The Moralization Corpus: Frame-Based Annotation and Analysis of Moralizing Speech Acts across Diverse Text Genres

Maria Becker, Mirko Sommer, Lars Tapken, Yi Wan Teh, Bruno Brocai
Department of German Linguistics, Heidelberg University
Department of Computational Linguistics, Heidelberg University
{maria.becker, bruno.brocai}@gs.uni-heidelberg.de
{sommer, tapken, teh}@cl.uni-heidelberg.de

Abstract

Moralizations – arguments that invoke moral values to justify demands or positions – are a yet underexplored form of persuasive communication. We present the Moralization Corpus, a novel multi-genre dataset designed to analyze how moral values are strategically used in argumentative discourse. Moralizations are pragmatically complex and often implicit, posing significant challenges for both human annotators and NLP systems. We develop a frame-based annotation scheme that captures the constitutive elements of moralizations – moral values, demands, and discourse protagonists – and apply it to a diverse set of German texts, including political debates, news articles, and online discussions. The corpus enables fine-grained analysis of moralizing language across communicative formats and domains. We further evaluate several large language models (LLMs) under varied prompting conditions for the task of moralization detection and moralization component extraction and compare it to human annotations in order to investigate the challenges of automatic and manual analysis of moralizations. Results show that detailed prompt instructions have a greater effect than few-shot or explanation-based prompting, and that moralization remains a highly subjective and context-sensitive task. We release all data, annotation guidelines, and code to foster future interdisciplinary research on moral discourse and moral reasoning in NLP.

Keywords: moralization, moral values, moral frames, corpus creation, annotation, LLMs, evaluation

1. Introduction

Recently, an increasing number of studies at the interface of Natural Language Processing (NLP) and Computational Social Science (CSS) have addressed the task of modeling morality in text, reflecting the growing interest in exploring moral phenomena through computational means. Most of this work has either focused on predicting moral values from text (e.g., Morteza Dehghani and Gratch, 2014; Zhang and Counts, 2016; Diakopoulos et al., 2014), or on analyzing moral biases in large language models (LLMs) (Schramowski et al., 2022; Hendrycks et al., 2021; Hämmerl et al., 2023; Jiang et al., 2021; Fraser et al., 2022, among others). However, the pragmatic patterns of moralizations – that is, how moral values are strategically used in argumentative contexts to justify demands or stances – have not yet been systematically modeled in NLP research.

We understand moralizations as persuasive strategies in which moral values are invoked to describe controversial topics and to demand specific actions or judgments (Felder and Müller, 2022; Becker et al., 2023). Three examples appear in Table 1. In moralizing practices, vocabulary associated with moral values (e.g., *freedom, justice, security, inequality*) serves to reinforce a demand by linking it to widely shared moral norms (Haidt et al., 2009; Graham et al., 2013). For instance, in the sentence *We should introduce a refugee cap in order to ensure the safety of Germans*, the term *safety* functions as a moral justification for a political demand. As the examples in Table 1 illustrate, moralizations can take many forms and occur in a broad range of contexts – political, social, religious, and even scientific – beyond explicitly populist or manipulative discourse. That said, our goal is not to assess moralizations in terms of being "good" or "bad," but rather to examine their linguistic realization and discourse functions.

(1) *We should all stop eating meat because it causes* ***unnecessary suffering to animals****.*
(2) *Women still earn less than men, even though* ***equality between men and women*** *is enshrined in the Basic Law.*
(3) *Immigrants are taking jobs from* ***hardworking citizens*** *and undermining* ***our values****.*

Table 1: Examples of moralizations from our dataset (moral phrases in bold) illustrate how moral values support explicit (Ex. 1) or implicit (Ex. 2–3) demands and occur in both populistic (Ex. 3) and non-populistic (Ex. 1–2) contexts.

While previous computational studies have primarily modeled morality through simplified categorical frameworks (e.g., one moral value per tweet or sentence), the complexity and heterogeneity of moralizations as a pragmatic phenomenon call for a more nuanced and structured approach. In this

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7069–7091
11–16 May 2026.

paper, we therefore propose a novel annotation framework for modeling moralization frames in text, which captures the interplay between moral values, demands, and discourse protagonists across multiple text genres. Annotating moral values and demands links values to concrete prescriptions, revealing how moral arguments serve as legitimation strategies. Identifying protagonists as moral agents, beneficiaries, or culprits uncovers the social dynamics behind such arguments. Thus, being able to identify and analyze moralizations in different text genres provides a valuable foundation for interesting research e.g. in linguistics, social and political sciences.[1]

Our framework is designed to operationalize moralizations in a way that makes their linguistic and pragmatic properties empirically accessible. It is holistic, in that it integrates moral, rhetorical, and argumentative dimensions within a unified frame structure; and flexible, in that it can be applied to various languages, genres, and segment sizes. The annotation process involves iterative refinement, combining qualitative and quantitative validation steps to ensure coherence and reliability.

By applying this framework to a German dataset comprising political debates, media reports, and online discussions, we show that certain types of moralizations and discourse roles become analytically visible only through our multidimensional annotation. In addition, our annotation studies highlight the inherent subjectivity of the task, revealing how differing conceptual understandings of moralization influence annotation consistency. We also probe several LLMs for their ability to detect moralizations automatically, providing both a feasibility study and a detailed error analysis that sheds light on which linguistic and contextual factors are decisive for successful moralization detection.

Our **contributions** are threefold: (1) We propose a novel, frame-based annotation framework for moralizations that captures moral values, demands, and protagonists and allows for a fine-grained analysis of moralizations; (2) We apply and refine this framework across diverse genres of German texts, demonstrating its analytical potential for investigating moral rhetoric and framing, e.g. in political and social discourse; and (3) We conduct and compare several manual annotation and LLM-based detection experiments together with extensive evaluations to explore the challenges of identifying moralizations. All resources developed in this work – including the annotated dataset, annotation manual, and code – are released publicly to support further research in this area.[2] In sum, our goal is to provide a methodological foundation for analyzing how moral rhetoric operates across discourses – a foundation that, we argue, enables new insights into moral communication that previous modeling approaches could not reveal.

The remainder of the paper is structured as follows: §2 provides an overview of prior work on (computational) modeling of morality, then §3 and 4 introduce our annotation framework and dataset. §5 and §6 report our experiments and evaluation for moralization detection, and §7 discusses implications and future directions.

2. Related Work

Morality has been extensively studied both within NLP and in other disciplines; however, the specific phenomenon of moralization and its computational modeling have so far received little attention.

Morality in Computational Social Science (CSS). As pointed out by Reinig et al. (2024), many approaches computationally model morality in order to investigate research questions from the political or social sciences. These studies predict moral attitudes or sentiments from newspaper or social media text, e.g. on discourses about abortion policies (Zhang and Counts, 2016), vaccine campaigns (Islam and Goldwasser, 2022) or climate change (Diakopoulos et al., 2014). For all studies of morality in the field of CSS, Twitter is by far the most prominent empirical basis (see Reinig et al. 2024). In contrast, our dataset captures moralizations across heterogeneous genres and communicative formats, including more subtle contexts such as non-fiction.

