Exploring LLMs' Ability to Spontaneously and Conditionally Modify Moral Expressions through Text Manipulation

Candida Maria Greco
DIMES Department
University of Calabria
Rende, Italy

Lucio La Cava
DIMES Department
University of Calabria
Rende, Italy

Lorenzo Zangari
DIMES Department
University of Calabria
Rende, Italy

Andrea Tagarelli *
DIMES Department
University of Calabria
Rende, Italy

{candida.greco,lucio.lacava,lorenzo.zangari,tagarelli}@dimes.unical.it

Abstract

Morality serves as the foundation of societal structure, guiding legal systems, shaping cultural values, and influencing individual self-perception. With the rise and pervasiveness of generative AI tools, and particularly Large Language Models (LLMs), concerns arise regarding how these tools capture and potentially alter moral dimensions through machine-generated text manipulation. Based on the Moral Foundation Theory, our work investigates this topic by analyzing the behavior of 12 LLMs among the most widely used Open and uncensored (i.e., "abliterated") models, and leveraging human-annotated datasets used in moral-related analysis. Results have shown varying levels of alteration of moral expressions depending on the type of text modification task and moral-related conditioning prompt.

1 Introduction

Morality serves a cornerstone in shaping societies, influencing legal systems, socio-cultural norms, and individual identities (Schwartz, 1992; Ellemers, 2018; Kádár et al., 2019; Hofmann et al., 2014).

Language is central to communicating morality, as moral values are conveyed through word choice, framing, and rhetoric, often embedded in subtle linguistic patterns (Kennedy et al., 2021b). Understanding this communication is essential for analyzing its societal impact, especially in the digital age due to the advent of generative AI, with large language models (LLMs) playing an increasingly dominant role in creating or editing textual content. Trained on world-scale crowdsourced corpora, these models learn lexical, semantic, and factual information that also capture a wide range of cultural and moral biases embedded within the language (Schramowski et al., 2022; Abdulhai et al., 2024; Hämmerl et al., 2023). While this enhances their ability to recognize moral aspects in text (Guo et al., 2023; Zhang et al., 2023; Preniqi et al., 2024), it also fosters implicit moral tendencies during training, influencing the language they generate.

Problem Statement, Research Questions and Hypothesis. Despite existing studies on moral-related tasks based on LLMs, there is however a lack of understanding regarding the way moral values are communicated through language that is subjected to a review or editing process by LLMs. Our research aims to fill this gap by exploring the following central question: how do generative LLMs influence moral expressions when modifying human-authored content? More specifically, we investigate the following research questions:

- **RQ0**: *Are LLMs aware of a psychological conceptualization of moral foundations?*
- **RQ1**: *How does model-generated text editing influence the moral expressions in the modified text?*
- **RQ2**: *How do LLMs behave when prompting them to emphasize the strength of any moral expressions detected in a text?*
- **RQ3**: *How do LLMs respond when prompted to amplify or weaken moral expressions associated with a specific moral dimension in a text?*

Assuming that the LLMs under evaluation are adequately informed about moral foundations (RQ0), we hypothesize that they are more likely to alter the strength of moral expressions in a text as the level of text manipulation increases (RQ1). This effect might be further amplified when the LLM is instructed to focus on the evidence of moral dimensions (RQ2), or even more when conditioned on particular moral dimensions (RQ3).

Contributions. We aim to explore the intrinsic ability of LLMs to (i) spontaneously or (ii) conditionally alter the expressions of moral dimensions in a text after manipulating the contents therein to some extent. To the best of our knowledge, this is

*Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18047–18070

the first study to pursue this objective, based on the set of RQs previously stated.

Our research builds on six key methodological components: the *Moral Foundation Theory* (MFT) as the reference theoretical backbone, *human-annotated datasets* from various domains, a comprehensive selection of *Open LLMS* (OLLMs) for machine/hybrid-generated text processing, a set of *text-manipulation prompt types*, *moral foundation prediction* models for the evaluation of the generated texts, and assessment criteria based on *moral shift measures*.

It should be noted that MFT is a well-known psychological framework for conceptualizing the core moral foundations that shape moral reasoning, which has been widely used to explore moral perspectives across cultures. Our choice of LLMs falls into the landscape of open models, aligning not only with our vision of cost-free accessibility and openness in research, but also with an additional criterion of *worldwide coverage* that cannot equally be provided by commercially-licensed models. Moreover, we include *abliterated LLMs*, i.e., uncensored models that bypass the refusal mechanism, for a total of 12 LLMs under examination. These are prompted using a set of instructions that differ in terms of types of text-modification and conditioning on the moral-targeted text-manipulation. Concerning the moral foundation scoring models, a key requirement is generalizability across domains, while our defined assessment criteria are designed to capture both magnitude change and rank-based change of the moral dimension importance in the generated texts. Finally, our selected evaluation datasets can be regarded as de-facto benchmarks for moral-related NLP tasks involving LLMs.

We release our resources at https://mlnteam-unical.github.io/resources/

2 Preliminaries

2.1 Moral Foundation Theory

Our study is grounded in the *Moral Foundations Theory (MFT)* (Haidt and Joseph, 2004; Atari et al., 2023, 2020),[1] which provides a theoretical framework for operationalizing the concept of human morality. The original framework of MFT identified five *foundations*, which are strongly supported by evidence across various cultures. These foundations, or dimensions, are expressed as vice/virtue dichotomies: *Care/Harm* (CH), focusing on empathy and protection versus infliction of suffering; *Fairness/Cheating* (FC), centered on upholding justice and integrity versus deceit and exploitation; *Loyalty/Betrayal* (LB), promoting allegiance to one's group versus acts of betrayal; *Authority/ Subversion* (AS), valuing obedience to societal norms and traditions versus challenges to authority; *Purity/Degradation* (PD), emphasizing the sanctity of what is considered sacred versus its defilement.

[1] https://moralfoundations.org/publications/

2.2 Related Work

MFT is the cornerstone of most of the existing works that analyze and detect moral dimensions through modern NLP tools. In recent years, research in this field has evolved along two main directions (Zangari et al., 2025a): (i) training models on MFT-based data for the task of moral foundation prediction, and (ii) analyzing the moral foundations reflected in model responses and embeddings.

Moral foundation prediction. Early approaches to moral foundation prediction are lexicon-based, using word lists associated with moral foundations, particularly the Moral Foundations Dictionary (MFD) (Graham et al., 2009), and its extensions (Hopp et al., 2021; Frimer, 2019). Moral-Strength (Araque et al., 2020) is one such approaches, which enhances the MFD by quantifying the relevance and strength of words associated with MFT. Distributed Dictionary Representations (DDR) (Garten et al., 2018) integrates the MFD with word embeddings, capturing moral concepts in semantic spaces. However, relying on predefined word lists, these methods often lack adaptability across varied linguistic contexts.

Approaches based on pre-trained language models have recently gained increased attention, typically relying on encoder-only architectures. DAMF (Guo et al., 2023) and MoralBERT (Preniqi et al., 2024) fine-tuned BERT on different types of data sources for moral foundation prediction. Their effectiveness has been shown mainly in in-domain scenarios. To overcome limitations in out-of-domain scenarios (i.e., where a significant domain-shift occurs between training and evaluation data), ME2-BERT (Zangari et al., 2025b) leverages events and emotions to align data from different domains into a shared vector space and to facilitate the detection of morally relevant text segments. ME2-BERT has also shown to achieve strong performance even when compared against recent LLMs.

Moral foundation assessment. A common approach is to administer LLMs with questionnaires that have been validated with human participants (e.g., (Graham et al., 2011)), mostly on politics (Abdulhai et al., 2024; Simmons, 2023) or controversial topics (He et al., 2024). Nunes et al. (2024) test LLMs in realistic moral dilemmas, assessing consistency between their learned abstract principles (via MFQ) and concrete moral decisions. Other studies probe the models' embeddings to examine how moral values are encoded (Fraser et al., 2022; Kennedy et al., 2021a; Xie et al., 2020).

Our work is a unique hybrid study bridging the above two lines of research, as we adopt moral-foundation prediction models to assess how text manipulation by LLMs influences the moral expressions therein. A key novelty of our work is examining the impact of different levels of text manipulations in various settings of moral-conditioning of a comprehensive set of generative LLMs.

3 Resources

Datasets. We employed *five human-annotated datasets* from various domains, including social media and news, which have been widely used in moral-related analysis tasks: **The Moral Foundations Twitter Corpus (MFTC)** (Hoover et al., 2020) consists of 32,218 tweets spanning various sociopolitical and cultural contexts, such as Black-Lives Matter and MeToo movements, Baltimore Protests, Hurricane Sandy, 2016 US presidential election, and hate-speech tweet collection. **The Moral Foundations Reddit Corpus (MFRC)** (Trager et al., 2022) consists of 16,123 English Reddit posts from 12 subreddits, categorized into US politics, French politics, and everyday moral life. **The Moral Foundations News Corpus (MNFC)** (Weber et al., 2021) consists of 35,935 news articles from major outlets published between 2013 and 2015, which was annotated based on the Moral Foundations Dictionary and an online annotation platform. **Moral Event (ME)** (Zhang et al., 2024) consists of 12,355 news articles, based on about 5.5K event annotations, published by different media outlets on US politics from 2012 to 2022, including abortion, gun control, and public health. **EMONA** (Lei et al., 2024) contains about 10,815 sentences annotated with event-level moral opinions. It integrates three datasets covering political and social issues at different levels of granularity.

Note that all datasets contain sentences originally annotated by experts with one or more MFT dimensions, or a *non-moral* label. Our data preprocessing is described in Appendix A.

From	Model	Abbrev.	Params
US	`Llama-3.1-8B-Instruct`	Llama3	8.03B
	`Phi-3.5-mini-instruct`	Phi	3.82B
EU	`Mistral-7B-Instruct-v0.3`	Mistral	7.25B
	`EuroLLM-9B-Instruct`	EuroLLM	9.15B
China	`Qwen2.5-7B-Instruct`	Qwen	7.62B
	`Yi-1.5-9B-Chat`	Yi	8.83B
UAE	`Falcon3-7B-Instruct`	Falcon	7.22B
World	`aya-expanse-8b`	Aya	8.03B

Table 1: LLMs selected for our study, annotated with their geographic "location" and number of parameters.

Model	Abbrev.	Params
`Meta-Llama-3.1-8B-Instruct-abl.`	Llama3-abl	8.03B
`NeuralDaredevil-8B-abl.`	NeuralDD	8.03B
`Qwen2.5-7B-Instruct-abl-v2`	Qwen-abl	7.62B
`Phi-3-mini-128k-instruct-abl-v3`	Phi-abl	3.82B

Table 2: Abliterated LLMs selected for our study, annotated with their ID (abl. stands for abliterated) and number of parameters.

Generative Models for Text Manipulation. We considered a representative selection of Open LLMs varying by sizes and architectures, for which we accessed their publicly available implementations on the HuggingFace Model Hub[2] at the end of 2024. Our rationale was to select models spanning various geographic areas to assess whether different "cultures" underlying the models might impact the moral dimensions after manipulation.

In addition, we included a batch of *abliterated* (or uncensored) models (Arditi et al., 2024), aiming at analyzing the impact the lack of safeguarding has on morality. Tables 1-2 summarize the main characteristics of the LLMs selected in this study.

We considered two *temperature* settings, 0.1 and 1.0, in order to test the LLMs under a less or more random generation setting. Also, we kept parameters *top_p* and *top_k* to their default values of 1 and 50, respectively, to ease reproducibility.

We used the `vllm` inference and serving library[3] on a 8x NVIDIA A30 GPU server with 24 GB of RAM each, 764 GB of system RAM, a Double Intel Xeon Gold 6248R with a total of 96 cores, and Ubuntu Linux 20.04.6 LTS as OS.

[2] https://huggingface.co/

[3] https://github.com/vllm-project/vllm

4 Methodology

Overview. Referring back to the components outlined in the Introduction, our analysis framework requires that our selected Open LLMs are fed with human-annotated datasets and prompted to manipulate them according to various text modification tasks. The generated outcomes are then annotated by moral-foundation-scoring models and assessed based on moral shift measures. Next we elaborate on each of these components and their flow of interaction.

