Less Is Moral: A CHARMing Framework for Moral Foundations Detection in Endorsement Behaviour

Anonymous ACL submission

Abstract

Moral language plays a central role in shaping online endorsement and the diffusion of information, yet existing moral foundation detection systems often suffer from poor cross-domain generalization, weak rationale grounding, and reliance on costly prompting-based large language models (LLMs). We introduce CHARM, a MA**C**- and **H**ate-speech-**A**ware **R**ationale-aligned **M**oral foundation detection framework built on a lightweight fine-tuned LLM, which integrates complementary moral grounding, rationale alignment, and polarity-aware hate speech signals to support more robust and faithful moral prediction. Unlike prior dictionary-, fine-tune-, or prompt-based detectors, which decouple computation from psychological theory, CHARM is built so that each component — MAC cross-attention, rationale alignment, and hate-speech modulation — operationalizes a distinct psychological construct. Using only 30% of the training data required by prior fine-tuned baselines, CHARM improves AUC by up to 15.3% in-domain, surpasses these baselines on every out-of-domain dataset in both AUC and F1, and offers a scalable, low-cost alternative to LLM-based detectors. We further apply CHARM to large-scale COVID-19 discourse on Twitter and show that moral value alignment is strongly associated with online endorsement behavior. By making moral framing measurable at scale, CHARM offers a practical tool for studying the spread of morally charged misinformation[1].

1 Introduction

In March 2021, the Center for Countering Digital Hate (2021) named a "Disinformation Dozen" responsible for up to 65% of anti-vaccine content on Facebook and Twitter — including Joseph Mercola, whose posts claiming hydrogen peroxide could treat COVID-19 were shared over 4,600 times on Facebook alone. *Why does such content attract so much endorsement?* Prior work links sharing to alignment between a post's moral framing and the audience's values (Abdurahman et al., 2025); Rao et al. (2025) further show that during COVID-19, pseudo-experts (the Disinformation Dozen among them) used markedly more negative moral framing than public health experts — especially along the care/harm and authority/subversion dimensions of Moral Foundations Theory (MFT) (Haidt, 2001; Graham et al., 2013). Two questions remain open: **(i) is endorsement driven by the moral framing of content creators**, and **(ii) does the resulting endorsement network exhibit moral homophily between users and the creators they endorse?**

Answering either at population scale requires a reliable way to detect moral framing in text. Computational moral foundation detection has evolved from lexicon-based methods (Graham et al., 2009b; Hopp et al., 2021) to supervised fine-tuning with pretrained language models (Nguyen et al., 2024; Preniqi et al., 2024), and more recently to prompting LLMs for moral classification (Chen et al., 2025; Skorski and Landowska, 2025). Three gaps persist. (1) **Detection decoupled from theory.** Moral cognition is grounded in two complementary psychological frameworks — MFT and Morality-as-Cooperation (MAC) (Curry et al., 2019a,b; Graham et al., 2011; Haidt, 2012) — yet fine-tuned detectors inherit only the MFT label vocabulary, collapsing this richer structure into flat classification (Araque et al., 2022; Reinig et al., 2024). LLM-based detectors substitute statistical pretraining patterns for human moral deliberation (Chen et al., 2025; Hendrycks et al., 2021; Jiang et al., 2025; Jain et al., 2024; Lyu et al., 2024), producing labels with little auditable evidence. (2) **Polarity and antisocial coupling ignored.** Negative moral framing elevates endorsement (Rao et al., 2025), and moral language is tightly coupled to antisocial expression online (Kennedy et al., 2023;

[1]Code and additional materials: https://anonymous.4open.science/r/CHARM-1B33/

Brady et al., 2019) — yet most detectors collapse virtue and vice into a single foundation label and treat hate-speech detection as an unrelated task. (3) **Robustness–efficiency tension.** Fine-tuned models achieve strong in-domain accuracy but generalize poorly; prompt-based LLMs are more robust but expensive and depend on closed APIs.

We address these gaps with **CHARM** (MAC- and **H**ate-speech-**A**ware **R**ationale-aligned **M**oral foundation): a two-stage framework on a LoRA-adapted LLaMA-3.1-8B backbone in which each architectural component operationalizes a distinct psychological construct — MAC cross-attention introduces cooperation-domain structure as a complementary inductive bias, rationale-aligned pooling grounds prediction in human moral-deliberation traces, and FiLM hate-speech modulation captures the documented moral–antisocial coupling. This work makes three contributions:

(1) **Theory-integrated architecture.** To our knowledge, CHARM is the first moral foundation detector to embed psychological structure in the model architecture rather than only the label space, departing from the dictionary-, fine-tune-, and prompt-based traditions that treat moral detection as classification disconnected from the theories whose labels they reuse. Ablations identify MAC grounding as the strongest cross-domain inductive bias, with the largest OOD F1 drops when removed (e.g., ARG .54→.43, VIG .67→.50).

(2) **Efficient, polarity-aware, and faithful.** CHARM is open-weight and trains on 30% of the data used by prior fine-tuned baselines, yet improves AUC by up to 29.5% in-domain and 5.1% out-of-domain, surpasses all fine-tuned baselines on every OOD dataset in both AUC and F1, and matches prompting-based LLMs at a fraction of per-inference cost. Its 10-dimensional polarity-aware head improves over the strongest polarity-aware baseline by 0.33 AUC. Token-level rationale spans align closely with human annotations and causally support predictions, supplying auditable evidence behind each decision.

(3) **Moral framing predicts endorsement and network homophily at scale.** On COVID-19 Twitter discourse, CHARM shows that producer-side moral framing predicts endorsement beyond behavioural and network features (author-only moral AUC .84 vs. retweeter-only .72), and that the network is morally assortative across all foundations, most strongly along loyalty ($r = 0.46$).

2 Related Work

2.1 Moral Foundations Theory (MFT) and Morality-as-Cooperation (MAC)

MFT organizes human moral reasoning around five evolved foundations (Haidt and Joseph, 2004; Graham et al., 2009a, 2013): *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, and *purity/degradation* — the first two termed *individualizing* and the latter three *binding* foundations — with a later extension adding *liberty/oppression* (Iyer et al., 2012).

MAC (Curry, 2016) grounds morality in evolutionary game theory, identifying seven cooperation-based domains: *family values* (kin altruism), *group loyalty* (coalition support), *reciprocity* (mutual exchange), *heroism* (costly altruism), *deference* (hierarchy), *fairness* (equitable allocation), and *property rights* (ownership). Although developed through distinct methodologies, the two frameworks are largely complementary: several MAC domains map onto MFT foundations (e.g., group loyalty onto loyalty/betrayal, deference onto authority/subversion), while MAC additionally captures cooperative moral types absent from MFT, such as kin altruism, heroism, and property rights (Curry et al., 2019a).

2.2 Moral Foundation Detection

Computational moral foundation detection spans three families. *Lexicon-based methods* rely on predefined moral dictionaries — MFD (Graham et al., 2009b), MFD 2.0 (Frimer et al., 2019), and eMFD (Hopp et al., 2021) — and scale well but struggle with contextual and implicit expressions. *Fine-tuned models* such as MoralBERT (Preniqi et al., 2024) and MFormer (Nguyen et al., 2024) improve contextual sensitivity but generalize poorly across domains; recent work addresses this through auxiliary supervision and domain adaptation, e.g., DAMF (Guo et al., 2023) and ME^2-BERT (Zangari et al., 2025). *Prompting-based LLMs* such as MoVa (Chen et al., 2025) improve transferability but depend on closed, proprietary APIs.

All three families treat moral detection as a classification task disconnected from the psychological theories whose labels they reuse. CHARM departs from this by embedding psychological constructs directly in the model architecture, allowing theory to constrain inference rather than merely supply categories.

3 CHARM Framework

Each component of CHARM operationalizes a distinct psychological construct: (i) the MFT foundation classifier targets the standard moral foundation taxonomy; (ii) the MAC cross-attention module introduces cooperation-domain structure as a complementary inductive bias; (iii) the rationale-alignment module grounds prediction in human moral-deliberation traces; and (iv) the hate-speech modulation captures the documented coupling between moral language and antisocial expression. As shown in Figure 1, CHARM consists of two stages: (i) a LoRA-based encoder adaptation stage for learning foundation-level moral representations, and (ii) a polarity-aware classification stage that integrates rationale supervision, MAC grounding, and hate-speech modulation.

3.1 Problem Statement

Given a dataset $\mathcal{D} = \{(x_i, y_i^{(m)}, y_i^{(h)}, r_i, g_i)\}_{i=1}^{N}$, where $y_i^{(m)} \in \{0,1\}^{10}$ denotes polarity-aware moral labels, $y_i^{(h)} \in \{0,1\}$ the hate-speech label, r_i the rationale mask, and $g_i \in \mathbb{R}^7$ the MAC vector, the goal is to learn a polarity-aware moral classifier.