Interdisciplinary Perspectives. Outside NLP, moralization has been studied in linguistics (Felder and Müller, 2022; Becker et al., 2023), communication studies (Kampf and Katriel, 2016), psychology (Rhee et al., 2019), and political science (Mooijman et al., 2018). These works emphasize moralization as a persuasive strategy – a phenomenon largely unexplored computationally. Our study addresses this gap by analyzing the potential of computationally modeling moralizations.

Morality and Argumentation. We examine moral values in argumentative contexts. Work at this intersection (e.g., Kobbe et al., 2020; Kiesel et al., 2023) shows that moral values contribute to argumentative quality and persuasion. The goal of the SemEval shared task ValueEval'23 (Kiesel

[1] Similar to Rehbein et al. (2025), our focus is not on aligning LLMs with human values or investigating moral biases in LLMs, but instead to use NLP approaches to analyze moralizations in different texts, thereby contributing to research on value-based reasoning.

[2] https://github.com/GS-Uni-Heidelberg/Paper-TheMoralizationCorpus

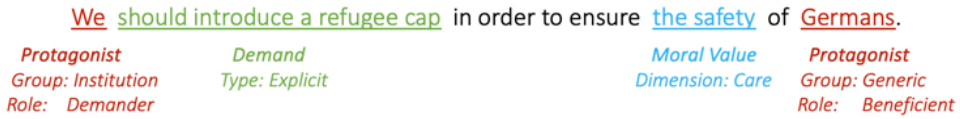

Figure 1: Fully annotated example of a moralization frame, labeled with the demand, the supporting moral value and the protagonists (all translations from German in this paper are by the authors).

et al., 2023) is to explore if it is possible to automatically uncover the values on which arguments draw; and Landowska et al. (2024) label political texts preannotated for argument structures with moral foundations in order to explain the strategies in the use of moral arguments.

Frame-Based Approaches. Few studies model morality in relation to entities or events. Roy et al. (2022, 2021) define morality frames as the moral foundation(s) invoked by a text, along with the sentiment toward mentioned entities. Similarly, Lei et al. (2024) and Zhang et al. (2024) combine moral foundations with entity and event information in order to learn morality-relevant text representations. The approach by Rehbein et al. (2025), most related to ours, provides fine-grained moral frame annotations including moral frame types, moral foundations, and narrative roles in parliamentary debates. We extend this line toward broader genres and pragmatic functions.

Subjectivity in Annotations. Moral labeling is inherently subjective (Falk and Lapesa, 2025; Chochlakis et al., 2025). Recent work proposes multi-annotator and justification-based methods (Weber-Genzel et al., 2024; Nogales and Araque, 2024). We follow this direction with a multi-step annotation pipeline designed to capture and justify subjective variation (see §3).

3. Dataset and Annotation

Our dataset of moralizing text passages was created in four steps: (1) development of a dictionary of morality-indicating words; (2) retrieval of text snippets from large corpora and web sources based on the dictionary entries; (3) creation of an annotation scheme capturing key components of moralizations; and (4) a multi-step annotation process ensuring data quality and consistency.

Dictionary Creation. To identify moralizing passages, we developed DIMI, a multilingual dictionary of morality-indicating words. Starting from a manually curated German seed list of 130 words (e.g., *freedom, fairness, guilt*) (Felder and Müller, 2022), we expanded it using co-occurrence profiles from the CCDB database (Belica, 2011). After manual cleaning, the dictionary comprised 3,000 entries, which we automatically translated into English, French, and Italian and manually verified.[3]

Data Collection. Next, we used DIMI to query large corpora and online sources for text passages containing at least one dictionary entry. For each language (German, English, French, Italian), we retrieved 2,000 five-sentence snippets from seven genres: letters to the editor, interviews (both from various newspapers), parliamentary debates (plenary minutes), commentaries (opinion articles), court reports (newspaper articles about legal cases), Wikipedia discussions (where users discuss how to improve an article), and non-fiction books (on history, parenting, cultural studies, etc.). The German data were drawn from DEREKO (IDS, 2022), while other languages were collected from publicly available web texts. Each dataset was split into training (70%), development (15%), and test (15%) sets, balanced across genres.

Category Development. Our annotation captures three interrelated layers, designed to capture the key pragmalinguistic features of moralizations and to enable consistent corpus-based analysis across different genres and contexts. Figure 1 illustrates a fully annotated example.

(1) **Moral values** (phrase level), mapped to the six Moral Foundations CARE/HARM, FAIRNESS/CHEATING, LOYALTY/BETRAYAL, AUTHORITY/SUBVERSION, PURITY/DEGRADATION, and LIBERTY/OPPRESSION according to the Moral Foundations Theory (MFT) (Graham et al., 2009, 2013).[4] Multi-label assignments are allowed, and multiple values in the same instance are annotated separately; (2) **Demands** (clause or sentence level), here annotators mark all explicit demands (e.g. *We should all stop eating meat*) in the texts. For implicit demands, annotators rephrase the claim in a simple sentence (e.g. *Women still earn less*

[3] Unlike existing moral dictionaries, our resource includes not only explicitly moral terms but also contextually moralizing words such as *guise*. Importantly, the lexicon itself does not distinguish between explicit and contextually moralizing terms; rather, this distinction is determined during annotation based on contextual use.

[4] The MFT assumes that moral reasoning is driven by a set of intuitive emotional responses, or "gut feelings", that underlie and rationalize moral judgment.

(a) *The mayor remains silent on topics such as child poverty, while publicly championing prestige projects. Millions are being wasted on the mistakes of the senators. It is time that politicians, too, are held accountable.* →**Moralization**	(b) *Researchers collected data on child poverty and other broader social issues, and how politicians respond to them in public discourse. The material was analyzed to identify recurring themes across different text types.* →**No Moralization/ NVI**

Table 2: Passages retrieved with DIMI, where (a) constitutes a moralization while (b) neutrally describes a research activity about child poverty.

than men is explicated as *Women should be paid equally*); and (3) **Protagonists** (phrase level), labeled by (a) group type: INDIVIDUALS (e.g. *Angela Merkel*), GENERIC (references to humans, such as *the people, men and women*), INSTITUTIONS/ORGANIZATIONS (e.g. *the democrats, the stakeholders*), and SOCIAL GROUPS (e.g. *parents, homeless people*); and (b) discourse role (role within the moralization): person who is moralizing (DEMANDER), target of the demand (ADRESSEE), person who would benefit (BENEFICIARY) or being disadvantaged (MALEFICIARY) from the demand.

Together, these layers form a **moralization frame**, which we define as the text span that links moral values, demands, and protagonists. These elements are constitutive of moralizations and encompass all central components necessary for the systematic analysis and interpretation of moralizations in discourse.[5]

Annotation Procedure. The annotation of moralizations is not only complex but also inherently subjective. To address this, we designed a multi-step annotation procedure that captures nuanced judgments without compromising the operationalization required for computational modeling. As outlined above, annotations (and analysis) were conducted only for the German dataset; extending the annotation framework to the English, French, and Italian data is currently in progress. The annotation proceeded in six stages:

(1) **Identification**: The mere occurrence of a moral word or phrase does not in itself constitute a moralization (see Table 2). In fact, we observe that in many contexts, moral terms may also appear in neutral or reportive texts without carrying any persuasive or argumentative intention (referred to as *Neutral Value-Referring Instances*, NVIs). Distinguishing between NVIs and moralizations was thus a crucial first step in our annotations. We therefore prepared a check list for moralizations together with positive and negative examples. Then, each retrieved text passage is annotated independently by two annotators (binary classification), both with a background in linguistics, yielding a high agreement of 0.71 (Cohen's Kappa). Passages with disagreement are adjudicated by an expert annotator (one of the authors).

