Computers in Human Behavior Reports 17 (2025) 100609

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Full length article

Large Language Models meet moral values: A comprehensive assessment of moral abilities☆

Luana Bulla [a,b,*], Stefano De Giorgis [b], Misael Mongiovì [a,b], Aldo Gangemi [c,b]

[a] University of Catania, Catania, Italy
[b] ISTC - National Research Council, Rome and Catania, Italy
[c] University of Bologna, Bologna, Italy

ARTICLE INFO

Keywords:
Large Language Models
Value detection
Natural language inference
Natural Language Processing

ABSTRACT

Automatic moral classification in textual data is crucial for various fields including Natural Language Processing (NLP), social sciences, and ethical AI development. Despite advancements in supervised models, their performance often suffers when faced with real-world scenarios due to overfitting to specific data distributions. To address these limitations, we propose leveraging state-of-the-art Large Language Models (LLMs) trained on extensive common-sense data for unsupervised moral classification. We introduce an innovative evaluation framework that directly compares model outputs with human annotations, ensuring an assessment of model performance. Our methodology explores the effectiveness of different LLM sizes and prompt designs in moral value detection tasks, considering both multi-label and binary classification scenarios. We present experimental results using the Moral Foundation Reddit Corpus (MFRC) and discuss implications for future research in ethical AI development and human–computer interaction. Experimental results demonstrate that GPT-4 achieves superior performance, followed by GPT-3.5, Llama-70B, Mixtral-8x7B, Mistral-7B and Llama-7B. Additionally, the study reveals significant variations in model performance across different moral domains, particularly between everyday morality and political contexts. Our work provides meaningful insights into the use of zero-shot and few-shot models for moral value detection and discusses the potential and limitations of current technology in this task.

1. Introduction

Automatic moral classification in text is a rapidly evolving field with significant implications for NLP, social sciences, and ethical decision-making. However, identifying moral values in textual documents is a persistent challenge, mainly due to the limitations of current supervised models. These models often struggle with overfitting problems, as they over-conform the specific training data's distribution (Bulla, Gangemi et al., 2023). Although these classifiers show good results for datasets with similar distributions, their performance suffers statistically significant degradation when dealing with data from a real scenario with an unknown data distribution (Liscio, Dondera, Geadau, Jonker, & Murukannaiah, 2022; Trager et al., 2022). This highlights the need for more versatile approaches to moral classification. Researchers can achieve this by addressing the limitations of domain-specific training data and implementing methods that adapt to diverse data distributions. Such advances promise to impact different areas, including content moderation, ethical decision-making frameworks, and the broader landscape of human–computer interaction.

A further key aspect of moral detection concerns the evaluation methodology. Current gold standards rely on annotations from multiple human coders, but this approach faces limitations due to low inter-annotator agreement for inherently subjective tasks. This inconsistency weakens the gold standard's reliability, hindering accurate assessment of model performance.

In addressing these challenges, our goal is to leverage the ability of state-of-the-art models to grasp abstract social concepts through training on extensive pools of common-sense data (Asprino et al., 2022). This strategy aims to mitigate the risk of overfitting specific moral datasets and develop models that generalize across real-world

☆ We acknowledge financial support from the H2020 projects TAILOR: Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization – EC Grant Agreement number 952215 – the the Next Generation EU Program with the Future Artificial Intelligence Research (FAIR) project, code PE00000013, CUP 53C22003630006, and by the Italian Research Center on High Performance Computing, Big Data and Quantum Computing (ICSC)

* Correspondence to: Piazza Università, 2, 95124 Catania (CT), Italy.
E-mail address: luana.bulla@phd.unict.it (L. Bulla).

https://doi.org/10.1016/j.chbr.2025.100609
Received 16 July 2024; Received in revised form 10 January 2025; Accepted 27 January 2025
Available online 8 February 2025

scenarios. To validate our approach, we propose an in-depth evaluation comparing the performance of state-of-the-art unsupervised models with human judgments. Our novel framework directly juxtaposes model outputs with human annotations. This head-to-head approach addresses subjectivity concerns by placing human and model ratings on equal footing, offering valuable insights into both.

Our methodology explores the use of state-of-the-art Large Language Models (LLMs) of different sizes (i.e. Mistral-7B, Llama-7B, Mixtral-8x7B, Llama-70B, GPT-3.5, and GPT-4) as zero-shot, pre-trained unsupervised multi-label classifiers for moral value detection. We consider with the term LLMs models based on a generative approach with a dimension greater than 1B. This approach allows us to identify moral values without the need for explicit training on the annotated data. To evaluate the technique's adaptability and effectiveness, we incorporate both zero-shot and few-shot settings. The latter uses a limited number of examples to guide predictions, enhancing the models' contextual understanding and performance. To evaluate the effectiveness and adaptability of our technique, we compare it with a smaller zero-shot method based on Natural Language Inference (NLI). Our assessment reveals GPT-4 as the top performer, followed by GPT-3.5, Llama-70B, Mixtral-8x7, Mistral-7B and Llama7B. Notably, all systems outperform the NLI-based model. Furthermore, we delve into the strengths and limitations of the models in different value domains and prompting configurations. Our exploration includes everyday morality and morality applied to political contexts, employing two distinct prompting approaches that leverage the models' contextual understanding capabilities.

For a comprehensive and unbiased evaluation, we design a strategy that divides moral detection tasks into two sub-tasks: (i) multi-label moral value detection and (ii) binary moral classification. The multi-label moral value detection subtask focuses on identifying moral values in a text, assuming the model recognizes the text as inherently moral. The binary classification sub-task determines whether a text contains moral content or not. This strategy allows for a nuanced and in-depth understanding of findings in the field of moral classification. Through separate analyses of these two aspects, we aim to provide a clearer perspective on model capabilities. This ensures that the evaluation of the simpler binary task does not overshadow the more complex multi-label task. Furthermore, we introduce a novel experimental framework that allows a direct comparison between human and model performance under identical conditions. This comparative analysis juxtaposes results from automated value detection with those derived from human evaluation, clarifying the moral dimensions in which models achieve superior or comparable performance to humans. In our research, we employ the Moral Foundation Reddit Corpus (MFRC) (Trager et al., 2022), comprising 16,123 Reddit comments. This corpus is divided into three separate subcorpora, each delineating distinct domains. Furthermore, annotations in the corpus adhere to Graham and Haidt's Moral Foundation Theory (MFT) (Graham et al., 2013).

This paper outlines several significant contributions:

- We investigate the impact of model size on moral value detection by evaluating the performance of different LLMs (i.e., GPT-4, GPT-3.5, Llama-70B, Mixtral-7x8, Mistral-7B, and Llama7B) with different numbers of parameters and different training sets.
- We investigate the effectiveness of prompt design by evaluating the models' capabilities in two distinct prompting scenarios. This analysis will explore how different prompt structures influence the models' performance across various difficulty levels in moral value detection tasks.
- We separately evaluate models on both multi-label moral value detection (i.e. identifying specific values) and binary moral classification (i.e. discerning the presence or absence of moral content) for a more nuanced assessment.
- We propose a direct model-human comparison framework to evaluate moral detection, overcoming subjectivity limitations in gold standards.

The paper is structured as follows. Section 2 provides an overview of the current state of the art in the field. In Section 3, we offer an overview of the theoretical grounding of MFT. Section 4 delves into the unsupervised methodologies we use, distinguishing between the techniques involving LLM (Section 4.1) and the NLI-based method (Section 4.2). In Section 5, we present a comprehensive overview of our experimental setting (Section 5.1), exploring an assessment of different linguistic models in the task of moral value detection (Section 5.2). We specifically focus on two key sub-tasks: multi-label identification of moral value dimensions (Section 5.2.1) and binary detection of text containing moral content (Section 5.2.2). Additionally, in Section 5.3, we introduce a direct human model comparison framework for evaluating both multi-label and binary moral detection that addresses the issue of subjectivity limitations in gold standards. Furthermore, Section 6 provides a detailed discussion of the results obtained across all task configurations and evaluation settings. Finally, in Section 7, we summarize the key findings of our work and discuss the implications of our methods for future research.[1]

2. Related works

Previous research in moral value detection has focused on two dominant paradigms: feature-based methods, leveraging word embeddings and sequences (Garten, Boghrati, Hoover, Johnson, & Dehghani, 2016; Kennedy et al., 2021), and methodologies focused on word count analysis (Fulgoni, Carpenter, Ungar, & Preoţiuc-Pietro, 2016). These paradigms can be further classified into supervised and unsupervised learning frameworks. Supervised methods require pre-annotated datasets that establish an external moral framework supporting the model training process, while unsupervised methods operate independently of such annotations. In Section 2.1, we detail the supervised methodologies. In Section 2.2, we explore unsupervised methods, with a particular emphasis on LLM-based approaches for tasks related to moral value analysis and detection.

2.1. Supervised methods

In the field of supervised learning from a cross-domain perspective, a notable contribution is the Tomea system (Liscio et al., 2023), which offers an explainable method for analyzing how a supervised text classifier represents moral rhetoric across various domains. This approach determines if text classifiers learn domain-specific expressions of moral language, highlighting differences and similarities in moral concepts across social domains. Tomea employs the SHapley Additive exPlanations (SHAP) method (Lundberg & Lee, 2017), using Shapley values to measure the contribution of each word towards predicting a moral element, thereby compiling domain-specific moral lexicons that allow for direct comparisons of linguistic cues for predicting morality across different domains. Among the studies focused on the cross-domain analysis of moral value in text through a supervised approach, we consider Liscio et al. (2022), van Luenen (2020), Huang, Wormley, and Cohen (2022), Guo, Mokhberian, and Lerman (2023), Nguyen et al. (2024), and Preniqi, Ghinassi, Kalimeri, and Saitis (2024). Liscio et al. (2022) evaluates the performance of supervised systems for cross-domain value detection using the MFTC. van Luenen (2020) investigate the performance of a supervised model trained on non-extremist Twitter data and tested on extremist forum data, demonstrating that cross-domain classification is feasible, though with some performance reduction. This study compares the efficacy of Word2Vec and BERT embeddings, finding that BERT offers slightly better generalization capabilities. Huang et al. (2022) present the Learning to

[1] All experiment results and the corresponding code for replication are available at the following link: https://osf.io/kw6rc/?view_only=73c3d8e2f7c9483ea012347d70d27b6f

Adapt Framework (L2AF), designed to address domain variations in morality classification when training and testing data come from different domains. L2AF includes a neural feature extractor, a prediction network for moral values, a weighting network for domain adaptation, and joint optimization. L2AF effectively adjusts to variations in data domains, demonstrating superior performance compared to existing methods when applied to cross-domain tasks using the MFTC dataset. Guo et al. (2023) explore moral sentiment recognition in text, highlighting challenges due to diverse annotated datasets in data collection methods, domains, topics, and annotation. Merging such heterogeneous datasets can hinder model generalization. To address this, they propose a data fusion framework using adversarial domain training to align datasets and a weighted loss function to balance label distributions, improving model generalizability. Nguyen et al. (2024) and Preniqi et al. (2024) both address the task of moral detection using Transformer-based models. Nguyen et al. train models on multiple data domains, while Preniqi et al. train on single, diverse datasets. Both studies then evaluate the models' performance in in-distribution and out-of-distribution scenarios.

The methods described above primarily rely on a supervised approach that requires a substantial amount of training data, making it impractical when such data is unavailable. Hence, our research prioritizes robust unsupervised techniques, which offer distinct advantages such as inherent domain adaptability and the mitigation of potential biases introduced by specific training datasets. Additionally, our study provides a comprehensive comparison of current unsupervised methods, including different LLMs and NLI approaches, focusing on the validation of these models in a novel scenario suited to tasks with a high degree of subjectivity.

Some work focuses on overcoming the limits of supervised methods, in particular when training data are insufficient or annotations are sparse. Indeed, moral value detection often involves highly subjective judgments, leading to limited training data and potential biases. Golazizian, Omrani, Ziabari, and Dehghani (2024) introduce a novel framework for efficient annotation collection and modeling. Their two-stage approach first uses a small group of annotators to build a multitask model. Then, they select and annotate a few samples per annotator to incorporate their unique perspective. To evaluate their framework at scale, they released a unique dataset, the Moral Foundations Subjective Corpus, containing 2000 Reddit posts annotated by 24 annotators for moral sentiment. Their findings demonstrate that the framework surpasses the previous state of the art in capturing individual annotator perspectives with only 25% of the usual annotation budget on two datasets. Additionally, their approach fosters more equitable models by reducing performance disparities among annotators. In a complementary approach, van der Meer, Falk, Murukannaiah, and Liscio (2024) propose Annotator-Centric Active Learning (ACAL). This framework incorporates an annotator selection strategy alongside data sampling. ACAL aims to achieve two objectives: efficiently capture the full spectrum of human judgments in subjective tasks, and evaluate model performance using annotator-centric metrics that prioritize minority perspectives over the majority. The authors experiment with various annotator selection strategies across seven subjective NLP tasks, using both traditional and newly developed human-centered evaluation metrics. Their results suggest that ACAL improves data efficiency and excels in capturing diverse human perspectives. However, its effectiveness depends on having a large and diverse pool of annotators for selection. Our research takes a distinct approach by focusing on unsupervised learning with LLMs. This eliminates the need for pre-labeled data, making it more scalable and adaptable to real-world scenarios with limited annotations. Additionally, we introduce a novel validation framework that goes beyond traditional approaches. Our framework assesses LLM performance in the moral domain without relying on supervised labels, providing a unique perspective on model effectiveness in subjective tasks.

