004}$	$0.20_{0.004}$	$0.17_{0.011}$	$0.23_{0.006}$
	$\text{Llama}_{\text{multi}}$	$0.28_{0.003}$	$0.31_{0.011}$	$0.20_{0.017}$	$0.14_{0.006}$	$0.18_{0.013}$	$0.13_{0.017}$	$0.01_{0.004}$
	$\text{Ministral}_{\text{single}}$	$0.24_{0.014}$	$0.25_{0.010}$	$0.15_{0.019}$	$0.21_{0.011}$	$0.09_{0.008}$	$0.10_{0.013}$	$0.08_{0.008}$
	$\text{Ministral}_{\text{multi}}$	$0.41_{0.010}$	$0.27_{0.006}$	$0.19_{0.005}$	$0.17_{0.009}$	$0.22_{0.012}$	$0.17_{0.009}$	$0.20_{0.012}$
5-shot	$\text{Llama}_{\text{single}}$	$0.43_{0.005}$	$0.21_{0.002}$	$0.21_{0.008}$	$0.14_{0.002}$	$0.20_{0.002}$	$0.12_{0.003}$	$0.26_{0.005}$
	$\text{Llama}_{\text{multi}}$	$0.33_{0.005}$	$0.33_{0.011}$	$0.14_{0.022}$	$0.18_{0.012}$	$0.19_{0.011}$	$0.14_{0.020}$	$0.01_{0.001}$
	$\text{Ministral}_{\text{single}}$	$0.26_{0.018}$	$0.33_{0.009}$	$0.13_{0.014}$	$0.21_{0.011}$	$0.00_{0.002}$	$0.21_{0.007}$	$0.20_{0.006}$
	$\text{Ministral}_{\text{multi}}$	$0.44_{0.010}$	$0.30_{0.010}$	$0.21_{0.013}$	$0.18_{0.019}$	$0.25_{0.010}$	$0.17_{0.017}$	$0.15_{0.004}$
PEFT	$\text{Llama}_{\text{single}}$	$0.54_{0.012}$	$0.51_{0.016}$	$0.28_{0.011}$	$0.32_{0.007}$	$0.34_{0.015}$	$0.22_{0.012}$	$0.27_{0.024}$
	$\text{Llama}_{\text{multi}}$	$0.53_{0.013}$	$0.52_{0.009}$	$0.28_{0.024}$	$0.40_{0.016}$	$0.36_{0.016}$	$0.31_{0.037}$	$0.25_{0.049}$
FFT	$\text{BERT}_{\text{single}}$	$\mathbf{0.62_{0.020}}$	$\mathbf{0.58_{0.030}}$	$\mathbf{0.37_{0.040}}$	$\mathbf{0.45_{0.040}}$	$\mathbf{0.40_{0.050}}$	$\mathbf{0.51_{0.070}}$	$\mathbf{0.39_{0.020}}$
	$\text{BERT}_{\text{multi}}$	$0.59_{0.020}$	$0.57_{0.030}$	$0.31_{0.050}$	$0.43_{0.040}$	$0.35_{0.040}$	$0.48_{0.070}$	$0.34_{0.040}$

Table 4: BERT Results on MFTC

Foundation	F1	Precision	Recall
Authority	0.65(0.02)	0.65(0.04)	0.65(0.05)
Care	0.75(0.02)	0.76(0.03)	0.74(0.03)
Fairness	0.82(0.01)	0.83(0.04)	0.81(0.03)
Loyalty	0.58(0.05)	0.66(0.09)	0.52(0.06)
Purity	0.54(0.04)	0.63(0.05)	0.48(0.05)

For now though, for predicting Fairness labels in the MFTC, the union of Proportionality and Equality labels are calculated based on the MFRC trained models, and the output of the union is then compared to the Fairness category in the MFTC. Similarly, to evaluate the MFTC Fairness models on the MFRC, both Proportionality and Equality are assigned the Fairness label predicted from the MFTC.

Table 5: BERT Models Trained on MFRC, Evaluated on MFTC

Foundation	F1	Precision	Recall
Authority	0.38	0.53	0.30
Care	0.53	0.58	0.48
Fairness	0.35	0.76	0.23
Loyalty	0.38	0.50	0.31
Purity	0.28	0.56	0.18

Cross-corpus results are presented in Tables 5 and 6. These preliminary results are indeed encouraging in that we demonstrate transferability between the two corpora in predicting out-of-domain distributions. Previous research training classifiers on the MFTC, and testing on Reddit (e.g., Atari et al., 2021) have shown similar levels of accuracies for cross-domain classification. Any cross-corpus investigation should take into account the different time periods in which these two corpora were compiled. This difference can potentially impact the topics, the sentiments expressed about the topics, and the type of justification and reasoning used for the expressed sentiments. We believe though that more advanced methods in knowledge capture and representation could use the two corpora together to further achieve more generalized and better performing models.

Table 6: BERT Models Trained on MFTC, Evaluated on MFRC

Foundation	F1	Precision	Recall
Authority	0.31	0.23	0.44
Care	0.43	0.53	0.35
Fairness	0.34	0.36	0.32
Loyalty	0.32	0.48	0.23
Purity	0.34	0.52	0.25

9. Discussion

Moral rhetoric and framing have been shown to be predictive of various important pro-social and anti-social behaviors. Several NLP methods have recently been proposed for capturing and categorizing moral sentiment based on textual data (for a review see Atari and Dehghani, 2021). However, this is a subjective sentiment analysis task for which training data plays a vital role. To facilitate further research in this domain, here we introduced the MFRC, a collection of 16,123 Reddit comments annotated for 8 categories of moral sentiment, and provided a number of baseline results for different NLP models trained to predict moral sentiment.

The MFTC was introduced in 2020, and so far this corpus has already facilitated multiple lines of research in both NLP and the social sciences. We believe Reddit's distinct linguistic and social struc-

Table 7: Model F1 (mean ± sd) by Moral Category on the three Subreddit Buckets (French Politics, Everyday, US Politics)

Model	Subreddit Bucket	Care	Equality	Prop	Loyalty	Authority	Purity	Thin
$BERT_{single}$	French Politics	$0.43_{0.04}$	$0.59_{0.03}$	$0.16_{0.05}$	$0.41_{0.08}$	$0.20_{0.12}$	$0.32_{0.15}$	$\mathbf{0.41_{0.04}}$
	Everyday	$\mathbf{0.72_{0.02}}$	$\mathbf{0.61_{0.06}}$	$0.37_{0.08}$	$0.47_{0.08}$	$0.34_{0.11}$	$\mathbf{0.51_{0.10}}$	$0.37_{0.05}$
	US Politics	$0.54_{0.05}$	$0.58_{0.04}$	$\mathbf{0.42_{0.07}}$	$\mathbf{0.51_{0.09}}$	$\mathbf{0.51_{0.05}}$	$0.37_{0.16}$	$0.36_{0.05}$
$BERT_{multi}$	French Politics	$0.33_{0.05}$	$0.51_{0.06}$	$0.06_{0.07}$	$0.33_{0.10}$	$0.26_{0.04}$	$0.20_{0.16}$	$0.37_{0.05}$
	Everyday	$0.66_{0.04}$	$0.56_{0.07}$	$0.34_{0.08}$	$0.30_{0.07}$	$0.21_{0.10}$	$0.45_{0.12}$	$0.25_{0.03}$
	US Politics	$0.45_{0.07}$	$0.55_{0.06}$	$0.40_{0.08}$	$0.44_{0.11}$	$0.43_{0.09}$	$0.34_{0.20}$	$0.25_{0.07}$

ture, along with MFRC's methodological and theoretical updates, allow for potential new research that can both improve and expand the applications of MFTC. Specifically, the increased character limits on Reddit compared to Twitter is important for the more naturalistic expressions of moral rhetoric and its potential impact on the performance of classification models. While social media often provides large amount of data needed for training NLP models, with respect to sentiment analysis of moral language, the paucity of moral rhetoric in some domains (Atari et al., 2022) makes it difficult to gather sufficient amounts of training data (Hoover et al., 2020). Given Reddit's longer posts, models trained on the MFRC may perform better in out of domain tasks, especially in longer documents (e.g., articles or speeches compared to tweets). Further, the distinct subreddit communities allow for the study of group linguistic dynamics. For example, shifts in moral language over time associated with a hashtag on Twitter may show the evolving general public opinion on a topic, while shifts in moral language within a particular subreddit may reflect the changing views of a specific community.

Our results also showed that model performance varied across subreddits, suggesting that Reddit's community-based structure represents a distinct linguistic and moral domain compared to other platforms. Recognizing these structural differences is important for interpreting the downstream group behaviors associated with moral language use on different platforms, such as voting and political mobilization. Moreover, the increased anonymity on Reddit can facilitate research into the gap between identity-linked and publicly-expressed moral concerns, such as on Facebook or Linkedin, and anonymous expressions of moral values, as on Reddit.

As discussed previously, another important feature of the MFRC is that it is based on the newly updated version of the Moral Foundations Theory (Atari et al., 2023) which breaks the Fairness concern into the distinct moral concerns of Proportionality and Equality. The MFRC can be used to further investigate these nuances of fairness across topics such as income inequality and excessive wealth (Trager and Atari, 2025) .

Similar to the MFTC, the MFRC has detailed meta-data on the corpus annotators. We hope that by providing demographics and several key psychological measurements of our annotators, MFRC can facilitate future research into how annotators background characteristics impact their annotations.

As shown in Table 3, fine-tuned models (BERT and PEFT-Llama) substantially outperform the zero-shot and few-shot baselines across all moral categories on MFRC. This pattern is consistent with the view that subjective, label-rich moral classification benefits most from domain-specific supervision, reinforcing the value of large, human-annotated corpora for both evaluation and downstream alignment work (Abdurahman et al., 2024).