(2) **Pilot phase**: We then conducted an initial exercise using a subset of 200 moralizations to train six annotators (all with a background in linguistics) for the task of moral value component detection and classification (values, demands and protagonists). Besides annotator training, the goal was to optimize the annotation manual by identifying sources of annotation variance, problematic categories, and ambiguous examples. We iteratively refined the manual through additional examples and special-case rules, before proceeding to the annotation of the full dataset.

(3) **Full annotation**: All instances that have been identified as moralizations in the identification step were then distributed across the six trained annotators who labeled moral values, demands, and protagonists based on our codebook using the INCEpTION platform (Klie et al., 2018).

(4) **Review**: Each file was secondarily reviewed by another annotator, allowing for corrections and additions, and generating a further set of discussion items.[6] Open questions or ambiguous cases were discussed and resolved with the team.

(5) **Re-annotation of NVIs**: While the review focused on moralizing instances, some entries labeled as NVI qualify, upon closer inspection, as moralizations. These NVIs were re-reviewed by three expert annotators, limited to the test set due to the time-intensive process.[7] Instances confirmed as moralizations were retroactively annotated with all moralization components, resulting in 278 additional annotated moralization instances.

(6) **Formal validation**: Final consistency checks ensured that the formal annotation rules had been respected (e.g., that each moralization contains a demand or spans were marked correctly). We correct and supplement the data accordingly and remove irrecoverable instances.

The resulting Moralization Corpus provides a rich, pragmatically grounded resource for studying moralizations across genres and serves as a benchmark for computational modeling of moralizing discourses.

[5]We acknowledge that retrieval may yield incomplete frames; this limitation is addressed in our evaluation, and future work will focus on methods for extracting and segmenting complete frames.

[6]This approach has been inspired by similar approaches such as Weber-Genzel et al. (2024) or Becker et al. (2024). Since moralization feature annotation is highly time-intensive, full parallel annotation of the dataset was not feasible.

[7]A refined test set suffices for the prompting experiments in this paper, automated re-annotation of the full dataset is underway for future finetuning experiments.

4. Data Statistics

This section provides an overview of the dataset and its main characteristics. Illustrative examples are provided throughout the text; additional examples are listed in §A.6.

Overview. The final dataset contains 11,503 instances with an average length of 83 tokens, evenly distributed across seven genres (for detailed statistics, see Table 5). The proportion of moralizations varies, mainly due to the particular consideration of the test set within our multi-step annotation process (18% in Dev and Train, and 45% in Test). Across all genres, however, NVIs clearly outnumber moralizations, confirming that moral terms are often used descriptively rather than strategically.

Moral Values. Within moralizations, an average of 1.6 moral values were annotated per instance, most often in Wikipedia discussions (1.7) and least frequently in online comments (1.4). Across the dataset, the CARE–HARM pair dominates (CARE 16%, HARM 22%), followed by FAIRNESS–CHEATING (16%/13%). AUTHORITY (3%) and SUBVERSION (2%) are rare. Genre variation aligns with communicative context (see Fig. 4): FAIRNESS–CHEATING is especially frequent in Wikipedia discussions (28%/19%), reflecting norms of equality and rule compliance in the metadiscourses about the editing of articles (see Example 1 and 2 in §A.6). In court reports, FAIRNESS and LIBERTY as foundational principles of law prevail, while non-fiction books show more OPPRESSION due to historical topics and war narratives.

Demands. Explicit and implicit demands are balanced overall (53% vs. 47%), but differ by genre (see Fig. 5). Explicit demands dominate in parliamentary debates (66%), while implicit ones prevail in commentaries (48%) and letters to the editor (26%), where addressees are diffuse publics rather than interlocutors. In such cases, moralizations serve less to direct action than to express evaluation or positioning, and accordingly, moral appeals often stay implicit, as in the following extract from a letter to the editor (see Example 3 in §A.6 for the full text): *It is outrageous how arrogantly our politicians ignore the needs of our children. From a political perspective, children do not pay off, but for society they certainly do.*

Protagonists. Across all instances, the **roles** BENEFICIARIES (which appear avg. 0.65 times within a moralization) and ADDRESSEES (0.64) occur most frequently, followed by demanders (0.42); MALEFICIARIES are rare (0.10) (see Fig. 6 for the distribution across different genres). Moralizations therefore tend to emphasize positive outcomes rather than blame (see Example 4 and 5 in §A.6). In most cases, not all protagonist slots within the moralization frame are explicitly filled and must be inferred from context or world knowledge, as in the following example from a parliamentary debate where the beneficiary stays implicit: *We need stricter laws to prevent racially motivated violence.* (see Example 6 in §A.6 for the full instance.) INSTITUTIONS (32%) and SOCIAL GROUPS (30%) are the most frequent protagonist **group** types, followed by INDIVIDUALS (20%) and GENERIC HUMAN references (15%) (for the distribution across genres, see Fig. 7). This suggests that moral demands often invoke collective actors, as group-level outcomes appear more socially relevant and persuasive than individual ones (see Example 7 in §A.6), consistent with previous findings on the social function of moralization (Becker, 2025).

Typical **role–group configurations** (see Fig. 8) show that individuals act as demanders, institutions as addressees, and social groups as beneficiaries. Moralizations thus reflect a characteristic pattern linking individual agency, institutional responsibility, and collective good.

These distributional patterns illustrate the analytical potential of the dataset for studying moral rhetoric and framing in discourse. By explicitly linking moral values, demands, and discourse protagonists, the annotation framework makes visible how moral arguments are structured and strategically deployed across communicative contexts. For instance, the prevalence of collective protagonists and beneficiary roles highlights the tendency of moralizations to frame issues in terms of collective welfare, while the balance between explicit and implicit demands shows that moral claims are often conveyed indirectly through evaluative or argumentative framing. These observations demonstrate how the corpus can be used not only to detect moral language but also to analyze the rhetorical mechanisms through which moral values shape argumentative discourse.

5. Experiments

Building on these observations, we then explore how computational methods can support the analysis of moralizations in text. To this end, we evaluate several LLMs using different prompt designs to assess their ability to detect moralizations. The central objective, however, is a systematic comparison of model and human judgments in light of the inherent subjectivity of the task, which is why our evaluation focuses in particular on both human–model and human–human agreement.

Prompt Engineering. Our prompts[8] are derived from our annotation manual and follow its structure[9]. Each defines moralization by three criteria: (1) the presence of moral values, (2) an explicit or implicit demand, and (3) an argumentative link between both. The prompts further guide the extraction and classification of moral values and protagonists. Output is generated in a standardized JSON format including all components and a short explanation: Moral phrases and their classification according to MFT, extracted or reconstructed demands, protagonists together with their assigned roles and group affiliations, the binary decision on whether the passage constitutes a moralization, and a short explanatory rationale. The underlying chain of thought requires the model to proceed step by step: first identifying the core components of moralizations, and then deciding whether the text qualifies as a moralization.