2.2. Unsupervised methods

In the field of unsupervised learning, Mokhberian, Abeliuk, Cummings, and Lerman (2020) and Priniski et al. (2021) offer an approach based on the Frame Axis technique (Kwak, An, Jing, & Ahn, 2021). This approach employs procedures that project words onto microframe dimensions, defined by contrasting sets of words. Furthermore, several unsupervised methods make use of the extended version of the Moral Foundation Dictionary (MFD) (Hopp, Fisher, Cornell, Huskey, & Weber, 2021), which includes terms associated with the general morality, vices, and virtues of the five MFT dyads. By connecting MFD entries to WordNet, Kobbe, Rehbein, Hulpus, and Stuckenschmidt (2020) contribute by expanding and clarifying the lexicon inside a dictionary-based framework. A graph-structured data model is included in Hulpuş, Kobbe, Stuckenschmidt, and Hirst (2020)'s unsupervised method to investigate the capture of moral values using Knowledge Graphs (KG). Their research assesses the WordNet 3.1, ConceptNet, and DBpedia entities' relevance to MFT. A knowledge graph of moral values is provided in De Giorgis, Gangemi, and Damiano (2022), connecting WordNet, FrameNet, BabelNet, ConceptNet, and several others graph-shape resources to linguistic and factual value-related knowledge for value extraction. Recently, Asprino et al. (2022) introduced two distinct unsupervised approaches for detecting moral content in natural language. The first method employs KG to identify moral values in text, aiming to capture explicit references to moral concepts and values. The second approach uses a zero-shot machine learning model based on an NLI system as a zero-shot classifier. This method leverages the commonsense knowledge of the NLI model, fine-tuned on a commonsense dataset for language inference (MNLI) (Williams, Nangia, & Bowman, 2017). The NLI model's indirect acquisition of commonsense knowledge enables it to detect main values in the input text based solely on the semantic interpretation conveyed by the MFT label taxonomy. Additionally, incorporating the emotional tone of the input sentence significantly enhances the classifier's performance. Both methods are evaluated using the Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020), a dataset of 35,000 items across seven domains covering different socio-political areas, annotated with the MFT value taxonomy. Bulla, Gangemi, and Mongiovì (2024) investigate moral value detection by comparing the performance of supervised models in cross-domain scenarios with the results of unsupervised methods, leveraging abstract concepts and common sense information from LLMs and NLI. In the present work, we follow a more rigorous approach that considers the annotators' agreement, explores the use of more effective prompts and compares the performance of other LLMs (GPT-3.5 turbo, Llama2 and Mistral).

Very recent approaches focus on analyzing morality through unsupervised methods that leverage the capabilities of LLMs. Researchers such as Khandelwal, Agarwal, Tanmay, and Choudhury (2024), Almeida, Nunes, Engelmann, Wiegmann, and de Araújo (2023), Dillion, Mondal, Tandon, and Gray (2024), and Scherrer, Shi, Feder, and Blei (2024) are exploring this new research area by investigating the moral reasoning abilities of different LLMs using different validation settings. It is important to note that these works operate outside the MFT paradigm, focusing on evaluating the moral value of actions within a text rather than identifying and analyzing the moral content of sentences. This distinction highlights a different aspect of studying morality in text using automated tools, which proves to be complementary to our work. Specifically, Khandelwal et al. (2024) assess the moral reasoning frameworks of LLMs (i.e. ChatGPT, GPT-4, Llama2 Chat −70B) using the Defining Issues Test, a standardized tool that evaluates moral reasoning through ethical dilemmas. Their findings reveal significant variations in performance across different languages (i.e. Chinese, Hindi, Russian, Spanish, and Swahili), with weaker performance observed in Hindi and Swahili compared to others. Almeida et al. (2023) investigate how LLM reasoning aligns with human moral and legal frameworks. Replicating established psychological studies with GPT-4, they find that

the model's responses are generally aligned with human responses. However, models tend to exaggerate inherent human biases. This highlights the need for caution when using LLMs in psychological research. Dillion et al. (2024) explore LLMs as moral experts, finding that GPT-4's advice is rated higher in morality, trustworthiness, and thoughtfulness compared to human advice. Their studies suggest that LLMs are effective in emulate the human moral sense. Scherrer et al. (2024) conducts a large-scale empirical investigation, presenting LLMs with hypothetical moral scenarios with different levels of ambiguity. This study uses Gert's framework of "common morality"(Gert, 2004), based on ten fundamental rules, to create the scenarios. Through a methodology that includes manipulating the wording of questions and measuring uncertainty in answers, the study analyzes LLMs' preferences in concrete decision-making contexts, providing new insights into the understanding of the moral values encoded in these models. Scherrer et al.'s approach is distinguished by its focus on the assessment of moral beliefs, rather than alignment with specific norms, and introduces innovative metrics for a more in-depth assessment. Therefore, this work offers a complementary analysis to MFT-based studies, providing an alternative for studying morality in LLMs by analyzing specific decision scenarios.

Several studies explore the moral behavior of LLMs in specific contexts and tasks. These works aim to shed light on LLMs' moral reasoning capabilities by analyzing their understanding of morality within the framework of MFT. Abdulhai et al. (2023), Nunes, Almeida, de Araujo, and Barbosa (2024), Tlaie (2024) and Ji et al. (2024) all leverage the Moral Foundations Questionnaire (MFQ), a well-established tool for measuring MFT's core psychological foundations of moral judgment. Abdulhai et al. (2023) study biases in LLMs using the MFQ, finding that models like GPT-3 often lean towards conservative biases influenced by their training data. These biases can be manipulated by adversarial prompts, impacting tasks like charity donations and posing risks for targeted political content. Nunes et al. (2024) investigate moral hypocrisy in LLMs using the MFQ and the Moral Foundation Vignettes dataset. Their findings underlying that models like GPT-4 and Claude 2.1 show internal consistency but exhibit contradictory behavior between abstract values and concrete moral evaluations. Tlaie (2024) compare the moral profiles of different LLMs, noting that proprietary models are largely utilitarian, while open-weight models align more with values-based ethics. They also identify a strong liberal bias in most models, except for Llama2. Ji et al. (2024) introduce MoralBench, a novel framework for evaluating the moral reasoning capabilities of LLMs. This benchmark uses two main datasets: MFQ-30-LLM, based on the MFQ, to evaluate sensitivity to the six moral dimensions of the MFT; and MFV-LLM, based on the Moral Foundations Vignettes, which analyzes realistic scenarios of moral dilemmas in contexts without politicization. MoralBench is divided into two main evaluation methodologies: Binary Moral Evaluation, where LLMs answer "I agree" or "I disagree" to moral statements, and Comparative Moral Evaluation, where LLMs choose the more morally acceptable statement between two options. Their scores are determined by alignment with the average of human responses. Experiments with models like Zephyr (Tunstall et al., 2023), LLaMA2-70B, Gemma-1.1 (Team et al., 2024), GPT-3.5, and GPT-4 reveal that LLaMA-2 and GPT-4 excel in binary evaluation, while GPT-3.5 and GPT-4 outperform in comparative evaluation. However, inconsistencies between the two approaches suggest models often recognize patterns or keywords without a deep understanding of moral principles. While these prior works provide valuable insights into LLM moral reasoning, they primarily focus on analyzing LLM responses through questionnaires. Our research takes a different approach, focusing on automatic moral classification in textual data. We leverage unsupervised learning with LLMs to address the limitations of supervised models in real-world scenarios, where data distributions can be unpredictable. Additionally, we introduce an evaluation framework that directly compares model outputs with human annotations, ensuring a more accurate assessment of model performance. Furthermore, our methodology explores the effectiveness of different LLM sizes and prompt designs for optimal performance in moral value detection tasks, considering both multi-label and binary classification scenarios.

3. Theoretical grounding

Building on the idea of universal moral foundations, our work focuses on Graham and Haidt's MFT (Graham et al., 2013) as a framework. According to MFT, moral judgments vary significantly throughout cultures and over time, but there are core moral dimensions that form the foundation of an "intuitive" ethical system across human societies. This theory integrates four key perspectives on the origins and nature of morality: nativism, cultural development, intuitionism, and pluralism. MFT acknowledges the potential role of neurophysiological bases for moral responses (nativism), recognizes the influence of environmental factors on moral development (cultural-evolutionary), suggests that moral judgments arise from various cognitive processes (intuitionism), and allows for the existence of multiple moral frameworks across cultures (pluralism) (Graham et al., 2013). At the core of MFT there are five distinct moral dimensions or dyads, each representing a fundamental dimension of human moral reasoning:

- **Care/Harm**: This domain highlights the importance of avoiding harm and promoting the well-being of others. It is associated with virtues such as kindness, nurturance, and compassion.
- **Fairness/Cheating**: This domain focuses on the concepts of justice, equity, and reciprocity. It supports the notion of fairness in social interactions and promotes cooperation.
- **Loyalty/Betrayal**: This domain highlights the importance of loyalty to one's group and disapproval of betrayal. It promotes social cohesion and cooperation within groups.
- **Authority/Subversion**: This domain promotes respect for legitimate authority, social order, and hierarchies. It promotes adherence to rules and social norms.
- **Purity/Degradation**: This domain focuses on ideas of cleanliness and contamination, often extending to moral concepts. It can be linked to notions as spiritual purity and adherence to social taboos.

The five core moral dimensions constitute the foundation of MFT, offering a comprehensive framework for analyzing the multifaceted nature of human morality. Subsequent refinements to MFT introduced the "Liberty/Oppression" dimension and differentiated the "Fairness" dyad into "Equality" and "Proportionality" (Atari, Haidt et al., 2023). We consider the original 5-dimensions version of MFT since the "Liberty/Oppression" dimension is missing in most available datasets.

4. Methodology

We leverage a set of pre-trained LLMs and an NLI model as zero-shot ready-made unsupervised multi-label moral value classifiers. Our goal is to evaluate their effectiveness in identifying moral content in textual data.

Section 4.1 provides an overview of the LLMs we employ, including a comprehensive analysis of the prompting techniques we use. Section 4.2 delves into the architecture of the NLI-based model detailing the different configurations we implement.

4.1. LLM-based models

We employ GPT-4 (Achiam et al., 2023), GPT-3.5 (Brown et al., 2020; Ouyang et al., 2022),[2] Llama3-70B (AI@Meta, 2024), Mistral-7B (Jiang et al., 2023), and Mixtral-8x7B[3] and Llama2-7B (Touvron et al., 2023) models in unsupervised zero-shot and few-shot settings to assess LLMs' capability in detecting moral values in text. GPT-3.5 is specifically designed to excel at interpreting and executing instructions, aiming to deliver consistent and contextually relevant responses. Its

[2] We employ the GPT-3.5-turbo-instruct version.

[3] We employ the NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO version.

training process integrates Reinforcement Learning of Human Feedback (RLHF) (Ouyang et al., 2022), which enhances the model's ability to follow instructions and minimize the generation of erroneous or harmful outputs. RLHF has applied Proximal Policy Optimization (PPO) (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017), a reinforcement learning algorithm demonstrably effective in training an agent's decision-making capabilities to tackle complex tasks.

GPT-4 represents a significant advancement in the GPT series, emerging as a multimodal LLM capable of processing both text and image inputs while generating text outputs. Trained on a massive scale, with the exact number of parameters undisclosed, GPT-4 exhibits human-level performance across different benchmarks, surpassing the capabilities of its predecessor, GPT-3.5.[4] The model's development involved an iterative training process, incorporating adversarial testing and continuous feedback integration. This iterative process has resulted in notable improvements in factual accuracy, steerability (the ability to control and guide the model's output through instructions), and adherence to safety constraints.

Llama is a suite of pre-trained generative text models ranging from 7 billion to 405 billion parameters. We employ both the chat version of the Llama2-7B model and Llama3-70 model optimized for dialogue. The chat version prioritizes conversational fluency through further fine-tuning on tens of thousands quick response pairs and additional datasets. To ensure data quality, its training data underwent a rigorous filtering process, which excluded websites with known privacy issues and unreliable sources. Llama training has included RLHF with PPO, enhanced by the strategic assimilation of Rejection Sampling, a technique for acquiring observations from targeted distributions. Llama2 shows little differences with respect to Llama3, with the latter demonstrating notable advancements in computational efficiency and contextual understanding, as well as enhanced multi-turn dialogue consistency. Additionally, Llama3 incorporates a more expansive training dataset and utilizes refined alignment techniques to improve output reliability.