In conclusion, we hope that the MFRC, along with this report, in addition to the other previously released corpora in this domain (e.g., Kennedy et al., 2022b; Hoover et al., 2020; Trager et al., 2025), can aid researchers by providing much-needed data and open new lines of research both in NLP and in the social sciences. In an age where political movements, grassroots activism, and plans for insurrections seem to take place in online environments, it is vital that we can better understand the moral dynamics of these online conversations. The intent of this project has been to further facilitate research into these timely topics.

10. Data Disclaimer

We acknowledge that the compiled dataset contains biases and is not representative of diverse moral concerns present in world populations. Potential biases in the data include: biases specific to English-speaking countries and the English language, biases inherent to `Reddit.com` and its user base, biases in the researchers' criteria for corpus curation as well as the underlying MFT itself, bias in the assessment of moral labels, and the fact that annotators were all undergraduate research assistants at a private academic institute. All of these factors, among others, likely influenced the annotations, as well as the performance of machine learning models trained on the corpus. Anyone using this corpus should be aware of these

limitations and should acknowledge and/or try to mitigate them to the extent possible.

11. Ethics Statement

This work adheres to ethical standards for research involving human annotations and publicly available data. All annotators participated voluntarily, were compensated fairly, and received extensive training on respectful engagement with sensitive moral and political content. Annotation sessions included monitoring for emotional well-being and debriefing opportunities to mitigate potential distress. The study protocol was reviewed and approved by the authors' institutional ethics board. We highlight that models trained on this corpus should be used only for scientific and educational purposes, and not for applications that could profile, surveil, or harm individuals or communities.

12. Bibliographical References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Suhaib Abdurahman, Nils K Reimer, Preni Golazizian, Elisa Baek, Yixuan Shen, Jackson Trager, Roshni Lulla, Jonas Kaplan, Carolyn Parkinson, and Morteza Dehghani. 2025. Targeting audiences' moral values shapes misinformation sharing. *Journal of Experimental Psychology: General*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Luigi Asprino, Luana Bulla, Stefano De Giorgis, Aldo Gangemi, Ludovica Marinucci, and Misael Mongiovì. 2022. Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 33–41.

Mohammad Atari, Aida Mostafazadeh Davani, Drew Kogon, Brendan Kennedy, Nripsuta Ani Saxena, Ian Anderson, and Morteza Dehghani. 2021. Morally homogeneous networks and radicalism. *Social Psychological and Personality Science*, page 19485506211059329.

Mohammad Atari and Morteza Dehghani. 2021. Language analysis in moral psychology. In Morteza Dehghani and Ryan L. Boyd, editors, *Handbook of language analysis in psychology*, pages 207–228. Guilford.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157.

Mohammad Atari, Matthias R. Mehl, Jesse Graham, John M. Doris, Norbert Schwarz, Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Elaine Gonzalez, Nikki Jafarzadeh, Alyzeh Hussain, Arineh Mirinjian, Annabelle Madden, Rhea Bhatia, Alexander Burch, Allison Harlan, David A. Sbarra, Charles L. Raison, Suzanne A. Moseley, Angelina J. Polsinelli, and Morteza Dehghani. 2022. The paucity of morality in everyday talk. *Scientific Reports*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Arnout B Boot, Erik Tjong Kim Sang, Katinka Dijkstra, and Rolf A Zwaan. 2019. How character limit affects language usage in tweets. *Palgrave Communications*, 5(1):1–13.

William J Brady, Molly J Crockett, and Jay J Van Bavel. 2020. The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, 15(4):978–1010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Luana Bulla, Stefano De Giorgis, Misael Mongiovì, and Aldo Gangemi. 2025. Large language models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609.

Jason William Burton. 2022. *Understanding and supporting belief accuracy in a digital world*. Ph.D. thesis, Birkbeck, University of London.

Cristian Candia, Mohammad Atari, Nour Kteily, and Brian Uzzi. 2022. Overuse of moral language dampens content engagement on social media. *Under review*.

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. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Ziyu Chen, Junfei Sun, Chenxi Li, Tuan Dung Nguyen, Jing Yao, Xiaoyuan Yi, Xing Xie, Chenhao Tan, and Lexing Xie. 2025. Mova: Towards generalizable classification of human morals and values. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33204–33248.

Calvin Yixiang Cheng and Scott A Hale. 2025. Beyond english: Evaluating automated measurement of moral foundations in non-english discourse with a chinese case study. *arXiv preprint arXiv:2502.02451*.

Georgios Chochlakis, Alexandros Potamianos, Kristina Lerman, and Shrikanth Narayanan. 2025a. Aggregation artifacts in subjective tasks collapse large language models' posteriors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5513–5528.

Georgios Chochlakis, Jackson Trager, Vedant Jhaveri, Nikhil Ravichandran, Alexandros Potamianos, and Shrikanth Narayanan. 2025b. Semantic f1 scores: Fair evaluation under fuzzy class boundaries. *arXiv preprint arXiv:2509.21633*.

Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Information Processing & Management*, 57(2):102034.

Srayan Datta and Eytan Adar. 2019. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, pages 146–157.

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

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

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Morteza Dehghani, Kate Johnson, Joe Hoover, Eyal Sagi, Justin Garten, Niki Jitendra Parmar, Stephen Vaisey, Rumen Iliev, and Jesse Graham. 2016. Purity homophily in social networks. *Journal of Experimental Psychology: General*, 145(3):366.

Peter DeScioli and Robert Kurzban. 2013. A solution to the mysteries of morality. *Psychological bulletin*, 139(2):477.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Naomi Ellemers, Jojanneke Van Der Toorn, Yavor Paunov, and Thed Van Leeuwen. 2019. The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, 23(4):332–366.

Benjamin Enke. 2019. Kinship, cooperation, and the evolution of moral systems. *The Quarterly Journal of Economics*, 134(2):953–1019.

Casey Fiesler. 2019. Ethical considerations for research involving (speculative) public data. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–13.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

JA Frimer, R Boghrati, J Haidt, J Graham, and M Dehghani. 2019. Moral foundations dictionary 2.0.

Akinori Fujino, Hideki Isozaki, and Jun Suzuki. 2008. Multi-label text categorization with model combination based on f1-score maximization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, 50(1):344–361.

Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019a. Incorporating demographic embeddings into language understanding. *Cognitive science*, 43(1):e12701.

Justin Garten, Brendan Kennedy, Kenji Sagae, and Morteza Dehghani. 2019b. Measuring the importance of context when modeling language comprehension. *Behavior research methods*, 51(2):480–492.

Tiana Gaudette, Ryan Scrivens, Garth Davies, and Richard Frank. 2021. Upvoting extremism: Collective identity formation and the extreme right on reddit. *New Media & Society*, 23(12):3491–3508.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Joshua D Greene. 2014. Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, 124(4):695–726.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 281–291.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Brenna Helm, Ryan Scrivens, Thomas J Holt, Steve Chermak, and Richard Frank. 2022. Examining incel subculture on reddit. *Journal of Crime and Justice*, pages 1–19.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Frederic R Hopp, Jacob T Fisher, Devin Cornell, Richard Huskey, and René Weber. 2021. The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Albert Jiang, Alexandre Abou Chahine, Alexandre Sablayrolles, Alexis Tacnet, Alodie Boissonnet, Alok Kothari, Amélie Héliou, Andy Lo, Anna Peronnin, Antoine Meunier, Antoine Roux, Antonin Faure, Aritra Paul, Arthur Darcet, Arthur Mensch, Audrey Herblin-Stoop, Augustin Garreau, Austin Birky, Avinash Sooriyarachchi, Baptiste Rozière, Barry Conklin, Bastien Bouillon, Blanche Savary de Beauregard, Carole Rambaud, Caroline Feldman, Charles de Freminville, Charline Mauro, Chih-Kuan Yeh, Chris Bamford, Clement Auguy, Corentin Heintz, Cyriaque Dubois, Devendra Singh Chaplot, Diego Las Casas, Diogo Costa, Eléonore Arcelin, Emma Bou Hanna, Etienne Metzger, Fanny Olivier Autran, Francois Lesage, Garance Gourdel, Gaspard Blanchet, Gaspard Donada Vidal, Gianna Maria Lengyel, Guillaume Bour, Guillaume Lample, Gustave Denis, Harizo Rajaona, Himanshu Jaju, Ian Mack, Ian Mathew, Jean-Malo Delignon, Jeremy Facchetti, Jessica Chudnovsky, Joachim Studnia, Justus Murke, Kartik Khandelwal, Kenneth Chiu, Kevin Riera, Leonard Blier, Leonard Suslian, Leonardo Deschaseaux, Louis Martin, Louis Ternon, Lucile Saulnier, Lélio Renard Lavaud, Sophia Yang, Margaret Jennings, Marie Pellat, Marie Torelli, Marjorie Janiewicz, Mathis Felardos, Maxime Darrin, Michael Hoff, Mickaël Seznec, Misha Jessel Kenyon, Nayef Derwiche, Nicolas Carmont Zaragoza, Nicolas Faurie, Nicolas Moreau, Nicolas Schuhl, Nikhil Raghuraman, Niklas Muhs, Olivier de Garrigues, Patricia Rozé, Patricia Wang, Patrick von Platen, Paul Jacob, Pauline Buche, Pavankumar Reddy Muddireddy, Perry Savas, Pierre Stock, Pravesh Agrawal, Renaud de Peretti, Romain Sauvestre, Romain Sinthe, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Soham Ghosh, Sylvain Regnier, Szymon Antoniak, Teven Le Scao, Theophile Gervet, Thibault Schueller, Thibaut Lavril, Thomas Wang, Timothée Lacroix, Valeriia Nemychnikova, Wendy Shang, William El Sayed, and William Marshall. 2024. Ministral-8b-instruct-2410. https://huggingface.co/mistralai/Ministral-8B-Instruct-2410. Mistral AI model release; accessed October 2025.