3.2 Foundation-Level Moral Encoding

Moral datasets differ substantially in annotation granularity: most provide only coarse-grained foundation labels, while a smaller subset additionally includes polarity annotations. To leverage all available supervision consistently, we adopt a two-stage training strategy. Stage 1 learns transferable foundation-level moral representations using supervision from the five MFT foundations, while Stage 2 extends the model with polarity-aware signals built on top of the frozen Stage 1 encoder. The encoder is fine-tuned using LoRA (Hu et al., 2022). A lightweight classifier predicts five foundation logits $\hat{y} \in \mathbb{R}^5$, optimized with:

$$\mathcal{L}_{\text{moral}}^{(5d)} = \frac{1}{5}\sum_{k=1}^{5} \text{BCE}(\hat{y}_k,\ y_k^{(5)}).$$

3.3 Polarity-Aware Moral Classification

Stage 2 builds on the frozen LoRA-adapted encoder from Stage 1 and performs polarity-aware moral classification using rationale alignment, MAC grounding, and hate-speech modulation.

3.3.1 Human Rationale Supervision

Rationale Selector. Given an input x, the LoRA encoder from Stage 1 produces text representations $H = \text{Enc}(x) \in \mathbb{R}^{L\times d}$, where L is the sequence length and d is the hidden dimension. A rationale selector predicts token-level rationale logit z_t and probability p_t, indicating whether the token belongs to a human-annotated rationale span. The selector is pretrained using token-level rationale supervision with binary cross-entropy:

$$\mathcal{L}_{\text{rat}} = \frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}} \text{BCE}(\sigma(z_t), r_t),$$

where $r_t \in \{0,1\}$ denotes the aligned rationale label and $\mathcal{V}$ represents non-padding positions. To encourage contiguous rationale predictions, we additionally apply a total variation regularizer:

$$\mathcal{L}_{\text{tv}} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} (p_t - p_{t-1})^2,$$

where $\mathcal{A}$ denotes adjacent valid token positions.

Rationale-Steered Attention Pooling. We combine the encoder attention distribution a^{base} with the predicted rationale distribution a^{rat} through a learnable interpolation $a = (1-\alpha)a^{\text{base}} + \alpha a^{\text{rat}}$, where $\alpha \in [0,1]$ is a learnable interpolation coefficient parameterized through a sigmoid function. The final text representation is obtained by weighted pooling over token representations, where $u^{\text{rat}} = \sum_{t=1}^{L} a_t h_t$. This encourages the model to focus on human-identified moral evidence while retaining contextual information from the encoder representations.

3.3.2 MAC-Theory Grounding

We introduce the Moral-As-Cooperation (MAC) grounding as an auxiliary supervision signal. Each sample is associated with seven MAC cooperation dimensions from eMACDscore (Malik et al., 2025), a lightweight dictionary-based moral mining tool that scores text over the seven MAC cooperation domains (see Appendix A.2 for details). Each dimension provides a probability score $\hat{p}_i$ and a sentiment-polarity score $\hat{s}_i$ that capture the strength and direction of the cooperative signal. After normalization, the two scores are multiplied as $g_i = \tilde{p}_i \cdot \tilde{s}_i$, forming a MAC vector $g \in \mathbb{R}^7$. Each dimension of g is then projected into the hidden space to produce seven MAC tokens $M \in \mathbb{R}^{7\times d}$. The MFT labels are represented as ten learnable queries $Q \in \mathbb{R}^{10\times d}$, corresponding to the virtue and vice poles of the five foundations. Cross-attention between Q and M produces ten moral representations:

$$C = \text{softmax}\left(\frac{QM^\top}{\sqrt{d}}\right) M \in \mathbb{R}^{10\times d},$$

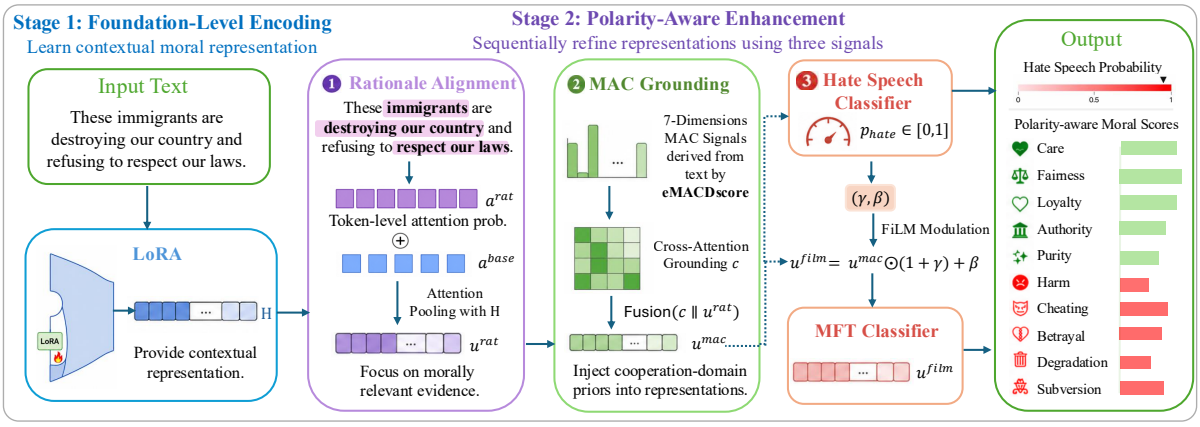


Figure 1: Overview of CHARM, a two-stage framework for polarity-aware moral foundation detection. Stage 1 trains a LoRA adapter to learn foundation-level moral semantics. Stage 2 refines the text representation using rationale alignment, MAC-guided cross-attention, and hate-speech-aware FiLM modulation. The model jointly predicts 10 polarity-aware moral dimensions and hate speech.

along with an attention map $W = \text{softmax}(QM^\top/\sqrt{d}) \in \mathbb{R}^{10\times 7}$, where W_{ij} denotes how strongly the i-th moral dimension attends to the j-th MAC dimension. The representations in C are then pooled into a single MAC vector c using learnable attention-based weighting. Specifically, each moral representation C_i is assigned a scalar attention score:

$$a_i = \frac{\exp(w^\top C_i)}{\sum_{k=1}^{10} \exp(w^\top C_k)},$$

where $w \in \mathbb{R}^d$ is a learnable parameter vector. The final pooled MAC representation is then computed as $c = \sum_{i=1}^{10} a_i C_i \in \mathbb{R}^d$. Finally, the MAC vector c is fused with the rationale-guided text representation u^{rat} to obtain the MAC-enhanced moral representation u^{mac}:

$$u^{\text{mac}} = \text{FusionMLP}([u^{\text{rat}} \,\|\, c]) \in \mathbb{R}^d.$$

This formulation injects cooperation-domain structure into the moral representation, enabling the model to capture complementary interactions between MAC signals and moral foundations.

3.3.3 Hate Speech Modulation

Prior work has shown that moral language and hate speech are closely related, with certain moral dimensions disproportionately associated with hateful rhetoric (Vargas et al., 2026). Motivated by this connection, we introduce hate speech as an auxiliary modulation signal for polarity-aware moral prediction. Given the MAC-grounded representation $u^{\text{mac}} \in \mathbb{R}^d$, a hate-speech classifier first predicts a hate probability $\hat{y}^{(h)} = \sigma(w_h^\top u^{\text{mac}})$. The hate speech head is trained using binary cross-entropy loss. Following Feature-wise Linear Modulation (FiLM) (Perez et al., 2018), the predicted hate probability is transformed into channel-wise scaling and shifting parameters:

$$\gamma = \tanh(f_\gamma(\hat{y}^{(h)})), \qquad \beta = f_\beta(\hat{y}^{(h)}),$$

where $\gamma \in [-1, 1]^d$ and $\beta \in \mathbb{R}^d$. We then modulate u^{mac} through $u^{\text{film}} = u^{\text{mac}} \odot (1 + \gamma) + \beta$, allowing hate speech information to act as a soft conditioning signal over the moral representation. The final polarity-aware moral prediction is computed as a 10-dimensional probability vector, $\hat{y}^{(m)} = \sigma(W_m u^{\text{film}}) \in (0, 1)^{10}$. The corresponding classification objective is:

$$\mathcal{L}_{\text{moral}}^{(10d)} = \frac{1}{10} \sum_{j=1}^{10} \text{BCE}(\hat{y}_j^{(m)},\ y_j^{(m)}).$$

4 Experiments

Datasets. We use nine datasets covering diverse forms of morally charged discourse. Dataset statistics and label distributions are provided in Table 1 and Appendix A.1. For all datasets, we precompute 7-dimensional MAC signals using eMACDscore. Training proceeds in two stages. Stage 1 jointly trains on four in-domain datasets using foundation-level supervision, while Stage 2 further fine-tunes on MFTCXplain with polarity, rationale, and hate-speech annotations. During Stage 1, polarity labels in MFTCXplain are collapsed into foundation categories, while the same data split is retained

Table 1: Full profile of datasets used for training and evaluation. # Inst is the original dataset size. †MFTCXplain contains four languages (Portuguese, Persian, Italian, English). ‡HateBR is in Brazilian Portuguese.