Mistral, an LLM with 7.3 billion parameters, addresses the challenges of processing extensive text and computational efficiency. Sharing a similar size to the previously employed Llama2-7B model, Mistral leverages two key techniques: Sliding Window Attention (SWA) (Beltagy, Peters, & Cohan, 2020; Child, Gray, Radford, & Sutskever, 2019) and Grouped Query Attention (GQA) (Ainslie et al., 2023). SWA tackles the computational demands of long sequences by segmenting them into manageable 4096-token windows, effectively reducing memory footprint and accelerating inference. This allows Mistral to handle sequences with a total context length of 32,768 tokens. GQA departs from the conventional attention mechanism by grouping hidden states, leading to significant improvements in efficiency compared to traditional methods. We employ an "instructed" version of Mistral,[5] which has been fine-tuned specifically for processing chat-style instructions, incorporating additional optimizations for excelling in conversational interactions.

Mixtral, a composite model combining 8 Mistral-7B models into an ensemble (Mixtral-8x7B), represents an innovative approach to leveraging model diversity for improved performance. Fine-tuned with Direct Preference Optimization (DPO) (Rafailov et al., 2024), Mixtral achieves notable improvements in both response quality and contextual adaptability. The ensemble model harnesses the strengths of its constituent Mistral models, employing sophisticated alignment and optimization techniques to ensure consistency and coherence in dialogue tasks. By distributing computational loads and combining insights across models, Mixtral not only enhances robustness but also achieves superior reliability in generating nuanced, contextually appropriate outputs. This design underscores its efficacy in managing complex and high-stakes conversational interactions.

To ensure consistency, we adopt the same hyperparameters configuration for all LLMs employed. We set a low-temperature score of 0.10 to promote deterministic outputs, favoring the most probable token selections at each generation step. We set top-p nucleus sampling with a threshold of 0.90. Finally, we establish a maximum output length of 200 tokens and leave all the other parameters with their default values.

To detect moral values from text, we leverage the LLMs' inherent knowledge and contextual comprehension through prompting. As shown in Fig. 1, we develop two prompting configurations. The first, labeled "single-prompt", assesses the model's capacity for comprehensive in-context reasoning. This configuration presents a single prompt requiring the model to identify whether a sentence conveyed moral content and, if so, to simultaneously identify the specific MFT moral dimensions. This approach tests the model's ability to handle both binary and multi-label moral value detection tasks jointly. The second configuration, labeled "multi-taks", presents a two-step approach with separate instructions to simplify the task. Initially, the model focuses on determining the presence or absence of moral content in the sentence, answering in a binary format (i.e. "yes" or "no"). If the answer is positive, a subsequent prompt asks the model to tackle the multi-label task, predicting the moral dimensions in the sentence. As shown in Fig. 1, we design the prompt to encourage the model to generate binary responses (positive or negative) for each moral dimension. This approach aims to reduce the variability in the model's textual output during the detection task.

Building on these configurations, we adapt the structure of the 0-shot single-prompt setting to create the 6-shot and 12-shot few-shot prompts. These few-shot prompts incorporate randomly selected examples from the training set, appended as context to the prompt to guide the model's predictions. The examples are formatted consistently with the single-prompt approach, providing both positive and negative cases for each moral dimension. This design ensures alignment with the 0-shot structure while leveraging in-context learning to improve performance. The random sampling of examples minimizes selection bias and introduces variability, enabling an evaluation of the model's sensitivity to few-shot learning across different moral dimensions. To ensure coverage of all labels, we exclude combinations that do not encompass all labels.

A post-processing step aligns the model's predicted moral dimensions with the MFT framework. Here, the model's identified dimensions are directly mapped to their corresponding MFT labels, except for "Proportionality" and "Equality", which are mapped into "Fairness" to conform with the original version of MFT.

4.2. NLI-RoBERTa-based model

Our investigation into unsupervised methodologies for moral value detection explores the potential of NLI systems as zero-shot ready-made classifiers (Bulla, Gangemi & Mongiovì, 2023). The NLI model's capacity focuses on determining the relationship between two inputs text, a premise and a hypothesis. This involves determining the degree to which the hypothesis is entailed, contradicts or is neutral w.r.t the premise. Understanding the contextual relationship between text pairs allows NLI models to make inferences about their logical connections. Following (Asprino et al., 2022), we formulate hypotheses for each potential moral value using the input text as the premise. To perform this classification, we leverage pre-trained checkpoints of MNLI-RoBERTa-large,[6] a model trained on the MNLI dataset (Williams et al., 2017).

Our method focuses on two distinct configurations. The first configuration (i.e.'single-prompt') follows the methodology detailed in Asprino et al. (2022) (crf. Fig. 2). Here, we evaluate the entailment score for

[4] The number of parameters in the GPT-4 model is estimated to be about 1.7 trillion.

[5] We employ the Mistral-7B-Instruct version.

[6] https://huggingface.co/FacebookAI/roberta-large-mnli

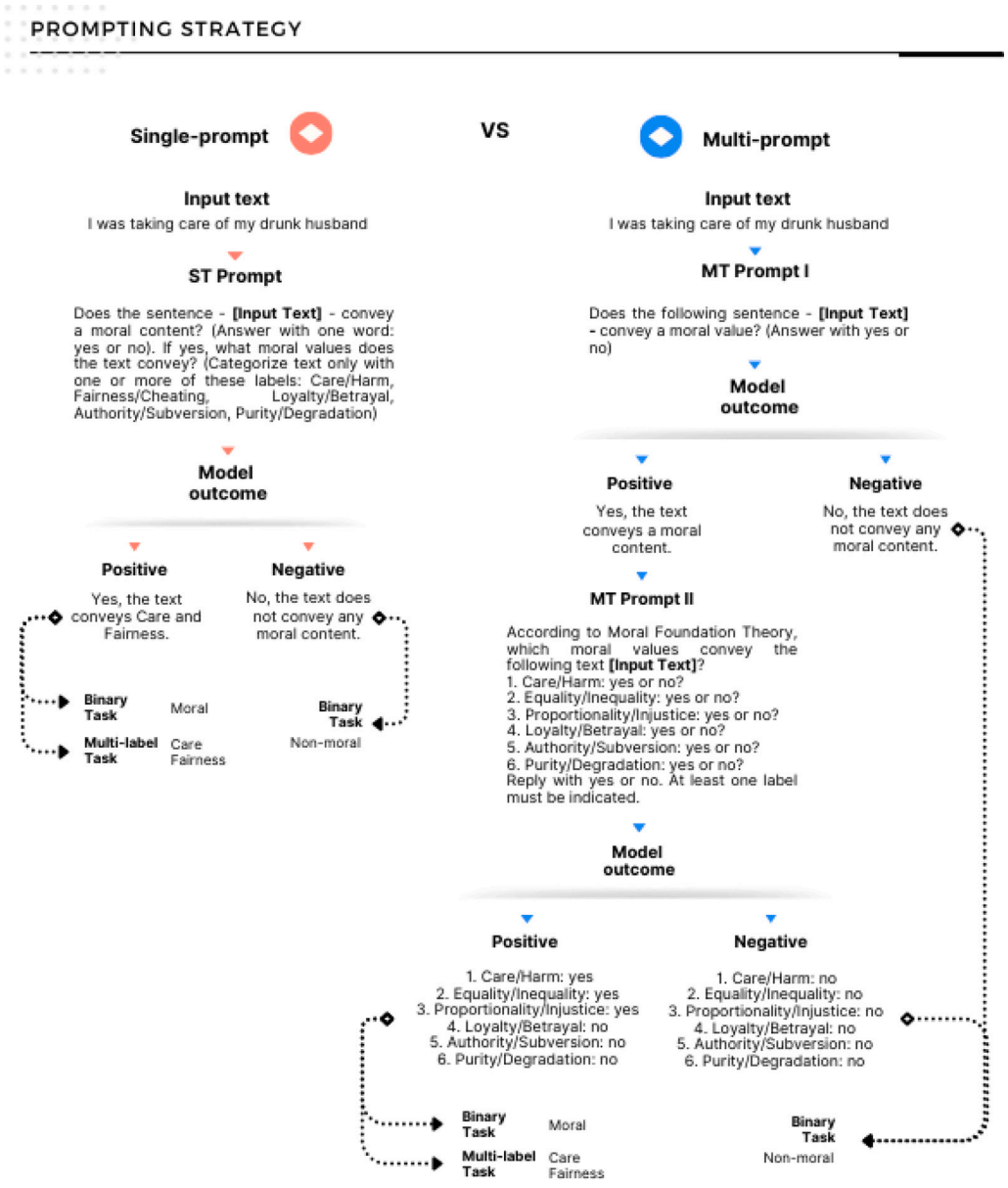


Fig. 1. Overview of the LLMs-based approach. The single-prompt configuration (left) leverages a single prompt to address both binary and multi-label tasks: identifying the presence of moral content and detecting the specific MFT dimension associated with the text. Conversely, the double-prompt configuration (right) adopts a two-step approach with separate prompts for each task. The first prompt focuses on a binary classification, which identifies whether the text conveys moral content or not. Text classified as moral then progresses to a second prompt that focuses on the detection of one or more MFT dimensions.

each value and violation using the pre-trained NLI-RoBERTa model. The entailment score is calculated by normalizing the model's predicted entailment and contradiction probabilities for a given moral value associated to the input text. We consider moral values with a normalized entailment score of 0.50 or higher as strongly associated with the text. Sentences with entailment scores below 0.50 for all moral values are classified as "Non-Moral", signifying a lack of discernible moral content.

The "double-prompt" configuration adopts a two-step approach. In the first stage, we leverage the methodology of Asprino et al. (2022) to perform a binary classification task. Here, the model assesses whether the input text expresses moral implications by comparing it with the term "moral", treated as a hypothesis. If the entailment score for the hypothesis exceeds a predefined threshold (i.e., 0.50), the text is classified as moral and moves to the second stage. Conversely, the sentence is classified as lacking moral content (i.e. "Non-Moral"), and the process ends. The second stage focuses on identifying moral foundations in text classified as moral in stage one. Following the approach of Bulla, Gangemi et al. (2023) we normalize entailment and neutrality scores to perform the classification.

5. Results and evaluation

In this section, we present an evaluation of the methods discussed in Section 4 in terms of their ability to detect and identify moral content. Specifically, our analysis aims to provide an answer to the following questions:

1. Which model exhibits the superior performance in recognizing moral dimensions within text? How the size of a model or different prompting strategies affect performances?
2. Considering the subjectivity of the task, can we consider the models' performance as satisfactory? How a human would perform on the same task?

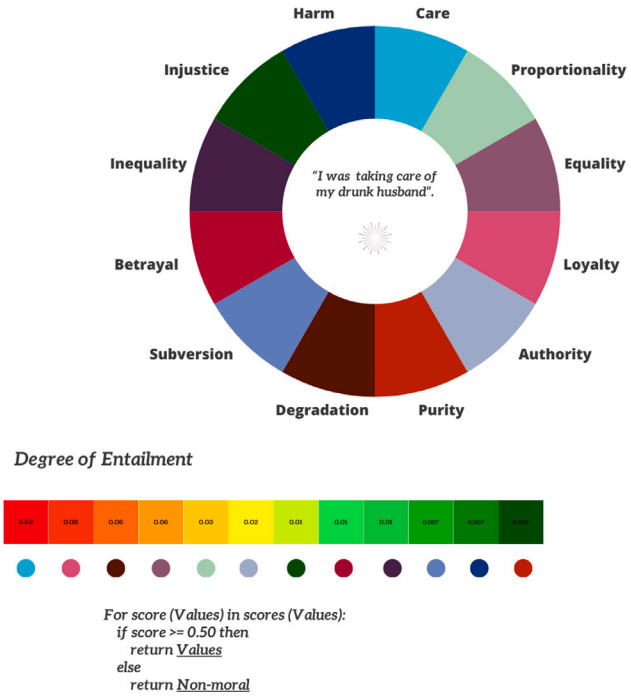


Fig. 2. Overview of NLI-based method: After processing the input sentence as a premise, the NLI model considers all of the MFT dimensions to be hypotheses, and assesses their connection with each other. As the classification outcomes, we choose labels having an entailment score of at least 0.50. If no label satisfies this requirement, the sentence is classified as lack of moral content ("Non-Moral").