Farzan Karimi-Malekabadi, Suhaib Abdurahman, Zhivar Sourati, Jackson Trager, and Morteza Dehghani. 2026. Theory trace card: Theory-driven socio-cognitive evaluation of llms. *arXiv preprint arXiv:2601.01878*.

Ian Keen. 2015. The language of morality. *The Australian Journal of Anthropology*, 26(3):332–348.

Brendan Kennedy, Ashwini Ashokkumar, Ryan L Boyd, and Morteza Dehghani. 2022a. Text analysis for psychology: Methods, principles, and practices. *Handbook of Language Analysis in Psychology*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. Moral concerns are differentially observable in language. *Cognition*, 212:104696.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022b. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Brendan Kennedy, Preni Golazizian, Jackson Trager, Mohammad Atari, Joe Hoover, Aida Mostafazadeh Davani, and Morteza Dehghani. 2023. The (moral) language of hate. *PNAS nexus*, 2(7):pgad210.

Blair Kidwell, Adam Farmer, and David M Hardesty. 2013. Getting liberals and conservatives to go green: Political ideology and congruent appeals. *Journal of Consumer Research*, 40(2):350–367.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pages 933–943.

Alex Gwo Jen Lan and Ivandré Paraboni. 2022. Text-and author-dependent moral foundations classification. *New Review of Hypermedia and Multimedia*, pages 1–21.

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024. Emona: Event-level moral opinions in news articles. *arXiv preprint arXiv:2404.01715*.

Leon Li and Michael Tomasello. 2021. On the moral functions of language. *Social Cognition*, 39(1):99–116.

Eneldo Loza Mencía, Moritz Kulessa, Simon Bohlender, and Johannes Fürnkranz. 2023. Tree-based dynamic classifier chains. *Machine Learning*, 112(11):4129–4165.

Morgan Marietta. 2008. From my cold, dead hands: Democratic consequences of sacred rhetoric. *The Journal of Politics*, 70(3):767–779.

Salma Moaz. 2020. *Using Moral Foundations Framing to Influence Partisan Attitudes Toward Latino Immigrants*. Ph.D. thesis, Loyola University Chicago.

Negar Mokhberian, Frederic R Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 147–154. IEEE.

Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE.

Ece Çiğdem Mutlu, Toktam Oghaz, Ege Tütüncüler, Jasser Jasser, and Ivan Garibay. 2020. Quantifying latent moral foundations in twitter narratives: The case of the syrian white helmets misinformation. *arXiv preprint arXiv:2004.13142*.

Tuan Dung Nguyen, Ziyu Chen, Nicholas George Carroll, Alasdair Tran, Colin Klein, and Lexing Xie. 2024. Measuring moral dimensions in social media with mformer. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 1134–1147.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Kyriaki Kalimeri, and Charalampos Saitis. 2024. Automatic detection of moral values in music lyrics. *arXiv preprint arXiv:2407.18787*.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004.

Benjamin Grant Purzycki, Anne C Pisor, Coren Apicella, Quentin Atkinson, Emma Cohen, Joseph Henrich, Richard McElreath, Rita A McNamara, Ara Norenzayan, Aiyana K Willard, et al. 2018. The cognitive and cultural foundations of moral behavior. *Evolution and Human Behavior*, 39(5):490–501.

Muhammad Reza Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2024. Moral association graph: A cognitive model for moral inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Roose. 2021. The gamestop reckoning was a long time coming. Posted 28-January-2021, The New York Times, www.nytimes.com/2021/01/28/technology/gamestop-stock.html.

Alexander Ruch, Ari Decter-Frain, and Raghav Batra. 2022. Millions of co-purchases and reviews reveal the spread of polarization and lifestyle politics across online markets. *arXiv preprint arXiv:2201.06556*.

Eyal Sagi and Morteza Dehghani. 2014. Measuring moral rhetoric in text. *Social science computer review*, 32(2):132–144.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Johanna Simfors and Rasmus Rudling. 2020. How does the degree of anonymity affect our morals?: A study examining behavioural changes in online communication.

Maciej Skorski and Alina Landowska. 2025. Beyond human judgment: A bayesian evaluation of llms' moral values understanding. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 17–26.

Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259–263.

Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 proceedings*.

Jackson Trager and Mohammad Atari. 2025. The immorality of too much money. *PNAS nexus*, 4(6):pgaf158.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K Ngueajio, Ameeta Agrawal, Flor Plaza-del Arco, Yalda Daryanai, and Farzan Karimi-Malekabadi. 2025. Mftcxplain: A multilingual benchmark dataset for evaluating the moral reasoning of llms through hate speech multi-hop explanation. *arXiv preprint arXiv:2506.19073*.

Anthony Henry Triggs, Kristian Møller, and Christina Neumayer. 2021. Context collapse and anonymity among queer reddit users. *new media & society*, 23(1):5–21.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Upvoted Staff. 2021. Reddit's 2020 year in review. Posted 8-December-2020, Upvoted: The Official Reddit Blog, www.redditinc.com/blog/reddits-2020-year-in-review/.

Michiel Van Der Meer, Piek Vossen, Catholijn M Jonker, and Pradeep K Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? *arXiv preprint arXiv:2310.15757*.

Pekka Väyrynen. 2016. Thick ethical concepts. *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*.

Jan G Voelkel, Mashail Malik, Chrystal Redekopp, and Robb Willer. 2022. Changing americans' attitudes about immigration: Using moral framing to bolster factual arguments. *The ANNALS of the American Academy of Political and Social Science*, 700(1):73–85.

Rong Wang and Wenlin Liu. 2021. Different pathways to identify moral framing from media content: A response to hopp and weber. *Communication Monographs*, 88(3):380–388.

Christopher Wolsko. 2017. Expanding the range of environmental values: Political orientation, moral foundations, and the common ingroup. *Journal of Environmental Psychology*, 51:284–294.

Yuexin Wu and Xiaolei Huang. 2022. Unsupervised reinforcement adaptation for class-imbalanced text classification. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 311–322, Seattle, Washington. Association for Computational Linguistics.

Hiroaki Yamane, Yusuke Mori, and Tatsuya Harada. 2021. Humor meets morality: Joke generation based on moral judgement. *Information Processing & Management*, 58(3):102520.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lorenzo Zangari, Candida M Greco, Davide Picca, and Andrea Tagarelli. 2025a. Me2-bert: Are events and emotions what you need for moral foundation prediction? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9516–9532.

Lorenzo Zangari, Candida Maria Greco, Davide Picca, and Andrea Tagarelli. 2025b. A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & SOCIETY*, pages 1–26.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Appendix: Appendix includes the moral foundations coding guide, additional model results, prompts, and Theory Trace Card (Karimi-Malekabadi et al., 2026). For additional material, see the huggingface[3].

A. Moral Foundations Coding Guide

Moral expressions in text serve as informationally rich indicators of individuals' moral values. Whether individuals are signaling their moral beliefs or concerns, framing particular issues or events in moral terms, or expressing a moral emotion, moral expressions are a domain of human language which can inform as to the nature of morality (Atari et al., 2022). Here, we describe a taxonomy and set of instructions for annotating moral content in natural language, based on Moral Foundations Theory. This taxonomy can be used for the annotation of individual Tweets, Facebook posts, other social media, transcribed speech, and other textual media.

In this coding guide, we describe the theoretical framework that we rely on to operationalize moral values, Moral Foundations Theory (MFT; Haidt and Joseph, 2004; Graham et al., 2013), describe how

[3] https://huggingface.co/datasets/USC-MOLA-Lab/MFRC

moral expressions are annotated, and provide detailed examples and procedures for the process of annotation. [4] We follow recent work which expands the original five moral foundations (Care, Fairness, Loyalty, Authority, and Purity) by partitioning Fairness into "Proportionality" and "Equality" (Atari et al., 2022).

A.1. Background: Morality, language analysis, and handling ambiguity

A.1.1. Moral Foundations Theory

Our theoretical framework for annotating morality in language is Moral Foundations Theory (MFT; Haidt and Joseph, 2004; Graham et al., 2013), a pluralistic, psychological model of moral values. MFT was developed in order to fill the need of a systematic theory of morality, explaining its evolutionary origins, developmental aspects, and cultural variations. MFT can be viewed as an attempt to specify the psychological mechanisms which allow for intuitive bases of moral judgments as well as moral reasoning. Care, Fairness, Loyalty, Authority, and Purity, according to the original conceptualization of MFT, are five "foundations" that are conceptualized to have contributed to solving adaptive problems over humans' evolutionary past, and are ubiquitous in current human populations (Graham et al., 2013).

Each of the five foundations in MFT is conceptualized as having solved different adaptive problems in humans' evolutionary past (Haidt, 2012). The Care foundation accounts for our nurturing of the young and caring for the infirm. The Fairness foundation accounts for the development of human cooperation, justice, and reciprocity. Loyalty is concerned with coalition-building with ingroup members, Authority is concerned with respecting high-status individuals in social hierarchies, and Purity is about physical cleanliness and spiritual sacredness of objects, humans, and groups.