Moral Foundation Scoring Models. We considered a selection of state-of-the-art tools for moral-foundation prediction, namely ME2-BERT, MoralBERT, MoralStrength, and DDR (cf. Sect. 2), and assessed their agreement with the human-annotated datasets. To this aim, we identified the data instances over all datasets that were human-annotated with at least one moral dimension, and counted the matchings achieved by each of the above models—since the models provide continuous scores in [0,1] for every moral dimension, while human annotations are binary, we applied a 0.5 threshold to determine whether a given moral dimension was present.

The results in Table 3 show that MoralStrength and DDR have the lowest matching, while MoralBERT performs best on MFTC and MFRC. However, this is not surprising, as the model was fine-tuned on these datasets. More interestingly, ME2-BERT achieves a relatively lower number of matched instances, despite not being trained on MFTC and MFRC, while turning out to be the best-matching model in two of the other datasets.

Overall, ME2-BERT should be regarded as preferable to MoralBERT due to its higher versatility, efficiency and out-of-domain abilities. In fact, MoralBERT was designed for single-label classification, thus requiring to execute a separate classifier for each moral dimension. Moreover, there is no version of MoralBERT designed to detect non-moral content, i.e., this could be inferred indirectly by checking that all moral-dimension classifiers' outcomes remain below a certain threshold. These MoralBERT's limitations are absent in ME2-BERT; furthermore, as shown in (Zangari et al., 2025b), ME2-BERT generally aligns more closely with human-assigned moral labels than MoralBERT across a range of test datasets.

Benchmark Data Selection. Having chosen ME2-BERT as our reference model for moral foundation scoring, our final step was to build a robust corpus of texts that will be administered to our generative models for text manipulation, with the outcomes evaluated by ME2-BERT.

Dataset	#ma	ME2BERT	MoralBERT	DDR	MoralStrength
MFTC	25,397	2972	3688	1804	463
MFRC	10,058	1186	1368	545	66
MFNC	35,935	5365	4824	1897	637
ME	4144	294	250	175	17
EMONA	5166	673	637	306	16
All	80,700	10,490	10,767	4727	1199

Table 3: Matchings between model-annotations and human-annotations. Column '#ma' contains the total number of instances that were human-annotated with at least one moral dimension.

From each of the datasets we selected those instances that were perfectly matched by ME2-BERT w.r.t. the original human-annotations. These correspond to the third column in Table 3. In addition, we included an equal number of instances that were human-annotated as "non-moral" and matched by ME2-BERT, for a total of **20,980 annotated instances** selected from all datasets. Appendix B provides statistics on the agreement of the models on this subset.

Text Modification Tasks. We examined three types of text modification, each representing a different degree of machine influence on the output:

- **Revision**: the LLM is asked to improve or refine a given text while correcting errors, inconsistencies, or awkward phrasing; it typically yields a text that reads better while retaining much of the original. Since this task requires the LLM to stay relatively closer to the original text, we identify it as the least impactful text manipulation.
- **Rewriting**: the LLM is asked to restructure the content, while potentially improving clarity, tone, and readability. Rewriting can result in a version of the text that substantially differs in tone and style from the original, while conveying the same purpose. Therefore, it is regarded as more impactful than revision.
- **Continuation**: the LLM is asked to generate a continuation of a given text, producing content that remains consistent with the given one. Unlike revision and rewriting, which focus on modifying existing text, continuation requires the model to predict and extend the content in a coherent, contextually appropriate manner. This can hence be highly impactful, as the generated text may introduce new ideas, expand arguments, or develop nar-

RQ1-RQ3 Prompting

RQ1 Prompt Message:
Please, $[revise/rewrite/continue]$ *the following text:*
`{text}`
Do not report the changes you made nor comments on them, just return the $[revised/rewritten/continued]$ *text.*

.....

RQ2 Prompt Message:
Please, $[revise/rewrite/continue]$ *the following text:*
`{text}`
If you detect any moral dimension, please emphasize it or make it more explicit in the revised text. If you do not detect any moral dimension, return the original text. Do not add any other text or comment.

.....

RQ3 Prompt Message:
You are given a text that contains `{a specific moral dimension}`*. Please,* $[revise/rewrite/continue]$ *the following text to* $[intensify/diminish]$ *it (i.e., make it* $[more/less]$ *prominent in the* $[revision/rewriting/continuation]$ *of the text). Do not add any other text or comment.*
`{text}`

Figure 1: **(RQ1-RQ3)** Prompt instructions for unconditioned (top), moral-blind conditioned (middle), and moral-informed conditioned (bottom) text editing.

ratives beyond what was originally present.

Prompt Types. To address our RQs, we designed different prompts for the selected LLMs.

The first prompt refers to RQ0, since it asks the models to describe the MFT, in order to assess their awareness of it. This is performed under a minimum-temperature setting only, to reduce the randomness in a model's response, thus ensuring consistency with its learned world-knowledge. RQ0 prompt is reported in Appendix C.

The main RQs, i.e., RQ1-RQ3, are addressed by developing three types of *text manipulation* prompts, which differ from each other in terms of model-conditioning approach:

• **Unconditioned**: the model is asked to perform the required text manipulation task without any moral-related bias. The goal is to understand the inherent ability of a model to spontaneously alter the moral expressions in a text (RQ1).

• **Moral-blind conditioned**: the model is asked to make it more explicit *any* moral dimension expressed in the original text; however, the model is not explicitly informed about which specific moral dimension to look for, i.e., it is left to detect any moral dimension it identifies on its own (RQ2).

• **Moral-informed conditioned**: in contrast to moral-bind conditioning, the prompt here specifies a particular moral dimension the model has to focus on. Moreover, this prompt type involves *moral intensity modulation*, since the model is required either to increase or diminish the strength of the given moral expression (RQ3).

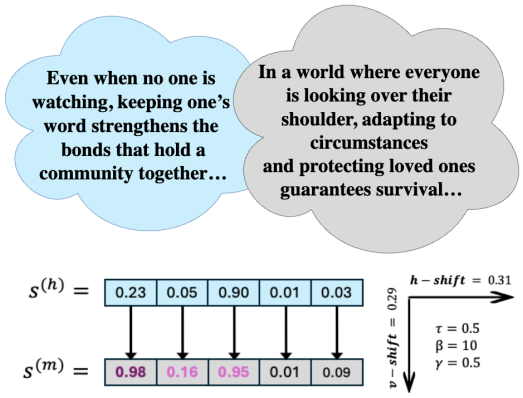


Figure 2: Example of moral changes after text manipulation according to our $h\text{-}shift$ and $v\text{-}shift$ criteria.

Figure 1 shows the prompt templates used for RQ1-RQ3. Note that the models are required not to add comments or explanations, to avoid introducing extra tokens that would impact moral scoring; the latter was semi-automatically checked before processing texts for the moral foundation scoring. Each of the prompts was repeated under both models' temperature settings (cf. Sect. 3).

Moral Shift Evaluation. Given a human-written text and an associated manipulated text, let $\mathbf{s}^{(h)} \in [0,1]^5$ and $\mathbf{s}^{(m)} \in [0,1]^5$ denote the vectors of scores assigned to the 5 MFT dimensions, provided by the reference human-annotator-aligned moral-foundation-scoring model and the reference LLM, respectively. Note that, since human annotations are binary indicators of presence for moral dimensions, we employ the continuous prediction provided by the reference moral-foundation-scoring model to enable the quantitative measurement of moral shifts; nonetheless, the alignment with human judgment is preserved, as we restrict the analysis to the subset of instances on which ME2-BERT achieves 100 % agreement with the annotators.

Our goal is to measure the *moral shifts*, i.e., the changes in moral expressions into a text after its manipulation by the model. This can be regarded as a twofold objective, since two complementary aspects are (i) the magnitude change in the scores associated to the moral dimensions, and (ii) how

	Llama3	Phi	Mistral	EuroLLM	Qwen	Yi	Falcon	Aya
WO	0.190	0.239	0.166	0.215	0.215	0.166	0.180	0.229
Sim	0.902	0.826	0.867	0.909	0.894	0.768	0.901	0.872

Table 4: **(RQ0)** Models' MFT awareness expressed via Word Overlap (WO) and Cosine Similarity (Sim).

much the ranking of the importance (scores) of the moral dimensions has changed. In the following, we provide our definitions for measuring both aspects, hereinafter referred to as *vertical shift* and *horizontal shift*, respectively. Appendix F discusses properties, while Fig. 2 shows an example of moral shifts.

Vertical Shift. To measure how the moral dimension scores have changed after text manipulation, we could simply compute the mean absolute difference over the dimension scores. However, this approach or others based on standard error measures (e.g., MAE, MSE) would be unable to capture both the *signed* change in intensity of each dimension and the impact of deviating from a certain threshold τ that might be used as a *decision boundary* to toggle on/off a moral dimension's signal.

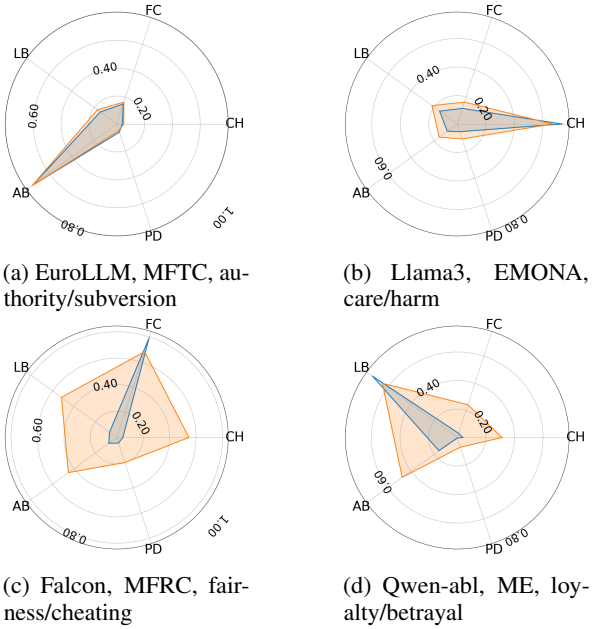
Let us define function $\phi(x_i, \gamma, \beta, \tau) = x_i + \gamma\sigma(\beta(x_i - \tau))$, where $\sigma(\cdot)$ is the logistic function, $\beta > 0$ is the logistic growth rate (i.e., it controls how sharply mid-range values are emphasized), τ is the value of the function's midpoint, and $\gamma > 0$ acts as a scaling factor. Intuitively, the above defined function augments a value x_i with a nonlinear transformation that emphasizes values above/below τ by a magnitude determined by γ. Note that, if $x_i \in [0, 1]$, $\phi(\cdot)$ ranges within $[0, 1 + \gamma]$. Since β, γ and τ are fixed, we simplify the notation by writing the function as $\phi(x_i)$.

Based upon function ϕ, we define the vertical shift of $\mathbf{s}^{(m)}$ w.r.t. $\mathbf{s}^{(h)}$ as follows:

$$v\text{-}shift(\mathbf{s}^{(m)}, \mathbf{s}^{(h)}) = \frac{1}{5}\sum_{i=1}^{5} \phi(s_i^{(m)}) - \phi(s_i^{(h)}).$$

This ranges within $[-1-\gamma, 1+\gamma]$, where positive, resp. negative, values indicate how much, on average, the moral dimension scores have increased, resp. decreased. $s_i^{(m)}$ and $s_i^{(h)}$ denote the score of the i-th moral dimension before and after the text manipulation process, respectively.

In our experiments, we will use $\tau = 0.5$, $\beta = 10$ for a moderate mid-range sensitivity, and $\gamma = 0.5$ to keep the transformation smooth yet effective in emphasizing mid-range shifts.



(a) EuroLLM, MFTC, authority/subversion

(b) Llama3, EMONA, care/harm

(c) Falcon, MFRC, fairness/cheating

(d) Qwen-abl, ME, loyalty/betrayal

Figure 3: **(RQ1)** Examples of changes in Revise (a), Rewrite (b), Continue (c-d) settings on moral-dimension-specific subsets. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.