Next we introduce the dataset utilized in our study (Section 5.1). Next we present and discuss a comparative analysis among various models, which aims to elucidate the strengths and weaknesses of each model in the context of moral content identification (Section 5.1). Last, in Section 5.2.1, we present an experimental framework designed to facilitate direct comparison between human and model performances.

5.1. Dataset

To assess the ability of the previously described models in detecting and classifying moral content, we use the MFRC dataset (Trager et al., 2022), which includes three sub-corpora with 5366, 5351, and 7169 items for the Everyday Moral Life, USA and French politics subcorpus, respectively. Reddit comments sourced from a comprehensive spectrum of 12 distinct subreddit communities. These comments have been annotated across the eight moral dimensions, under the theoretical framework detailed in Atari, Haidt et al. (2023) (crf. Section 3). Diverging from Atari et al. MFT version, we consider the MFT version of Graham et al. (2013), Haidt (2012), in which the dyads of "Equality" and "Proportionality" are grouped into the singular dimension "Fairness". This aggregation allows for a more concise and focused exploration of moral foundations, mitigating redundancy and maintaining the crucial nuances associated with fairness. Furthermore, we exclude the label "Thin Morality", as conceptualized by Atari, Mehl et al. (2023), due to its ambiguous nature. This exclusion arose from the inability to map "Thin Morality" onto any of the five established moral dimensions. Additionally, we take into account comments identified by the majority of annotators as devoid of moral content, designated with the label "Non-Moral". Finally, we exclude entries that have been both labeled as "Non-Moral" and attributed to one or more moral dimensions by annotators, to mitigate potential ambiguity in the corpus. The dataset includes comments from various subreddits, that recall topics about the USA politics, French politics, and everyday moral life. The Everyday Moral Life sub-corpus (i.e. EveryDay) focuses on moral judgments and expressions of moral emotions in non-political contexts, sourced from comments coming from four subreddits. The USA Politics sub-corpus (i.e. USA Politics) comprises comments from three subreddits, encompassing political moral language from diverse perspectives. The French Politics sub-corpus contains comments from subreddits associated with the French presidential race. Our analysis highlighted a low inter-annotator agreement for this sub-corpus , therefore we exclude it from our assessment. To reduce subjectivity and possible noise, we exclude items classified by a single annotator and by annotators with uncertain confidence levels. Eventually, we obtain 2837 items labeled by two to six different annotators for the Everyday Moral Life sub-corpus and 2757 items labeled by two to five different annotators for the USA Politics sub-corpus. To establish baseline performance metrics, we aggregated Reddit annotations by majority vote for each moral value, using a threshold of 50%.

5.2. Comparison among language models

We evaluate the performance of GPT-3.5, Mistral, Llama2, and NLI-RoBERTa models, on the "Everyday" and "USA Politics" sub-corpora of the MFRC. Our primary focus is to assess their ability to identify moral content and its associated dimensions based on MFT. We conduct a comparative analysis using two different prompting scenarios (crf. Section 4). In the single-prompt prompt configuration, we discriminate between moral and non-moral content and predict the corresponding dimensions with a single prompt. In the double-prompt prompt configuration, we use a two-stage approach: first, we classify items as moral or non-moral, and then, we associate these items with one or more moral dimensions.

We perform the evaluation from two distinct perspectives. In Section 5.2.1, we explore the models' capability to identify specific moral value dimensions. Conversely, Section 5.2.2 focuses on the ability to detect text that conveys moral content. Although the tools considered address both tasks simultaneously, we discuss the two aspects separately for the sake of clarity. All experiments have been performed on the same dataset, discussed in Section 5.1.

5.2.1. Identification of moral value dimensions

Tables 1 to 6 present the performances of the models in the identification of moral value detection in both single-prompt and double-prompt settings. Each model has been evaluated on the dataset introduced in Section 5.1.

Table 1 summarizes the overall performance of the models in terms of weighted precision, recall, and F1 score. The metrics have been averaged across the five dimensions of the MFT. Unlike other studies, we do not include a "Non-Moral" dimension to avoid bias. This decision stems from the observation that a significant number of items in the dataset do not convey any moral value and are easier to classify, which could falsely suggest superior performance.

The GPT-4 model demonstrates the highest overall performance, achieving an F1 score of 0.55 across all MFT dyads in the double-prompt configuration. This represents a significant improvement of over 7 percentage points compared to GPT-3.5, which exhibits competitive performance in both single- and double-prompt settings, attaining F1 scores of 0.50 and 0.48, respectively. A prompt analysis reveals that while GPT-4 shows performance comparable to GPT-3.5 in the single-prompt scenario, it demonstrates a substantial performance gain in the double-prompt setting, achieving an increase of over 5 percentage points and surpassing the 0.50 F1 score threshold. The performance decline is due to the binary classification, as discussed in Section 5.2.2. In the single-prompt configuration, GPT-4 exhibits a propensity to classify a greater proportion of content as moral rather than non-moral. This is likely because the model is less sensitive to more complex prompts, potentially interpreting the dyads as constraints that limit the number of positive predictions it can generate.

Table 1
Models overall performance in multi-label moral value detection under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.18	0.24	0.19	0.21	0.18	0.21
	Recall	0.44	0.17	0.44	0.11	0.44	0.14
	F1	0.24	0.20	0.23	0.13	0.24	0.16
Llama-2 (7B)	Precision	0.23	0.25	0.18	0.17	0.20	0.20
	Recall	0.68	0.63	0.65	0.48	0.66	0.55
	F1	0.34	0.35	0.28	0.24	0.30	0.29
Mistral (7B)	Precision	0.31	0.49	0.21	0.35	**0.56**	0.42
	Recall	0.43	0.54	0.30	0.35	0.17	0.45
	F1	0.32	0.51	0.19	0.34	0.24	0.43
Mixtral (8 × 7)	Precision	0.42	0.36	0.27	0.24	0.34	0.28
	Recall	0.73	0.83	0.63	0.74	0.68	0.78
	F1	0.51	0.49	0.24	0.36	0.43	0.41
Llama-3 (70B)	Precision	0.37	0.39	0.26	0.26	0.32	0.32
	Recall	**0.86**	0.85	0.76	**0.80**	0.81	**0.83**
	F1	0.51	0.52	0.38	0.39	0.44	0.45
GPT-3.5 (175B)	Precision	**0.57**	0.54	0.41	0.38	0.49	0.42
	Recall	0.63	0.63	0.48	0.47	0.55	0.55
	F1	0.58	0.51	0.42	0.42	0.50	0.48
GPT-4 (Unknown)	Precision	0.46	0.55	0.33	**0.42**	0.39	0.48
	Recall	0.81	0.77	0.73	0.64	0.77	0.70
	F1	0.58	**0.61**	0.42	**0.49**	0.50	**0.55**

Llama-70B and Mixtral exhibit F1 scores of 0.45 and 0.43, respectively, under their optimal prompting scenarios. Despite having significantly fewer parameters than GPT-based models and Llama-70B, Mixtral demonstrates competitive performance, achieving F1 scores of 0.43 and 0.41 in the single- and double-prompt settings, respectively. In contrast, Llama-70B achieves slightly higher F1 scores of 0.44 and 0.45 in the single- and double-prompt configurations, respectively.

Notably, Mistral and Llama2 achieve F1 scores of 0.43 and 0.30, respectively, under their optimal prompting configurations. Although Mistral has a significantly lower parameter count compared to Llama-70B, Mixtral, and the GPT-based models, it exhibits competitive performance relative to Llama-7B and is comparable to Mixtral in the single-prompt context. In the double-prompt configuration, Mistral achieves an F1 score of 0.43, although its performance decreases to 0.24 in the single-prompt scenario. These results highlight the importance of prompt complexity in influencing model performance, as models with limited in-context understanding capabilities, as smaller LLMs, may benefit more from the double-prompt scenario. Conversely, Mixtral, Llama-70B and GPT-3.5 demonstrate greater adaptability across both prompt configurations, suggesting its versatility in tackling this task efficiently. GPT-4, characterized by a substantially larger parameter count and enhanced generalization capabilities relative to GPT-3.5, demonstrates superior performance metrics. This improvement is particularly evident in dual-prompt configurations, where it achieves an F1 score that is 5 percentage points higher than GPT-3.5's best result.

Llama-based models exhibit similar performance across both scenarios. Llama-7B achieves F1 scores of 0.30 and 0.29 in the single- and double-prompt configurations, respectively, while Llama-70B attains F1 scores of 0.44 and 0.45 under the same conditions. In a comparison of models with equivalent parameter configurations, such as Llama-7B and Mistral, both demonstrate analogous performance in the single-prompt configuration. However, in the double-prompt scenario, Mistral exhibits superior performance, surpassing Llama-7B, which showed a notable decline in efficacy under this condition. Furthermore, while Mistral achieves higher precision, indicating its ability to accurately identify relevant moral dimensions, Llama-7B displayed significantly higher recall, suggesting a tendency to identify a broader range of moral dimensions, even potentially irrelevant ones, as shown by a lower precision. This aspect aligns closely with the performance characteristics observed in the Llama-70B model. This trade-off in precision and recall can render Llama-7B less effective, leading to inconsistent results and potentially introducing a positive bias, where the model tends to associate moral values with sentences even for neutral elements. In contrast, Mistral's strength in precision suggests greater selectivity in identifying moral dimensions, resulting in fewer but more accurate outputs, potentially improving the overall quality of predictions.

The NLI-RoBERTa-based model exhibited inferior performance compared to the LLMs, achieving F1 scores of 0.24 and 0.16 in the single- and double-prompt configurations, respectively.

Across the sub-corpora, all models performed significantly better on "EveryDay" content compared to "USA Politics" content. This discrepancy is likely due to the greater difficulty and subjectivity of the tasks in the "USA Politics" sub-corpus, which is supported by the lower level of inter-annotator agreement. Consequently, the subjectivity of these tasks makes it challenging to establish an objective gold standard, increasing the likelihood of mismatches between the model predictions and the gold standard. This observation is further supported by the analysis of individual annotators, detailed in Section 5.3. Among the models, GPT-4 and GPT-3.5 demonstrated the strongest overall performance in MFT detection. GPT-4 achieved F1 scores of 0.61 and 0.49 for the "EveryDay" and "USA Politics" sub-corpora, respectively. Similarly, GPT-3.5 exhibited F1 scores of 0.58 and 0.42 for the "Everyday" and "USA Politics" domains, respectively. Llama-70B, Mixtral, and Mistral exhibit lower, yet comparable performance relative to GPT-based models, considering their reduced parameter configurations. Specifically, under the optimal prompting scenario for the EveryDay subcorpus, these models attain F1 scores of 0.52, 0.51, and 0.51 for Llama-70B, Mixtral, and Mistral, respectively. In the USA Politics subcorpus, the F1 scores are 0.39, 0.36, and 0.34 for Llama-70B, Mixtral, and Mistral, respectively. In contrast, both NLI-RoBERTa and Llama-7B exhibited limitations in this task. This is evident also in the sub-corpora analysis. Specifically, although Llama-7B maintained consistent performance across different prompting scenarios, its F1 scores remained consistently lower than other models.