Prompt versions were refined through iterative evaluations. Key adjustments improved recall and precision: (a) clearer definition of positive and negative values, (b) stricter rules for identifying implicit demands, (c) stronger emphasis on the argumentative link, (d) explicit description of NVIs, and (e) integration of borderline cases into the examples.

Prompt Configurations. We experimented with seven configurations varying in level of detail, reasoning requirement, and example inclusion: (1) `basic-0shot`: minimal instruction; (2) `cot-0shot`: stepwise reasoning with detailed instructions but without examples; (3) `cot-10shot`: same, plus ten examples; (4) `cot-explain-0shot`: here the model must explicitly verbalize its reasoning (an explanatory step), no examples; (5) `cot-explain-10shot`: explanation plus examples; (6) `manual-0shot`: a detailed configuration based on our annotation manual; and (7) `manual-explain-0shot`: manual plus explanatory step. Each of the five models was tested with all seven configurations, resulting in 35 outputs per instance.

Models. We tested five state-of-the-art instruction-tuned LLMs differing in architecture, scale, and context window, allowing us to assess prompt performance across models with varying memory and reasoning capacities: LLaMA-4-Scout-17B-16E-Instruct (109B), C4AI-Command-a-03-2025 (111B), Mistral-Small-3.2-24B-Instruct-2506 (24B), GPT-5-mini-2025-08-07, and Claude-3.5-Haiku-20241022. Model parameters are displayed in Table 6.

[8] https://github.com/GS-Uni-Heidelberg/Paper-TheMoralizationCorpus/tree/main/prompts

[9] https://github.com/GS-Uni-Heidelberg/Paper-TheMoralizationCorpus/tree/main/manual

6. Evaluation and Analysis

We evaluated annotation quality and model performance for (a) binary moralization detection and (b) component extraction and classification (values, demands, and protagonists).

6.1. Binary Moralization Classification

Fig. 2 summarizes results on the test set across all prompting conditions for the binary moralization classification task (detailed classification results for each model, including precision and recall scores, are displayed in the Appendix, Table 7 to 12). Performance differences are generally small; detailed prompts yield the most consistent gains above the basic version, underscoring that a clear, structured definition of moralization is crucial. Among models, Cohere attains the best F1 scores, followed by the ensemble model (majority vote of all five models); Claude and Mistral lag behind. Notably, few-shot examples and forced explanations do not consistently improve performance, suggesting that moralization requires deeper pragmatic reasoning than these techniques capture. Error analysis shows a precision-recall trade-off: detailed prompts reduce false positives (↑ precision, ↓ recall), while example-enriched prompts reduce false negatives (↑ recall, ↓ precision). In practical applications, the choice therefore depends on whether recall (e.g. for monitoring or detection systems) or precision (e.g. for analytical research tasks) is prioritized.

6.2. Moralization Component Detection and Classification

Moral values & Protagonists. Automatically evaluating whether a model has identified and classified all relevant moral values and protagonists within a moralization is particularly challenging, since precise boundary detection and overlapping labels, among others, limit automatic agreement with gold annotations.

We evaluate moral value and protagonist spans with the SemEval-2013 NER-style setup (Segura-Bedmar et al., 2013) using strict and partial matching (see §A.4.2 for details). **Moral values** achieve low F1 scores (strict $\leq$ 0.20; partial up to 0.22; see Fig. 9, 10, 11, and 12), indicating difficulties with span boundaries and context-sensitive, often implicit value expressions. Detailed prompts (opposed to basic descriptions) boost performance most, while examples and explanation generation yield small but consistent gains.

Protagonists perform higher (strict F1 of 0.20–0.28; +0.03–0.05 pp under partial, see Fig. 13, 14, 15, 16, 17, and 18 for details). Here, detailed prompts and examples help recall while precision remains limited, reflecting overgeneration

Figure 2: Binary moralization classification across models and prompting conditions (macro F1, test set).

and multi-label ambiguity.

Across models, Cohere and Mistral perform best for values, while GPT leads for protagonists.

Methodologically, results are constrained by our selective annotation scheme (focus on morally relevant values/actors) and the task's subjectivity, so metrics should be read as indicative rather than definitive. Nevertheless, the consistent relative ranking across models and strategies provides a first indication of system behavior and points to directions for targeted fine-tuning and nuanced evaluation.

Demands. Next, we evaluate the models' ability to extract or generate demand formulations. We employ a combination of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), both measuring lexical overlap; and BERTScore (Zhang et al., 2020) which leverages embeddings to estimate semantic similarity.

The results show only minimal variation between models and prompt configurations. BERTScore ranges from 73 (Mistral, `manual-explain`) to 78 (Cohere, `cot-0shot`). As expected, explicit demands yield substantially higher performance (up to 82) compared to implicit ones (maximum 75). Overlap metrics are lower, as expected for free text generation tasks.

Given the limitations of reference-based evaluation in this setting (cf. Becker et al., 2021), we additionally conducted a manual evaluation of extracted (in case of explicit) or generated (in case of implicit) demands. Therefore, two team annotators assessed a subset of 73 moralizations (see §6.3 for the selection process). For each instance, the demands extracted or generated by the three best-performing configurations (see §6.1) of each of the five models were collected. In total, 1,095 demands were rated on a five-point Likert scale for semantic correctness, i.e., how accurately the demand conveyed the intended moral claim. Agreement was substantial (Cohen's Kappa = 0.63), and remaining disagreements were resolved by an expert adjudicator.

Results appear in Fig. 19 and show that models achieved high average ratings of 4.27, indicating that generated demands largely captured the intended moral argumentation. Differences across configurations were minimal, and no notable differences appeared between implicit and explicit demands. Cohere performed best (up to 4.6), while Mistral scored lowest (3.8).

Overall, these results confirm that moralization is a complex linguistic and conceptual phenomenon challenging for automated detection and interpretation. Although detailed prompts and strong models improve performance slightly, overall scores remain moderate, highlighting the need for both model adaptation and refined evaluation methods.

6.3. Agreement between Models and Humans

Next, we take a closer look at these results to explore moralization patterns across genres and to interpret common deviations between human and model-based annotation – with the ultimate goal of informing linguistic and social-scientific analysis of moralizations in discourse.

Annotation Setup. To assess moralization detection challenges for both humans and LLMs, we selected a genre-balanced subset of 150 test instances (Test-150), consisting of 73 moralizations and 77 NVIs. The subset was subsequently annotated by five annotators with varying levels of expertise: Expert 1 (project lead, >2 years), Experts 2 & 3 (doctoral researchers, >1 year), and Student Assistant 1 & 2 (few months of experience). This setup enabled analysis of how project familiarity – and thus understanding of our definitions – affects annotation consistency. All annotators received the same detailed prompts as the models and made binary yes/no moralization judgments (on the instance level).

Findings. Moralization rates increased with project familiarity: Students labeled 23–24% as moralizations, while Experts averaged 38%. This suggests that familiarity broadens detection, as everyday notions are narrower than our operational definition. LLMs labeled 59%, indicating a more liberal classification tendency (see §6.4, §6.5).