Split	Dataset	# Inst.	Text Type	Content Focus
In-domain	MFTC (Hoover et al., 2020)	34,987	Twitter posts	Socio-political discourse
	MFRC (Trager et al., 2022)	17,886	Reddit comments	Politics and everyday morality
	News (Hopp et al., 2021)	34,262	News articles	Diverse news topics from GDELT
	†MFTCXplain (Trager et al., 2025)	3,245	Multilingual tweets	Moral reasoning over hate speech
OOD	SC (Forbes et al., 2020)	29,239	Rules-of-thumb	Everyday social and moral norms
	VIG (Clifford et al., 2015)	132	Psychology vignettes	Foundation-violating scenarios
	ARG (Kobbe et al., 2020)	320	Debate arguments	Pro/con stances on 16 topics
	MIC (Ziems et al., 2022)	11,375	Chatbot RoTs	Human–chatbot moral biases
	‡HATEBR (Vargas et al., 2026)	5,040	Instagram comments	Brazilian political hate speech

across both stages to prevent leakage. Following MFORMER, we retain the original train/test splits and use 30% of the training data for the remaining in-domain datasets. We exclude liberty/oppression due to inconsistent supervision across datasets. The five out-of-domain datasets are reserved for evaluation under a shared validation setting.

Backbone selection. We compare LLAMA-3.1-8B[2] and QWEN3-8B[3], and find that LLAMA-3.1-8B consistently performs better across settings. We therefore use it as the backbone in all main experiments. Additional experimental details and QWEN3-8B results are provided in Appendix A.5 and Appendix B.2, respectively.

Baselines. We compare against four baselines described in Sec. 2.2: (1) MFORMER (Nguyen et al., 2024), five binary RoBERTa-based classifiers fine-tuned on multi-domain corpora; (2) MOVA (Chen et al., 2025), an LLM prompting framework with an *all@once* strategy that does not model moral polarity; (3) a fine-tuned GPT-4o mini trained on 30K examples using MOVA's prompting strategy (Chen et al., 2025); (4) MORALBERT (Preniqi et al., 2024), BERT-base classifiers fine-tuned on multi-platform social media corpora to predict moral polarities. Details are provided in Appendix A.3.

Metrics. We evaluate the model along two axes: classification performance and rationale quality. For classification, we report F1 and AUC on both moral foundation prediction and hate speech detection. For rationale quality, we follow the ERASER protocol (DeYoung et al., 2020) to assess *plausibility* and *faithfulness*. Definitions of all evaluation metrics are provided in Appendix A.4.

[2] https://huggingface.co/meta-llama/Llama-3.1-8B

[3] https://huggingface.co/Qwen/Qwen3-8B

4.1 Moral Detection Performance

In-domain performance. CHARM achieves strong in-domain performance despite using only 30% of the training data required by MFORMER. It substantially outperforms MFORMER on MFRC and News, with AUC improvements of 0.07 and 0.11 (a relative gain of up to 15.3%), respectively, suggesting better adaptation to datasets with different linguistic styles and label distributions. On MFTC, MFORMER achieves slightly higher F1, while CHARM maintains comparable AUC (.87 vs. .89). A finer-grained analysis further shows that this small gap is primarily driven by lower performance on the authority foundation.

Generalization under distribution shift. The advantage of CHARM becomes more pronounced under distribution shift. Across OOD datasets, CHARM surpasses MFORMER on every dataset in both AUC and F1. Against Tuning-30K GPT-4o mini, the same-scale tuning baseline, CHARM matches or exceeds it on AUC across OOD datasets, with VIG as the only exception, indicating competitive robustness under limited supervision.

Compared with MOVA, CHARM achieves higher F1 on ARG (.54 vs. .49), matches MOVA's AUC on ARG, and slightly outperforms it on SC, while MOVA performs better on MIC and VIG. CHARM also consistently surpasses MOVA on the care foundation across SC and MIC, suggesting that hate-speech-aware supervision improves sensitivity to harm-related cues. One possible explanation lies in the different learning paradigms. Prompting-based methods rely on fixed pretrained representations and in-context instructions, which work well for the short and standardized inputs in SC and MIC (14 and 17 tokens on average). In contrast, ARG contains much longer and more complex inputs (69.6 tokens), where moral cues

Table 2: F1 and AUC across evaluation datasets. Best results per dataset are shown in **bold**. Baseline AUC values (†) are taken from Chen et al. (2025); F1 scores are independently reproduced by running each model. "–" indicates results not reported and not reproducible. Per-foundation breakdowns are provided in Appendix A.8.

		MFTC	MFRC	News	MFTCXpl.	ARG	SC	MIC	VIG	HateBR
AUC	CHARM	.87	**.91**	**.83**	**.79**	**.87**	**.77**	.73	.89	**.82**
	MoVa†	.77	.80	.68	.75	**.87**	.76	**.75**	**.96**	.78
	Mformer†	**.89**	.84	.72	.61	.86	.70	.70	.81	.62
	MoralBERT†	.75	.79	.61	.60	.77	.70	.69	.80	.54
	Tuning-GPT4o-mini†	–	–	–	–	.85	.73	.73	.92	–
F1	CHARM	.72	**.72**	**.55**	**.59**	**.54**	.46	.50	.67	**.38**
	MoVa	.57	.48	.37	.52	.49	**.52**	**.51**	**.69**	.36
	Mformer	**.77**	.66	.53	.50	.51	.43	.44	.52	.31
	MoralBERT	.56	.49	.33	.38	.47	.37	.41	.32	.24

are distributed across the context. CHARM's task-specific fine-tuning appears better suited to capturing these dispersed signals. In addition, MoVa incurs inference-time API costs of approximately $0.066 and $0.065 per 1K input tokens on SC and MIC, respectively, with charges applied to every inference pass, causing costs to grow with dataset size. By contrast, CHARM is trained once on open-source models and introduces no additional per-example inference cost.

Moral polarity classification. As shown in Figure 2, CHARM clearly outperforms MoralBert on both datasets, achieving AUC improvements of 0.33 on MFTCXplain and 0.21 on HateBR, alongside even larger gains in F1. Notably, CHARM also performs well on the Portuguese HateBR dataset, even though it was not designed for multilingual learning. This suggests that training on the multilingual MFTCXplain dataset may provide an unintended cross-lingual benefit.

4.2 Ablation Studies

To evaluate the contribution of each component, we perform ablations on MAC grounding, rationale supervision, and hate-speech supervision. Results are shown in Table 3.

MAC provides the strongest structural inductive bias for cross-domain generalization. Removing MAC causes the largest overall degradation across OOD evaluation datasets, with especially severe drops on ARG (.54→.43) and VIG (.67→.50). Notably, compared with the other signals, SC and MIC are more strongly affected by removing MAC (SC: .46→.41; MIC: .50→.44). Both datasets primarily involve socially contextualized norm judgments and rule-of-thumb reasoning rather than explicit moral declarations, suggesting that MAC provides useful structural grounding for implicit moral interpretation under distribution shift.

Rationale supervision improves evidence grounding and prediction faithfulness. Removing rationale supervision leads to larger performance drops on datasets with relatively longer and more compositional inputs, particularly MFRC (.69→.65) and VIG (.67→.60), while having comparatively smaller effects on shorter-text datasets such as SC. Since rationale supervision is provided at the span level, this pattern suggests that rationale alignment is especially beneficial when morally relevant evidence is distributed across contextual spans rather than concentrated in a single local cue. This effect is further supported by the rationale evaluation results in Table 4. Compared with the variant without rationale supervision, CHARM consistently improves performance across all rationale evaluation metrics, indicating that rationale alignment enhances both plausibility and faithfulness, while also improving the causal relevance of the identified rationale spans for prediction.

Hate-speech and moral supervision reinforce each other. Removing hate-speech supervision consistently reduces performance across multiple datasets, including MFTCXplain_{10d} (.47→.45), SC (.46→.44), and VIG (.67→.65). The degradation on MFTCXplain_{10d} suggesting that hate-speech supervision can help the model better distinguish fine-grained moral polarity. Table 5 further shows that CHARM achieves the strongest hate-speech detection performance, outperforming both prompted GPT-4o mini and a fine-tuned LLaMA model that is trained without moral supervision. These results indicate a mutually beneficial relationship between moral reasoning and harmful-content identification.

Table 3: Ablation study reporting F1 scores across evaluation datasets. The baseline applies a task-specific classification head on top of the frozen Stage 1 LoRA encoder using only the 10-dimensional MFT supervision. AUC results show consistent trends and are reported in Appendix A.7.

Variant	MFTCXpl.$_{10d}$	MFTCXpl.$_{5d}$	MFTC	MFRC	News	ARG	SC	MIC	VIG
baseline	.41	.56	.68	.56	.50	.45	.30	.43	.47
w/o hate speech	.45	.59	.64	.55	.55	.53	.44	.45	.65
w/o MAC	.46	.59	.66	.53	.47	.43	.41	.44	.50
w/o rationale	.42	.57	.68	.65	.55	.53	.46	.47	.60
CHARM	**.47**	**.59**	**.72**	**.69**	**.55**	**.54**	**.46**	**.50**	**.67**

Table 4: Rationale quality evaluation on MFTCXplain. Arrows indicate whether higher (↑) or lower (↓) is better. The *w/o rationale* baseline removes rationale supervision while keeping all other signals and the architecture unchanged.