Horizontal Shift. A well-suited measure for the horizontal, i.e., rank-based, changes in the moral dimensions is the difference between the two ranking orderings $\mathbf{r}^{(h)}$ and $\mathbf{r}^{(m)}$ derived from $\mathbf{s}^{(h)}$ and $\mathbf{s}^{(m)}$, respectively. This can be formalized as the Kendall tau distance between $\mathbf{r}^{(h)}$ and $\mathbf{r}^{(m)}$, which is the number of pairwise disagreements (i.e., discordant pairs), normalized to scale within [0,1]:

$$h\text{-}shift(\mathbf{s}^{(m)}, \mathbf{s}^{(h)}) = \\ = \frac{1}{10}\sum_{i<j\in[1..5]} \mathbb{1}\Big(\big[r_i^{(h)} - r_j^{(h)}\big]\,\big[r_i^{(m)} - r_j^{(m)}\big] < 0\Big),$$

where $\mathbb{1}$ is the indicator function. Values closer to 1 correspond to greater changes in the ranking.

5 Results

5.1 RQ0: LLM's awareness of MFTs

We first assessed the lexicon used by LLMs to describe MFT by measuring the *word overlap* (WO) between the reference MFT description [4] and the ones generated (based on the prompt reported in Appendix C). We define this overlap as $WO(t_1, t_2) = \frac{|t_1 \cap t_2|}{\min(|t_1|, |t_2|)}$, where t_1 and t_2 denote two texts. In addition, we measured their semantic similarity by encoding the reference and the

[4] https://moralfoundations.org/

models' descriptions using the *all-mpnet-base-v2* [5] sentence-embedding model. Table 4 shows low overlap values, which indicate a different lexicon by LLMs than our reference source for MFT. However, the high semantic similarities suggest that, despite the jargon differences, all models grasp MFT and capture its core traits.

5.2 RQ1: Unconditioned Manipulation

The unconditioned text-manipulation tasks yielded outcomes that highlight the following key findings. Under revise or rewrite, the models tend to produce texts that substantially keep the original text's overall level of moral intensity, as shown by negligible $v\text{-}shift$ values (ranging between -0.03 and 0.05, cf. Table 5), although the relative importance of the moral dimensions shows more evident changes, with average $h\text{-}shift$ in (0.08, 0.22) for revise, and in (0.12, 0.25) for rewrite. This generally holds regardless of the temperature setting, while news data are slightly more affected by $h\text{-}shift$ changes.

When prompted to continue a text, the models tend to emphasize not only the most important moral dimension, but also, in most cases, the other moral dimensions (cf. Figure 3), making them closer in scores to the dominant one, with average $v\text{-}shift$ ranging in (0.22, 0.45). This also causes an evident $h\text{-}shift$ change, on average within (0.28, 0.31). Higher temperature slightly influences $h\text{-}shift$ values, with increments up to 0.1.

For texts originally labeled as non-moral, revise and rewrite tasks do not introduce particular evidence of moral dimensions, regardless of the model and temperature settings. In the continue setting, we notice a tendency of all models to produce texts with evidence for some moral dimensions, mostly Authority/Subversion and Loyalty/Betrayal.

Qwen, Qwen-abl and EuroLLM are the models that most preserve the original text's moral expressions (i.e., lowest $h\text{-}shift$ and absolute $v\text{-}shift$ values in Table 5), whereas the highest changes (in the continue setting) correspond to Llama3-abl, Phi-abl and Falcon in $h\text{-}shift$ and to Phi, Mistral and NeuralDD in $v\text{-}shift$.

5.3 RQ2: Moral-blind Conditioned Manipulation

When explicitly prompted to emphasize moral expressions, LLMs consistently amplify or introduce the moral content across all evaluated settings. Most models exhibit values of $h\text{-}shift$ above 0.3 and $v\text{-}shift$ above 0.5. Particularly, the continue operation leads to significantly amplify the moral expressions, with peaks around or above 1 for Phi, Mistral and Aya (see Fig. 4 as an example on MFTC). By contrast, Qwen-based models tend to behave much closely to the unconditioned manipulation setting (cf. Table 5). Moreover, news texts appear to be more impacted than social media texts, with an increase in $v\text{-}shifts$ up to 0.15.

A common trait to all models is that, when moral cues are present in the original text, all models tend to emphasize them. Also, when the text appears morally neutral, the tendency is to introduce moral expressions extensively. While the continue setting is the most altering of the moral expressions, revise and rewrite have comparable effects but with the former being slightly more impacting, unlike what observed for the unconditioned prompting. This might be explained since, while keeping the original structure and meaning intact, a revision might refine the wording in a way that makes that moral dimension(s) even more pronounced than a rewrite, which might dilute or shift the tone, and hence, the moral expression in the text.

Similarly to our observations for the unconditioned prompting, with the exception of LLama3-abl, the abliterated models tend to alter the moral expressions less than the other models; this particularly holds for Qwen-abl and Phi-abl vs. their non-abliterated counterparts.

5.4 RQ3: Moral-informed Conditioned Manipulation

Looking at Table 5, we notice generally lower values than those observed for the moral-blind conditioning. This is actually not surprising since, when prompted to alter the expressions of a particular moral dimension, regardless of the intensify or diminish instruction, a model would produce a modified text where the $h\text{-}shift$, and especially the $v\text{-}shift$ depend on a localized effect, rather than a more generalized effect on moral dimensions.

Revision, rewriting and continuation correspond to increasing impact on the alteration of moral expressions, with a general tendency for 'intensify' to create a larger gap than 'diminish' (i.e., absolute $v\text{-}shift$ for 'intensify' higher than for 'diminish'), as shown in the summary of Table 5 and in Fig. 5.

The abliterated models exhibit the most significant decrease of the moral tone when explicitly prompted; particularly, Llama3-abl, Phi-abl, and

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

	Task	Llama3	Phi	Mistral	EuroLLM	Qwen	Yi	Falcon	Aya	Llama3-abl	NeuralDD	Qwen-abl	Phi-abl
RQ1	Rev.	.112/.018	.199/-.014	.198/.005	.095/.001	.105/.012	.151/.009	.124/.005	.168/**.039**	.152/.018	.144/.019	.082/.011	**.216**/-.026
	Rew.	.212/.015	.221/.022	.209/.021	.147/.015	.158/.007	.213/.025	.179/.011	.232/**.049**	**.245**/.012	.213/.020	.123/.004	.210/-.006
	Cont.	.304/.386	.303/**.454**	.302/.430	.284/.405	.289/.388	.293/.428	.307/.400	.300/.339	**.313**/.392	.305/.429	.277/.223	.308/.340
RQ2	Rev.	.265/.505	**.329**/**.853**	.325/.656	.310/.571	.201/.304	.288/.345	.240/.209	.309/.526	.291/.601	.296/.622	.172/.177	.298/.327
	Rew.	.273/.520	**.325**/**.724**	.321/.584	.273/.549	.201/.283	.285/.332	.245/.187	.305/.539	.294/.575	.297/.613	.193/.190	.312/.207
	Cont.	.347/.691	.351/**1.112**	.356/1.030	.337/.938	.308/.590	.346/.956	.343/.094	.354/1.003	.360/.717	.359/.862	.285/.352	**.370**/.500
RQ3	Rev. ↑	.185/.469	.178/.419	.196/.365	.129/.242	.158/.350	.198/.455	.181/.384	.207/.398	**.215**/**.704**	.196/.579	.129/.196	.192/.363
	Rew. ↑	.196/.462	.184/.417	.195/.318	.136/.242	.165/.400	.209/.550	.186/.351	.212/.483	**.222**/**.705**	.202/557	.137/.215	.190/.310
	Cont. ↑	.238/.667	.220/**.884**	.220/.793	.204/.652	.188/.546	.233/.856	.218/.791	.238/.824	**.250**/.759	.224/.742	.167/.330	.245/.809
	Rev. ↓	.207/-.074	.219/-.120	.227/-.081	.109/.059	.152/-.089	.190/-.033	.174/-.063	.200/.055	**.252**/-.115	.227/-.134	.105/-.056	.244/-**.207**
	Rew. ↓	.228/-.094	.232/-.138	.236/-.122	.115/.051	.179/-.120	.205/-.053	.191/-.085	.210/.043	**.265**/-.116	.241/-.145	.126/-.080	.241/-**.213**
	Cont. ↓	.279/.094	.269/.432	.264/.317	.206/.533	.211/-.066	.267/.615	.229/.129	.257/**.616**	**.296**/.093	.268/.038	.141/-.070	.260/.080

Table 5: Summary of per-model $h\text{-}shift$ (left) and $v\text{-}shift$ (right) values averaged over all datasets. ↑, ↓ denote intensify and diminish, respectively. Bold values correspond to highest $h\text{-}shift$ and (absolute) $v\text{-}shift$ in each row.

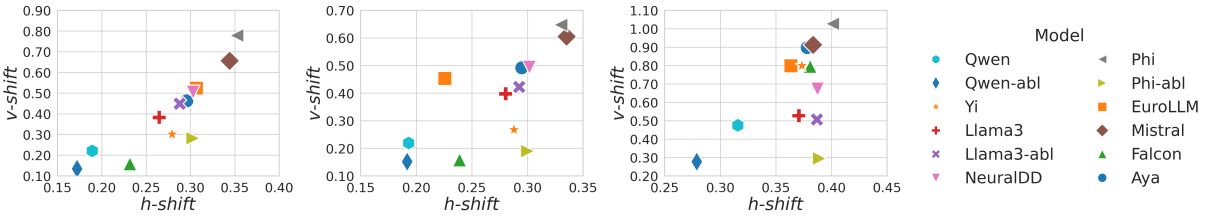


Figure 4: **(RQ2)** From left to right: Revise, Rewrite, and Continue, on MFTC.

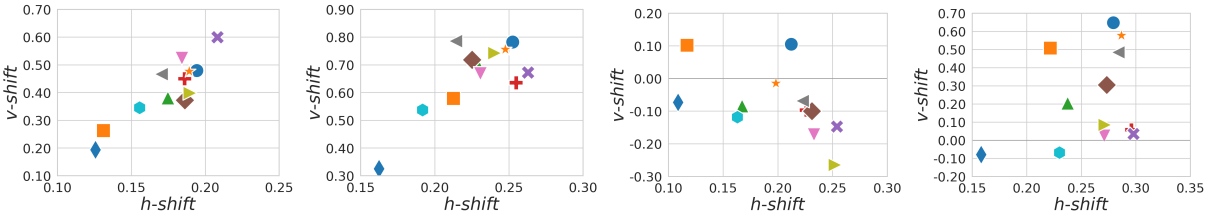


Figure 5: **(RQ3)** From left to right: Revise and intensify, Continue and intensify, Revise and diminish, Continue and diminish, on MFTC.

NeuralDD show the largest negative $v\text{-}shift$ values along with the highest $h\text{-}shift$. Regardless of the instructions, Llama3-abl, resp. Qwen-abl and EuroLLM, show the highest, resp. lowest changes in both moral $h\text{-}shift$ and $v\text{-}shift$.

5.5 Remark on Refusals

It is worth noting that the examined LLMs exhibited a low *refusal rate* (Pasch, 2025) which, across all text modification tasks, is around 6%, 10% and 3% w.r.t. the unconditioned (RQ1), moral-blind (RQ2) and moral-informed conditioned (RQ3) text-manipulation, respectively. In all cases, approximately 55% of these refusals correspond to texts that human-annotated as non-moral, mostly from social media data, particularly MFTC.

The observed low refusal rate could be regarded as a proxy of lack of safety warnings or violations due to the moral-targeted text manipulations. However, this should be taken with caution since some prompts that might appear free of safety risks at a first glance may still lead to model's output with unintended biases or ethical concerns. Further investigation on both the prompts and the manipulated texts is needed and left as a future work.

5.6 Remark on Comparison with GPT-4o

While our study focuses on open LLMs, we also considered whether a non-open or larger model, such as GPT-4o, would behave consistently with the examined models. To explore this, we tested GPT-4o by replicating our RQ-related tasks. Preliminary results, shown in Appendix I, suggest that GPT-4o's impact closely aligns with that of the open, smaller models. This particularly holds for the unconditioned manipulation and moral-blind conditioned manipulation tasks.