Next, human-human, model-model, and human-model agreement were compared using Fleiss' Kappa and PABAK (which adjusts Kappa for prevalence and bias). Results (see Fig. 3, for Fleiss' Kappa which follows the same tendencies cf. Fig. 20) show that experts agree more with each other than students, and student labels align more with LLMs than experts do. Models agree with each other to a degree comparable to human–human agreement, but expert–expert consistency is highest. Interestingly, the weakest models in binary classification (Claude, Mistral) show the highest consistency across prompts. Explanation prompts increase agreement between models, suggesting that they foster more consistent interpretations, even if they do not improve overall predictive performance, as shown in §6.1.

6.4. Analysis of (Dis)Agreement

To understand both the limits of model-based moralization detection and the characteristic patterns in human versus model reasoning, we analyzed cases of agreement and disagreement within Test-150 and collected the following statistics, focusing on the most clearly classified cases (for a tabular overview, see Table 13): How often, and in which cases, do (1) ...all annotators agree (5/5) or at least 80% ($\geq$ 80% 4/5)? (2) ...do all models/configurations agree (35/35) or at least 80 80% ($\geq$ 80% 29/35)? (3) ...do all models and humans agree (40 identical decisions), or at least 80% of each group coincide? (4) ...do models and humans diverge fundamentally (i.e., $\geq$ 80%80% of models vs. $\geq$ 80%80% of humans make opposite decisions)?

Results show that models agree more on moralizations (43%) than on NVIs (24%); humans show the opposite pattern (22% vs. 57%). Instances with total agreement among all humans and models are displayed in Example 8 and 9 in §A.6. In total, only twelve cases can be identified in which human and model decisions fundamentally diverge, see Example 10 in §A.6 for an illustrative instance.

Overall, the results of our (dis-)agreement analysis suggest that models rely strongly on surface cues, whereas humans capture more subtle or implicit cases. To explore this observation further, we examine lexical cues in the data in the next subsection.

6.5. Linguistic Indicators of Moralization

To test the hypothesis that specific text features function as the primary drivers for model (and human) decisions, a selected range of linguistic indicators was examined which, according to Becker (2025), may serve as cues for the classification of moralizations. Examples for each category appear in Table 3.

Categories. (1) Text genre: Moralizations may be easier to identify in certain genres (e.g., opinionated texts); (2) Moral vocabulary: A high frequency of moral words (≥ 5 per instance) may impel both humans and models to label a text as a moralization; (3) Explicit demands: are likely easier to recognize than implicit ones; (4) Modal verbs: As markers of deontic modality, they often co-occur with explicit demands and thus can serve as cues for moralization; (5) Subjunctive mood: Since moralizations often describe future scenarios and subsequent actions, subjunctive forms can work as signals; and (6) Instance length: Very short fragments might lack sufficient context for classification and are more likely labeled as NVIs.

Focusing on cases with $\geq$80% model–human agreement, we found no genre effects, but other indicators showed trends largely confirming our hypotheses: We find high moral-term density in 90% of agreed moralizations vs. only 25% of agreed NVIs; modal verbs in 71% of moralizations vs. 22% in NVIs; subjunctive is also more frequent in moralizations (23% vs. 8%); and explicit demands occur in 81% of unanimously moralizations. Finally, the prevalence of short snippets among agreed NVIs points to the need for dataset refinement, as instance length stems from data extraction artifacts rather than linguistic content. A systematic overview of the results is provided in Table 14.

6.6. Deviation Analysis (Models Only)

Finally, we compared model predictions with the human majority vote for Test-150 to identify typical divergences. The analysis distinguishes between False Positives (FP – model predictions of moralization where the human majority vote is NVI) and False Negatives (FN – model predictions of NVI where the human majority vote is moralization).[10]

Our analysis and annotations reveal three main sources for **FPs**. (i) Neutral uses of moral vocabulary: Many models label passages as moralizing whenever words such as *duty, moral, responsibility, right,* or *wrong* appear, even when used descriptively or within quotations, see Example 11 in §A.6;

[10] For reasons of space, we summarize the key findings below; further details on the annotation procedure and results can be found in the Appendix, §A.5.4.

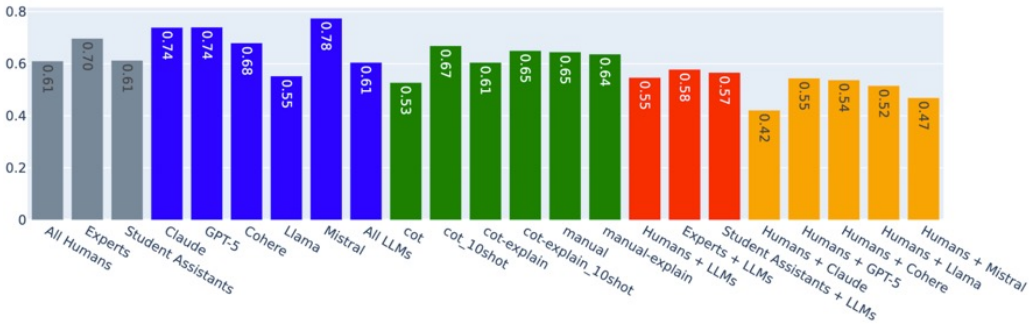

Figure 3: Mean PABAK Scores for different comparisons of agreement between and within humans and models. Fleiss' scores (on avg. 1–2 pp lower) follow precisely the same tendencies, cf. Fig. 20.

(4) *It is not enough to simply stand up and say: We reject* ***all forms of violence*** *with* ***disgust*** *and* ***indignation****. We must not only show that we do not tolerate* ***racism*** *and* ***violence****, but also convey the* ***democratic val- ues*** *by which we want to convince our children. Therefore, I believe that under no circumstances should we restrict* ***civil rights****.* (Parliamentary debates)

(5) *Attac* ***should*** *advocate for a Global Marshall Plan for developing countries. Poverty reduction* ***can*** *only succeed if infrastructure problems are addressed: expanding education systems, enforcing women's rights, and ensuring access to energy and water. In addition, international institutions such as the IMF, WTO, and World Bank* ***must*** *be democratized.* (Letters to the editor)

(6) *Politically, however, resistance from the left* ***would*** *not only jeopardize the currency reform but also the constitutional basis for the Solidarity Foundation. What is needed now is a truly Swiss-style compromise.* (Interviews)

Table 3: Examples for lexical cues of moralizations (in bold): density of moral words (4), modal verbs (5), and subjunctives (6).

(ii) missing context: In some cases, the surrounding context is missing, making it impossible to determine whether moralization is present (Example 12); and (iii) borderline cases: Since moralization detection is a subjective task, certain instances can reasonably be interpreted either way. These ambiguous cases are those where moralization could plausibly be argued even though the majority human label is NVI (Example 13). Better-performing models (LLaMA, Cohere, GPT) show more borderline FPs, suggesting finer sensitivity, while weaker models (Mistral, Claude) more often mislabel neutral passages. Dominant cause for **FNs** are missed demands, especially implicit ones. Less frequent causes are missed moral terms and missed value-demand link across sentences, as well as figurative language, irony, and negated demands.