	Plausibility			Faithfulness	
	IoU↑	TF1↑	AUP↑	Suf↓	Cmp↑
CHARM	**.62**	**.63**	**.60**	**.02**	**.25**
w/o rationale	.14	.12	.42	.20	.04

Table 5: Hate speech detection results on MFTCXplain.

	Bin. F1	Mac. F1	AUC
CHARM	**.66**	**.70**	**.79**
TUNING-LLAMA-3.1-8B	.61	.64	.71
GPT-4o mini (0-shot)	.43	.58	.61
GPT-4o mini (4-shot)	.47	.58	.61

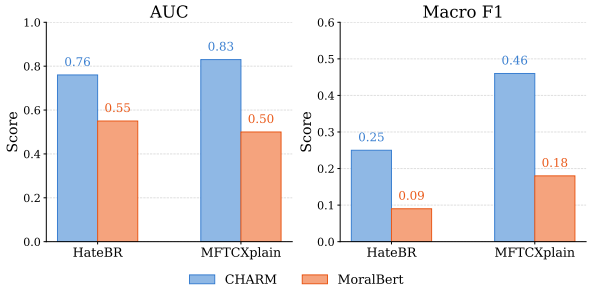


Figure 2: AUC and Macro-F1 comparison between CHARM and MORALBERT for 10-dimension moral polarity classification on MFTCXplain and HateBR (only datasets with 10-dimensional moral annotations).

Table 6: Ablation study for endorsement prediction using interaction, behavioral, and moral-profile features.

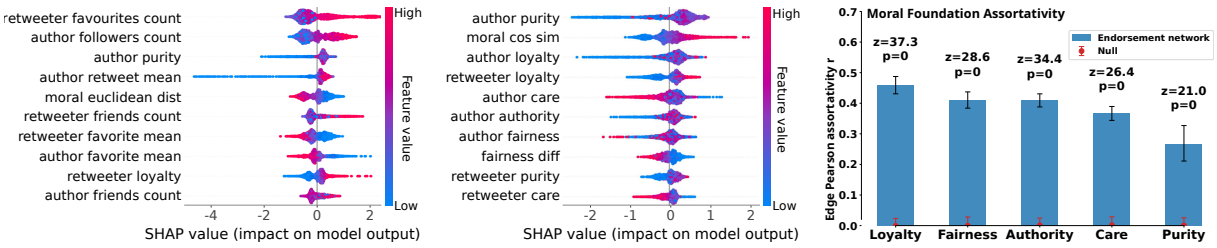
Feature Setting	AUC	Macro F1	Micro F1
full	**0.92**	**0.84**	**0.84**
w/o moral features	0.90	0.82	0.82
w/o interaction features	0.85	0.77	0.77
author moral only	0.84	0.76	0.76
retweeter moral only	0.72	0.66	0.66

Overall, all three supervision signals contribute complementary gains.

5 Moral Alignment and Endorsement Behaviour

We use CHARM to test, at scale, the two questions raised in Section 1: *whether producer-side moral framing predicts endorsement, and whether the resulting endorsement network exhibits moral homophily between users and creators* (Dehghani et al., 2016; Van Bavel et al., 2021). We use a subset of the COVID-19 Twitter dataset introduced by Chen et al. (2020), covering discourse from March 25 to April 16, 2020. Each user's moral profile is constructed by aggregating tweet-level moral scores inferred by CHARM. Although CHARM produces polarity-aware predictions over ten moral dimensions, we collapse scores into five foundation-level representations for endorsement analysis because several polarity dimensions are highly sparse in large-scale Twitter discourse. The full polarity distribution is provided in Appendix B.1. To reduce noise from weak moral expressions, we retain only tweets above the 90th percentile of overall moral intensity. We then construct a directed endorsement network based on retweet behavior, which is commonly treated as a proxy for endorsement (Metaxas et al., 2015). Mentions and quote tweets are excluded due to their ambiguous endorsement semantics.

Finding 1: Moral content predicts endorsement, with content producer framing carrying stronger signal than consumer-side. To evaluate moral alignment in endorsement behavior, we formulate endorsement prediction as a binary classification task over directed user pairs from the retweet network. Positive samples correspond to repeated retweet interactions occurring more than five times, while negative samples are constructed using a two-hop non-endorsement constraint. Endorsed pairs exhibit higher moral cosine



(a) (b) (c)

Figure 3: Endorsement analysis. (a) SHAP feature importance for the full prediction model; (b) SHAP for the moral-only model; (c) moral assortativity coefficients compared with degree-preserving random networks (edges rewired while preserving each node's in- and out-degree). Error bars in (c) indicate bootstrap confidence intervals.

similarity than non-endorsed pairs (mean 0.934 vs. 0.892, $\Delta = 0.042$), with both Mann–Whitney and KS tests indicating significant differences ($p < 10^{-119}$); the separation is consistent across all five foundations, with loyalty and fairness showing the sharpest distributional gap (Appendix B.6). We further test whether this structure provides predictive utility beyond network and behavioral signals using a LightGBM classifier with moral-profile, interaction, behavioral, and semantic features. More training details are provided in Appendix B.5. As shown in Table 6, removing moral features consistently degrades performance, whereas models using only moral representations still retain substantial predictive power. Moreover, author-only moral features substantially outperform retweeter-only features, indicating that endorsement behavior is more strongly associated with the moral framing of content producers than with the inferred moral profiles of endorsers.

SHAP analysis. In the full model (Figure 3a), behavioral and network features dominate overall, yet author purity (rank 3) and moral Euclidean distance (rank 5) appear among the top ten, confirming that moral framing adds predictive signal beyond engagement covariates. In the moral-only model (Figure 3b), author-side features hold seven of the top ten positions, with author purity and moral cosine similarity ranking first and second. Notably, author purity is the strongest individual moral predictor of endorsement despite purity having the lowest network assortativity ($r = 0.27$), suggesting that high purity framing in content attracts broad endorsement rather than congruent followers.

Finding 2: Endorsement networks exhibit robust moral homophily. We next test whether endorsement structure exhibits moral homophily. Following Newman (2002), we measure moral assortativity for each foundation as the edge-level Pearson correlation between the foundation scores of the two endpoints of each endorsement tie. The full computation is provided in Appendix B.7. As shown in Figure 3c, all five foundations are significantly positively assortative (all $z > 21$, $p \approx 0$ against a degree-preserving null), indicating that users systematically endorse others with similar moral profiles. Assortativity is highest for loyalty ($r = 0.46$), followed by fairness and authority ($r = 0.41$), care ($r = 0.37$), and purity ($r = 0.27$). Per-foundation moral alignment distributions are provided in Appendix B.6.

6 Conclusions

We introduced CHARM, a lightweight moral foundation detection framework that unifies rationale alignment, MAC-based cooperative signals, and hate-speech-aware polarity modeling within a fine-tuned LLM architecture. Across multiple datasets, CHARM demonstrates strong robustness, faithfulness, and sample efficiency, while remaining substantially more efficient than prompt-based LLM approaches. Using CHARM, we further show that online endorsement behavior is closely associated with latent moral alignment, particularly the moral framing of content authors. Notably, author purity is the strongest individual predictor of endorsement yet exhibits the weakest network assortativity ($r = 0.27$), suggesting that morally charged content can attract broad endorsement beyond moral congruence. By grounding model components in explicit psychological construct, CHARM bridges social-psychological theory and LLM practice.

Limitations

Several limitations should be noted. First, polarity-level moral annotations remain limited, and most existing datasets do not cover the liberty/oppression foundation. As a result, training and evaluation of our 10-dimensional setting were largely restricted to MFTCXplain and HateBRMoralXplain. Future work could improve this by developing larger datasets with polarity-level annotations and broader foundation coverage.

Second, although CHARM shows strong out-of-domain performance on several datasets, results are weaker on VIG and MIC compared to prompt-based approach, whose content focuses more on abstract social norms. In contrast, the in-domain training data mainly consists of concrete, opinion-rich social media and news discourse. Since fine-tuning tends to adapt models toward the training distribution, this mismatch in content type may limit generalization to more abstract forms of moral reasoning.

Finally, MAC supervision relies on **eMACD-score** automatically generated labels because no publicly available human-annotated MAC dataset currently exists. Although the same process was applied consistently across all samples, the generated labels may still introduce noise into downstream training.

Ethics Statement

This work studies moral reasoning discourse using publicly available datasets collected from social media, news, and online discussion platforms. Since moral annotations can reflect cultural and annotator-specific biases, model predictions should not be interpreted as objective judgments of morality. Our goal is to support the analysis of moral framing and moral conflict in language, rather than to determine whether individuals or viewpoints are morally right or wrong.

The proposed framework may also inherit biases present in the training data, especially in politically or culturally sensitive contexts. We therefore encourage careful use of such models and emphasize that they should not be deployed as standalone systems for high-stakes moderation or decision-making.