6 Conclusions

As generative AI grows, understanding how LLMs modify moral dimensions—either spontaneously

or through conditioning—is crucial. This study investigates 12 among the most widely used Open and uncensored LLMs from different regions and cultures, analyzing their influence on moral expressions using Moral Foundation Theory and human-annotated datasets. Our findings reveal diverse levels of alterations of moral expressions across text modification tasks, and the impact of conditioning prompts on selectively shift moral expressions. Notably, models exhibit consistent behavior across the study, suggesting distinct moral footprints.

As future work, it would be interesting to investigate the influence of morality on model refusals (Arditi et al., 2024), the impact of human personalities (La Cava and Tagarelli, 2025; Ge et al., 2024) in shaping moral expressions, as well as the characterization of moral expressions in machine-generated texts (La Cava et al., 2024).

Acknowledgements

AT is partly supported by project "Future Artificial Intelligence Research (FAIR)" spoke 9 (H23C22000860006), under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU. CMG and LLC are supported by project SERICS (PE00000014), under the MUR National Recovery and Resilience Plan funded by the EU - NextGenerationEU. LZ is supported by project PRIN2022 "AWESOME" (H53D23003550006).

Limitations

Moral issues are inherently complex aspects of human behavior. In this regard, human annotations reflect the subjectivity of moral judgments, which can lead to annotator disagreement driven by personal beliefs, as well as to challenges arising from textual ambiguity and human errors (Mokhberian et al., 2022). In this respect, one potential limitation of our work is that we relied on human annotations for selecting the target data subset. Nonetheless, our findings have shown that the relatively large set of examined LLMs—all open with wide coverage, but also to a limited extent GPT4o (cf. Sect. 5.6)— yield a consistent behavior across the different tasks related to our RQs. Despite this perception of robustness of our study, concerns still remain regarding the most appropriate procedures for human annotations related to fundamental aspects of human behavior, such as morality.

Furthermore, although we endeavored to be inclusive in our choice of models—particularly by considering multilingual approaches from various continents, as shown in Tables 1 and 2—we acknowledge the need for broader language coverage in our selected datasets. In this regard, note that we follow the established practices to refer to high-resource languages, particularly English, while acknowledging the inherent risk of cultural biases. Nevertheless, expanding the language coverage would enable us to assess the impact of LLMs in cross-linguistic scenarios, to enhance the generalizability of our findings as well as to validate the observed patterns across different cultural contexts. We leave this direction of research for future work.

We based our work on MFT, which is a widely established framework used by most of existing works in studying AI and morality (Zangari et al., 2025a; Jiang et al., 2021). Nevertheless, another area of improvement would be to explore other morality theories as well. In particular, it remains an open question whether the psychological differences observed across various moral theories also manifest in machine-based analyses. This "inter-morality-theory" investigation is also left for future work.

References

Marwa Abdulhai, Gregory Serapio-García, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17737–17752. Association for Computational Linguistics.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowl. Based Syst.*, 191:105184.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083.

Mohammad Atari, Jesse Graham, and Morteza Dehghani. 2020. Foundations of morality in iran. *Evolution and Human Behavior*, 41(5).

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*.

Naomi Ellemers. 2018. Morality and Social Identity. In Martijn van Zomeren and John F. Dovidio, editors, *The Oxford Handbook of the Human Essence*, page 0. Oxford University Press.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A. Association for Computational Linguistics.

Jeremy Frimer. 2019. Moral foundations dictionary 2.0.

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: Distributed dictionary representation. *Behavior research methods*, 50:344–361.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

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.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media, ICWSM 2023*, pages 281–291. AAAI Press.

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

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindrich Libovický, Constantin A. Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2137–2156. Association for Computational Linguistics.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024. Whose emotions and moral sentiments do language models reflect? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6611–6631, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wilhelm Hofmann, Daniel C Wisneski, Mark J Brandt, and Linda J Skitka. 2014. Morality in everyday life. *Science*, 345(6202):1340–1343.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 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*, 53:232–246.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

B. Kennedy, M. Atari, A. M. Davani, J. Hoover, A. Omrani, J. Graham, and M. Dehghani. 2021a. Moral concerns are differentially observable in language. *Cognition*, 212:104696.

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

Dániel Z. Kádár, Vahid Parvaresh, and Puyu Ning. 2019. Morality, moral order, and language conflict and aggression: A position paper. *Journal of Language Aggression and Conflict*, 7(1):6–31.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Is Contrasting All You Need? Contrastive Learning for the Detection and Attribution of AI-generated Text. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3179–3186.

Lucio La Cava and Andrea Tagarelli. 2025. Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models. In *Proceedings of the AAAI 2025*, pages 1355–1363.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5239–5251. Association for Computational Linguistics.

Negar Mokhberian, Frederic R. Hopp, Bahareh Harandizadeh, Fred Morstatter, and Kristina Lerman. 2022. Noise audits improve moral foundation classification. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022, Istanbul, Turkey, November 10-13, 2022*, pages 147–154. IEEE.

José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araújo, and Simone D. J. Barbosa. 2024. Are large language models moral hypocrites? A study based on moral foundations. *CoRR*, abs/2405.11100.

Stefan Pasch. 2025. Llm content moderation and user satisfaction: Evidence from response refusals in chatbot arena. *Preprint*, arXiv:2501.03266.

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*, New York, NY, USA. Association for Computing Machinery.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297, Toronto, Canada. 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, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The moral foundations reddit corpus. *CoRR*, abs/2208.05545.

René Weber, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2021. Extracting latent moral information from text narratives: Relevance, challenges, and solutions. In *Computational Methods for Communication Science*, pages 39–59. Routledge.

Jing Yi Xie, Graeme Hirst, and Yang Xu. 2020. Contextualized moral inference. *CoRR*, abs/2008.10762.

Lorenzo Zangari, Candida M Greco, Davide Picca, and Andrea Tagarelli. 2025a. A Survey on Moral Foundation Theory and Pre-Trained Language Models: Current Advances and Challenges. *AI & Society*.

Lorenzo Zangari, Candida M. Greco, Davide Picca, and Andrea Tagarelli. 2025b. 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.

Xinliang Frederick Zhang, Winston Wu, Nicholas Beauchamp, and Lu Wang. 2024. MOKA: moral knowledge augmentation for moral event extraction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4481–4502. Association for Computational Linguistics.

Yuyan Zhang, Jiahua Wu, Feng Yu, and Liying Xu. 2023. Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Sciences*, 13(2):181.

A Data Pre-processing

Following (Guo et al., 2023; Preniqi et al., 2024; Zangari et al., 2025b), we removed URLs, hashtags, and non-ASCII characters, replaced user mentions with "@user", and converted emojis to their textual equivalents. In MFRC, Equality/Inequality and Proportionality/Disproportionality were treated as Fairness/Cheating (Preniqi et al., 2024; Zangari et al., 2025b).

B Agreement of Moral Foundation Scoring Models

Table 6 shows Cohen's κ statistics for each pair of moral foundation scoring models— namely, ME2BERT, MoralBERT, MoralStrength and DDR—on the selected subset (cf. Sect. 4). ME2BERT shows particularly strong agreement with MoralBERT across all of the moral dimensions: Cohen's κ score ranges from 0.54 to 0.67, indicating a generally higher level of consistency compared to the other pairs. Therefore, this further reinforces our decision to use ME2BERT as annotator, given its strong alignment with MoralBERT and its greater versatility and scalability compared to MoralBERT (cf. Sect. 4). By contrast, DDR and MoralStrength show low agreement both with ME2BERT, MoralBERT and among themselves.

Model	Vs.	Authority	Care	Fairness	Loyalty	Purity
ME2-BERT	MoralBERT	0.54	0.58	0.59	0.67	0.58
	MoralStrength	0.08	-0.07	0.06	0.06	0.03
	DDR	0.33	0.26	0.22	0.18	0.42
MoralBERT	MoralStrength	0.05	0.02	0.03	0.04	0.00
	DDR	0.33	0.31	0.27	0.21	0.35
MoralStrength	DDR	0.07	0.04	0.01	0.04	0.02

Table 6: Per-dimension Cohen's Kappa agreement between all pairs of moral-foundation-scoring models, averaged over all datasets.

C Prompt for RQ0

Figure 6 shows the prompt we used to assess models' awareness of the Moral Foundations Theory.

RQ0 Prompting

Prompt Message:
Please, provide me with a detailed overview of the Moral Foundations Theory.

Figure 6: **(RQ0)** Prompt instructions for assessing MFT awareness of LLMs.

D Task-pair Shift Significance Analysis

We present a statistical significance analysis regarding the impact of the shifts between any pairs of tasks. We first retrieved all punctual values of $h\text{-}shift$ and of v-shift over all instances, for each dataset, task, and model. Given a dataset, a model, and a criterion (i.e., either $h\text{-}shift$ or $v\text{-}shift$), we selected one pair of tasks at a time and considered two task-related variables. These variables correspond to vectors of shift measurements for the same criterion, generated on the dataset according to the two selected tasks. Then, we carried out a paired two-sided t-test under the null hypothesis of no mean difference between the two variables. Overall, we performed a total of 1440 tests, considering all pairs of tasks, models, criteria, and datasets.

In Table 7, we report a summary of the results of the above tests. Each of these percentage values corresponds to the percentage of times (averaged across datasets) the null hypothesis was rejected, indicating that there is sufficient statistical evidence to conclude that there is a significant difference between the two groups being compared (i.e., revise vs. rewrite, revise vs. continuation, rewrite vs. continuation). We observe that in the totality of cases regarding continuation vs. revise or rewrite, the results produced by a model are statistically different, i.e., continuation brings bigger shifts in moral

	Task pair	Llama3	Qwen	Phi	Qwen-abl	Llama3-abl
RQ1	Rev. vs Cont.	100/100	100/100	100/100	100/100	100/100
	Rev. vs Rew.	100/40	100/80	100/60	100/80	100/60
	Rew. vs Cont.	100/100	100/100	100/100	100/100	100/100
RQ2	Rev. vs Cont.	100/100	100/100	60/100	100/100	100/80
	Rew. vs Cont.	100/100	100/100	60/100	100/100	100/80
	Rew. vs Rev.	40/40	60/60	20/100	60/40	30/80
RQ3	Rev.↓ vs Cont. ↓	100/100	100/100	80/100	100/40	80/100
	Rev.↓ vs Rew. ↓	100/60	100/100	40/40	80/80	60/20
	Rew. ↓ vs Cont. ↓	80/100	100/100	80/100	100/40	80/100
	Rev. ↑ vs Cont. ↑	100/100	80/100	100/100	100/100	60/60
	Rev.↑ vs Rew. ↑	80/20	40/100	60/20	60/80	60/30
	Rew. ↑ vs Cont.↑	80/100	60/100	100/100	100/100	60/60

Table 7: $h\text{-}shift$ / $v\text{-}shift$ (left / right) percentage of paired two-sided t-tests (significance level at 0.05) that reject the null hypothesis of equal means between the two shift distributions, averaged over the five datasets. Larger percentages indicate stronger statistical evidence that the two compared tasks induce different moral-value shifts for the model.