One strength of MFT, as formulated by Graham et al., is its openness to new foundations, with the idea that plurality is the most important concept for understanding human morality, and that the specific set of five foundations originally proposed in MFT are just one proposed set of foundations. Recently, Atari et al. (2022) proposed breaking Fairness into two more narrowly-defined foundations, "Equality" and "Proportionality". Equality describes people's concern with similar outcomes or status (e.g., a violation of Equality is systematic racial inequality, the state of individuals of different races having different access to resources and opportunities). Proportionality describes people's desire for balance between actions and responses. Proportionality concerns are typically centered around the ideas of meritocracy and deservingness. For example, cheaters should be punished, hard workers should be rewarded, and slackers should be excluded relative to the extent of their contribution.

A.1.2. Moral Foundations in Language

While explicitly moral language is not common in everyday interactions (Atari et al., 2022), moral values, whether implicit or explicit, do play an important role in social functioning (Li and Tomasello, 2021). They influence our judgments and behaviors (Ellemers et al., 2019; Greene, 2014; Haidt, 2012) and help coordinate complex large-scale cooperation (DeScioli and Kurzban, 2013; Enke, 2019; Dehghani et al., 2016; Purzycki et al., 2018).

When people express their moral attitudes, emotions, and concerns about people, actions, events, concepts, and ideas, they employ diverse rhetorical strategies (Keen, 2015). Often, these strategies rely on words that are explicitly normative, such as "right," "wrong," "good," or "bad" (which we call "thin morality"; see Atari et al., 2022); however, in many cases, people communicate a moral attitude by communicating the relevance of a moral domain (e.g., Care). For example, "I can't believe that happened. It's so harmful!", "People should be compassionate," or "This decision hurts so many people!" position the discussed entity or topic as either aligned or misaligned with "good" morality, by assuming the virtue or desirability of care centered actions, people, or things.

The six moral foundations (i.e., Care, Equality, Proportionality, Loyalty, Authority, and Purity) have a natural mapping to language, which can be instantiated by identifying words which are used by speakers to communicate their attitudes with respect to each moral foundation. In the above examples, speakers' attentiveness to the Care foundation is apparent from their usage of the words "harmful," "compassionate," and "hurt." Moral foundations in language were first studied in this way by Graham et al. (2009), which provided the Moral Foundations Dictionary (MFD) (for a detailed discussion of language analysis in moral psychology, see Atari and Dehghani, 2021). The MFD considers each moral foundation, including the "vice" and "virtue" poles of each foundation, as a collection of related words, the usage of which indicates a concern with the given foundation. In early work, Graham et al. (2009) used the MFD to measure differences in moral values sentiment between conservative and liberal sermons. More recently, researchers have shown that moral value annotation based on the MFT taxonomy can fruitfully be applied to a range of applications and domains (Dehghani et al., 2016;

[4] Please note this guide is based on the the original Moral Foundations Coding Guide which can be found in the Appendix of (Hoover et al., 2020).

Sagi and Dehghani, 2014). Additionally, extensions and improvements on the initial MFD have been performed by Hopp et al. (2021) (extended MFD) and Frimer et al. (2019) (MFD 2.0).

While the above findings, facilitated by the MFD, have shed light on the nature of morality in "the wild" by analyzing observational text data, there are reasons to question the use of the dictionary approach for measuring moral phenomena in text. There is an implicit assumption that the frequency of certain explicit moral words indicates an underlying concern with the corresponding moral domain. For example, MFD findings articulate implicitly that using moral words implies a higher concern with morality at the individual level. Recently, however, Kennedy et al. (2021) tested this assumption using responses to the Moral Foundations Questionnaire (MFQ; Graham et al., 2011) and participants' Facebook status updates. Among other findings, this work established that the existence and size of the relationship between explicitly moral language (i.e., the MFD) and individuals' moral concerns was inconsistent across foundations. Specifically, Care, Fairness, and Purity language had a positive correlation with the corresponding moral domain in language (e.g., high Care concerns implied the usage of Care words), while Authority and Loyalty did not. Moreover, other techniques such as topic modeling, which represent all words (and not just explicitly moral ones) predicted significantly more variance in individual-level moral concerns than did methods based on explicitly moral language.

As far as the present coding guide is concerned, these findings imply that the domain of language which is related to moral concerns is far wider than purely explicit moral language. In terms of the annotation of moral phenomena in text, we will take the view that moral concerns can commonly be communicated without the presence of explicitly moral words, e.g., "No matter what, it's [my team] forever! 10-0 or 0-10, nothing changes for me" is an expression of loyalty to a sports team without explicitly using words like "loyalty."

A.1.3. Thin Morality

Not all moral language falls within the scope of the six domains indicated by MFT. In fact, philosophers have denoted two types of morally evaluative language:

Evaluative terms and concepts are often divided into "thin" and "thick". We don't evaluate actions and persons merely as good or bad, or right or wrong, but also as kind, courageous, tactful, selfish, boorish, and cruel. The latter are examples of thick concepts . . . [which] stand in contrast to those we typically express when we use thin terms such as *right*, *bad*, *permissable*, and *ought* (emphasis in original; Väyrynen, 2016)

In our above section on moral concerns in language, we have been discussing thick morality. In our annotation, we will attempt to comprehensively cover all types of moral language by also annotating thin morality.

A.1.4. Annotator Uncertainty and Authorial Inferences

There is unavoidable ambiguity that affects text annotation, which is key to understand when conducting annotation-based studies of moral concerns in language. Our approach to ambiguity in this guide is to use background context to inform annotations, but also to report the level of uncertainty for a given annotation.

The major source of ambiguity is caused by the difficulty of inferring the moral content intended by an author. For example, a social media message might simply state that the author thinks "Everything that is going on with abortion these days is reprehensible." In this case, it is clear that this is likely a morally relevant statement, but it is less clear what foundation this statement is relevant to. If we knew that the author was concerned with civil rights, we might assume that the author is concerned about violations of women's reproductive rights (i.e., an instance of Equality). In contrast, if we knew that the author was a conservative Christian, we might assume that the author was expressing an anti-abortion sentiment, perhaps associated with Purity.

These ambiguities present considerable challenges for human annotators who must strike an acceptable balance between exploiting often weak signals of moral sentiment while also avoiding unfounded speculation about author's intent. In this guide, we recommend that annotators tend to focus on objective sources of confidence for resolving ambiguities of intent. However, since so much in the domain of moral language is not objective (i.e., relying on author assumptions), in this guide we propose the usage of a "Confidence" label, which is designed to allow annotators to assign a label based on an inference about intent, but to also indicate a lower confidence in the label.

A.2. Instructions for Annotators: Annotating moral concerns in language

Annotating moral concerns in language involves determining whether a given text's author is communicating a moral attitude, emotion, judgment, or moral issues toward particular persons, groups, questions or problems, or event. In this section, we provide instructions for annotators for the identification of moral concerns in text. We emphasize the

six domains of moral language based on MFT; detail the *target* and *vice/virtue* components of moral concerns in text; detail instructions for annotators to assign a confidence rating to each annotation; and describe the particular approaches annotators ought to use for different language types (e.g., social media versus transcribed speech).

A.2.1. Annotation Task

For moral concern annotation in text, annotators should complete, in order, the four subtasks outlined in Figure 2. Below, we will explain each subtask.

Annotating Moral Domains We first annotate text by categorizing text into non-mutually exclusive "domains" of moral concerns, which are the six foundations of MFT.

In Table 8, we list the six foundations, giving their name, a definition, and an example item from the recently developed MFQ-2 (Atari et al., 2022). Items were presented to participants with the prompt, "Please indicate how well each statement describes you or your opinions."

Even with clarity as to the conceptual domains described by each of the six foundations, it is not straight-forward to map these categories to language. Below, we describe three distinctions, or components, of a voiced moral concern: (1) the explicit/implicit distinction, (2) the concrete versus abstract target distinction, and (3) the vice/virtue (positive/negative) distinction.

Explicit and implicit expressions of moral concerns Each of these six concerns can be invoked in language in ways that can be either implicit or explicit. Explicit invocations will use words that map clearly to the domain in question; a sample of such words are given in Table 9.

The presence of these words in a document/sentence/utterance hints at the presence of a moral expression, but does not necessarily confirm it. An actual invocation of a given moral domain will use these words *in a particular way*. Differentiating what we might call "moral" uses of these words from "non-moral" uses is similar to the challenge of "word sense disambiguation" in Natural Language Processing. Word sense disambiguation acknowledges that words can have multiple uses or "senses" depending on the context in which they were used. For example, the word "fairly" can be used in a Care sense, e.g., "Humanity is we treat every person *fairly*, even when we're threatened," but can also be used in a non-Care sense, e.g., "There's *fairly* universal protocol on how to treat anyone who makes dumb decisions".

Other times, moral concerns can be invoked implicitly (i.e., without using explicitly moral language). This category of moral expression is harder to specify a priori due to its myriad forms; here, annotators should rely on their understanding of the *concepts* underlying each moral domain, as denoted in the flowchart in Figure 2 and in Table 8.

For example, it is possible to express a concern about Equality without explicitly naming the Equality domain: "AT_USER it's a shame Skin color and beliefs fuel hatred" communicates a concern regarding the fair treatment of people based on skin color or beliefs. Similarly, it is possible to invoke the Loyalty domain without using explicit loyalty words: "Rep voter [suppression] efforts in Florida a disgrace to Americans my Dad who fought in WWI for freedom & democracy" describes the virtue of loyalty to American veterans without using the word "loyalty" (etc.). The takeaway from these examples is that moral domains can be invoked by understanding the meaning of the text in question, and is not limited to the presence or non-presence of explicitly moral words.

In summary, there are two types of language which signal the presence of a moral concern:

1. Explicitly moral words used in a moral way
2. Any language used to express the speaker's moral concerns, attitudes, emotions, behavior, or beliefs about some persons, actions, objects, events, ideas, etc.

Next, we detail two sets of distinctions for moral expressions that can help annotators to identify different types of moral language.