References

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*, 154(4):935–957.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, GoodIT '22, page 154–160, New York, NY, USA. Association for Computing Machinery.

William J. Brady, Julian A. Wills, Dominic Burkart, John T. Jost, and Jay J. Van Bavel. 2019. An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, 148(10):1802–1813.

Center for Countering Digital Hate. 2021. The disinformation dozen. `https://counterhate.com/research/the-disinformation-dozen/`. Accessed: 2026-05-24.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

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 33216–33260, Suzhou, China. Association for Computational Linguistics.

Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198.

Oliver Scott Curry. 2016. *Morality as Cooperation: A Problem-Centred Approach*, pages 27–51. Springer International Publishing, Cham.

Oliver Scott Curry, Matthew Jones Chesters, and Caspar J Van Lissa. 2019a. Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. *Journal of Research in Personality*, 78:106–124.

Oliver Scott Curry, Daniel Austin Mullins, and Harvey Whitehouse. 2019b. Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1):47–69.

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):366–375.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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

Jeremy A. Frimer, Reihane Boghrati, Jonathan Haidt, Jesse Graham, and Morteza Dehghani. 2019. Moral foundations dictionary 2.0. https://osf.io/ezn37.

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. 2009a. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009b. Moral foundations dictionary. *PsycTESTS Dataset*.

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):366–385.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):281–291.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books, New York.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations (ICLR)*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, and Madelyn Mendlen. 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.

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 *International Conference on Learning Representations (ICLR)*.

Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLOS ONE*, 7(8):1–23.

Shomik Jain, Dan Calacci, and Ashia Wilson. 2024. "As an AI Language Model, Yes I Would Recommend Calling the Police": Norm inconsistency in LLM decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 624–633.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T. Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, and Jack Hessel. 2025. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, 7(1):145–160.

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.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Musa Malik, Sungbin Youk, Frederic R. Hopp, Oliver Scott Curry, Marc Cheong, Mark Alfano, and René Weber. 2025. The extended morality as cooperation dictionary (emacd): A crowd-sourced approach

via the moral narrative analyzer platform. *Communication Methods and Measures*, 19(3):201–231.

Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. 2015. What do retweets indicate? results from user survey and meta-review of research. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 658–661.

M. E. J. Newman. 2002. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701.

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.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: a fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 433–442.

Ashwin Rao, Nazanin Sabri, Siyi Guo, Louiqa Raschid, and Kristina Lerman. 2025. Public health messaging on twitter during the covid-19 pandemic: Observational study. *Journal of Medical Internet Research*, 27.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A survey on modelling morality for text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand. Association for Computational Linguistics.

Maciej Skorski and Alina Landowska. 2025. The moral gap of large language models. *arXiv preprint arXiv:2507.18523*.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K. Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi Malekabadi, and Flor Miriam Plaza-del Arco. 2025. MFTCXplain: A multilingual benchmark dataset for evaluating the moral reasoning of LLMs through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China. Association for Computational Linguistics.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, and Melissa Reyes. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Jay J. Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. 2021. How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11):913–916.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-explaining hate speech detection with moral rationales. *arXiv preprint arXiv:2601.03481*.

Lorenzo Zangari, Candida M. Greco, Davide Picca, and Andrea Tagarelli. 2025. 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, Abu Dhabi, UAE. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773.

A Additional Experimental Details

A.1 Experimental Dataset Details

Tables 7 and 8 report label prevalence for the in-domain training datasets, while Table 9 summarizes label prevalence for the out-of-domain evaluation datasets.

A.2 Details of eMACDscore for MAC Grounding

Our MAC grounding signal is derived from **eMACDscore**, a Python-based moral mining tool introduced by Malik et al. (2025) as part of the extended Morality-as-Cooperation Dictionary (eMACD). Built upon the Morality-as-Cooperation framework (Curry, 2016), eMACD provides lexicon-based scores across seven cooperation domains: family, group, reciprocity, heroism, deference, fairness, and property. The lexicon itself was constructed through crowd-sourced annotation using the Moral Narrative Analyzer (MoNA) platform. We use eMACDscore to automatically generate MAC supervision signals for each training instance. Specifically, for every sample and MAC domain, the tool outputs a probability score $\hat{p}_i$ and a sentiment-polarity score $\hat{s}_i$, which are combined into the MAC vector $g \in \mathbb{R}^7$. We adopt eMACDscore because, to the best of our knowledge, there is currently no publicly available human-annotated MAC benchmark dataset.

Importantly, eMACD is developed independently of Moral Foundations Theory (MFT). Rather than being derived from MFT labels or annotations, it is grounded in a distinct theoretical framework and constructed through an independent crowd-sourcing process. As a result, the MAC supervision provides a complementary moral signal instead of simply repackaging the MFT information already available to the model. We refer readers to Malik et al. (2025) for full details regarding the dictionary construction process, annotation protocol, and validation experiments.

A.3 Baseline Implementation Details

For baseline comparisons, we reproduce each baseline under its original released configuration, including matching the random seed specified in their code, to ensure a fair comparison.

For evaluation, we follow the metrics commonly used in the original studies of each dataset, reporting macro-AUC and macro-F1. For F1, per-label decision thresholds are tuned on a shared validation set and applied unchanged to the test set. We release the trained checkpoint so that all reported scores can be exactly reproduced by running inference on the released model.

MFORMER. MFORMER (Nguyen et al., 2024) is a transformer-based moral foundation classification framework built on ROBERTA-base. The model trains five independent binary classifiers, each corresponding to one moral foundation in a one-vs-rest setting. It is trained on a mixture of annotated datasets spanning social media, news, and online discussion domains. During inference, predictions from the five classifiers are combined to produce the final multi-label moral output.

MOVA. MOVA (Chen et al., 2025) is a prompting-based framework for moral value classification using instruction-tuned GPT-4o mini. It formulates moral foundation detection as a multi-label generation task, where a single prompt is used to predict all moral dimensions jointly. The framework does not require task-specific fine-tuning and instead relies on prompt engineering to guide structured moral predictions. In our experiments, we use the official prompt templates released by the authors for reproduction, and treat MOVA as our primary zero-shot baseline for LLM-based moral reasoning.

Fine-tuned GPT-4o mini. Following Chen et al. (2025), we include a supervised fine-tuned variant of GPT-4o mini trained on 30K labeled instances sampled from the MFORMER training corpus. The model is fine-tuned to generate structured moral labels directly from input text without rationale supervision or auxiliary moral signals. This baseline serves as a decoder-only LLM adapted for supervised moral classification. This setup allows us to evaluate whether fine-tuning a LLM improves over prompting-based inference under comparable supervision data.

MORALBERT. MORALBERT (Preniqi et al., 2024) is a BERT-based framework for polarity-aware moral classification across ten virtue–vice dimensions. Unlike standard MFT classifiers that predict only foundation presence, MORALBERT explicitly models both positive (virtue) and negative (vice) polarities using a multi-label classification objective. Its explicit polarity modeling provides a relevant comparison for evaluating fine-grained moral representation learning.

Table 7: Distribution of moral virtue dimensions and input length statistics across in-domain training datasets. The five moral dimensions correspond to virtue–vice axes. Mean token and P90 token denote the average and 90th-percentile token lengths, respectively.

Dataset	Care	Fairness	Loyalty	Authority	Purity	Avg. Tokens	P90. Tokens	Total Size
MFTC	40.7%	41.8%	37.0%	41.5%	29.7%	31.9	50	13,995
MFRC	35.5%	44.2%	19.4%	32.5%	17.4%	47.5	92	7,155
News	26.5%	32.4%	33.1%	35.4%	30.4%	31.1	50	12,920
MFTCXplain	37.3%	36.0%	27.2%	23.2%	18.6%	49.6	85	3,245
Overall	**39.9%**	**40.3%**	**33.4%**	**38.3%**	**27.0%**	**36.2**	**63**	**37,315**

Table 8: Label distribution of MFTCXplain across moral polarity dimensions and hate speech annotations.

Dimension	Percentage
Care	8.91%
Harm	29.34%
Fairness	13.50%
Cheating	24.35%
Loyalty	15.04%
Betrayal	13.50%
Authority	10.29%
Subversion	13.22%
Purity	7.95%
Degradation	11.31%
Hate Speech	38.92%
Total size	**3,245**

A.4 Evaluation Metric Definitions

We report AUC and F1 scores for both hate speech detection and moral classification tasks.

Hate Speech Detection. For hate speech classification, we use the standard binary F1 score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where precision and recall are computed with respect to the hate speech class. We additionally report AUC, which measures the area under the receiver operating characteristic (ROC) curve and evaluates the model's ranking quality across all classification thresholds.