	Task	Llama3	Phi	Mistral	EuroLLM	Qwen	Yi	Falcon	Aya	Llama3-abl	NeuralDD	Qwen-abl	Phi-abl
CH	Rev. ↑	0.449	0.375	0.309	0.243	0.337	0.387	0.360	0.383	0.672	0.533	0.204	0.318
	Rew. ↑	0.443	0.363	0.256	0.245	0.384	0.477	0.312	0.474	0.677	0.509	0.214	0.261
	Cont. ↑	0.660	0.839	0.754	0.669	0.551	0.827	0.773	0.815	0.752	0.720	0.351	0.783
FC	Rev.↑	0.488	0.421	0.337	0.284	0.373	0.490	0.397	0.405	0.718	0.587	0.198	0.345
	Rew. ↑	0.459	0.408	0.275	0.279	0.422	0.570	0.351	0.498	0.707	0.549	0.219	0.295
	Cont.↑	0.716	0.876	0.775	0.724	0.580	0.885	0.805	0.852	0.805	0.769	0.344	0.756
LB	Rev. ↑	0.486	0.394	0.384	0.243	0.397	0.440	0.411	0.385	0.679	0.587	0.247	0.345
	Rew.↑	0.481	0.387	0.340	0.250	0.453	0.524	0.376	0.461	0.679	0.573	0.273	0.296
	Cont. ↑	0.682	0.870	0.806	0.662	0.600	0.847	0.788	0.795	0.758	0.761	0.398	0.806
AS	Rev. ↑	0.454	0.416	0.406	0.244	0.310	0.468	0.379	0.334	0.678	0.567	0.142	0.392
	Rew. ↑	0.443	0.423	0.363	0.241	0.361	0.573	0.356	0.442	0.679	0.544	0.165	0.344
	Cont. ↑	0.697	0.897	0.851	0.707	0.541	0.899	0.806	0.871	0.778	0.752	0.294	0.856
PD	Rev. ↑	0.468	0.490	0.391	0.195	0.335	0.488	0.375	0.485	0.770	0.621	0.190	0.415
	Rew. ↑	0.481	0.503	0.356	0.196	0.379	0.608	0.360	0.543	0.782	0.608	0.202	0.354
	Cont. ↑	0.579	0.939	0.782	0.498	0.459	0.822	0.783	0.788	0.704	0.709	0.264	0.844

Table 8: Per-dimension $v\text{-}shift$ scores for the RQ3–Intensify setting. Rows are grouped by moral dimension—**CH** (care–harm), **FC** (fairness–cheating), **LB** (loyalty–betrayal), **AS** (authority–subversion), **PD** (purity–degradation).

values than revise or rewrite. Also, with no surprise as already discussed in the paper, the comparison between revise-rewrite pairs generally results in a smaller number of cases (ranging from 20% to 80%) of statistical difference across the datasets.

E Analysis of Individual MFT Dimensions

To obtain more granular insights into the behavior of models with respect to each individual moral dimension, Table 8 reports the $v\text{-}shift$ score of each dimension in the context of the RQ3-intensify setting—note that, since each dimension is considered in isolation, the $h\text{-}shift$ is zero and therefore omitted. It can be noticed that all models in the RQ3 tasks generally tend to exhibit a consistent behavior across all moral dimensions, not showing a particular "preference" for any specific dimension.

F Properties of v-shift Measure

In the following, we outline the main properties of the $v\text{-}shift$ signal defined in Sect. 4.

Property 1 (Anti-symmetry). *For any vectors* $\mathbf{a}, \mathbf{b} \in [0,1]^N$*, and scalars* $\beta, \gamma, \tau \in \mathbb{R}^+$*, it holds that* $v\text{-}shift(\mathbf{a}, \mathbf{b}) = -\, v\text{-}shift(\mathbf{b}, \mathbf{a})$.

Proof. By definition, swapping $\mathbf{a}$ and $\mathbf{b}$, we obtain:

$$v\text{-}shift(\mathbf{b}, \mathbf{a}) \;=\; \frac{1}{N}\sum_{i=1}^{N}\big[\phi(a_i) - \phi(b_i)\big]. \tag{1}$$

Then, considering that γ, β and τ are constant, and expanding each difference we obtain the following:

$$\begin{aligned}
&\phi(a_i) - \phi(b_i) = \\
&= \big[a_i + \gamma\,\sigma(\beta\,(a_i - \tau))\big] - \big[b_i + \gamma\,\sigma(\beta\,(b_i - \tau))\big] = \\
&= -\Big(\big[b_i - a_i\big] + \gamma\,\big[\sigma(\beta\,(b_i - \tau)) - \sigma(\beta\,(a_i - \tau))\big]\Big),
\end{aligned}$$

from which we have $\phi(a_i) - \phi(b_i) \;=\; -\big(\phi(b_i) - \phi(a_i)\big)$. Therefore, by replacing this term on Eq. 1, we conclude that:

$$\begin{aligned}
v\text{-}shift(\mathbf{b}, \mathbf{a}) &= -\frac{1}{N}\sum_{i=1}^{N}\big[\phi(b_i) - \phi(a_i)\big] \\
&= -v\text{-}shift(a, b).
\end{aligned}$$

□

Property 2 (Boundedness). *For any vectors $\mathbf{a}, \mathbf{b} \in [0,1]^N$, and scalars $\gamma, \beta, \tau \in \mathbb{R}^+$, we have:*

$$v\text{-}shift(\mathbf{a},\, \mathbf{b}) \;\in\; \big[-(1+\gamma),\, (1+\gamma)\big].$$

Proof. Since the logistic function $\sigma(x) \in (0,1)\ \forall x \in \mathbb{R}$, we have $0 \le \sigma\big(\beta\,(x-\tau)\big) \le 1$, for any choice of β. In particular, if $\gamma > 0$ and $x \in [0,1]$, then:

$$\phi(x) \;=\; x + \gamma\,\sigma\big(\beta\,(x-\tau)\big) \;\in\; \big[x,\; x+\gamma\big].$$

Hence, for each component $a_i, b_i \in [0,1]$, both $\phi(a_i)$ and $\phi(b_i)$ lie in $[\,0,\, 1+\gamma\,]$, which implies

$$\phi(b_i) \;-\; \phi(a_i) \;\in\; \big[-(1+\gamma),\, 1+\gamma\big].$$

Summing such terms over $i = 1, \ldots, N$ and dividing by N preserves this interval, yielding

$$-(1+\gamma) \;\le\; \frac{1}{N}\sum_{i=1}^{N}\big[\phi(b_i) - \phi(a_i)\big] \;\le\; 1+\gamma.$$

Therefore

$$v\text{-}shift(\mathbf{a}, \mathbf{b}) \;\in\; \big[-(1+\gamma),\, 1+\gamma\big],$$

as claimed.

□

Property 3 (Dimension-wise Monotonicity). *Let $\mathbf{a} = (a_1, \ldots, a_N)$ and $\mathbf{b} = (b_1, \ldots, b_N)$ be two vectors in $[0,1]^N$ and $\gamma, \beta, \tau \in \mathbb{R}^+$ positive real values. For any $i \in 1, \ldots, N$, given $\mathbf{b}' = (b_1, \ldots, b_{i-1}, b'_i, b_{i+1}, \ldots, b_N)$, with $b_i \le b'_i$, it holds that:*

$$v\text{-}shift(\mathbf{a}, \mathbf{b}) \;\;\le\;\; v\text{-}shift(\mathbf{a}, \mathbf{b}').$$

Thus, increasing one single coordinate does not decrease the function value.

Proof. First, we show that $\phi(x)$ is a monotonically non-decreasing function. Specifically, let

$$\phi(x) \;=\; x \;+\; \gamma\,\sigma\big(\beta\,(x-\tau)\big),$$

where $\sigma(z) = \frac{1}{1+e^{-z}} \in (0,1)$ for all $z \in \mathbb{R}$ and $\gamma > 0$, $\beta > 0$, $\tau \in \mathbb{R}$. We compute its derivative:

$$\phi'(x) \;=\; 1 \;+\; \gamma\,\beta\,\sigma\big(\beta\,(x-\tau)\big)\,\big[1-\sigma\big(\beta\,(x-\tau)\big)\big].$$

Since $\sigma(\cdot) \in (0,1)$, the product $\gamma\,\beta\,\sigma(\cdot)\,[1-\sigma(\cdot)]$ is strictly positive. Thus,

$$\phi'(x) \;>\; 1 \;>\; 0,$$

implying that ϕ is strictly increasing. Intuitively, $\phi(x)$ transitions smoothly from x (when $x \ll \tau$) to $x+\gamma$ (when $x \gg \tau$), while its derivative remains strictly positive over the entire range of x.

Then, recalling that

$$v\text{-}shift(\mathbf{a},\mathbf{b}) \;=\; \frac{1}{N}\sum_{j=1}^{N}\Big[\phi\big(b_j\big)-\phi\big(a_j\big)\Big],$$

By hypothesis, all terms in the sum are fixed except the one corresponding to $j=i$. Hence,

$$\begin{aligned} v\text{-}shift(\mathbf{a},\mathbf{b}) =& \frac{1}{N}\Big(\sum_{\substack{j=1\\ j\neq i}}^{N}\big[\phi(b_j)-\phi(a_j)\big]+ \\ &+\;\big[\phi(b_i)-\phi(a_i)\big]\Big). \end{aligned}$$

By the same definition, if we form $\mathbf{b}'$ by replacing b_i with b_i', we get

$$\begin{aligned} v\text{-}shift(\mathbf{a},\mathbf{b}') \;=& \frac{1}{N}\Big(\sum_{\substack{j=1\\ j\neq i}}^{N}\big[\phi(b_j)-\phi(a_j)\big]+ \\ &+\;\big[\phi(b_i')-\phi(a_i)\big]\Big). \end{aligned}$$

Since ϕ is a monotonically non-decreasing function, the condition $b_i \le b_i'$ yields

$$\phi\big(b_i'\big) \;\geq\; \phi\big(b_i\big).$$

Therefore,

$$\phi(b_i')-\phi(a_i) \;\;\geq\;\; \phi(b_i)-\phi(a_i).$$

This implies:

$$\begin{aligned} &v\text{-}shift(\mathbf{a},\mathbf{b}') \;-\; v\text{-}shift(\mathbf{a},\mathbf{b}) \;= \\ &= \frac{1}{N}\Big(\big[\phi(b_i')-\phi(a_i)\big]-\big[\phi(b_i)-\phi(a_i)\big]\Big)\geq 0, \end{aligned}$$

and hence

$$v\text{-}shift(\mathbf{a},\mathbf{b}) \;\;\leq\;\; v\text{-}shift(\mathbf{a},\mathbf{b}').$$

□

G Hyper-parameters Selection for the v-shift Measure

We discuss the settings of parameters τ, β and γ for the $v\text{-}shift$ measure.

The choice of $\tau = 0.5$ refers to a conventional threshold widely adopted in machine learning to define balanced decision boundaries between classes. In our setting—-as well as in the settings of other methods, including those in our related work—this means that a value above, resp. below, this threshold might indicate the presence, resp. absence of that moral foundation.

Regarding γ, we first would like to emphasize that (non-negative) values of γ below 1 concentrates the evaluation range to the original moral scores, while still ensuring that the activation or deactivation of moral expressions after the manipulation of LLMs are properly handled. Additionally, we have

experimentally observed that the choice of $\gamma = 0.5$ provides a balanced midpoint for smoothly controlling the sensitivity around the threshold τ, thereby reducing excessive amplification of minor differences and supporting a more stable and interpretable assessment of moral shifts. Conversely, values of γ greater than 1 tend to overly penalize the activation or deactivation of moral dimensions, resulting in less stable and interpretable scores.

The parameter β controls the steepness of the logistic function around the threshold value τ. Note that a higher value of β would result in a logistic curve exhibiting a sharper transition around τ, which could potentially amplify small fluctuations in moral scores, thus introducing noisy outcomes. By contrast, a lower value of β would yield a smoother transition near τ, which may diminish sensitivity, thus compromising the ability of the $v\text{-}shift$ score to properly detect moral shifts. Therefore, we selected $\beta = 10$ as a mid-range value of steepness, for capturing variations in moral expressions without introducing excessive instability.

H Additional Results on RQs

Figures 7-15 provide additional insights into our results. In particular, Fig. 8, Fig. 12 and Fig. 13 provide results under both temperature settings, revealing the negligible impact of temperature on models' overall behavior. This supports our decision to present only the low-temperature results in the main paper—for the sake of readability.

Figure 7 presents radar charts across different moral dimensions, unless otherwise specified. Each chart presents a single visualization encompassing all texts and moral dimensions, with values corresponding to averaged scores of any specific dimension across all texts. Interestingly, as already observed in Table 8, no single moral dimension stands out as being systematically manipulated across different tasks or models.

To gain more fine-grained insights into cases where a single moral dimension is initially predominant, Figs. 9, 11, 12, and 13 present illustrative radar charts for each research question (RQ). Each figure includes examples of Revise (top), Rewrite (middle), and Continue (bottom), obtained by fixing the dataset and moral dimension, while varying the model. Note that, for RQ1 in the Revise (Fig. 9 (a–c)) and Rewrite (Fig. 9 (d–f)) settings, the choice of the models is here arbitrary since no significant moral shifts were detected across models, as discussed in the main text. In the Continue setting (Fig. 9 (g–i)), two models are selected, i.e., Qwen-abl and Phi, to reflect different scores from the average patterns in $v\text{-}shift$ and/or $h\text{-}shift$ identified in Fig. 8, and Falcon, which instead exhibit a similar behavior to other LLMs, i.e., it lies near the center of the observed distribution in Fig. 8.