Summary. Overall, the analyses show that moralization detection is inherently subjective, with models relying on linguistic cues and humans – especially experts – capturing more implicit moralizations. Our findings highlight the challenges of evaluating and modeling nuanced moral reasoning, shaped by subjectivity, context, and linguistic variability.

7. Conclusion

In this paper, we introduced the Moralization Corpus, a novel, frame-based resource for analyzing how moral values are strategically employed in argumentative discourse. The empirical patterns observed in the dataset – including the distribution of moral values, demands, and discourse roles – illustrate how the annotation framework enables systematic analysis of moral framing strategies in discourse. Our framework operationalizes moralization as the interplay between moral values, demands, and discourse participants, allowing for a fine-grained analysis of how moral rhetoric functions across genres. The annotation procedure and resulting data shed light on the pragmatics of moralizing communication, demonstrating that moralizations often rely on implicit reasoning and situated inference rather than (only) overt moral vocabulary, and that many moralizing demands are realized implicitly and require contextual interpretation. Experimental results with several LLMs show that detailed task definitions are essential for reliable moralization detection, whereas few-shot examples and explanation generation do not consistently improve performance. Human-model comparisons further reveal that both groups face similar challenges – particularly regarding implicitness, subjectivity, and the pragmatic boundaries of moral speech acts. Taken together, this work provides an empirical and methodological foundation for future research on moral communication, argumentation, and persuasion. Beyond linguistic and social-scientific applications, our results also inform computational modeling of complex, subjective, and pragmatically grounded language phenomena.

Limitations

While the Moralization Corpus constitutes a unique resource for studying moralizing speech acts across genres, several limitations remain. First, both the automatic and manual evaluation of moral value and protagonist classification can be further improved. The current evaluation setup provides initial insights into model behavior, but more fine-grained semantic and boundary-sensitive measures are needed to better capture the nuanced character of moral references and role assignments. Second, although our experiments included configurations that prompted models to verbalize explanations, we have not yet systematically analyzed the content, coherence, or validity of these explanations. Future work will therefore include a dedicated investigation into how explanation quality correlates with model accuracy and human interpretability.

Third, no model fine-tuning has yet been performed on our dataset. Since the annotation scheme introduces new task-specific concepts such as moral frames and pragmatic roles, fine-tuned models might substantially improve the detection and classification of moralizations. A fourth limitation concerns the contextual scope of our instances: the current dataset is based on five-sentence snippets, which, while sufficient for local pragmatic analysis, may not always capture the full discursive context in which moralizations unfold. Expanding the contextual window or including paragraph-level annotations will thus be an important step toward a more comprehensive understanding of moral reasoning in discourse.

Furthermore, although our dataset is multilingual in structure, detailed annotations have so far only been carried out for the German data, therefore language-specific lexical, grammatical, and pragmatic conventions may influence how moralization is expressed, potentially limiting the direct transferability of the results to other languages. Future work will extend the annotation framework to English, French, and Italian, enabling cross-linguistic comparisons and broader generalization. Finally, due to the high complexity and time intensity of the task, parallel double annotation was conducted only for selected subsets rather than for the entire corpus. While our multi-step adjudication ensured consistency and reliability, a fully parallel annotation process would further strengthen inter-annotator agreement and improve the overall robustness of the dataset.

Ethics Statement

The Moralization Corpus was constructed using publicly available texts and copyright-compliant material (e.g., parliamentary debates, news articles, online discussions) and does not include private or sensitive data. Given the inherently normative character of moral discourse, annotators were trained to focus on linguistic and pragmatic aspects rather than moral evaluation or agreement with the content. We acknowledge that subjectivity is an integral part of moral interpretation; our multi-step annotation protocol and adjudication procedures were designed to minimize bias while preserving interpretive diversity. We further emphasize that the goal of this research is analytical and descriptive, not prescriptive: the dataset and models are not intended for moral judgment or behavioral prediction, but to support interdisciplinary research on the intersection of communication, argumentation, and moral framing.

Acknowledgements

We would like to thank all student assistants involved in the annotation of the Moralization Corpus for their dedication and careful work. We are also grateful to our colleagues for their insightful discussions on annotation methodology and moral discourse. This research was supported by institutional funding and by the interdisciplinary research project *Moralisierungen in der Wissenschaftskommunikation (MoWiKo)*, funded by the Federal Ministry of Research, Technology and Space (BMFTR), Germany. In addition, the authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).

8. Bibliographical References

Maria Becker. 2025. Die Rolle der Diskursgrammatik bei der Detektion und Analyse sprachlicher Praktiken. In M. Müller, M. Reisigl, M. Becker, M. Bender, and E. Felder, editors, *Diskursgrammatik*. De Gruyter, Berlin/Boston.

Maria Becker, Swetha Ananth, and Carina Kiemes. 2023. The moral dimensions of health: pilot study.

Maria Becker, Kanyao Han, Haejin Lee, Antonina Werthmann, Rezvaneh Rezapour, Jana Diesner,

and Andreas Witt. 2024. Detecting impact relevant sections in scientific research. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4744–4749, Torino, Italia. ELRA and ICCL.

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online. Association for Computational Linguistics.

Cyril Belica. 2011. *Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen*. Bozen University Press.

Georgios Chochlakis, Niyantha Maruthu Pandiyan, Kristina Lerman, and Shrikanth Narayanan. 2025. Larger language models don't care how you think: Why chain-of-thought prompting fails in subjective tasks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Nicholas A. Diakopoulos, Amy X. Zhang, Dag Elgesem, and Andrew Salway. 2014. Identifying and analyzing moral evaluation frames in climate change blog discourse. *Proceedings of the International AAAI Conference on Web and Social Media*.

Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921.

Ekkehard Felder and Marcus Müller. 2022. Diskurs korpuspragmatisch: Annotation, Kollaboration, Deutung am Beispiel von Praktiken des Moralisierens. *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge*, pages 241–262.

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

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Jonathan Haidt, Jesse Graham, and Conrad Joseph. 2009. Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, 20(2-3):110–119.

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

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

IDS. 2022. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2022-I (release vom 08.03.2022). PID: 00-04B6-B898-AD1A-8101-4.

Tunazzina Islam and Dan Goldwasser. 2022. Understanding covid-19 vaccine campaign on facebook using minimal supervision. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 585–595.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny T Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574.

Zohar Kampf and Tamar Katriel. 2016. Political condemnations. *The Handbook of Communication in Cross-cultural Perspective*, page 312.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych.

2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

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

Alina Landowska, Katarzyna Budzynska, and He Zhang. 2024. Quantitative and qualitative analysis of moral foundations in argumentation - argumentation.

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024. EMONA: Event-level moral opinions in news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5239–5251, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Sonya Sachdeva Morteza Dehghani, Kenji Sagae and Jonathan Gratch. 2014. Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the “ground zero mosque”. *Journal of Information Technology & Politics*, 11(1):1–14.

Anny D. Alvarez Nogales and Oscar Araque. 2024. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING*, pages 67–77.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ines Rehbein, Lilly Brauner, Florian Ertz, Ines Reinig, and Simone Ponzetto. 2025. Moral reckoning: How reliable are dictionary-based methods for examining morality in text? In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 232–250, Albuquerque, USA. Association for Computational Linguistics.