Moral Classification. Moral classification is formulated as a multi-label prediction task, where each sample may express multiple moral foundations simultaneously. Let L denote the set of moral labels. For each label $l \in L$, we compute a binary F1 score by treating the presence or absence of l as a separate binary classification problem:

$$F1_l = \frac{2 \cdot \text{Precision}_l \cdot \text{Recall}_l}{\text{Precision}_l + \text{Recall}_l}.$$

The reported Macro-F1 is then obtained by averaging over all moral labels:

$$\text{Macro-F1} = \frac{1}{|L|} \sum_{l \in L} F1_l.$$

Similarly, AUC is computed as the average AUC across all moral labels. Because moral prediction is inherently multi-label and many instances activate only a small subset of moral dimensions, Macro-F1 can sometimes overestimate overall performance. In particular, labels with sparse positive instances or easier decision boundaries may inflate the averaged score despite limited holistic understanding of moral content. We therefore report both Macro-F1 and AUC to provide a more comprehensive assessment of model behaviour.

Unlike AUC, which is threshold-free, F1 depends on a decision threshold. To ensure a fair comparison, we adopt a unified threshold-selection protocol across all models. Using a held-out validation set shared by all methods, we select, for each label l, the threshold that maximizes the validation (F1_l), and apply it unchanged to the test set. This isolates differences in model quality from differences in threshold choice.

Rationale Plausibility We compare the model-selected rationale tokens R_k against the human-annotated rationale spans H_k for each instance k. We report IoU-F1 and Token-F1:

$$\text{IoU-F1} = \frac{1}{N} \sum_{k=1}^{N} \frac{|R_k \cap H_k|}{|R_k \cup H_k|},$$

$$\text{Token-F1} = \frac{1}{N} \sum_{k=1}^{N} \frac{2|R_k \cap H_k|}{|R_k| + |H_k|}.$$

We also report AUPRC to evaluate how well token-importance scores rank gold rationale tokens under severe class imbalance.

Table 9: Distribution of moral virtue dimensions and input length statistics across out-of-domain evaluation datasets. The five moral dimensions correspond to virtue–vice axes. Avg. token and P90. token denote the average and 90th-percentile token lengths, respectively.

Dataset	Authority	Care	Fairness	Loyalty	Purity	Avg. Tokens	P90. Tokens	Total Size
VIG	14.78%	27.83%	14.78%	13.91%	14.78%	17.0	19	132
SC	9.72%	44.44%	16.92%	17.80%	6.79%	12.1	17	2,9239
ARG	14.37%	41.56%	16.56%	8.12%	19.69%	69.6	120	320
MIC	17.15%	51.19%	20.72%	19.43%	10.86%	10.2	14	1,1375
HateBR	6.90%	28.97%	22.88%	4.21%	22.30%	23.5	47	5,040

Rationale Faithfulness Following DeYoung et al. (2020), we report *comprehensiveness* and *sufficiency*. Let $p(\cdot)_y$ denote the model probability for the gold label y, and let e_k denote the extracted rationale for instance k.

Comprehensiveness measures the confidence drop after removing rationale tokens:

$$\text{Comp} = \frac{1}{N}\sum_{k=1}^{N}\big(p(x_k)_y - p(x_k \setminus e_k)_y\big).$$

Sufficiency evaluates whether the rationale alone preserves enough evidence:

$$\text{Suff} = \frac{1}{N}\sum_{k=1}^{N}\big(p(x_k)_y - p(e_k)_y\big).$$

Higher comprehensiveness indicates that the rationale contributes substantially to the prediction; lower sufficiency indicates that the rationale alone retains most of the predictive signal.

A.5 Experimental Setup

All experiments were conducted on two NVIDIA A40 GPUs using PyTorch 2.10.0 (CUDA 12.8), Hugging Face Transformers 5.0.0, and PEFT 0.18.1. CHARM is built on LLaMA-3.1-8B and trained in two stages. In Stage 1, we attach LoRA adapters (rank 16, $\alpha = 32$, dropout 0.1) to all attention and feed-forward projection matrices, and train the moral classifier and token scorer with a binary cross-entropy moral loss. The LoRA adapters introduce 0.56% trainable parameters of the full backbone, keeping the model lightweight. In Stage 2, we freeze the Stage 1 LoRA encoder and train the fusion heads, optimizing a combined objective that sums the moral loss with hate-speech, rationale, and total-variation (TV) smoothness terms, weighted by λ_{hate}, λ_{rat}, and λ_{TV} (initialized to 0.1, 0.2, and 0.05, respectively; the moral term has a fixed weight of 1.0). These three λ values are learnable via log parameterisation. Both stages use the AdamW optimizer with a learning rate of 2×10^{-4} and a maximum sequence length of 128; we adopt these values following common practice without extensive tuning. Stage 1 is trained for 3 epochs and Stage 2 for 5 epochs, with a total training time of approximately one hour.

Stage 1 trains on MFTCXplain together with a 30% subsample of the MFRC, MFTC, and News training pools, while Stage 2 trains on MFTCXplain only. The 30% subsampling uses a fixed seed (42) with distribution-aware weighting toward the MFTCXplain label distribution, followed by a 9:1 stratified train/validation split and the same train/test split as MFORMER. The same MFTCXplain train/validation/test partition is used in both stages, so that no test instance is seen during either stage of training.

A.6 Cross-Validation Robustness

To assess the stability of CHARM, we additionally perform cross-validation on the two smallest datasets, reporting the mean and standard deviation of macro-F1 and AUC across folds. This is an independent robustness check using a different partitioning from the shared-split protocol in Table 2, and is therefore not directly comparable to the main results.

Table 10: Stability of CHARM over 5-fold cross-validation on the small-sample datasets (mean $\pm$ std).

Dataset	Macro-F1	AUC
ARG	0.489 ± 0.079	0.864 ± 0.026
VIG	0.537 ± 0.048	0.889 ± 0.033

AUC remains stable across runs (std ≤ 0.033), consistent with our observation that threshold-free ranking is robust under variation. Macro-F1 exhibits larger variance, as expected given the small

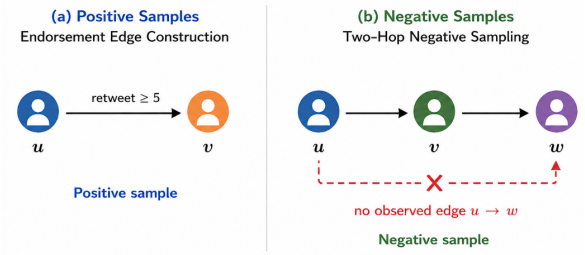


Figure 4: Construction of the endorsement prediction graph. Positive pairs are defined as repeated retweet interactions occurring more than five times, while negative pairs are sampled using a two-hop non-endorsement constraint.

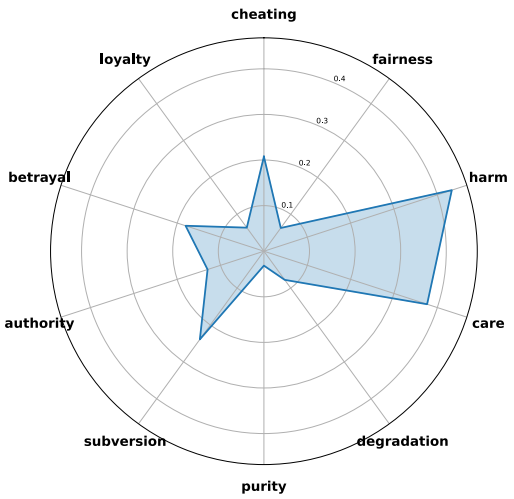


Figure 5: Distribution of users across the ten polarity-aware moral dimensions.

sample sizes (ARG: n=320; VIG: n=132) and F1's sensitivity to threshold selection.

A.7 Ablation Results on AUC

Table 11 reports AUC-based ablation results across both in-domain and out-of-domain datasets. Overall, removing rationale supervision consistently degrades performance on most datasets, particularly under cross-domain evaluation, suggesting that rationale-guided representations improve robustness beyond token-level alignment.

The removal of MAC-enhanced representations leads to the largest performance drops on socially grounded datasets such as SC and MIC, indicating that cooperation-oriented moral structure contributes more strongly in open-ended social reasoning settings. In contrast, the hate-speech auxiliary objective primarily improves performance on politically charged or toxicity-related datasets, especially MFTC and MFRC.

These results suggest that the three components contribute complementary inductive biases: rationale supervision improves interpretability and robustness, MAC representations improve social-generalization capacity, and hate-speech supervision improves sensitivity to morally salient harmful content.

A.8 Per-Foundation Performance Analysis

Figure 6 presents dataset-level Macro-F1 and Figure 7 AUC comparisons under both in-domain and out-of-domain settings. Across datasets, CHARM maintains relatively stable degradation under distribution shift, with larger robustness gaps emerging primarily on short-text and socially ambiguous datasets such as SC and MIC.

Compared with in-domain evaluation, out-of-domain performance decreases are generally moderate, suggesting that the model captures transferable moral representations rather than relying solely on dataset-specific lexical cues. Notably, AUC scores remain comparatively stable even when F1 declines, indicating that the model preserves ranking consistency under domain shift despite threshold-sensitive classification degradation.

B Moral Alignment and Endorsement Prediction

B.1 Endorsement Network Construction

We construct a directed endorsement graph from the COVID-19 Twitter dataset introduced by Chen et al. (2020). Retweets are treated as endorsement signals following prior work (Metaxas et al., 2015), while mentions and quote tweets are excluded because they may reflect disagreement or contextual commentary.