Tables 2 to 6 present the detailed performance of all models for every dimension in terms of precision, recall, and F1 score. Notably, GPT-4 and GPT-3.5 consistently demonstrate strong performance, particularly in detecting dimensions like "Care" (i.e. Table 2) and "Fairness" (i.e. Table 3). Under their optimal prompting scenario, GPT-4 attains F1 scores of 0.69 and 0.53 for these respective dimensions,

Table 2
Models performance for predicting the MFT "Care" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.21	0.33	0.17	0.22	0.19	0.21
	Recall	0.55	0.22	0.64	0.18	0.58	0.10
	F1	0.31	0.27	0.27	0.20	0.29	0.13
Llama-2 (7B)	Precision	0.34	0.37	0.22	0.20	0.28	0.30
	Recall	0.78	0.73	**0.78**	0.52	0.78	0.64
	F1	0.47	0.49	0.34	0.29	0.41	0.41
Mistral (7B)	Precision	0.48	0.63	0.35	0.36	0.75	0.52
	Recall	0.51	0.67	0.26	0.43	0.17	0.58
	F1	0.49	0.65	0.30	0.40	0.28	0.55
Mixtral(8 × 7B)	Precision	0.58	0.49	0.33	0.26	0.47	0.37
	Recall	0.75	0.89	0.51	0.82	0.66	0.86
	F1	0.66	0.63	0.40	0.40	0.55	0.52
Llama-3 (70B)	Precision	0.51	0.53	0.36	0.35	0.45	0.45
	Recall	0.91	**0.92**	0.67	0.77	0.82	**0.86**
	F1	0.65	0.67	0.47	0.48	0.58	0.59
GPT-3.5 (175B)	Precision	**0.76**	0.66	0.55	0.40	**0.69**	0.54
	Recall	0.69	0.72	0.38	0.56	0.57	0.66
	F1	0.72	0.69	0.45	0.47	0.63	0.60
GPT-4 (Unknown)	Precision	0.62	**0.76**	0.45	**0.56**	0.55	**0.69**
	Recall	**0.92**	0.80	0.61	0.54	0.80	0.70
	F1	0.74	**0.78**	0.51	**0.55**	0.65	**0.69**

Table 3
Models performance for predicting the MFT "Fairness" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.18	0.19	0.25	0.25	0.21	0.21
	Recall	0.33	0.13	0.31	0.08	0.32	0.14
	F1	0.23	0.15	0.27	0.12	0.25	0.17
Llama-2 (7B)	Precision	0.17	0.18	0.22	0.20	0.20	0.19
	Recall	0.66	0.69	0.72	0.57	0.70	0.62
	F1	0.27	0.28	0.34	0.30	0.31	0.29
Mistral (7B)	Precision	0.14	0.48	0.15	0.45	0.54	0.46
	Recall	0.07	0.48	0.06	0.34	0.16	0.40
	F1	0.09	0.48	0.09	0.39	0.25	0.43
Mixtral (8 × 7B)	Precision	0.34	0.26	0.34	0.29	0.34	0.27
	Recall	0.71	0.87	0.72	0.86	0.71	0.87
	F1	0.46	0.40	0.46	0.43	0.46	0.41
Llama-3 (70B)	Precision	0.28	0.30	0.25	0.28	0.26	0.29
	Recall	**0.90**	0.81	**0.96**	0.86	**0.93**	0.84
	F1	0.43	0.44	0.39	0.43	0.41	0.43
GPT-3.5 (175B)	Precision	**0.47**	0.38	0.44	**0.46**	**0.45**	0.42
	Recall	0.63	0.63	0.60	0.48	0.62	0.54
	F1	**0.54**	0.47	0.51	0.47	0.52	0.47
GPT-4 (Unknown)	Precision	0.33	0.38	0.31	0.43	0.32	0.40
	Recall	0.86	0.81	0.90	0.76	0.88	0.78
	F1	0.48	0.52	0.46	**0.54**	0.47	**0.53**

whereas GPT-3.5 achieves F1 scores of 0.63 and 0.52 for the same dimensions. Furthermore, GPT-based models exhibit a performance difference between the "Everyday" and "USA Politics" sub-corpora. In the "EveryDay" sub-corpus, which primarily consists of non-politically charged content and benefits from the higher inter-annotator agreement, GPT-4 and GPT-3.5 achieve F1 scores of 0.78 and 0.72 for "Care", respectively. This contrasts with the lower F1 scores of 0.55 and 0.47 obtained in the "USA Politics" sub-corpus. This discrepancy might be attributed to the lower consistency in human judgments in the "USA Politics" sub-corpus compared to the "Everyday" sub-corpus. For the "Fairness" dimension, GPT-based models demonstrate more balanced performance, with GPT-4 and GPT-3.5 achieving F1 scores of 0.52 and 0.54 for the EveryDay subcorpus, and 0.54 and 0.51 for the USA Politics subcorpus, respectively.

Llama-70B, Mixtral and Mistral also show promising results, achieving an overall F1 score of 0.59, 0.55, and 0.55 for the "Care" dimension, with a higher score (i.e. 0.67, 0.66, and 0.65) in the "EveryDay" sub-corpus compared to "USA Politics" (i.e. 0.48, 0.40, and 0.40). The "Fairness" dimension reflects the trend observed in GPT-based models, with Llama-70B and Mixtral exhibiting balanced performance, achieving F1 scores of 0.44 and 0.46 for the EveryDay subcorpus, and 0.43 and 0.46 for the USA Politics subcorpus, respectively. In contrast, Mistral demonstrates superior performance in the EveryDay subcorpus, attaining an F1 score of 0.48, compared to 0.39 in the USA Politics subcorpus. Notably, the double-prompt prompting scenario proves to be beneficial for Llama-70B, Mixtral and Mistral in most cases.

Table 4
Models performance for predicting the MFT "Purity" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.09	0.09	0.09	0.08	0.09	0.09
	Recall	0.50	0.13	0.53	0.07	0.51	0.10
	F1	0.15	0.11	0.15	0.08	0.15	0.09
Llama-2 (7B)	Precision	0.06	0.07	0.07	0.07	0.06	0.07
	Recall	0.16	0.11	0.17	0.08	0.16	0.09
	F1	0.08	0.09	0.10	0.07	0.09	0.08
Mistral (7B)	Precision	0.13	0.18	0.12	0.14	0.41	0.16
	Recall	0.60	0.35	0.46	0.27	0.11	0.31
	F1	0.21	0.24	0.19	0.18	0.17	0.21
Mixtral (8 × 7B)	Precision	0.14	0.23	0.09	0.16	0.11	0.20
	Recall	**0.77**	0.62	0.58	0.34	0.68	0.49
	F1	0.24	0.33	0.15	0.22	0.19	0.28
Llama-3 (70B)	Precision	0.21	0.19	0.18	0.18	0.20	0.18
	Recall	0.74	0.75	**0.66**	0.59	**0.70**	0.66
	F1	0.33	0.30	0.28	0.27	0.31	0.29
GPT-3.5 (175B)	Precision	**0.29**	0.28	0.23	0.24	0.27	0.26
	Recall	0.35	0.38	0.22	0.21	0.29	0.30
	F1	0.32	0.32	0.22	0.22	0.28	0.28
GPT-4 (Unknown)	Precision	0.28	0.29	**0.32**	0.31	**0.30**	**0.30**
	Recall	0.27	0.58	0.18	0.38	0.23	0.49
	F1	0.27	**0.39**	0.23	**0.34**	0.26	**0.37**

Table 5
Models performance for predicting the MFT "Loyalty" dimension in terms of precision. recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model. while bold values highlight the overall best-performing models.

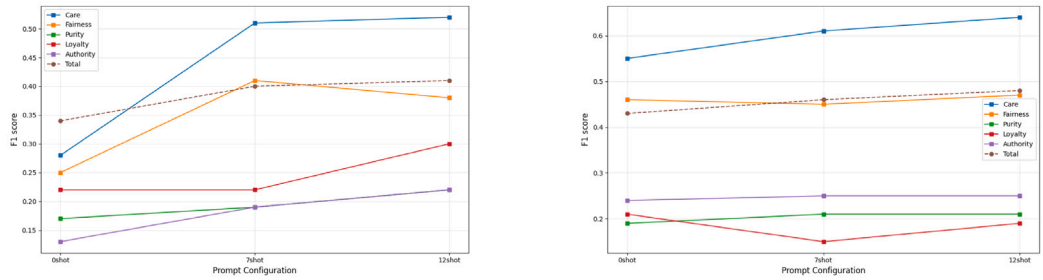
Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.21	0.15	0.30	**0.32**	0.26	0.20
	Recall	0.08	0.09	0.12	0.06	0.10	0.07
	F1	0.11	0.11	0.17	0.11	0.14	0.11
Llama-2 (7B)	Precision	0.05	0.05	0.05	0.06	0.05	0.05
	Recall	0.90	0.72	**0.83**	0.59	**0.86**	0.65
	F1	0.10	0.09	0.10	0.10	0.10	0.10
Mistral (7B)	Precision	0.11	0.19	0.11	0.18	0.17	0.18
	Recall	0.79	0.46	0.67	0.30	0.32	0.38
	F1	0.19	0.27	0.19	0.22	0.22	0.25
Mixtral (8 × 7B)	Precision	0.13	0.13	0.12	0.13	0.13	0.13
	Recall	0.79	0.73	0.59	0.54	0.69	0.63
	F1	0.22	0.23	0.20	0.21	0.21	0.22
Llama-3 (70B)	Precision	0.16	0.15	0.16	0.13	0.16	0.14
	Recall	0.82	**0.84**	0.60	0.76	0.71	0.80
	F1	0.27	0.26	0.26	0.22	0.26	0.24
GPT-3.5 (175B)	Precision	0.16	0.23	0.18	0.28	0.17	0.26
	Recall	0.67	0.53	0.62	0.44	0.64	0.49
	F1	0.26	0.32	0.28	0.34	0.27	0.33
GPT-4 (Unknown)	Precision	**0.34**	0.24	0.28	0.28	**0.30**	0.26
	Recall	0.64	0.73	0.73	0.65	0.69	0.69
	F1	**0.44**	0.36	**0.40**	0.39	**0.42**	0.38

Identifying moral values like "Purity", "Loyalty", and "Authority" presented a greater challenge for all models compared to other dimensions (crf. Tables 4–6). GPT-4 achieves F1 scores of 0.37, 0.42, and 0.30 for these dimensions, respectively, under its optimal prompting scenario, while GPT-3.5's scores are 0.28, 0.33, and 0.29, respectively. Further analysis of GPT-4, the best-performing model, revealed that for "Purity" and "Loyalty", the model achieved its strongest performance on the "Everyday" sub-corpus. While, for "Authority" moral dimension, the model achieved its strongest performance on the "USA Politics" sub-corpus. This suggests that references to moral foundations like "Authority" might be more explicitly expressed and frequently encountered in political contexts.

To evaluate the impact of few-shot prompting configurations on the task of moral values detection, we assess the performance of competitive open models of different dimensions in the zero-shot configuration (cf. Table 1), specifically Mistral-7B and Mixtral-7x8B. The analysis involves three prompting configurations: zero-shot (no examples in the prompt), 6-shot (six examples from the training set), and 12-shot (twelve examples from the training set). We measure performance variations for each model across individual moral dimensions in terms of F1 score (cf. Fig. 3). The analysis reveals distinct trends in the performance of Mistral and Mixtral across the three prompting configurations, with Mixtral generally outperforming Mistral across most dimensions and configurations. Considering overall performance (i.e. "Total"), Mistral

Table 6
Models performance for predicting the MFT "Authority" dimension in terms of precision, recall and F1 score under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.03	0.03	0.08	0.07	0.05	0.05
	Recall	0.47	0.14	0.48	0.09	0.47	0.10
	F1	0.06	0.05	0.13	0.08	0.10	0.07
Llama (7B)	Precision	0.05	0.06	0.14	0.11	0.10	0.10
	Recall	0.25	0.18	0.28	0.28	0.27	0.26
	F1	0.08	0.09	0.18	0.16	0.14	0.14
Mistral (7B)	Precision	0.07	0.12	0.13	0.26	0.30	0.22
	Recall	0.70	0.14	0.73	0.24	0.09	0.21
	F1	0.12	0.13	0.22	0.25	0.13	0.21
Mixtral (8 × 7B)	Precision	0.11	0.09	0.16	0.16	0.14	0.14
	Recall	0.55	0.51	0.70	0.55	0.66	0.54
	F1	0.18	0.16	0.26	0.25	0.24	0.22
Llama-3 (70B)	Precision	0.12	0.10	0.15	0.13	0.14	0.12
	Recall	0.42	0.64	0.58	0.85	0.54	0.79
	F1	0.18	0.18	0.24	0.22	0.22	0.21
GPT-3.5 (175B)	Precision	**0.16**	0.12	0.25	**0.26**	**0.22**	0.20
	Recall	0.32	0.36	0.43	0.41	0.40	0.39
	F1	0.21	0.18	**0.32**	**0.32**	0.29	0.27
GPT-4 (Unknown)	Precision	0.10	**0.16**	0.14	0.20	0.13	0.19
	Recall	**0.68**	**0.68**	**0.87**	0.67	**0.82**	0.67
	F1	0.18	**0.26**	0.23	0.31	0.22	**0.30**



(a) Mistral (7B) (b) Mixtral (7x8B)

Fig. 3. Performance comparison of Mistral-7B and Mixtral-3x8B models in few-shot prompting configurations for moral values detection. The analysis evaluates the impact of zero-shot, 7-shot, and 12-shot configurations on the models' performance across individual moral dimensions in terms of F1 score.

improves from an F1 score of 0.34 in the 0-shot configuration to 0.41 in the 12-shot configuration. Mixtral, however, demonstrates a higher starting point, achieving an F1 score of 0.43 in zero-shot and improving to 0.48 in the 12-shot configuration, underscoring its robustness and stronger overall performance. For the "Fairness"dimension, Mistral exhibits notable improvement, increasing its F1 score from 0.25 in the 0-shot configuration to 0.41 in the 6-shot configuration and 0.38 in the 12-shot configuration. Mixtral, by contrast, starts at a substantially higher F1 score of 0.46 in the 0-shot configuration and maintains stable performance across all configurations, achieving F1 scores between 0.45 and 0.47. In the "Care" dimension, both models benefit significantly from few-shot prompting. Mistral improves from an F1 score of 0.28 in 0-shot to 0.51 in 6-shot and 0.52 in 12-shot. However, Mixtral demonstrates superior performance, starting with an F1 score of 0.55 in zero-shot and further improving to 0.61 and 0.64 in the 6-shot and 12-shot configurations, respectively. The more complex dimensions, such as "Purity"and "Authority", pose challenges for both models. In "Purity", Mistral records F1 scores ranging from 0.17 in 0-shot to 0.22 in 12-shot, while Mixtral performs slightly better, achieving F1 scores between 0.19 and 0.21 across configurations. For "Authority", Mistral starts at 0.13 in 0-shot and improves to 0.22 in 12-shot, whereas Mixtral consistently outperforms Mistral, with F1 scores between 0.24 and 0.25. The "Loyalty"dimension follows a similar pattern. Mistral shows progressive improvement from 0.22 in 0-shot to 0.30 in 12-shot. In contrast, Mixtral starts at 0.21 in 0-shot but exhibits a slight decline in the few-shot configurations, with scores of 0.15 and 0.19 in 6-shot and 12-shot, respectively. Overall, these results demonstrate that Mixtral generally achieves better performance than Mistral, particularly in the more interpretable dimensions such as "Fairness" and "Care". Both models, however, benefit from few-shot prompting configurations, with improvements observed in the more challenging dimensions compared to the 0-shot baseline. Notably, Mistral exhibits a more substantial relative improvement across configurations, with a 7-point gain in the 12-shot configuration compared to its 0-shot baseline, compared to a 5-point gain for Mixtral. This highlights the responsiveness of smaller models, such as Mistral, to few-shot learning strategies.