Concrete versus abstract targets of moral expressions People often use moral values when they are expressing a judgment about someone or something — i.e., the *object*. The object of a moral judgment can be either *someone* (e.g., a person or a social group) or *something* (e.g., a behavior, an event, an abstract concept, or even a physical object). The object of a moral judgment can be *concrete* (e.g., judging a specific behavior or person) or *abstract* (e.g., judging a general value or opinion). When annotating moral foundations in natural language, we are not, per se, interested in the object of a moral judgments. However, identifying the object of a moral judgment can sometimes help clarify whether the moral judgment in question is related to one of the moral foundations.

Vice and virtue expressions The types of moral judgments individuals make about people or things can be either *positive* or *negative*. For example, a person might praise someone for engaging in moral behavior or condemn someone for engaging in immoral behavior. That is, a moral judgment entails

Figure 2: Order of operations for annotating moral values in text: (1) Determine whether one or more moral domains are present (i.e., *thick* morality; (2) Determine whether the text contains *thin* morality; (3) indicate a non-moral text if (1) and (2) categories are not present; and (4) assess confidence in label.

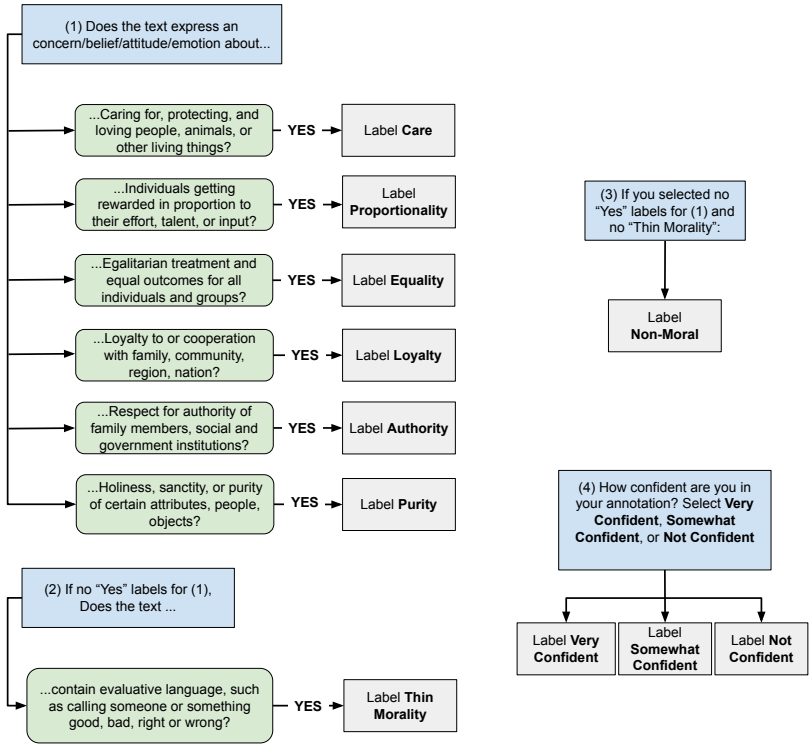


Table 8: Six moral foundations, which map to six domains of moral language

Foundation	Description	Example Item
Care	Intuitions about avoiding emotional and physical damage to another individual. It underlies virtues of kindness, gentleness, and nurturing.	*I believe that compassion for those who are suffering is one of the most crucial virtues.*
Equality	Intuitions about egalitarian treatment and equal outcome for all individuals and groups. It underlies virtues of social justice and equality.	*The world would be a better place if everyone made the same amount of money*
Proportionality	Intuitions about individuals getting rewarded in proportion to their merit (e.g., effort, talent, or input). It underlies virtues of meritocracy, productiveness, and deservingness.	*The effort a worker puts into a job ought to be reflected in the size of a raise they receive*
Loyalty	Intuitions about cooperating with ingroups and competing with outgroups. It underlies virtues of patriotism and self-sacrifice for the group.	*I believe the strength of a sports team comes from the loyalty of its members to each other*
Authority	Intuitions about deference toward legitimate authorities and high-status individuals. It underlies virtues of leadership and respect for tradition.	*I think obedience to parents is an important virtue*
Purity	Intuitions about avoiding bodily and spiritual contamination and degradation. It underlies virtues of sanctity, nobility, and cleanliness.	*It underlies the widespread idea that the body is a temple that can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions)*

a positive or negative evaluation of the object of the moral judgment. In broad terms, an expression of virtue communicates that "good should happen" while an expression of vice communicates that "bad should not happen"—what is "good" and "bad" depends, of course, on which moral concern is being evoked.

An evaluation is positive when it calls for moral actions, praises people for moral behavior, or lauds a moral value or opinion. An evaluation is negative when it decries immoral actions, criticizes people for immoral behavior, or condemns an immoral value or opinion. While we are not, per se, interested in distinguishing between positive and negative moral judgements, this distinction can sometimes help clarify what moral foundation is being invoked.

A.2.2. Annotating Thin Morality

For our purposes, thin morality is a moral judgment or concern which is voiced without clearly referring to one of the six moral domains. For example, the Tweet "Why does [he] reply to profane and disrespectful tweets from rude constituents **he's a good guy**" (emphasis added) makes a statement about the goodness of an individual but does not describe the individual's goodness on account of a particular moral domain.

Table 9: Example words that illustrate each moral language domain including both virtue and vice.

Foundation	Virtue	Vice
Care	compassion kindness	cruel exploit
Equality	equal fairly	discriminate injustice
Proportionality	proportional deserve	disproportional favoritism
Loyalty	collective family	betray disloyal
Authority	duty tradition	dissident rioter
Purity	sacred chast	sin disgust

We note that Thin morality *is* in fact mutually exclusive with thick morality, and thus if a document is, for example, annotated with Loyalty, it cannot also be labeled as Thin morality. Unlike our definition of thick morality (i.e., the presence of one of the six MFT domains), the presence of thin morality is most often marked by the presence of words (e.g., right, wrong, better, worse, good, bad)

A.2.3. Annotator Confidence

After completing annotation of moral domains and Thin morality, and regardless of whether or not Non-Moral was selected, annotators should select one of three confidence labels: Very Confident, Somewhat Confident, and Not Confident. These are fully defined in Table 10.

A.3. Language Domains and Appropriate Annotation Strategies

The goal of this coding guide is to be domain-agnostic. That is, rather than a guide for specifically Twitter (e.g., Hoover et al., 2020) or Facebook data (e.g., Atari et al., 2022), we aim to provide a framework that is flexible enough to guide annotations for any textual domain, including social-media posts, comments on online posts or articles, published text, literary text, historical pieces, and transcribed speech.

Here, we note particular strategies annotators should take for each language domain.

Social media (Twitter, Facebook, Instagram, etc.) Social media text is typically short form, containing incomplete sentences, abbreviations, hashtags and 'at'-mentions, and hyperlinks. For our annotation, we ask that annotators ignore at-mentions and hyperlinks (including media) and to infer as much as possible from the available context. In some cases, this might include references to current events; for example, the Tweet "No Social Security number should mean no claim to any benefits or credits. #takeastand" references the process in the United States whereby a social security number grants access to certain government-provided services. In other cases, abbreviations (e.g., BLM) can be used in ways that add meaningful information. For example, the sentence "We have endured too much!" might be labeled as Loyalty; however, with the inclusion of the BLM hashtag — "We have endured too much! #BLM", this might additionally be considered Care or Equality, given that the Black Lives Matter movement is focused on harms and systematic inequalities directed toward Black persons. Annotators are asked to look up unfamiliar abbreviations that occur frequently, or words that seem to have a unique use in a particular online platform or group; however, if the abbreviation is not obvious or frequent, it can be ignored. Hashtags, particularly if they are themselves abbreviations, can be used to resolve context. However, they should not be used as the *sole* reason for labeling a document as a given moral domain. For example, the text "Having some fries with my drink #Equality" contains a relevant hashtag that does not help to resolve ambiguity, and thus should be ignored. Lastly, hashtags that are used fluidly in a sentence (e.g., "#Dreamers play a vital role in our communities") can be treated as normal words.

In addition to general social media considerations, each platform has specific components that inform annotation strategies. For Twitter, posts are embedded in a networked context with sharing ("Retweeting") and conversational components. Currently, we do not support the ability to view conversational context when annotating a Tweet, though this might change in the future. Retweets, marked with "RT" at the beginning of the Tweet, and Quote Tweets, marked by a preceding remark followed by the main Tweet in quotes, should be viewed as endorsing the message contained in the original tweet, and annotated accordingly. Lastly, most of the considerations that apply to Twitter apply to other short-form social media, such as Instagram.

For Facebook, posts can be longer (i.e., multi-sentence or multi-paragraph), requiring more time than Tweets and other short form text. However, Facebook posts are typically less ambiguous than Tweets, as they contain more context. The moral concerns voiced in a Facebook posts will likely be contained in one or two select sentences. Additionally, it is more likely for a Facebook post to contain multiple moral concerns than it shorter media like Twitter.

Online comments (Reddit) The distinguishing characteristic of online comment language is its referencing of original posts. For example, a comment

Table 10: Three possible confidence scores to assign a given annotation, with explanations.