For endorsement prediction, positive samples correspond to directed retweet interactions occurring more than five times between two users. Negative samples are constructed using a two-hop non-endorsement constraint to avoid trivially disconnected user pairs. The construction process is illustrated in Figure 4.

Figure 5 shows the distribution of users across the ten polarity-aware moral dimensions. Positive moral poles are substantially denser and more stable than negative poles, while several vice dimensions remain highly sparse in large-scale Twitter discourse. To reduce sparsity and improve robustness, we therefore collapse virtue and vice polarity scores into five foundation-level moral representa-

Table 11: AUC across in-domain and out-of-domain datasets in ablation studies. Best result per dataset in **bold**.

Variant	MFTCXplain$_{10d}$	MFTCXplain$_{5d}$	MFTC	MFRC	News	ARG	SC	MIC	VIG
baseline	0.81	0.77	0.85	0.84	0.75	0.78	0.69	0.68	0.82
w/o hate speech	**0.83**	**0.79**	**0.88**	**0.91**	**0.84**	0.84	0.75	0.71	0.87
w/o MAC	**0.83**	**0.79**	0.84	0.84	0.73	0.80	**0.79**	**0.75**	0.82
w/o rationale	0.82	0.78	0.87	0.90	0.82	0.86	0.77	0.73	**0.89**
CHARM	**0.83**	**0.79**	0.87	**0.91**	0.83	**0.87**	0.77	0.73	**0.89**

Table 12: Performance comparison between Qwen3-8B and CHARM across datasets in terms of Macro-F1 and AUC.

Dataset	Qwen3-8B		CHARM	
	F1	AUC	F1	AUC
MFTC	0.57	0.76	0.72	0.87
MFRC	0.46	0.76	0.72	0.91
News	0.39	0.66	0.55	0.83
ARG	0.48	0.79	0.54	0.87
SC	0.38	0.70	0.46	0.77
MIC	0.40	0.67	0.50	0.73
VIG	0.63	0.90	0.67	0.89

tions when constructing user-level moral profiles for endorsement analysis.

B.2 Backbone Comparison Results

Table 12 compares CHARM with Qwen3-8B across both in-domain and out-of-domain datasets. CHARM achieves consistently stronger performance on most datasets, with particularly large gains on MFTC, MFRC, and News. The improvements are also stable under out-of-domain evaluation, especially on ARG and SC, suggesting that the combination of rationale supervision and MAC grounding provides more robust moral representations than direct prompting alone. While Qwen3-8B remains competitive on VIG, overall results indicate that CHARM offers a stronger balance between efficiency and predictive performance.

B.3 User Moral Profile Construction

User-level moral profiles are constructed by aggregating tweet-level moral predictions inferred by CHARM over the five MFT foundations. Virtue and vice polarity scores are collapsed into unified foundation-level moral intensity scores. For each user u, moral scores are aggregated over high-intensity moral tweets:

$$m_u^{(k)} = \frac{1}{|T_u^{(k)}|} \sum_{t \in T_u^{(k)}} s_t^{(k)},$$

where $s_t^{(k)}$ denotes the predicted moral intensity for foundation k, and $T_u^{(k)}$ contains tweets whose intensity exceeds the user-specific 90th percentile threshold. Figure 5 visualizes the distribution of users across the ten polarity-aware moral dimensions.

B.4 Feature Construction

Table 13 summarizes the features used for endorsement prediction.

Pairwise moral similarity is computed using cosine similarity:

$$\text{cos_sim}(u, v) = \frac{\mathbf{m}_u^\top \mathbf{m}_v}{\|\mathbf{m}_u\| \|\mathbf{m}_v\|},$$

and Euclidean distance:

$$\text{euclidean_dist}(u, v) = \|\mathbf{m}_u - \mathbf{m}_v\|_2.$$

Per-foundation moral differences are defined as:

$$\text{diff}^{(k)}(u, v) = m_u^{(k)} - m_v^{(k)}.$$

Behavioral statistics are aggregated using:

$$\text{activity_mean}(u) = \frac{1}{|T_u|} \sum_{t \in T_u} a_t,$$

where a_t denotes tweet-level engagement statistics such as replies, favorites, quotes, and retweets.

B.5 Endorsement Prediction Model Training Details

We train a binary LightGBM classifier using moral-profile, interaction, behavioral, and semantic features. The model is trained with a binary logistic objective and evaluated across five random seeds.

We use a learning rate of 0.05, maximum tree depth of 6, 200 boosting rounds, and 64 leaves per tree. Early stopping is applied based on validation-set AUC with a patience of 20 rounds. To reduce overfitting, we apply both feature and row subsampling with rates of 0.8 together with L2 regularization coefficient 1.0.

Table 13: Feature definitions for endorsement prediction.

Feature Category	Feature	Description
Moral Features		
Author / Retweeter	Care, Fairness, Loyalty, Authority, Purity	User-level moral intensity scores aggregated over the five moral foundations using polarity-collapsed representations.
Moral Interaction Features		
User Pair	Moral Similarity	Cosine similarity between author and retweeter moral representations.
User Pair	Moral Distance	Euclidean distance between author and retweeter moral representations.
User Pair	Foundation-Level Moral Difference	Per-foundation differences between retweeter and author moral scores.
Behavioral Features		
User-Level	Social Connectivity	User engagement statistics including follower count, friend count, and favourites count.
User-Level	Interaction Activity	Average frequencies of replies, quotes, and retweets across authored tweets.
User Pair	Relative Influence Gap	Differences in user influence and activity between author and retweeter.
Semantic Features		
User Pair	Biography Similarity	Cosine similarity and Euclidean distance between user biography embeddings.

All continuous features are z-score normalized before training. Semantic similarity features are computed using cosine similarity between user-level sentence embeddings, while moral similarity features are computed using cosine similarity between aggregated user moral profiles inferred by CHARM. Reported results are averaged across all runs.

B.6 Per-Foundation Moral Alignment Distributions

Figure 8 shows per-foundation distributions of moral divergence (|retweeter – author|) and overall cosine similarity for endorsed and non-endorsed pairs. Across all five foundations, endorsed pairs exhibit tighter moral alignment — smaller author–retweeter divergence — than non-endorsed pairs, confirming that homophily holds at the individual-foundation level. The separation is most pronounced for loyalty and fairness, consistent with the assortativity coefficients in Figure 3c.

B.7 Moral Assortativity Computation

We measure moral homophily using Newman assortativity for continuous attributes. Given a directed network with edges $(i \to j) \in E$, each edge is associated with source and target moral scores for a specific foundation. The assortativity coefficient is computed as the Pearson correlation across edges:

$$r = \frac{\sum_{(i,j)\in E}(x_i - \bar{x}_s)(y_j - \bar{y}_t)}{\sqrt{\sum_{(i,j)\in E}(x_i - \bar{x}_s)^2}\sqrt{\sum_{(i,j)\in E}(y_j - \bar{y}_t)^2}}$$

where $\bar{x}_s$ and $\bar{y}_t$ denote the edge-weighted means of source and target attributes. We compute assortativity separately for each moral foundation. Statistical significance is assessed using a degree-preserving null model in which edges are randomly rewired while preserving each node's in-degree and out-degree distributions.