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

Joshua J Rhee, Chelsea Schein, and Brock Bastian. 2019. The what, how, and why of moralization: A review of current definitions, methods, and evidence in moralization research. *Social and Personality Psychology Compass*, 13(12):e12511.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards few-shot identification of morality frames using in-context learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196, Abu Dhabi, UAE. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269.

Amy X. Zhang and Scott Counts. 2016. Gender and ideology in the spread of anti-abortion policy. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 3378–3389, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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

A. Appendix

A.1. Dataset Statistics and Visualizations

Table 5 provides additional statistical details on our dataset presented in §4.

Visualizations and detailed information on the analysis of the data statistics (§4) are displayed in Fig. 4, 5, 6, 7, and 8.

A.2. Prompts

Our prompts are divided into four sections. Part A defines the concept of moralization and specifies the three necessary criteria for its identification: (1) the presence of one or more moral values (based on the Moral Foundations Theory), (2) the presence of an explicit or implicit demand to act, refrain from acting, or adopt a stance, and (3) an argumentative link between the moral value and the demand. Part B outlines the extraction of protagonists, Part C their categorization into predefined social or institutional classes, and Part D the assignment of moralization roles (e.g., Demander, Addressee, Beneficiary). Examples illustrate each step, and the prompt concludes with a standardized JSON output format for computational processing. All prompt configurations can be found here: https://github.com/GS-Uni-Heidelberg/Paper-TheMoralizationCorpus/tree/main/prompts

A.3. Models

Table 6 summarizes the statistics of the models used in our experiments (§5).

A.4. Results and Evaluation

A.4.1. Binary Moralization Classification

Tables 7, 8, 9, 10, 11, and 12 provide detailed results of our binary moralization classification experiments (see §6.1). The tables report model-specific performance metrics, including precision and recall, allowing for a comprehensive comparison across different model architectures.

A.4.2. Metrics for Moral Value and Protagonist Evaluation

For the automatic evaluation of moral value detection and classification as well as protagonist detection and classification (see §6.2), we adopted the evaluation framework for Named Entity Recognition (NER) models as defined in the SemEval 2013 - 9.1 task (Segura-Bedmar et al., 2013). Following this setup, we used the Python library nervaluate[11] to compute token-level and span-level agreement between system output and gold-standard annotations. The evaluation distinguishes between five outcome categories: COR (correct, exact match), INC (incorrect, mismatch between system and gold annotation), PAR (partial overlap), MIS (missed gold annotation), and SPU (spurious system output).

Based on these, possible items (POS) are defined as COR + INC + PAR + MIS (true positives and false negatives), and actual items (ACT) as COR + INC + PAR + SPU (true positives and false positives). We report both strict and partial match scores. Under strict matching, precision and recall are calculated as

$$\text{Precision} = COR/ACT = TP/(TP + FP)$$
$$\text{Recall} = COR/POS = TP/(TP + FN)$$

Under partial matching, overlapping annotations are weighted by 0.5 to account for near matches:

$$\text{Precision} = (COR + 0.5PAR)/ACT$$
$$\text{Recall} = (COR + 0.5PAR)/POS$$

Since our annotation scheme allows multiple MFT labels per moral value, we counted COR, INC, PAR, SPU occurrences for six virtue/vice pairs separately, then aggregated the counts across MFT class labels to compute precision and recall.

[11] https://github.com/MantisAI/nervaluate

Split	Instances	Percentage	Proportion Moralizations	AVG tokens
test	1,584	13.8	44.7	82.7
Test-150	150	1.3	48.5	77.6
dev	1,953	17	17.7	84.0
train	7,816	67.9	17.8	83.2
Court reports	1,337	11.6	12.1	91.2
Letters to the editor	1,553	13.5	27.8	84.7
Parliamentary debates	1,583	13.8	36.1	74.4
Wikipedia discussions	1,659	14.4	7.3	79.0
Commentaries	1,807	15.7	33.6	97.3
Interviews	1,855	16.1	32.5	62.7
Non-fiction books	1,709	14.9	5.5	95.1
Total/Medium	**11,503**	**100**	**21.2**	**83.2**

Table 5: Data split across train, development, and test sets and distribution of genres within the complete dataset, including the number of instances, proportion of the entire dataset, ratio of moralizations, and the average number of tokens per instance. Differences in the proportion of moralizations across the splits, particularly between the test set and the training and development sets, result from the refinement of annotations, as described in §4.

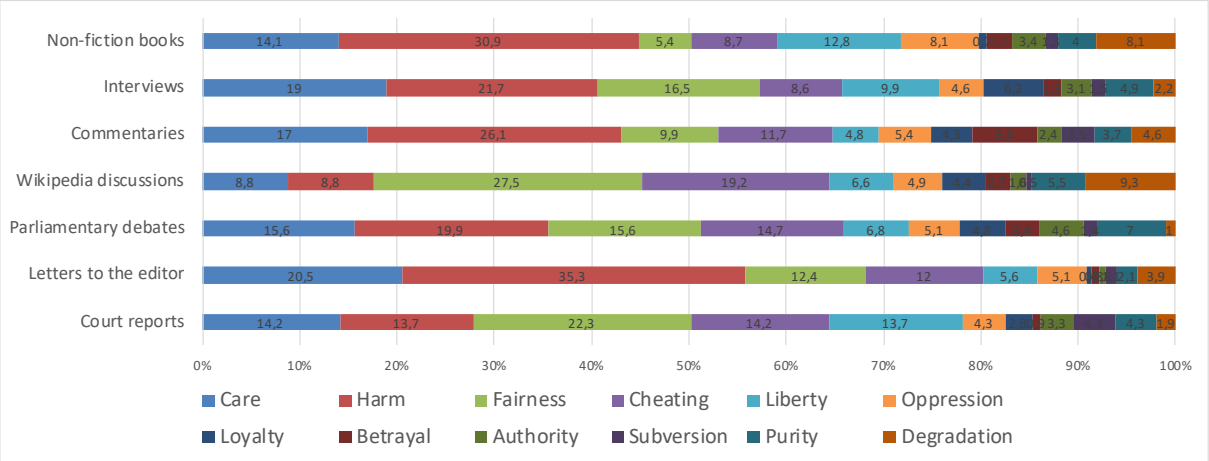

Figure 4: Distribution of moral values (according to MFT) across genres, numbers in percentage.

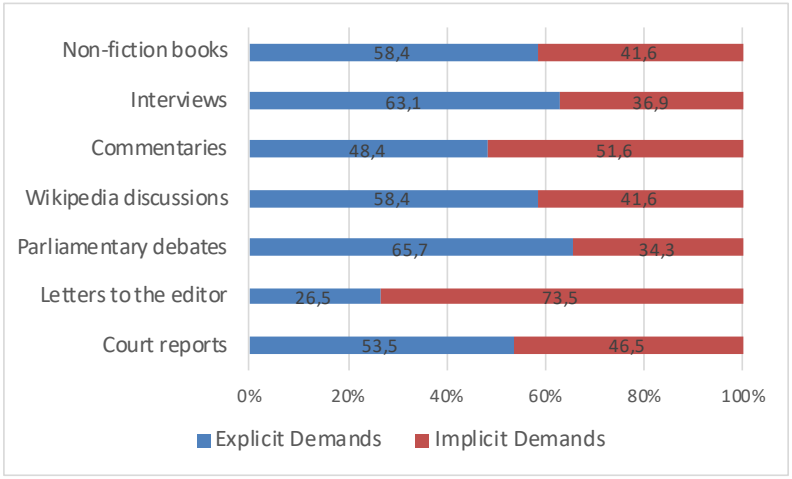

Figure 5: Distribution of explicit vs. implicit demands across genres, numbers in percentage.