The radar charts for RQ2 and RQ3 follow the same approach: for each scenario of Revise, Rewrite and Continue, two models are selected such that $v\text{-}shift$ and/or $h\text{-}shift$ scores diverge from the average patterns, along with one model that corresponds near to the center of the distribution observed in Figs. 10, 12, and 13.

I Insights into GPT-4o

Figure 16 presents insights into RQ1 and RQ2 for the GPT-4o model, which was prompted with a Rewrite task on the EMONA and MFRC datasets for news and social data, respectively. Also in this case, each radar chart is a single visualization for all moral dimensions and samples, with each value corresponding to the average moral score across all samples. When the model is asked to only rewrite the text (RQ1), it preserves the original moral expressions (Fig. 16 (a)-(c)). However, when instructed to identify and emphasize existing moral dimensions (RQ2), it slightly increases moral intensity on news samples, yet remains largely conservative for the social ones (Fig. 16 (e)-(g)). Overall, GPT-4o better adheres to existing moral trends, not introducing moral expressions in the presence of neutral text. However, it does not significantly amplify moral distributions when instructed to do this, contrary to the other models discussed in Sect. 5.3.

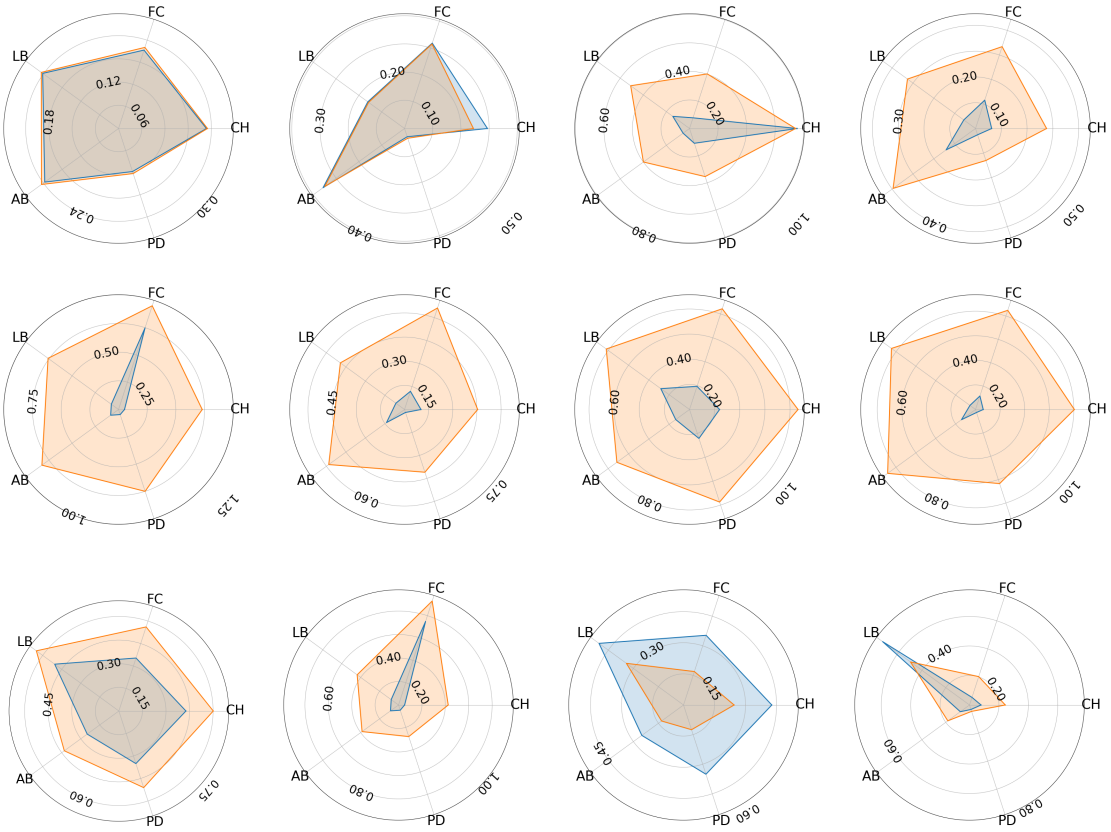


Figure 7: Radar charts for RQ1 (top row), RQ2 (middle row) and RQ3 (bottom row). Top row: Revise of NeuralDD on the "moral" and "non-moral" texts of MFNC, Rewrite of Qwen on ME, Continue of LLama-3 on MFNC considering only the CH dimension, and Continue of Aya on the "moral" and "non-moral" texts of EMONA. Middle row: Revise of Phi on MFNC considering only the FC dimension, Rewrite of EuroLLM on ME, Continue of Aya on MFTC on the "moral" and "non-moral" texts of MFTC, and Continue of Yi on EMONA. Bottom row: Revise-Intensify of EuroLLM on MFTC, Continue-Intensify of Qwen-Abl on ME considering only the FC dimension, Revise-Diminish of Phi on MFNC, and Continue-Diminish of LLama on MFRC considering only the LB dimension.

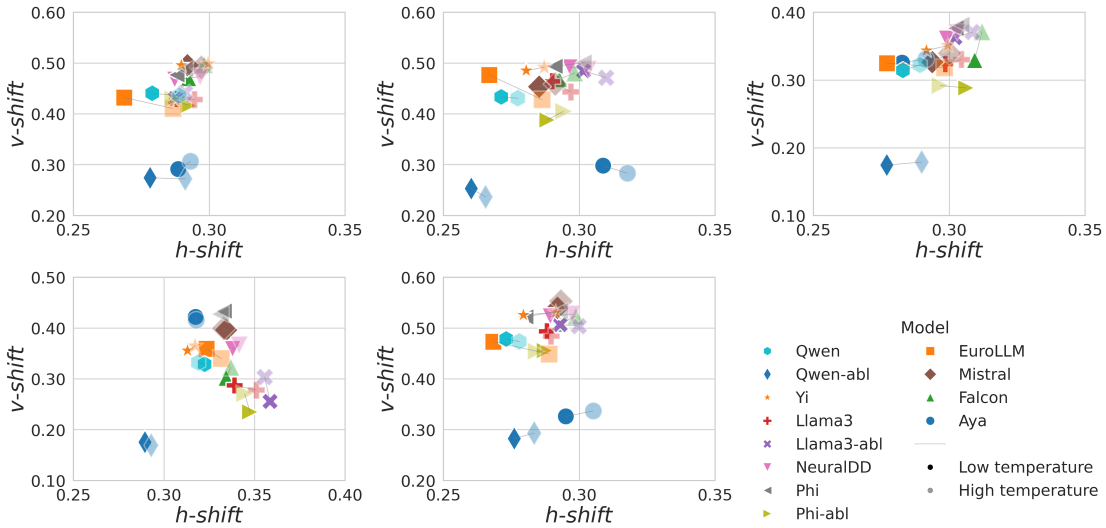
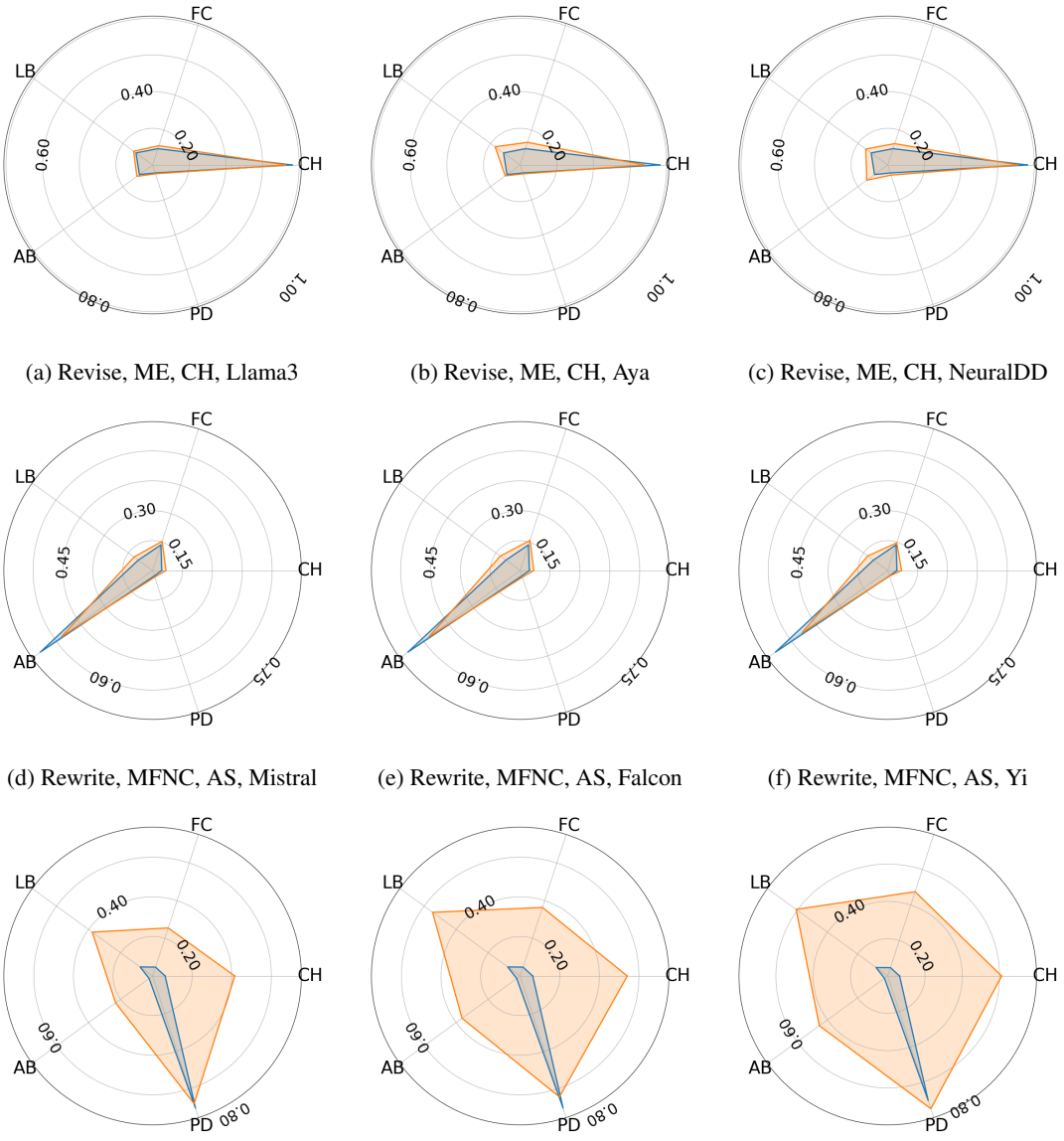


Figure 8: (**RQ1**) From top-left to bottom-right: Moral shifts due to Continue on ME, MFNC, MFRC, MFTC, and EMONA datasets, respectively.



(a) Revise, ME, CH, Llama3 (b) Revise, ME, CH, Aya (c) Revise, ME, CH, NeuralDD

(d) Rewrite, MFNC, AS, Mistral (e) Rewrite, MFNC, AS, Falcon (f) Rewrite, MFNC, AS, Yi

(g) Continue, MFTC, PD, Qwen-abl (h) Continue, MFTC, PD, Falcon (i) Continue, MFTC, PD, Phi

Figure 9: **(RQ1)** Examples of radar charts of the moral dimension scores under the Revise (top), Rewrite (middle), and Continue (bottom) settings. The dataset and the dominant moral dimension are fixed for each setting (row). The caption of each subfigure lists: setting, dataset, dominant moral dimension and model. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.