5.2.2. *Detection of text with moral content*

We assess the effectiveness of all models in discerning whether a text conveys or not moral content (as a binary variable). We evaluate both LLMs and the NLI-RoBERTa model performance for both the "Everyday" and "USA Politics" sub-corpora of the MFRC. Items annotated with at least one moral dimension by the majority of human annotators are considered positive cases (they convey moral content), while the remaining are considered as negative cases (lacking moral content).

Table 7 summarizes the overall performance in terms of accuracy, weighted precision, recall, and F1 score.

Table 7
Models overall performance in binary moral value detection under single- and double-prompt prompt configurations. Underlined values indicate the best-performing prompt configuration for each model, while bold values highlight the overall best-performing models.

Model	Metric	EveryDay		USA Politics		Total	
		Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Nli (355M)	Precision	0.33	0.47	0.35	0.42	0.34	0.45
	Recall	0.64	0.25	0.71	0.18	0.68	0.21
	F1	0.43	0.32	0.47	0.25	0.45	0.29
	Accuracy	0.38	0.63	0.43	0.62	0.41	0.62
Llama (7B)	Precision	0.37	0.44	0.35	0.37	0.36	0.41
	Recall	**0.98**	0.72	0.96	0.60	**0.97**	0.66
	F1	0.54	0.55	0.51	0.46	0.52	0.50
	Accuracy	0.38	0.56	0.36	0.48	0.37	0.52
Mistral (7B)	Precision	0.86	0.82	0.67	0.68	0.78	0.76
	Recall	0.46	0.62	0.24	0.43	0.35	0.53
	F1	0.60	0.71	0.35	0.53	0.49	0.62
	Accuracy	0.78	0.81	0.69	0.73	0.73	0.77
Mixtral (8 × 7B)	Precision	0.93	0.90	0.90	0.87	0.92	0.89
	Recall	0.66	0.68	0.48	0.52	0.56	0.59
	F1	0.77	0.77	0.63	0.65	0.70	0.71
	Accuracy	0.81	0.80	0.67	0.63	0.74	0.71
Llama (70B)	Precision	**0.99**	0.95	**0.99**	0.91	**0.99**	0.93
	Recall	0.53	0.62	0.39	0.44	0.45	0.52
	F1	0.69	0.75	0.56	0.59	0.62	0.67
	Accuracy	0.68	0.77	0.45	0.56	0.57	0.67
GPT-3.5 (175B)	Precision	0.82	0.90	0.65	0.79	0.73	0.85
	Recall	0.84	0.70	0.76	0.57	0.80	0.64
	F1	0.83	0.79	0.70	0.66	0.76	0.73
	Accuracy	0.87	0.86	0.77	0.80	0.82	0.83
GPT-4 (Unknown)	Precision	**0.99**	0.88	**0.99**	0.78	**0.99**	0.83
	Recall	0.69	0.87	0.48	0.71	0.57	0.79
	F1	0.81	**0.87**	0.65	**0.74**	0.72	**0.81**
	Accuracy	0.83	**0.90**	0.62	**0.81**	0.73	**0.86**

Among all models, GPT-4 achieves the highest performance under its optimal prompting configuration. It exhibits F1 and accuracy scores of 0.81 and 0.86, respectively. GPT-3.5 follows closely, achieving F1 and accuracy scores of 0.76 and 0.83, respectively. Llama2 achieves the lowest score among the LLMs, with F1 and accuracy scores of 0.52 for both metrics. Excepting GPT-3.5, all models exhibit a performance increase with the double-prompt prompting configuration. This configuration significantly improves Llama-70B, Mixtral and Mistral's performance, resulting in an accuracy of 0.67, 0.74 and 0.77 and an F1 score of 0.76, 0.67 and 0.62. The NLI-RoBERTa-based model exhibits the lowest overall performance among the evaluated models, achieving F1 and accuracy scores of 0.45 and 0.62, respectively, even under its optimal prompting configuration. Interestingly, the model demonstrates a slight improvement in moral content prediction accuracy in the double-prompt setting compared to the single-prompt configuration. This discrepancy might be attributed to the presence of non-moral content impacting the single-prompt metric calculations, potentially suggesting the model's enhanced ability to identify moral content under the double-prompt framework.

Subcorpus-level analysis reveals a significant performance disparity across all models, with superior results consistently observed in the "Everyday" subcorpus compared to the "USA Politics" subcorpus. GPT-4 demonstrates robust performance across both subcorpora. Under optimal prompting conditions, it achieves accuracy scores of 0.90 and 0.81, alongside F1 scores of 0.87 and 0.74, for the "Everyday" and "USA Politics" subcorpora, respectively. Among the smaller models, GPT-3.5 outperforms Llama-70B by 10 percentage points in accuracy, and by 6 percentage points when compared to Mistral and Mixtral. In terms of F1 score, GPT-3.5 leads by 8, 6, and 12 points for the "Everyday" subcorpus. For the "USA Politics" subcorpus, GPT-3.5 maintains this trend, achieving an accuracy of 0.80, compared to 0.56, 0.67, and 0.73 for Llama-70B, Mixtral, and Mistral, respectively, and an F1 score of 0.70, compared to 0.59, 0.65, and 0.53 for the same models.

Among the less parameterized models, Mistral consistently outperforms Llama-7B. In the "Everyday" corpus, the F1 score and accuracy differences are 16% and 25% points, respectively, while in the "USA Politics" subcorpus, the differences are 2% and 25% points for the F1 score and accuracy, respectively. In general, the overall disparity between F1 score and accuracy for each model is noteworthy. This is due to the unbalance between moral and non-moral items. F1 score summarizes precision and recall in predicting positive (moral) items, whereas accuracy considers both moral and non-moral items (the latter more prevalent in the dataset). The NLI-RoBERTa-based model exhibits consistent performance across both subcorpora; however, its scores remain significantly lower compared to the other models.

5.3. Human vs. models comparison

We present a framework for the evaluation of machine learning models in tasks characterized by high subjectivity and low inter-annotator agreement. This framework fosters a direct comparison between human and model performance by juxtaposing annotators' responses with model predictions. In short, we exclude an annotator from the calculation of the gold-standard annotations and consider them as a human predictor to be compared with model performance. In this way we ensure that the "human predictor" does not influence the gold standard. We repeat this process for each annotator and present both individual and aggregated results. The MFRC dataset employs six annotators for the "EveryDay" subcorpus and five for the "USA Politics" subcorpus. Since the "USA Politics" subcorpus provides comparable insights, we exclusively report the performance of the "EveryDay" subcorpus. To address variations in annotator coverage across dataset sentences, we employ six distinct versions of the "EveryDay" subcorpus. Each version focuses on items tagged by a single annotator, ensuring a minimum of three annotators per item. Within each version, each element is associated with two values: a distinct value assigned by the specific annotator (human predictor) and an aggregate value from the remaining annotators. The aggregate value reflects the majority consensus, determined by ratings exceeding or equaling a 50% threshold and it represents the gold standard (or ground truth) for that element.

Our results are summarized in three tables: Table 8 shows the results of comparing human annotators and models in a multi-label value detection scenario. In Table 9, the performance attained by the best annotator for each moral foundation is juxtaposed with the model's performance. Finally, Table 10 outlines the results of our comparison in a binary value detection task, focused on identifying the morality in the sentence.

Table 8 presents the performance of all models in a multi-label moral value detection task based on the five MFT dimensions. We analyze the results for each dyad in terms of weighted F1 score, which takes into account annotator support, and average F1 score, which disregards such support. The results indicate that GPT-4 and GPT-3.5 consistently outperform human annotators in three of the five moral dimensions ("Care", "Fairness", and "Purity"). Additionally, GPT-4 demonstrates superior performance in one additional dimension ("Loyalty"), highlighting its broader capability in this context. For the "Care" dyad, GPT-4 achieves the highest weighted F1 score (i.e. 0.83), whereas GPT-3.5 attains the best average F1 score (i.e. 0.77) compared to the human average (i.e. 0.73).[7] These results translate into a performance advantage of 4-10 percentage points. Similarly, GPT-3.5 shows superior performance in the "Fairness" (4-5 percentage point improvement) while GPT-4 outperforms across the "Purity"and "Loyalty" dimensions, achieving 9–10 and 5 percentage point improvements, respectively, as measured by F1 scores. Conversely, human annotators exhibit superior performance compared to GPT-based models in the "Authority" dimension regarding average F1 scores, with humans scoring 0.22 compared to GPT-4's 0.18. However, in terms of weighted F1, GPT-4 surpasses human performance (i.e. 0.25 vs. 0.20). In general, the results suggest that GPT-3.5 may be more proficient in the single-prompt prompting configurations, where it achieves comparable performance to human judgment in classifying individual value dyads. GPT-4, however, achieves superior performance across both single- and double-prompt configurations, matching human performance on three out of five dyads and yielding significant improvements in the double-prompt setting for the remaining two. Notably, GPT-3.5 excels in well-defined domains such as "Care" and "Fairness", demonstrating statistically significant improvements over human evaluation. However, the model has a drop in performance in more ambiguous domains, mirroring the complexities faced by human judgment in these areas. This is further supported by F1 scores falling below 0.50 for these MFT dimensions. Conversely, GPT-4 demonstrates consistent robustness, surpassing both GPT-3.5 and human annotators in these less well-defined domains, particularly for "Purity", "Loyalty", and "Authority". Among smaller models, Llama-70B exhibits competitive performance relative to Mixtral in the double-prompt configuration, with weighted average F1 scores of 0.57 and 0.54, respectively, and an identical average F1 score of 0.48. Despite this, both models underperform relative to human evaluators, with Llama-70B trailing human performance by a margin of 2–5 percentage points in weighted F1 scores. Similarly, Mistral achieved proficiency in the double-prompt prompting configuration, obtaining a weighted average F1 score of 0.54 and an average F1 score of 0.47. However, its performance falls short of human evaluation by a margin of 5 percentage points in the overall weighted F1 score. Conversely, Llama2 and NLI-RoBERTa models underperform, with weighted average and average F1 scores of 0.38 and 0.27, respectively, compared to the human benchmark of 0.59.

Table 9 provides a comparison between the top-performing human annotators and all models in terms of precision, recall, and F1 scores. Each MFT dimension is assessed in comparison with the best-performing human annotator. GPT-4 achieves an F1 score within 2 percentage points of the best human, indicating its proficiency in this task. GPT-3.5 follows closely, achieving an F1 score within 5 percentage points. In contrast, Llama-70B exhibits a 13 percentage point gap, while Mixtral and Mistral lag by 16 and 15 percentage points, respectively. Llama2 and NLI-RoBERTa show significantly larger gaps, trailing by 33 and 44 percentage points, respectively. These results are based on the optimal prompting configuration for each model. Specifically, Table 9 reveals notable variations in performance across MFT dimensions. In dyads considered to be more readable, such as "Care" and "Fairness", the performance of the best model (i.e. GPT-4) more closely reflects human evaluation. The model achieves F1 scores of 0.87 and 0.62 on these moral dimensions, compared to human averages of 0.87 and 0.61. In contrast, GPT-4 achieves F1 scores of 0.39, 0.25, and 0.29 in "Loyalty", "Authority", and "Purity", respectively. These scores are lower than human performance in these categories, which average 0.49, 0.67, and 0.40, respectively. GPT-3.5 shows a marginally better performance than GPT-4 on "Authority"and "Purity", with F1 scores higher by 2–4 percentage points, respectively. However, it underperforms relative to GPT-4 across the remaining dimensions. When comparing smaller models, Llama-70B outperforms Mixtral by 3–4 percentage points in "Care", "Fairness", and "Loyalty". Conversely, Mixtral demonstrates superior performance in "Authority"and "Purity", with gains of 4–9 percentage points over Llama-70B. Notably, Mistral shows comparable performance to Mixtral in the "Care"and "Authority"dyads, with respective gaps of 2% and 15%. The double-prompt configuration enhances Mistral's performance, enabling it to outperform in certain contexts. Additionally, we revisit a previously observed trend (crf. Sections 5.2.1 and 5.2.2) in model behavior, which reflects their inherent characteristics. Llama2 exhibits a bias towards high recall but lower precision, while Mistral demonstrates a tendency to predict fewer labels with greater precision. Notably, Mistral achieves a remarkably high overall precision score (i.e. 0.54 in the double-prompt configuration). However, the double-prompt prompting configuration in Mistral fails to yield optimal results, as it predicts fewer labels despite maintaining high precision.