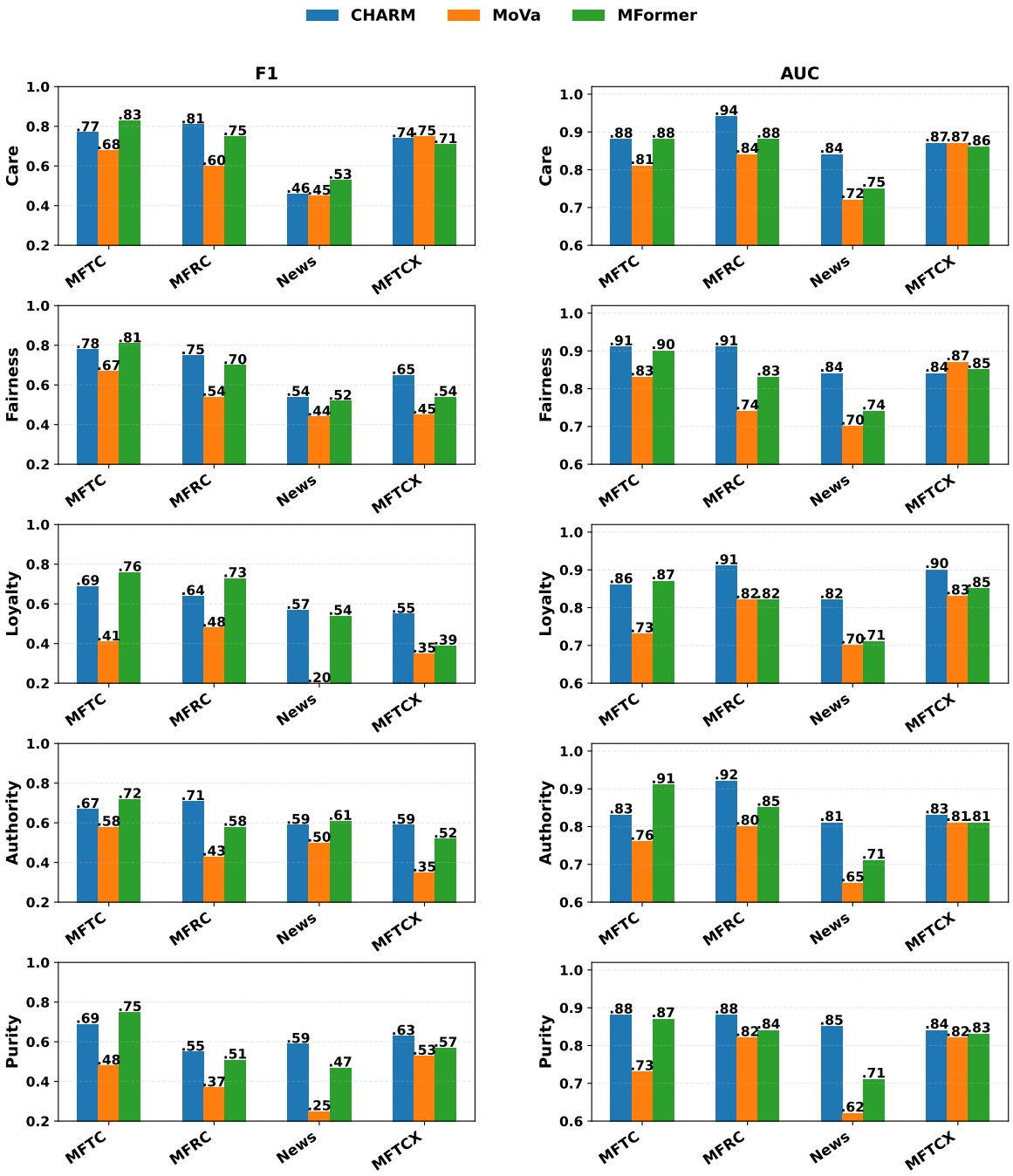


Figure 6: In-domain evaluation: F1 and AUC across datasets.

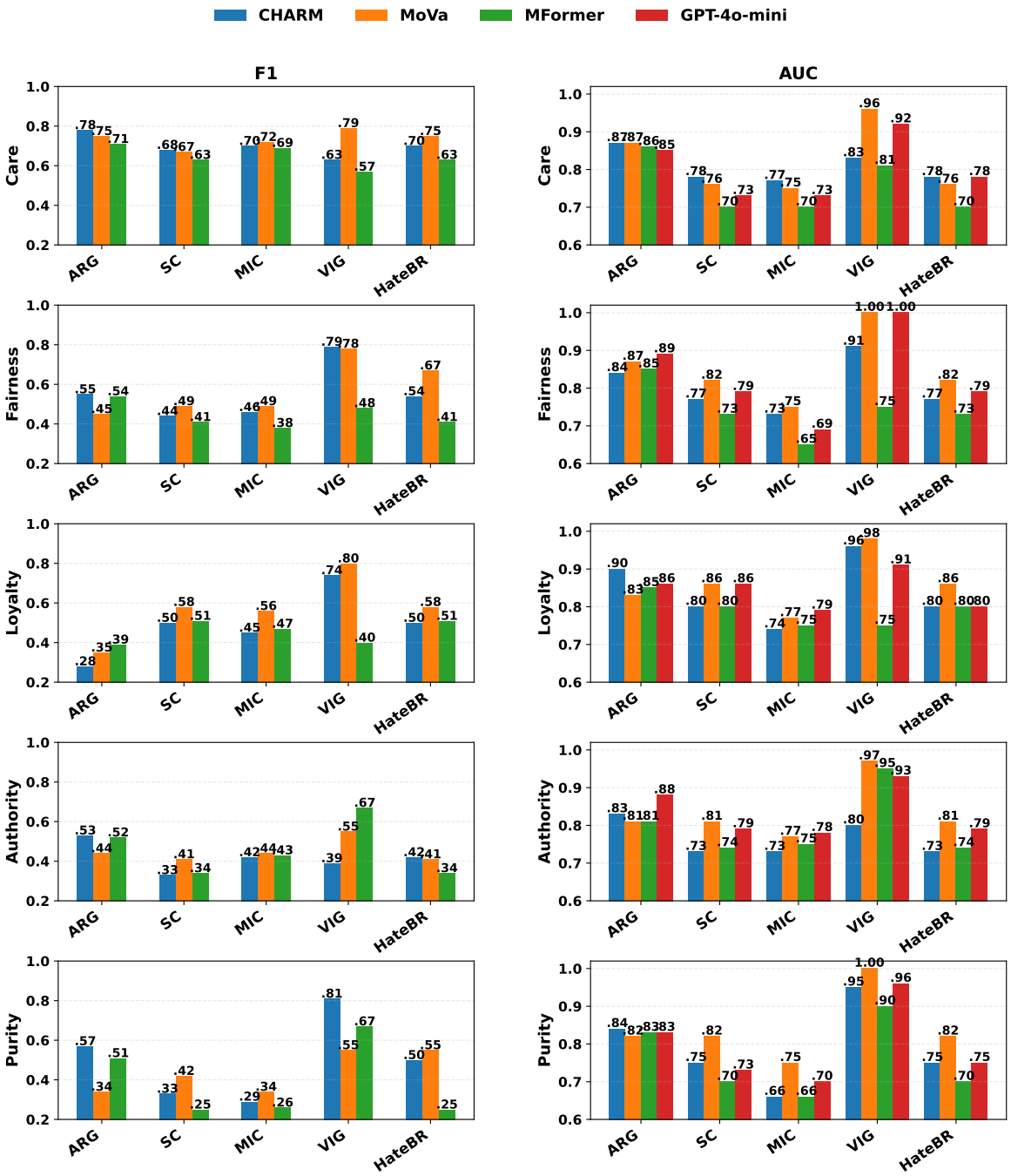


Figure 7: Out-of-domain evaluation: F1 and AUC across datasets.

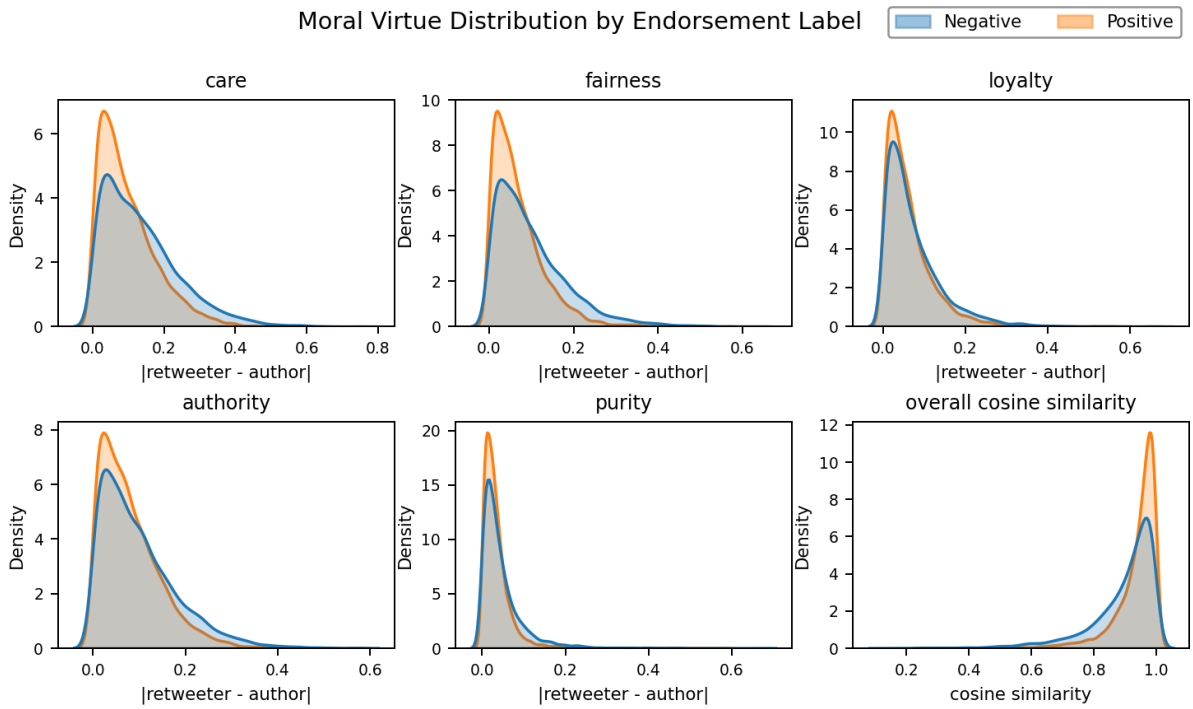


Figure 8: Per-foundation KDE distributions of |retweeter moral score − author moral score| for endorsed (positive) and non-endorsed (negative) pairs, across the five MFT foundations and overall cosine similarity (bottom right). Endorsed pairs consistently exhibit smaller per-foundation divergence and higher cosine similarity.