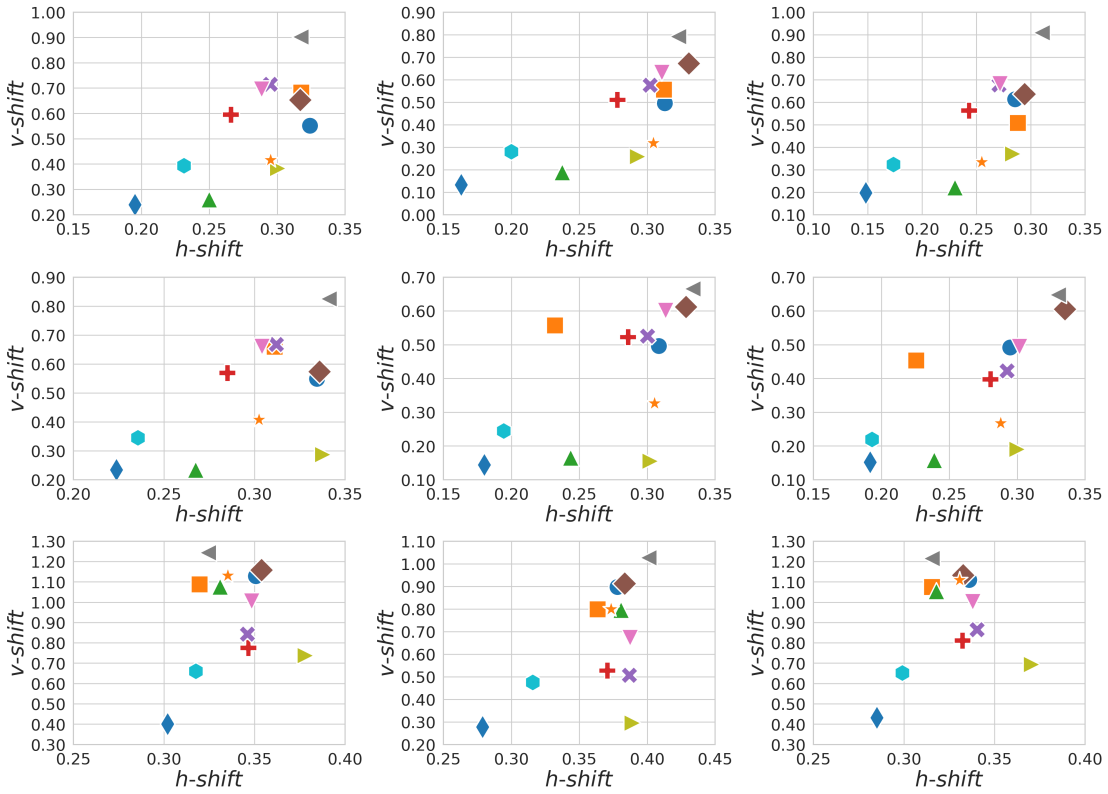
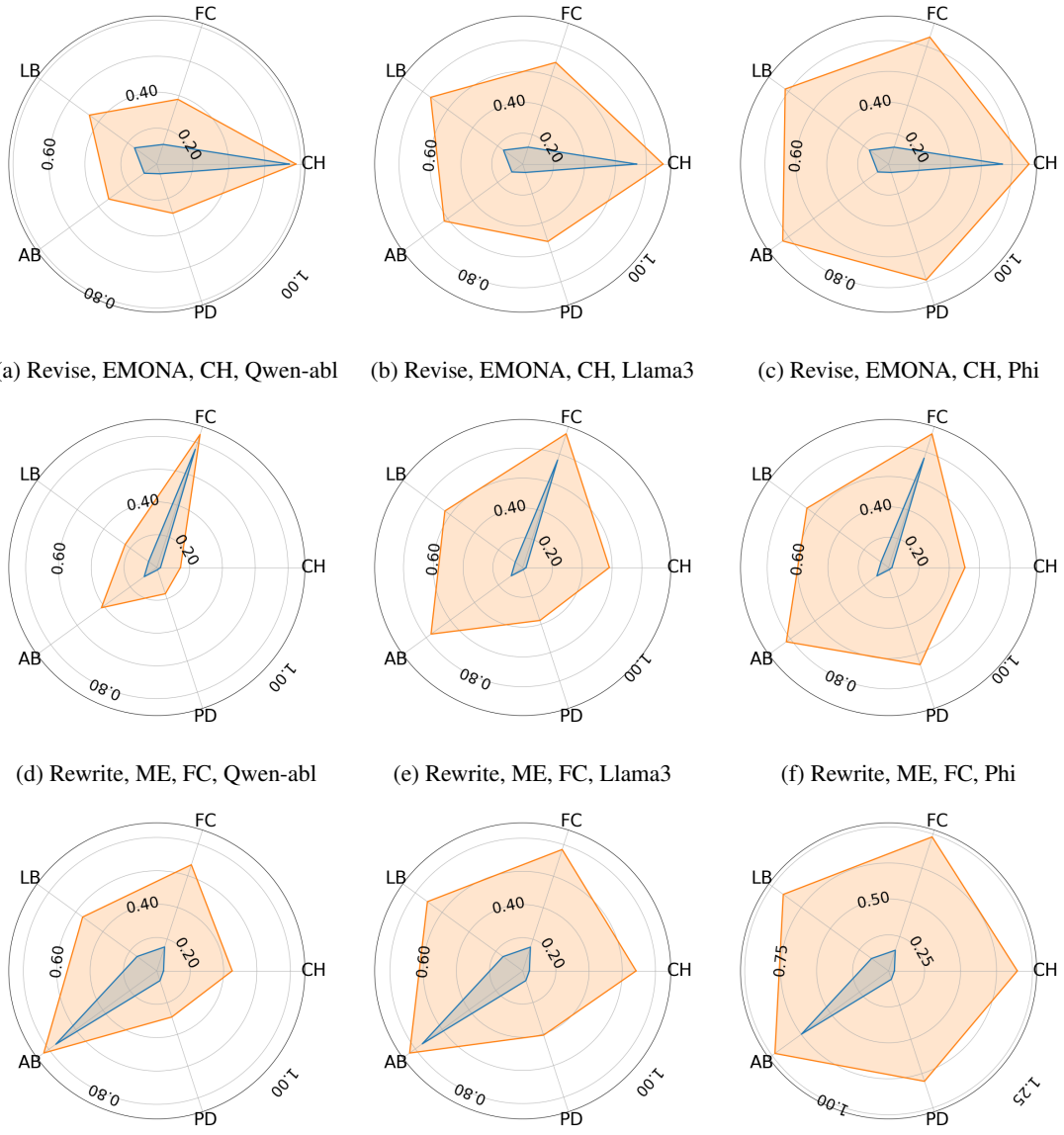


Figure 10: **(RQ2)** Moral shifts due to Revise on EMONA, MFRC and MFNC (top), Rewrite on ME, MFRC and MFTC (middle), Continue on ME, MFTC and EMONA (bottom).



(a) Revise, EMONA, CH, Qwen-abl (b) Revise, EMONA, CH, Llama3 (c) Revise, EMONA, CH, Phi

(d) Rewrite, ME, FC, Qwen-abl (e) Rewrite, ME, FC, Llama3 (f) Rewrite, ME, FC, Phi

(g) Continue, MFTC, AS, Qwen-abl (h) Continue, MFTC, AS, Llama3 (i) Continue, MFTC, AS, Phi

Figure 11: **(RQ2)** Examples of radar charts of the moral dimension scores under the Revise (top), Rewrite (middle), and Continue (bottom) settings. The dataset and the dominant moral dimension are fixed for each setting (row). The caption of each subfigure lists: setting, dataset, dominant moral dimension and model. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.

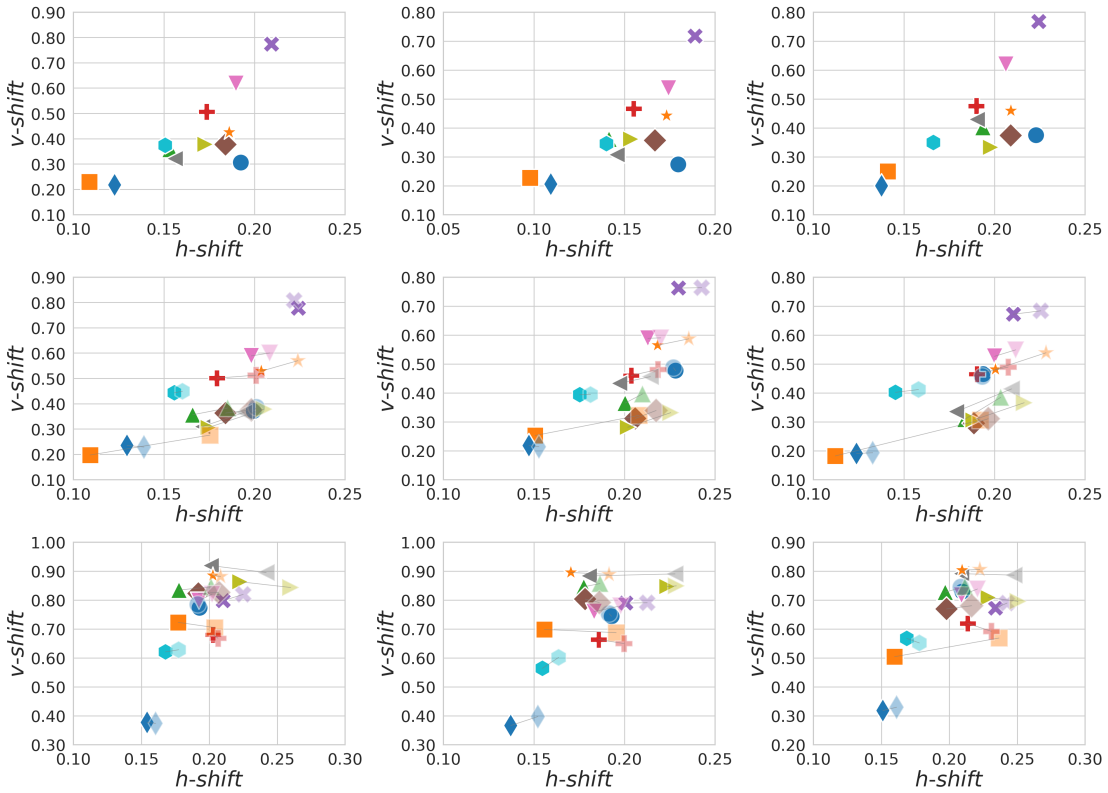


Figure 12: **(RQ3 - Intensify)** Moral "increase" due to Revise on EMONA, ME, and MFNC (top); Rewrite on EMONA, MFNC, and MFRC (center); Continue on EMONA, ME, and MFRC (bottom).