Table 10 shows the results of the binary moral detection task (crf. Section 5.2.2) in terms of moral F1 score and accuracy. Notably, GPT-4 exhibits high performance, achieving scores closer to human annotators across most cases for both F1 score and accuracy. GPT-4 obtained a noteworthy overall F1 score of 0.81 and an overall accuracy of 0.94, closely following human performance of 0.84 F1 score and 0.88 accuracy. Similarly, GPT-3.5 demonstrates competitive performance with F1 and accuracy scores of 0.79 and 0.91, respectively. It is worth noting that the best-performing human annotator (i.e. "Annotator01") achieves F1 and accuracy scores of 0.92 and 0.95, respectively. This outperforms the best-performing model (i.e. GPT-4 in its best-prompting scenario) by a narrow margin of 2 and 1 percentage points, respectively. Finally, the analysis of individual annotators reveals a specific pattern associated with "Annotator00". This annotator predominantly assigns labels indicating the absence of moral content and leaves most of the items with moral content without annotations. Therefore the associated dataset is unbalanced towards items that do not convey moral content. As a result the accuracy is very close to 1 (approximated to 1 in Table 10) while the F1-score is significantly lower (0.67) since this metric is influenced by the low recall.

6. Discussion

Our study investigates the performance of LLMs in identifying moral content and its corresponding MFT dimensions in both zero-shot and few-shot settings. We compare GPT-4, GPT-3.5, Mitral, Mistral, Llama-70B, Llama-7B, and NLI-RoBERTa under different prompting configurations (i.e. single-prompt vs. double-prompt) and evaluate them against human annotators. The zero-shot approach ensures generalizability by preventing the models from overfitting to a specific training dataset. Consequently, the evaluation focuses on the LLMs' inherent ability to recognize and correctly contextualize moral concepts. In contrast, the

[7] we discuss this counter-intuitive result in Section 6

Table 8
Direct comparison of human and model performance in multi-label value detection task based on the MFT dimensions. Performance is evaluated using both weighted F1 score (i.e. considering annotator support) and average F1 score (i.e. disregarding annotator support). Performance refers to the MFRC's EveryDay subcorpus.

Dyad	Metrics	Annotator	NLI (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8 × 7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Care	Weigthed avr.	0.73	0.31	0.30	0.47	0.53	0.39	0.71	0.72	0.68	0.73	0.73	0.78	0.73	0.80	**0.83**
	Avg.	0.73	0.27	0.26	0.41	0.46	0.39	0.64	0.62	0.58	0.62	0.62	**0.77**	0.67	0.68	0.75
Fairness	Weigthed avr.	0.53	0.26	0.18	0.26	0.30	0.28	0.45	0.48	0.46	0.47	0.49	**0.58**	0.46	0.54	0.54
	Avg.	0.44	0.23	0.15	0.21	0.24	0.25	0.37	0.39	0.36	0.38	0.39	**0.48**	0.38	0.43	0.44
Loyalty	Weigthed avr.	0.35	0.27	0.15	0.12	0.10	0.16	0.20	0.26	0.28	0.30	0.28	0.31	0.30	**0.40**	0.39
	Avg.	0.29	0.22	0.11	0.10	0.09	0.14	0.17	0.21	0.22	0.25	0.23	0.26	0.25	**0.34**	0.33
Authority	Weigthed avr.	0.21	0.06	0.08	0.05	0.13	0.20	0.16	0.25	0.22	0.21	0.23	0.19	0.18	**0.25**	0.23
	Avg.	**0.22**	0.05	0.06	0.04	0.08	0.18	0.16	0.17	0.12	0.13	0.15	0.16	0.14	0.17	0.18
Purity	Weigthed avr.	0.29	0.13	0.07	0.06	0.05	0.13	0.16	0.21	0.28	0.29	0.26	0.24	0.35	0.13	**0.39**
	Avg.	0.24	0.11	0.06	0.05	0.05	0.12	0.13	0.17	0.23	0.25	0.21	0.22	0.29	0.12	**0.33**
Total	Weigthed avr.	0.59	0.27	0.23	0.34	0.38	0.31	0.54	0.49	0.54	0.49	0.57	0.63	0.57	0.56	**0.66**
	Avg.	0.57	0.24	0.20	0.29	0.33	0.31	0.47	0.48	0.45	0.48	0.48	**0.60**	0.52	0.54	**0.60**

Table 9
Direct comparison between the top performing human annotator and models' performance for each MFTs dimensions. Performance is evaluated in terms of precision. recall and F1 score. Performance refers to the MFRC's EveryDay subcorpus.

Dyad	Metrics	Annotator	Nli (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8 × 7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Care	Precision	0.86	0.23	0.40	0.34	0.41	**0.89**	0.73	0.64	0.53	0.62	0.61	0.88	0.71	0.73	0.86
	Recall	0.88	0.63	0.31	0.74	0.76	0.28	0.76	0.84	0.92	**0.98**	0.97	0.80	0.80	0.98	0.88
	F1	**0.87**	0.33	0.35	0.46	0.53	0.42	0.75	0.73	0.67	0.76	0.75	0.83	0.75	0.84	**0.87**
Fairness	Precision	0.85	0.22	0.28	0.19	0.23	**0.93**	0.53	0.40	0.33	0.35	0.39	0.56	0.47	0.44	0.49
	Recall	0.48	0.33	0.15	0.59	**0.73**	0.17	0.47	0.65	**0.89**	0.84	0.84	0.62	0.70	0.84	0.84
	F1	0.61	0.27	0.19	0.29	0.35	0.29	0.50	0.50	0.49	0.49	0.53	0.59	0.56	0.58	**0.62**
Loyalty	Precision	**0.47**	0.21	0.38	0.06	0.04	0.15	0.11	0.14	0.14	0.17	0.16	0.22	0.22	0.32	0.27
	Recall	0.50	0.39	0.17	**0.94**	0.56	0.33	0.28	0.83	0.78	0.78	0.83	0.83	0.56	0.50	0.72
	F1	**0.49**	0.27	0.23	0.11	0.08	0.20	0.16	0.23	0.24	0.28	0.27	0.35	0.32	0.39	0.39
Authority	Precision	**0.50**	0.04	0.05	0	0.10	**0.50**	0.29	0.13	0.05	0.10	0.11	0.18	0.12	0.15	0.14
	Recall	**1**	0.50	0.25	0	0.25	0.25	0.50	0.50	0.25	0.25	0.50	0.50	0.50	0.75	0.75
	F1	**0.67**	0.07	0.08	0	0.14	0.33	0.36	0.21	0.09	0.14	0.17	0.27	0.19	0.25	0.23
Purity	Precision	**0.45**	0.06	0.10	0	0	0.33	0.12	0.10	0.17	0.11	0.10	0.14	0.25	0.14	0.21
	Recall	0.36	0.43	0.14	0	0	0.07	0.29	**0.86**	0.64	0.50	0.50	0.14	0.50	0.14	0.50
	F1	**0.40**	0.11	0.12	0	0	0.12	0.17	0.18	0.27	0.18	0.16	0.14	0.33	0.14	0.29
Total	Precision	**0.73**	0.19	0.32	0.23	0.28	0.69	0.54	0.46	0.39	0.45	0.45	0.66	0.52	0.55	0.63
	Recall	0.71	0.50	0.25	0.66	0.66	0.23	0.60	0.79	0.88	**0.90**	0.89	0.71	0.72	0.85	0.84
	F1	**0.71**	0.27	0.27	0.33	0.38	0.32	0.56	0.55	0.52	0.58	0.58	0.66	0.59	0.65	0.69

Table 10
Direct comparison of human and model performance in the binary moral detection task in terms of accuracy, precision, recall, and F1 score. Performance refers to the MFRC's EveryDay subcorpus.

Annotators	Metrics	Annotator	Nli (355M)		Llama (7B)		Mistral (7B)		Mixtral Nous Hermes (8 × 7B)		Llama (70B)		GPT-3.5 (175B)		GPT-4(Unknown)	
			Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt	Single-Prompt	Double-Prompt
Annotator00	F1	**0.67**	0.02	0.04	0.01	0.03	0.29	0.18	0.04	0.05	0.02	0.04	0.29	0.40	0.06	0.20
	Accuracy	**1**	0.18	0.84	0.04	0.56	0.98	0.97	0.85	0.87	0.69	0.83	0.97	0.99	0.91	0.97
Annotator01	F1	**0.92**	0.40	0.38	0.49	0.55	0.64	0.78	0.77	0.78	0.69	0.78	0.87	0.83	0.84	0.90
	Accuracy	**0.95**	0.36	0.68	0.35	0.61	0.82	0.87	0.82	0.83	0.72	0.82	0.92	0.90	0.89	0.94
Annotator02	F1	0.89	0.44	0.37	0.53	0.56	0.64	0.76	0.80	0.81	0.73	0.80	**0.90**	0.85	0.88	0.90
	Accuracy	**0.93**	0.38	0.66	0.37	0.58	0.81	0.85	0.84	0.85	0.74	0.83	**0.93**	0.91	0.90	0.93
Annotator03	F1	0.68	0.47	0.31	0.53	0.58	0.58	0.72	0.76	0.78	0.67	0.77	**0.83**	0.76	0.86	0.90
	Accuracy	0.83	0.40	0.63	0.38	0.60	0.78	0.83	0.83	0.85	0.73	0.83	**0.88**	0.86	0.91	0.94
Annotator04	F1	0.73	0.50	0.38	0.59	0.61	0.66	0.78	0.81	0.80	0.75	0.80	**0.86**	0.84	0.84	0.89
	Accuracy	0.70	0.44	0.62	0.43	0.60	0.77	0.83	0.81	0.80	0.72	0.79	**0.88**	**0.88**	0.84	0.91
Annotator05	F1	**0.90**	0.49	0.35	0.66	0.67	0.65	0.82	0.88	0.90	0.83	0.86	0.88	0.86	0.89	0.94
	Accuracy	**0.90**	0.42	0.57	0.51	0.64	0.74	0.85	0.87	0.90	0.81	0.85	0.89	0.88	0.88	0.94
Total	F1	0.84	0.40	0.32	0.47	0.51	0.59	0.69	0.70	0.71	0.63	0.70	**0.79**	0.77	0.76	0.81
	Accuracy	0.88	0.37	0.67	0.34	0.60	0.82	0.86	0.83	0.84	0.73	0.82	**0.91**	0.90	0.89	0.94

few-shot approach assesses the models' potential to improve performance with minimal guidance.

GPT-4 emerges as the strongest performer, closely followed by GPT-3.5, achieving the highest accuracy in both moral content detection and multi-label classification of MFT dimensions. This likely stems from its larger size and superior contextual understanding capabilities. These features allow GPT-4 to grasp complex moral concepts without specific training and make subtle distinctions between different MFT dimensions. Interestingly, GPT-4's performance in the double-prompt configuration outperforms that of a single human annotator, on average, suggesting its potential to predict the aggregate human assessment of moral content. GPT-3.5 also outperforms human annotators, albeit to a lesser degree than GPT-4, particularly in the single-prompt configuration. The single-prompt configuration, while computationally less demanding, demonstrates robust performance, especially with smaller models. However, the double-prompt configuration proves more effective with larger models such as GPT-4, leveraging their advanced capabilities to achieve superior results.