	Example Cases
Very Confident	Clearly no moral expression in the text
	A moral domain is clearly in the text, and it is clear that there are no others
	Multiple domains are clearly in the text, and all are clearly present
Somewhat Confident	No moral expression in the text, but possibility the speaker could be implying a moral concern
	A moral expression in the text, but possibility that the speaker is using sarcasm or similar
	One or more moral domains are clearly present, but at least one is vague or uncertain
Not Confident	No moral label, but with more context it might be possible to establish that the author did intend to communicate a moral concern
	One or more moral labels, but with more context it might be possible to establish that the author did not intend to communicate anything moral
	Two or more moral domains are equally present, but there is no way to resolve either confidently

on a Reddit post in the forum "Am I the Asshole" will be making judgments or comments about a post in which the original author explained a personal story, asking anonymous ethical judges to pronounce judgment on the individuals in the story. For our purposes, we will not have annotators read original posts in large part due to the length of original posts. Instead of relying on this context, we will ask annotators to label comments using only the language contained in the given comment. This has limitations with regard to resolving ambiguities, and thus annotators should take care to report annotator confidence when additional context would be needed to label an ambiguous comment.

Transcribed speech Spoken language is altogether different from written language, due to the difference between spontaneous conversation and the premeditated nature of written text. Transcriptions of speech capture all the artifacts of speech, including "er" and "ah" sounds, short sentences, and incomplete sentences. Also, the types of moral concern voiced in spoken language tends to be more concrete than abstract. The following examples illustrate the types of language in (transcribed) spoken text: "That's good. I'm happy you're taking care of you [*sic*] mom" (Care); "I will never leave your side" (Loyalty); "I'm your dad! You need to respect me!" (Authority); "John is a good man" (Thin Morality).

Published text/articles Lastly, published text or articles, whether sampled at the sentence, paragraph, or document level, require annotators to read carefully and to consider as much external context as possible. Similarly to Facebook posts, annotators should attempt to identify sentences or sequences of sentences that contain a give moral concern. Additionally, given the nature of published text and articles, more is known about the speaker, or at least the speaker's objectives, which might be to persuade readers about a certain point or to describe a story or event. This context can be used to resolve ambiguities in text.

A.4. Examples

A.4.1. *Examples of each Label from the MFTC*

See Table 11.

A.4.2. Examples of Thin Morality

See Table 12.

A.4.3. Examples of Equality and Proportionality

Special attention is given in this coding guide to the annotation of Equality and Proportionality. Table 13, we give illustrative examples of Tweets from the MFTC which were previously annotated as Fairness(Vice/Virtue), but in the present coding guide are either Equality or Proportionality.

B. Personalization

In this section, we focus on disaggregated performance and present results on each annotator, or the personalization capabilities that the dataset affords. Results are aggregated across labels, and therefore multi-label metrics are presented: Sample Semantic F1 (Chochlakis et al., 2025b; the same similarities as in the original paper are used for evaluation), Jaccard score, Macro F1 and Micro F1 (Fujino et al., 2008; Loza Mencía et al., 2023). In this section, we use Llama3-8B and Llama3-70B (Dubey et al., 2024), Qwen3-30B (Yang et al., 2025), and GPT-OSS-20B (Agarwal et al., 2025).

We present results in the few-shot setting (10, 20, and 30 shots), as well as more explicit prompting that incorporates the annotation manual to precisely define terms and derivation process. The prompts are shown in Table 18. Few-shot learning attempts to personalize the predictions to the annotator's choices, whereas definitions (which are 0-shot for GPT-OSS and 1-shot otherwise) leverage

Table 11: Examples from MFTC of each label in our taxonomy

Example	Foundation	Explanation
please remember to watch for frightened lost injured pets	Care	Asks others to care for (virtue) injured pets (object)
If hurricane Sandy hurts anyone I love She will be cunt punted	Care	Threatens violence (vice) if loved ones are hurt (object/event)
I'm rooting for equality #iamAME #BlackLivesMatter #AllLivesMatter	Equality	Positive support (virtue) for equality (object)
Why is no one worried about disenfranchisement caused by lack of electricity From a NJ voter election voting rights no electric Sandy	Equality	Expressing about worry about inequality (vice) due to lack of electricity (object)
Winning fair share of Wealth & Power will be key to any lasting change	Proportionality	Calls for fair share (virtue) of wealth and power (object)
USER should have asked if Springsteen was going to stop taking advantage of farmer loopholes and pay his fair share of taxes	Proportionality	Implies negative attitude (vice) toward someone not paying their fair share (object)
Solidarity Sunday. #blacklivesmatter. #icantbreathe	Loyalty	Expresses solidarity (virtue) with the Black Lives Matter movement
@LindseyGrahamSC Be a true patriot & speak up	Loyalty	Telling someone (object) to be a true patriot (virtue)
No to illegal immigrants-they need to follow the process, obey the law @realDonaldTrump #EndDACA	Authority	Expresses that people should obey (virtue) the US president (object)
@realDonaldTrump "If you love me, obey my commandments." -Jesus John 14:15	Authority	Telling some (virtue) to obey God's commandments (object)
Glad to see a reformation going on to restore sanctity of free speech	Purity	Praising (virtue) a restoration of the sanctity of free speech (object)
It's absolutely disgusting how every retailer exploits a serious storm situation by peddling their crap	Purity	Expression of disgust (vice) regarding exploiting a storm situation to sell wares (object)

Table 12: Examples of Thin Morality as well as Non-Moral examples.

Example	Label	Notes
John is a good man.	Thin Morality	-
Yes I think that's correct.	Non-Moral	Agreeing with someone, not expressing some moral evaluation.
What he did was absolutely wrong. Unacceptable!	Thin Morality	-
Mother Theresa's goodness won her a Nobel Prize.	Thin Morality	Praising on account of goodness.
I have no idea what to say. Hmmmm...	Non-Moral	-

Table 13: Examples of changes in labeling for the new, six-foundation taxonomy

Text	Original Labels	New Labels
#AllLivesMatter is a cop out. A convenient way to dismiss oppression and inequality in this country.	Harm, Cheating	Inequality
If you choose to be a police officer, you have a responsibility to uphold justice and treat everyone equal. #AllLivesMatter	Fairness	Equality
RT @CNN: "No justice, no peace." Crowds protest the death of #FreddieGray in #Baltimore	Non-Moral	Proportionality
To hear the police union president say the officers did nothing wrong breaks my heart. I can't even use the angry emotion.	Cheating	Proportionality
RT @chescaleigh: talking about injustice shouldn't upset you. the injustice should. #BlackLivesMatter	Cheating	Propotionality
I cancelled my direct debit and I'm going to refuse to pay! This is fraud by O2	Cheating	Disproportionality

the general semantics to ground the predictions to the manual.

Thorough results, including standard deviations, are presented in Tables 14, 15, 16, 17 respectively, both for the aggregate and for each annotator. Overall we see that reasoning models perform best with the definitions, while the rest achieve better performance with few-shot learning instead. We see that reasoning models outperform the rest (irrespective of setting) in micro and macro F1, which are more sensitive to thin rationality, and instruction-tuned models better capture the morality of comments with respect to Semantic F1 and Jaccard Score. Depending on the metric, therefore, personalization models perform better or worse than models guessing predictions based on the manual, proving the personalization study inconclusive. We note that sample-based evaluations, like Sample Semantic F1 and Jaccard Score, might overestimate the performance of models given the sparsity of the labels.

Similar to previous evaluations (Chochlakis et al., 2025a), we also find that the aggregate tends to have lower performance compared to annotators when measuring performance with micro and macro F1.

C. Additional Baseline Model Results

See Tables 19 - 20.

D. Theory Trace Card

See theory trace card below.

Table 14: Sample Semantic F1 comparison across annotators.

Method	Model	Aggregate	0	1	2	3	4	5
10-shot	Qwen3 30B Instruct	$0.72_{0.015}$	$0.70_{0.010}$	$0.64_{0.011}$	$0.69_{0.007}$	$0.73_{0.011}$	$0.61_{0.007}$	$0.65_{0.019}$
	Llama3 8B Instruct	$0.67_{0.009}$	$0.65_{0.009}$	$0.58_{0.012}$	$0.67_{0.026}$	$0.72_{0.009}$	$0.61_{0.016}$	$0.68_{0.032}$
	GPT-OSS 20B	$0.62_{0.002}$	$0.57_{0.017}$	$0.61_{0.008}$	$0.60_{0.007}$	$0.63_{0.004}$	$0.59_{0.003}$	$0.72_{0.035}$
20-shot	Qwen3 30B Instruct	$0.75_{0.014}$	$0.72_{0.029}$	$0.67_{0.001}$	$0.69_{0.019}$	$0.77_{0.013}$	$0.63_{0.015}$	$0.66_{0.049}$
	Llama3 8B Instruct	$0.72_{0.006}$	$0.69_{0.017}$	$0.62_{0.006}$	$0.70_{0.015}$	$0.75_{0.009}$	$0.63_{0.007}$	$0.70_{0.020}$
	GPT-OSS 20B	$0.64_{0.012}$	$0.61_{0.018}$	$0.62_{0.002}$	$0.66_{0.002}$	$0.66_{0.013}$	$0.60_{0.009}$	$0.69_{0.006}$
30-shot	Qwen3 30B Instruct	$0.76_{0.004}$	$0.75_{0.009}$	$0.68_{0.016}$	$0.73_{0.010}$	$0.78_{0.006}$	$0.62_{0.010}$	$0.71_{0.030}$
	Llama3 8B Instruct	$0.74_{0.010}$	$0.72_{0.009}$	$0.66_{0.007}$	$0.72_{0.008}$	$0.77_{0.008}$	$0.64_{0.003}$	$0.70_{0.026}$
w/ defs	Qwen3 30B Instruct	$0.54_{0.010}$	$0.52_{0.011}$	$0.61_{0.009}$	$0.54_{0.008}$	$0.56_{0.010}$	$0.56_{0.013}$	$0.63_{0.033}$
	Llama3 8B Instruct	$0.62_{0.009}$	$0.58_{0.001}$	$0.61_{0.009}$	$0.59_{0.008}$	$0.63_{0.011}$	$0.59_{0.010}$	$0.62_{0.048}$
	Llama3 70B Instruct	$0.47_{0.011}$	$0.45_{0.013}$	$0.55_{0.014}$	$0.46_{0.006}$	$0.47_{0.011}$	$0.52_{0.006}$	$0.67_{0.013}$
	GPT-OSS 20B	$0.68_{0.011}$	$0.66_{0.005}$	$0.66_{0.005}$	$0.68_{0.008}$	$0.70_{0.008}$	$0.62_{0.013}$	$0.66_{0.023}$

Table 15: Jaccard score comparison across annotators.