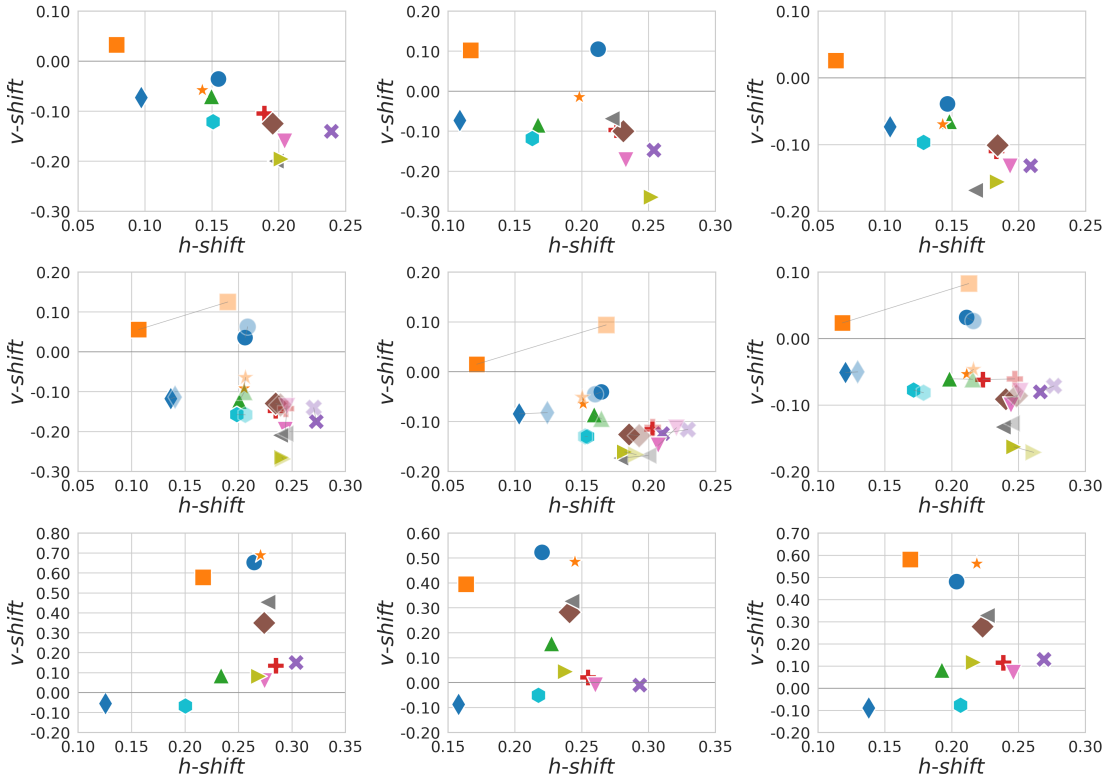
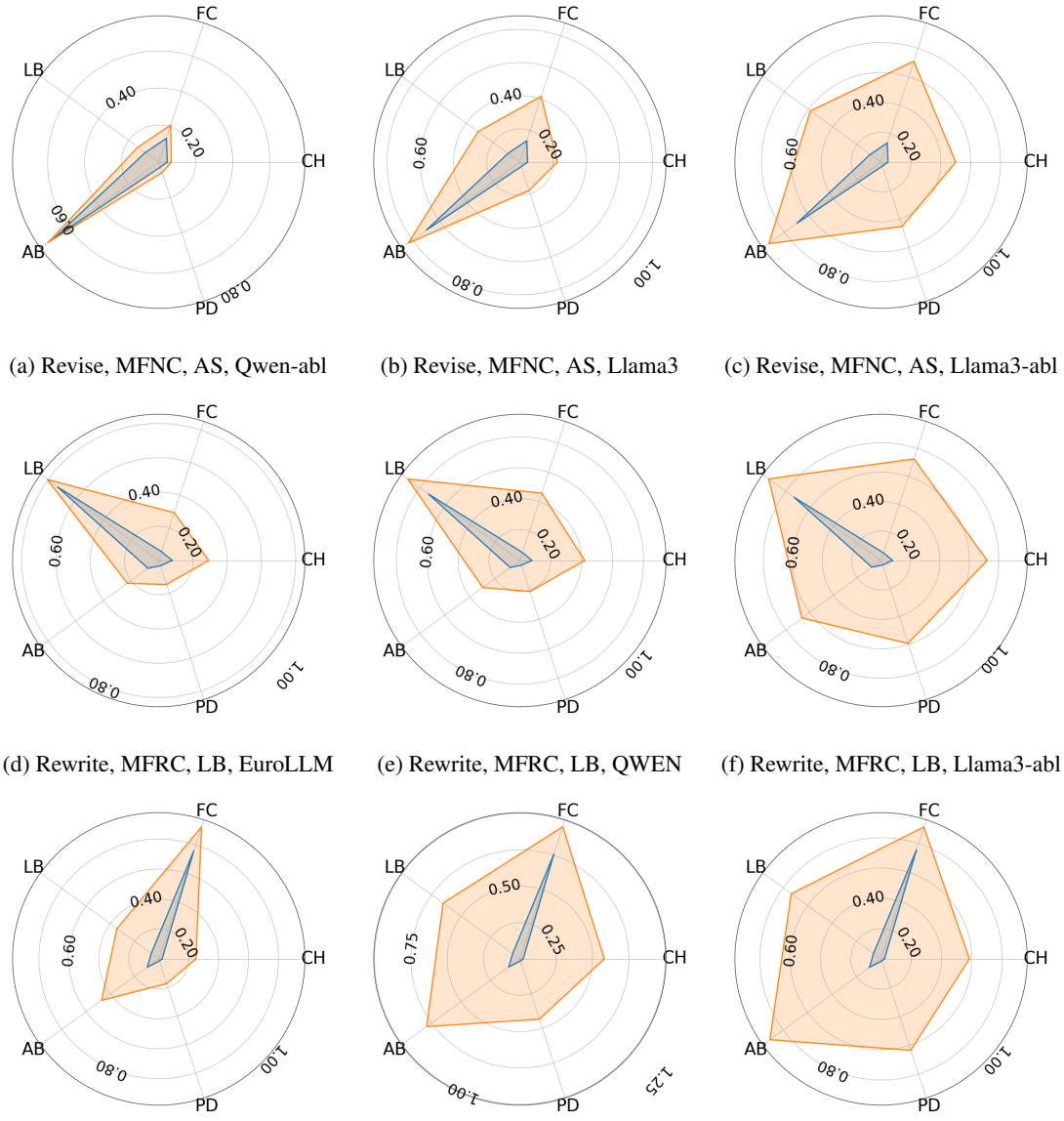


Figure 13: **(RQ3 - Diminish)** Moral "diminish" due to Revise on EMONA, MFTC and ME (top), Rewrite on MFRC, ME and MFNC (middle), Continue on MFNC, MFRC and EMONA (bottom).

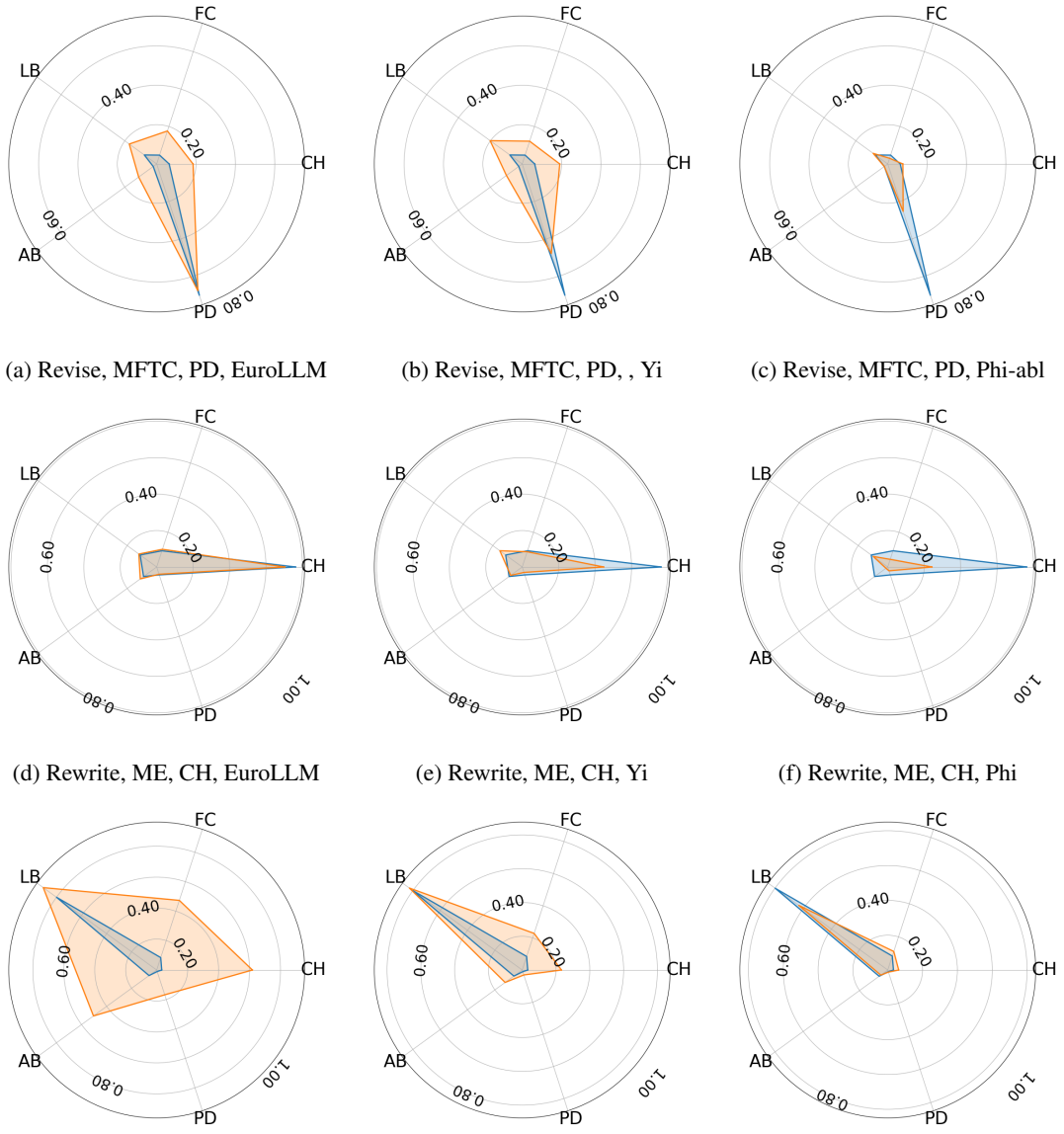


(a) Revise, MFNC, AS, Qwen-abl (b) Revise, MFNC, AS, Llama3 (c) Revise, MFNC, AS, Llama3-abl

(d) Rewrite, MFRC, LB, EuroLLM (e) Rewrite, MFRC, LB, QWEN (f) Rewrite, MFRC, LB, Llama3-abl

(g) Continue, ME, FC, Qwen-abl (h) Continue, ME, FC, Eurollm (i) Continue, ME, FC, Phi-abl

Figure 14: **(RQ3 - Intensify)** Examples of radar charts of the moral dimension scores under the Revise (top), Rewrite (middle), and Continue (bottom) settings. The dataset and the dominant moral dimension are fixed for each setting (row). The caption of each subfigure lists: setting, dataset, dominant moral dimension and model. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.

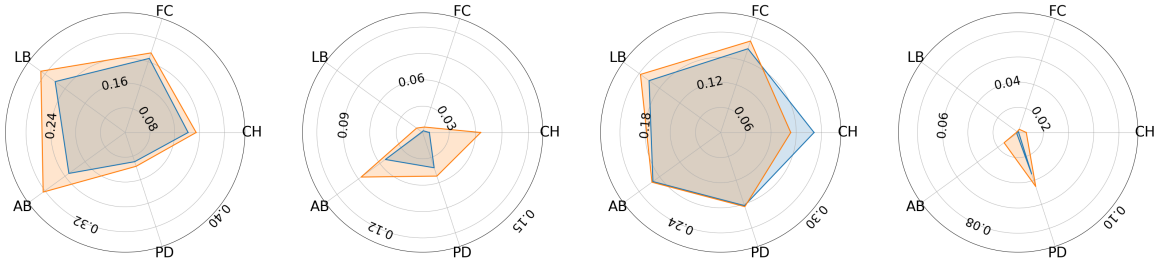


(a) Revise, MFTC, PD, EuroLLM (b) Revise, MFTC, PD, , Yi (c) Revise, MFTC, PD, Phi-abl

(d) Rewrite, ME, CH, EuroLLM (e) Rewrite, ME, CH, Yi (f) Rewrite, ME, CH, Phi

(g) Continue, EMONA, LB, Yi (h) Continue, EMONA, LB, Falcon (i) Continue, EMONA, LB, Qwen-abl

Figure 15: **(RQ3 - Diminish)** Examples of radar charts of the moral dimension scores under the Revise (top), Rewrite (middle), and Continue (bottom) settings. The dataset and the dominant moral dimension are fixed for each setting (row). The caption of each subfigure lists: setting, dataset, dominant moral dimension and model. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.



(a) RQ1, EMONA, moral (b) RQ1, EMONA, non-moral (c) RQ1, MFRC, moral (d) RQ1, MFRC, non-moral

(e) RQ2, EMONA, moral (f) RQ2, Emona, non-moral (g) RQ2, MFRC, moral (h) RQ2, MFRC, non-moral

Figure 16: Examples of moral changes in GPT-4o within the RQ1-Rewrite and RQ2-Rewrite settings on the EMONA and MFRC datasets. For each dataset, the radar chart of morally relevant texts and morally neutral texts is shown. Blue, resp. orange, charts denote the moral scores before resp. after manipulation.