Human annotator analysis yields surprising results. GPT-4 performs remarkably similarly to human annotators. The model exhibits internal consistency across all sub-corpora derived from human data, with minimal variability for each category. Furthermore, GPT-based models' overall performance seems to be superior to that of a single human annotator, especially for GPT-4 model. It might seem counter-intuitive that a tool outperforms humans in moral value detection since human values are defined by us. Therefore, we expect the human judgment to be "correct" by definition. However, given the subjectivity of the task, the gold standard is defined by averaging human judgments, therefore the superior performance of GPT-based models might suggest that its prediction resembles the average judgment of humans more closely than the judgment of a single average human. Notably, models excel at predicting high-level value dimensions like "Care" and "Fairness". However, they struggle with more specialized concepts like "Purity", "Loyalty", and "Authority". Interestingly, even humans face similar difficulties in these areas, suggesting these dimensions may require theoretical reconsideration due to their inherent ambiguity. As shown for the multi-label scenario, GPT-based models remain the top performer even in the binary moral content detection task, achieving results that match or surpass those of humans.

Among the smaller LLMs, Llama-70B demonstrates comparable performance to Mixtral in the single-prompt scenario and slightly surpasses Mixtral in the double-prompt scenario. This superior performance is likely attributable to Llama-70B's larger size relative to Mixtral. Regarding the GPT-3.5 model, the single-prompt configuration appears to be more effective than the double-prompt configuration, with the exception of Llama-70B, which exhibits similar performance across both prompting strategies. This observation highlights that these models can leverage their in-context learning capabilities to perform well even with a single prompt, thereby obviating the need for double interrogation. In contrast, the double-prompt configuration significantly enhances performance for larger models such as GPT-4 and smaller models like Mixtral and Llama-7B, underscoring the varying dependencies on prompting strategies based on model size and architecture.

Mistral outperforms Llama2 in both single-prompt and double-prompt settings, despite having the same size and training conditions. Additionally, the double-prompt configuration benefits Mistral more, potentially mitigating limitations in its in-context understanding capabilities. Mistral's responses are more concise and accurate, achieving high performance compared to humans in both binary and multi-label scenarios. In contrast, Llama2 appears not to be suitable for this task, exhibiting a bias towards predicting many labels and generating overly verbose and inconsistent responses. As expected, smaller models underperform both GPT-4 and GPT-3.5 across all tasks and prompting configurations. This is likely due to GPT-based models' larger size, granting them access to a broader range of implicit knowledge for understanding moral nuances. However, Mistral emerges as an alternative for binary and multi-label detection of moral values due to its good performance with a smaller size (7B) and lower computational requirements compared to GPT-4 and GPT-3.5 (175B). NLI-RoBERTa consistently underperforms, likely due to its smaller size and training focused on a specific task (entailment).

Subcorpus analysis reveals that all models perform better on the "Everyday" subcorpus compared to the "USA Politics" subcorpus. This is likely due to the higher level of inter-annotator agreement and the more straightforward moral content in the "Everyday" corpus. The "USA Politics" subcorpus, on the other hand, presents greater challenges due to its complexity and the presence of more ambiguous moral concepts.

The evaluation of LLM performance in a few-shot context reveals significant benefits associated with this configuration, particularly for smaller models such as Mistral. Notably, performance improves progressively as the number of examples increases, transitioning from a zero-shot setting to configurations incorporating 6 and 12 examples within a single-prompt scenario. These results underscore the models' ability to effectively utilize in-context learning, with smaller-scale models demonstrating substantial performance gains when provided with a limited set of task-specific examples.

Our findings have several important implications. First, they demonstrate the potential of LLMs for moral assessment, suggesting that they can be used reliably to identify and understand moral content in text. Second, they highlight the importance of model size and in-context understanding capabilities for moral content detection. Third, they suggest that the double-prompt configuration can be a beneficial prompting approach for smaller LLMs. Finally, they emphasize the need for further research on moral content detection in more complex and challenging contexts, such as the "USA Politics" subcorpus. Additionally, the performance on more ambiguous dimensions (i.e., "Purity", "Loyalty", and "Authority") of both models and humans, highlights these concepts' inherent ambiguity. The low inter-annotator agreement among humans suggests a potential need for a theoretical reevaluation of these dimensions.

7. Conclusion

In this paper, we focus on automatically classifying moral values in text in a real-world scenario. Our approach employs state-of-the-art LLMs (i.e. GPT-4, GPT-3.5, Mixtral, Mistral, Llama-70B and Llama-7B) trained on large pools of common-sense data. We consider models as zero-shot, unsupervised moral value classifiers and compare their performance with a smaller NLI-based model, which serves as a baseline. Furthermore, we extend our analysis to include a few-shot configuration, examining the impact of incorporating a limited number of task-specific examples on classification performance. Our research delves into the effectiveness of different LLMs in discerning moral content based on the MFT taxonomy across two distinct scenarios: multi-label and binary classification. We consider the identification of moral foundations as a multi-label task and the detection of sentences that convey moral content as a binary task. We evaluate different prompt configurations and design a comprehensive framework to compare the model's performance with human.

Significantly, GPT-4 emerges as the best performer in all tasks, underlining its potential as a reliable moral detector thanks to its contextual understanding capabilities. GPT-3.5 follows closely, further highlighting the strength of larger LLMs. Among smaller models, Llama-70B exhibits marginally better performance compared to Mixtral, with both models surpassing the smaller-scale LLMs, Mistral and Llama-7B. Interestingly, Mistral outperforms both Llama2 and NLI, indicating its potential for moral detection in contexts where computational resources are limited.

Overall, our results show the differences in performance among different domains, revealing that morality in expressions of everyday

situations is detected more effectively than morality expressed in sentences related to political situations. The analysis of human annotators underscores the challenges in achieving agreement, particularly in morally ambiguous dimensions, thus indicating a necessity for a theoretical reassessment. Furthermore, this analysis highlights the model's ability to match or even surpass human performance in predicting the average moral assessment, on specific moral dyads.

In conclusion, our research helps advance the field of automatic moral classification by demonstrating the potential of LLMs for reliable classification of moral content. Our results highlight the importance of model size, prompting strategies to improve model performance. Even at the same conditions, there is a substantial variability among models, probably due to different pre-training data and strategies. Furthermore, we highlight the need for further exploration of more opaque moral dimensions and a theoretical reconsideration of these ambiguous moral dimensions. Overall, our study significantly enhances the understanding of moral detection in text by addressing key challenges, such as developing versatile and generalizable models that are not overfitted to domain-specific training data. By leveraging state-of-the-art LLMs and creating a comprehensive evaluation framework, we offer valuable insights into the capabilities and limitations of these models across various moral contexts. Comparing model outputs with human judgments provides a novel perspective on the alignment between automated systems and human annotators, highlighting areas of strong performance as well as those requiring further refinement.

CRediT authorship contribution statement

Luana Bulla: Writing – original draft, Software, Resources, Methodology, Investigation, Conceptualization. **Stefano De Giorgis:** Writing – review & editing, Supervision. **Misael Mongiovì:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Aldo Gangemi:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge financial support from the H2020 projects TAILOR:Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization – EC Grant Agreement number 952215 – and the Next Generation EU Program with the Future Artificial Intelligence Research (FAIR) project, code PE00000013, CUP 53C22003630006, and by the Italian Research Center on High Performance Computing, Big Data and Quantum Computing (ICSC).

Data availability

Data will be made available on request.

References

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2023). Moral foundations of large language models. arXiv preprint arXiv:2310.15337.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

AI@Meta (2024). Llama 3 model card. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245.

Almeida, G. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2023). Exploring the psychology of GPT-4's moral and legal reasoning. arXiv preprint arXiv:2308.01264.

Asprino, L., De Giorgis, S., Gangemi, A., Bulla, L., Marinucci, L., Mongiovi, M., et al. (2022). Uncovering values: Detecting latent moral content from natural language with explainable and non-trained methods. In *Proceedings of deep learning inside out (DeeLIO 2022): the 3rd workshop on knowledge extraction and integration for deep learning architectures* (pp. 33–41). Association for Computational Linguistics.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2023). Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Atari, M., Mehl, M. R., Graham, J., Doris, J. M., Schwarz, N., Davani, A. M., et al. (2023). The paucity of morality in everyday talk. *Scientific Reports*, *13*(1), 5967.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bulla, L., Gangemi, A., & Mongiovì, M. (2023). Do language models understand morality? Towards a robust detection of moral content. In *Proceedings of VALE 2023: the 1st workshop on value engineering in AI*.

Bulla, L., Gangemi, A., & Mongiovì, M. (2024). Do language models understand morality? Towards a robust detection of moral content. arXiv preprint arXiv:2406.04143.

Bulla, L., Gangemi, A., et al. (2023). Towards distribution-shift robust text classification of emotional content. In *Findings of the association for computational linguistics: ACL 2023* (pp. 8256–8268).

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

De Giorgis, S., Gangemi, A., & Damiano, R. (2022). Basic human values and moral foundations theory in valuenet ontology. In *International conference on knowledge engineering and knowledge management* (pp. 3–18). Springer.

Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2024). Large language models as moral experts? GPT-4o outperforms expert ethicist in providing moral guidance.

Fulgoni, D., Carpenter, J., Ungar, L., & Preoţiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 3730–3736).

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on computational modeling of attitudes*.

Gert, B. (2004). *Common morality: Deciding what to do*. Oxford University Press.

Golazizian, P., Omrani, A., Ziabari, A. S., & Dehghani, M. (2024). Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. arXiv preprint arXiv:2402.14101.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., et al. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *vol. 47*, In *Advances in experimental social psychology* (pp. 55–130). Elsevier.

Guo, S., Mokhberian, N., & Lerman, K. (2023). A data fusion framework for multi-domain morality learning. *Vol. 17*, In *Proceedings of the international AAAI conference on web and social media* (pp. 281–291).

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (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.

Huang, X., Wormley, A., & Cohen, A. (2022). Learning to adapt domain shifts of moral values via instance weighting. In *Proceedings of the 33rd ACM conference on hypertext and social media* (pp. 121–131).

Hulpuş, I., Kobbe, J., Stuckenschmidt, H., & Hirst, G. (2020). Knowledge graphs meet moral values. In *Proceedings of the ninth joint conference on lexical and computational semantics* (pp. 71–80).

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., & Zhang, Y. (2024). MoralBench: Moral evaluation of LLMs. arXiv preprint arXiv:2406.04428.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., et al. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

Kennedy, B., Atari, M., Davani, A. M., Hoover, J., Omrani, A., Graham, J., et al. (2021). Moral concerns are differentially observable in language. *Cognition*, *212*, Article 104696.

Khandelwal, A., Agarwal, U., Tanmay, K., & Choudhury, M. (2024). Do moral judgment and reasoning capability of LLMs change with language? A study using the multilingual defining issues test. arXiv preprint arXiv:2402.02135.

Kobbe, J., Rehbein, I., Hulpus, I., & Stuckenschmidt, H. (2020). Exploring morality in argumentation. In *Proceedings of the 7th workshop on argument mining*. Association for Computational Linguistics, ACL.

Kwak, H., An, J., Jing, E., & Ahn, Y.-Y. (2021). FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, *7*, Article e644.

Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C., Kalimeri, K., et al. (2023). What does a text classifier learn about morality? An explainable method for cross-domain comparison of moral rhetoric. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 14113–14132).

Liscio, E., Dondera, A., Geadau, A., Jonker, C., & Murukannaiah, P. (2022). Cross-domain classification of moral values. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 2727–2745).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. In *International conference on social informatics* (pp. 206–219). Springer.

Nguyen, T. D., Chen, Z., Carroll, N. G., Tran, A., Klein, C., & Xie, L. (2024). Measuring moral dimensions in social media with mformer. *Vol. 18*, In *Proceedings of the international AAAI conference on web and social media* (pp. 1134–1147).

Nunes, J. L., Almeida, G. F., de Araujo, M., & Barbosa, S. D. (2024). Are large language models moral hypocrites? A study based on moral foundations. arXiv preprint arXiv:2405.11100.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Preniqi, V., Ghinassi, I., Kalimeri, K., & Saitis, C. (2024). MoralBERT: Detecting moral values in social discourse. arXiv preprint arXiv:2403.07678.

Priniski, J. H., Mokhberian, N., Harandizadeh, B., Morstatter, F., Lerman, K., Lu, H., et al. (2021). Mapping moral valence of tweets following the killing of george floyd. arXiv preprint arXiv:2104.09578.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*.

Scherrer, N., Shi, C., Feder, A., & Blei, D. (2024). Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, *36*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., et al. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Tlaie, A. (2024). Exploring and steering the moral compass of large language models. arXiv preprint arXiv:2405.17345.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Trager, J., Ziabari, A. S., Davani, A. M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., et al. (2022). The moral foundations reddit corpus. arXiv preprint arXiv:2208.05545.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., et al. (2023). Zephyr: Direct distillation of lm alignment. arXiv preprint arXiv:2310.16944.

van der Meer, M., Falk, N., Murukannaiah, P. K., & Liscio, E. (2024). Annotator-centric active learning for subjective NLP tasks. arXiv preprint arXiv:2404.15720.

van Luenen, A. F. (2020). Recognising moral foundations in online extremist discourse: A cross-domain classification study.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.