Method	Model	Aggregate	0	1	2	3	4	5
10-shot	Qwen3 30B Instruct	$0.67_{0.013}$	$0.64_{0.006}$	$0.50_{0.011}$	$0.63_{0.002}$	$0.68_{0.010}$	$0.48_{0.014}$	$0.42_{0.024}$
	Llama3 8B Instruct	$0.64_{0.009}$	$0.59_{0.012}$	$0.44_{0.012}$	$0.60_{0.023}$	$0.66_{0.004}$	$0.47_{0.012}$	$0.48_{0.026}$
	GPT-OSS 20B	$0.54_{0.001}$	$0.49_{0.026}$	$0.47_{0.006}$	$0.51_{0.001}$	$0.55_{0.008}$	$0.44_{0.001}$	$0.49_{0.015}$
20-shot	Qwen3 30B Instruct	$0.69_{0.015}$	$0.66_{0.026}$	$0.55_{0.006}$	$0.64_{0.021}$	$0.72_{0.012}$	$0.50_{0.008}$	$0.47_{0.038}$
	Llama3 8B Instruct	$0.67_{0.006}$	$0.64_{0.017}$	$0.50_{0.009}$	$0.63_{0.011}$	$0.70_{0.006}$	$0.51_{0.005}$	$0.52_{0.007}$
	GPT-OSS 20B	$0.58_{0.012}$	$0.54_{0.019}$	$0.50_{0.009}$	$0.57_{0.001}$	$0.59_{0.018}$	$0.45_{0.008}$	$0.46_{0.005}$
30-shot	Qwen3 30B Instruct	$0.72_{0.001}$	$0.70_{0.006}$	$0.57_{0.015}$	$0.68_{0.010}$	$0.74_{0.009}$	$0.51_{0.004}$	$0.51_{0.037}$
	Llama3 8B Instruct	$0.69_{0.007}$	$0.67_{0.009}$	$0.55_{0.009}$	$0.66_{0.011}$	$0.72_{0.007}$	$0.53_{0.009}$	$0.52_{0.037}$
w/ defs	Qwen3 30B Instruct	$0.46_{0.010}$	$0.44_{0.011}$	$0.44_{0.010}$	$0.44_{0.006}$	$0.48_{0.011}$	$0.39_{0.011}$	$0.42_{0.028}$
	Llama3 8B Instruct	$0.54_{0.007}$	$0.50_{0.002}$	$0.45_{0.008}$	$0.50_{0.005}$	$0.55_{0.011}$	$0.45_{0.010}$	$0.41_{0.050}$
	Llama3 70B Instruct	$0.40_{0.012}$	$0.37_{0.012}$	$0.39_{0.009}$	$0.36_{0.006}$	$0.40_{0.010}$	$0.34_{0.008}$	$0.40_{0.005}$
	GPT-OSS 20B	$0.63_{0.012}$	$0.60_{0.008}$	$0.57_{0.006}$	$0.61_{0.010}$	$0.65_{0.005}$	$0.49_{0.009}$	$0.50_{0.025}$

Table 16: Macro F1 comparison across annotators.

Method	Model	Aggregate	0	1	2	3	4	5
10-shot	Qwen3 30B Instruct	$0.27_{0.019}$	$0.21_{0.013}$	$0.31_{0.004}$	$0.27_{0.027}$	$0.26_{0.017}$	$0.31_{0.028}$	$0.26_{0.024}$
	Llama3 8B Instruct	$0.25_{0.016}$	$0.19_{0.019}$	$0.25_{0.008}$	$0.21_{0.027}$	$0.21_{0.014}$	$0.32_{0.018}$	$0.32_{0.038}$
	GPT-OSS 20B	$0.22_{0.002}$	$0.20_{0.019}$	$0.31_{0.002}$	$0.21_{0.024}$	$0.20_{0.016}$	$0.32_{0.004}$	$0.29_{0.039}$
20-shot	Qwen3 30B Instruct	$0.28_{0.015}$	$0.21_{0.018}$	$0.34_{0.005}$	$0.26_{0.031}$	$0.29_{0.008}$	$0.33_{0.008}$	$0.27_{0.045}$
	Llama3 8B Instruct	$0.25_{0.022}$	$0.20_{0.014}$	$0.26_{0.017}$	$0.17_{0.012}$	$0.26_{0.012}$	$0.33_{0.007}$	$0.36_{0.027}$
	GPT-OSS 20B	$0.25_{0.017}$	$0.23_{0.015}$	$0.32_{0.029}$	$0.25_{0.005}$	$0.24_{0.016}$	$0.32_{0.015}$	$0.31_{0.007}$
30-shot	Qwen3 30B Instruct	$0.30_{0.023}$	$0.25_{0.007}$	$0.33_{0.014}$	$0.29_{0.002}$	$0.27_{0.026}$	$0.32_{0.008}$	$0.26_{0.014}$
	Llama3 8B Instruct	$0.25_{0.012}$	$0.22_{0.013}$	$0.29_{0.012}$	$0.22_{0.007}$	$0.23_{0.033}$	$0.35_{0.019}$	$0.35_{0.061}$
w/ defs	Qwen3 30B Instruct	$0.25_{0.006}$	$0.23_{0.009}$	$0.31_{0.010}$	$0.25_{0.006}$	$0.25_{0.005}$	$0.32_{0.009}$	$0.30_{0.018}$
	Llama3 8B Instruct	$0.23_{0.005}$	$0.20_{0.005}$	$0.29_{0.009}$	$0.22_{0.005}$	$0.20_{0.012}$	$0.30_{0.006}$	$0.27_{0.034}$
	Llama3 70B Instruct	$0.26_{0.011}$	$0.20_{0.010}$	$0.31_{0.010}$	$0.25_{0.013}$	$0.23_{0.011}$	$0.32_{0.006}$	$0.36_{0.012}$
	GPT-OSS 20B	$0.35_{0.027}$	$0.32_{0.018}$	$0.39_{0.021}$	$0.31_{0.019}$	$0.30_{0.011}$	$0.28_{0.010}$	$0.35_{0.053}$

Table 17: Micro F1 comparison across annotators.

Method	Model	Aggregate	0	1	2	3	4	5
10-shot	Qwen3 30B Instruct	$0.32_{0.019}$	$0.26_{0.013}$	$0.35_{0.011}$	$0.32_{0.021}$	$0.30_{0.015}$	$0.36_{0.028}$	$0.31_{0.024}$
	Llama3 8B Instruct	$0.29_{0.012}$	$0.24_{0.011}$	$0.29_{0.009}$	$0.24_{0.016}$	$0.25_{0.008}$	$0.35_{0.018}$	$0.39_{0.048}$
	GPT-OSS 20B	$0.24_{0.004}$	$0.24_{0.020}$	$0.35_{0.001}$	$0.22_{0.021}$	$0.22_{0.018}$	$0.34_{0.002}$	$0.31_{0.026}$
20-shot	Qwen3 30B Instruct	$0.34_{0.020}$	$0.28_{0.027}$	$0.37_{0.002}$	$0.32_{0.029}$	$0.35_{0.013}$	$0.38_{0.004}$	$0.33_{0.040}$
	Llama3 8B Instruct	$0.31_{0.014}$	$0.26_{0.013}$	$0.31_{0.013}$	$0.24_{0.015}$	$0.29_{0.012}$	$0.38_{0.010}$	$0.41_{0.020}$
	GPT-OSS 20B	$0.29_{0.018}$	$0.27_{0.017}$	$0.38_{0.013}$	$0.28_{0.003}$	$0.27_{0.013}$	$0.34_{0.012}$	$0.31_{0.009}$
30-shot	Qwen3 30B Instruct	$0.36_{0.009}$	$0.31_{0.016}$	$0.37_{0.009}$	$0.36_{0.016}$	$0.34_{0.013}$	$0.37_{0.006}$	$0.34_{0.017}$
	Llama3 8B Instruct	$0.31_{0.009}$	$0.28_{0.015}$	$0.36_{0.016}$	$0.28_{0.015}$	$0.29_{0.019}$	$0.39_{0.015}$	$0.42_{0.047}$
w/ defs	Qwen3 30B Instruct	$0.25_{0.006}$	$0.24_{0.009}$	$0.32_{0.009}$	$0.25_{0.007}$	$0.23_{0.005}$	$0.33_{0.010}$	$0.35_{0.021}$
	Llama3 8B Instruct	$0.23_{0.003}$	$0.20_{0.005}$	$0.30_{0.008}$	$0.22_{0.004}$	$0.21_{0.011}$	$0.32_{0.004}$	$0.28_{0.036}$
	Llama3 70B Instruct	$0.27_{0.009}$	$0.24_{0.013}$	$0.37_{0.008}$	$0.24_{0.007}$	$0.25_{0.005}$	$0.34_{0.007}$	$0.33_{0.011}$
	GPT-OSS 20B	$0.40_{0.021}$	$0.37_{0.015}$	$0.44_{0.018}$	$0.34_{0.018}$	$0.37_{0.011}$	$0.32_{0.008}$	$0.34_{0.043}$

Few-shot prompt

Classify the following inputs into none, one, or multiple the following moral foundations per input: authority, care, equality, loyalty, proportionality and purity.

Input: ‘Or maybe her ’picker’ is broken. This guy seems really desperate to justify things that don’t need justification, and by doing so sets my spidey sense a’tingling.‘
{"label": ["none"]}

Definition prompt

Assign ZERO OR MORE labels from this set (multi-label allowed): authority, care, equality, loyalty, proportionality and purity

KEY IDEA: WHAT “THICK MORALITY” MEANS HERE
“Thick morality” means a moral judgment that is tied to a specific kind of moral reason, i.e., it clearly maps onto at least one of the six foundation domains below.
So: thick morality = (moral evaluation) + (domain-specific basis).
If the text is moral but the basis is not identifiable (just generic good/bad/right/wrong), that is NOT thick morality for this task.

CORE MEANINGS (THE DOMAIN-SPECIFIC BASES)
- Authority: respect for legitimate roles, rules, and tradition; duty, obedience, order, insubordination, disrespect.
- Care: preventing or responding to suffering; compassion, cruelty, harm, protection, neglect.
- Equality: equal treatment and equal standing; discrimination, oppression, civil rights, unfair exclusion.
- Loyalty: commitment to an in-group; solidarity, patriotism, betrayal, disloyalty, “us vs them” obligations.
- Proportionality: merit-based fairness; deservingness, earned reward, freeloading, cheating, corruption, unfair advantage.
- Purity: sanctity and contamination concerns; cleanliness, chastity, disgust, “degrading” acts, sacred/profane framing.

DECISION PROCEDURE (DO THIS IN ORDER)
Step 1: Check for thick morality.
Does the text express a moral concern/judgment AND is the moral basis clearly one or more of the six domains?
- If yes: output all applicable domain labels and STOP.
- If no: continue.

Step 2: Otherwise output no-label.
This includes:
- generic moral language with no identifiable domain (e.g., “that’s wrong”, “immoral”, “people should be better” with no clear reason)
- non-moral content (facts, preferences, jokes, logistics, neutral descriptions)

WHAT COUNTS AS “EXPRESSES A DOMAIN”
A domain is present if the text contains a moral evaluation, emotion, norm, obligation, blame/praise, or call-to-action grounded in that domain.
- Virtue or vice both count (praise or condemnation).
- Multiple domains can co-occur; label all that clearly apply.

AVOID THESE FAILURE MODES
- Mind-reading: Don’t invent intent or moral meaning that isn’t supported by the text. If it’s underdetermined, don’t guess.
- Keyword anchoring: Moral concerns can be implicit. Moral-sounding words can be non-moral. Label from meaning, not wordlists.
- Thin vs domain mix-up: If a foundation domain fits, use it. Generic “right/wrong” with no clear domain is different and only applies when no domain fits.
- Skipping the procedure: Always check domains → generic morality → non-moral. Don’t jump straight to non-moral.
- Uncertainty blind spots: If missing context or sarcasm could flip the stance or the domain, don’t overcommit.
- Multi-domain handling: Domains can co-occur. Label all that clearly apply. If multiple domains are equally plausible and you can’t resolve, don’t force it.

Input: ‘Make no mistake, Le Pen was a friend to jews and the gay community.
Her loss will lead to an increase in murders and hate-crime against those communities, inevitably. ‘
{"label": ["care", "equality"]}

Table 18: Example prompts

Table 19: Model Precision (mean ± sd) by Moral Category on the full MFRC dataset.

Model	Care	Equality	Prop	Loyalty	Authority	Purity	Thin
BERT	0.58 (.05)	0.60 (.05)	0.44 (.07)	0.48 (.05)	0.39 (.04)	**0.69 (.12)**	0.36 (.03)
ML-BERT	0.61 (.02)	0.61 (.04)	0.40 (.06)	0.54 (.08)	0.42 (.04)	0.65 (.10)	0.41 (.06)
Llama-0-shot-single	0.41 (.01)	0.29 (.01)	0.23 (.01)	0.18 (.01)	0.20 (.01)	0.18 (.01)	0.24 (.01)
Llama-0-shot-multi	0.29 (.01)	0.32 (.01)	0.21 (.02)	0.15 (.01)	0.19 (.01)	0.14 (.02)	0.02 (.01)
Ministral-0-shot-single	0.25 (.01)	0.26 (.01)	0.16 (.02)	0.22 (.01)	0.10 (.01)	0.11 (.01)	0.09 (.01)
Ministral-0-shot-multi	0.42 (.01)	0.28 (.01)	0.20 (.01)	0.18 (.01)	0.23 (.01)	0.18 (.01)	0.21 (.01)
Llama-5-shot-single	0.44 (.01)	0.22 (.01)	0.22 (.01)	0.15 (.01)	0.21 (.01)	0.13 (.01)	0.27 (.01)
Llama-5-shot-multi	0.34 (.01)	0.34 (.01)	0.15 (.02)	0.19 (.01)	0.20 (.01)	0.15 (.02)	0.02 (.01)
Ministral-5-shot-single	0.27 (.02)	0.34 (.01)	0.14 (.01)	0.22 (.01)	0.01 (.00)	0.22 (.01)	0.21 (.01)
Ministral-5-shot-multi	0.45 (.01)	0.31 (.01)	0.22 (.01)	0.19 (.02)	0.26 (.01)	0.18 (.02)	0.16 (.01)
Llama-PEFT-single	**0.55 (.01)**	0.52 (.02)	0.29 (.01)	0.33 (.01)	0.35 (.02)	0.23 (.01)	0.28 (.02)
Llama-PEFT-multi	0.54 (.01)	**0.53 (.01)**	**0.29 (.02)**	**0.41 (.02)**	**0.37 (.02)**	0.32 (.04)	**0.26 (.05)**

Table 20: Model Recall (mean ± sd) by Moral Category on the full MFRC dataset.

Model	Care	Equality	Prop	Loyalty	Authority	Purity	Thin
BERT	**0.66 (.04)**	**0.56 (.05)**	**0.33 (.04)**	**0.41 (.04)**	**0.41 (.06)**	**0.42 (.09)**	**0.44 (.05)**
ML-BERT	0.58 (.04)	0.53 (.05)	0.26 (.06)	0.37 (.06)	0.30 (.05)	0.39 (.06)	0.30 (.05)
Llama-0-shot-single	0.36 (.01)	0.27 (.01)	0.20 (.01)	0.16 (.01)	0.18 (.01)	0.16 (.01)	0.22 (.01)
Llama-0-shot-multi	0.26 (.01)	0.30 (.01)	0.19 (.02)	0.13 (.01)	0.17 (.01)	0.12 (.02)	0.01 (.00)
Ministral-0-shot-single	0.22 (.01)	0.24 (.01)	0.14 (.02)	0.20 (.01)	0.08 (.01)	0.09 (.01)	0.07 (.01)
Ministral-0-shot-multi	0.38 (.01)	0.25 (.01)	0.18 (.01)	0.16 (.01)	0.21 (.01)	0.15 (.01)	0.18 (.01)
Llama-5-shot-single	0.40 (.01)	0.20 (.01)	0.20 (.01)	0.13 (.01)	0.18 (.01)	0.11 (.01)	0.25 (.01)
Llama-5-shot-multi	0.31 (.01)	0.31 (.01)	0.13 (.02)	0.16 (.01)	0.17 (.01)	0.13 (.02)	0.01 (.00)
Ministral-5-shot-single	0.25 (.02)	0.31 (.01)	0.12 (.01)	0.20 (.01)	0.00 (.00)	0.19 (.01)	0.19 (.01)
Ministral-5-shot-multi	0.42 (.01)	0.28 (.01)	0.20 (.01)	0.17 (.02)	0.23 (.01)	0.15 (.02)	0.14 (.01)
Llama-PEFT-single	0.52 (.01)	0.49 (.02)	0.27 (.01)	0.31 (.01)	0.32 (.02)	0.20 (.01)	0.26 (.02)
Llama-PEFT-multi	0.51 (.01)	0.50 (.01)	0.27 (.02)	0.39 (.02)	0.34 (.02)	0.30 (.04)	0.24 (.05)

Theory Trace Card (Karimi-Malekabadi et al., 2026) for the *Moral Foundations Reddit Corpus*

1. Theory

- **Framework:** Revised Moral Foundations Theory (MFT) (Atari et al., 2023), a pluralistic account of moral cognition identifying multiple distinct moral domains.
- **Core components:**
  - Care/Harm.
  - Equality/Inequality.
  - Proportionality/Disproportionality.
  - Loyalty/Betrayal.
  - Authority/Subversion.
  - Purity/Degradation.
  - Thin Morality (pragmatic category for moral judgment not clearly grounded in a specific foundation).

2. Components Exercised

- Detection of foundation-specific moral sentiment in text.
- Multi-label recognition of co-occurring moral foundations.
- Identification of Thin Morality.

3. Task Operationalization

- **Task:** Given a Reddit comment, the model predicts the presence or absence of each moral foundation and Thin Morality.
- **Key specs:** English-language Reddit comments sampled from political and everyday discourse. Each comment labeled by $\geq$3 trained annotators using a revised MFT coding manual. Majority vote aggregation determines final labels. Multi-label classification setting.
- **Scoring Criterion:** Standard classification metrics (e.g., F1, precision, recall) computed against majority human annotations.

4. Inference and Limitations

- **Inference:** Performance supports moral sentiment classification in online discourse under revised MFT.
- **Limitations:** Does not evaluate moral reasoning, normative correctness, or cross-theoretical moral competence. Operationalizes morality strictly within revised MFT; competing moral theories are outside scope. English-language Reddit data and annotator demographics may limit cultural generalizability.