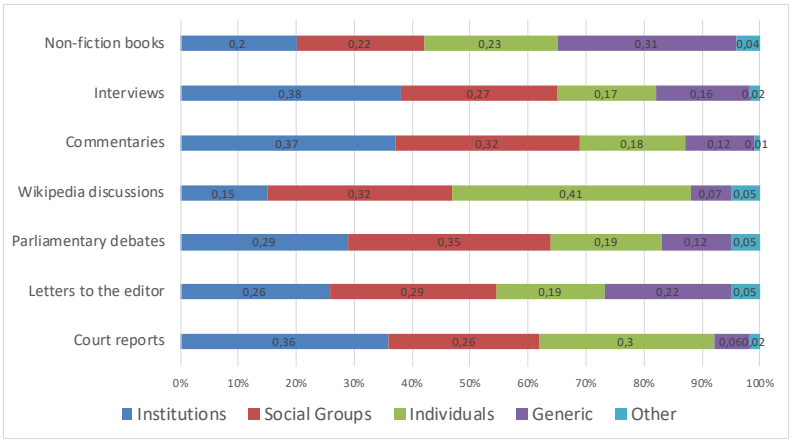

Figure 6: Distribution of groups across genres, numbers in percentage.

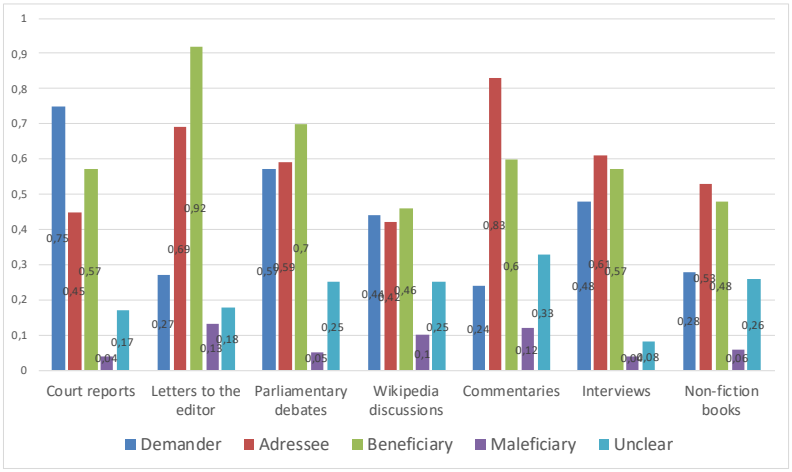

Figure 7: Distribution of moralizing roles across genres, numbers in total.

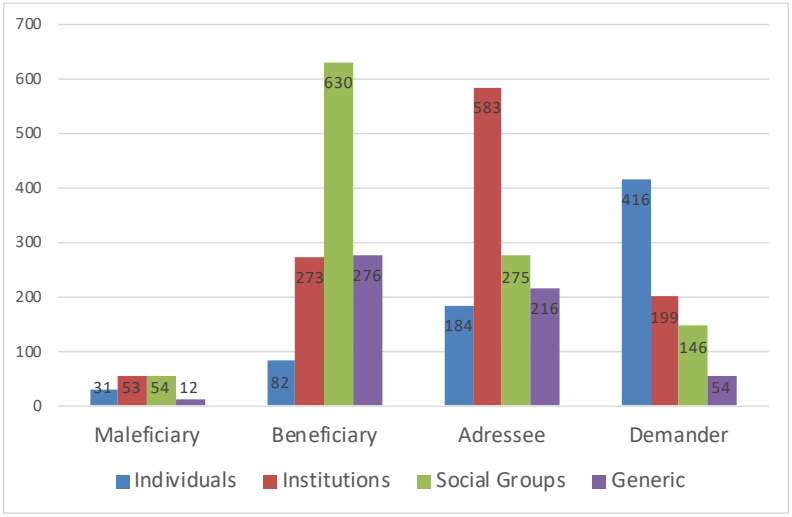

Figure 8: Co-occurrences of protagonist groups and roles, numbers in total.

	llama	cohere	mistral	gpt-5	claude
model size	109B	111B	24B	not public	not public
context size	10M	256K	128K	400k	200k

Table 6: Statistics of the models used in our experiments.

Claude	acc	pre	rec	f1
basic_0shot	0.3889	0.6635	0.5587	0.3568
cot_0shot	0.6788	0.7193	0.7500	0.6752
cot_10shot	0.6701	0.7200	0.7474	0.6675
cot-explain_0shot	0.7089	0.7313	0.7690	0.7029
cot-explain_10shot	0.7031	0.7300	0.7664	0.6978
manual_0shot	0.5527	0.6891	0.6711	0.5517
manual-explain_0shot	0.5839	0.6965	0.6920	0.5839

Table 7: Binary classification results for Claude.

GPT-5	acc	pre	rec	f1
basic_0shot	0.6094	0.7047	0.7099	0.6093
cot_0shot	0.7639	0.7324	0.7598	0.7398
cot_10shot	0.7645	0.7329	0.7602	0.7403
cot-explain_0shot	0.7506	0.7228	0.7533	0.7287
cot-explain_10shot	0.7679	0.7370	0.7653	0.7445
manual_0shot	0.7477	0.7425	0.7849	0.7364
manual-explain_0shot	0.7448	0.7428	0.7854	0.7343

Table 8: Binary classification results for GPT.

Cohere	acc	pre	rec	f1
basic_0shot	0.5486	0.6844	0.6666	0.5476
cot_0shot	0.7963	0.7634	0.7873	0.7718
cot_10shot	0.7604	0.7376	0.7736	0.7423
cot-explain_0shot	0.7917	0.7584	0.7819	0.7666
cot-explain_10shot	0.7812	0.7515	0.7823	0.7597
manual_0shot	0.7789	0.7535	0.7895	0.7605
manual-explain_0shot	0.7980	0.7674	0.7970	0.7764

Table 9: Binary classification results for Cohere.

Llama	acc	pre	rec	f1
basic_0shot	0.6262	0.6964	0.7125	0.6251
cot_0shot	0.7789	0.7738	0.6738	0.6915
cot_10shot	0.7656	0.7394	0.7731	0.7458
cot-explain_0shot	0.7980	0.7673	0.7397	0.7504
cot-explain_10shot	0.7656	0.7366	0.7673	0.7437
manual_0shot	0.7743	0.7418	0.7678	0.7497
manual-explain_0shot	0.7506	0.7321	0.7696	0.7342

Table 10: Binary classification results for Llama.

A.4.3. Moral Value Evaluation

Fig. 9, 10, 11, and 12 show the detailed results for automatic moral value evaluation with the SemEval-2013 NER-style setup (Segura-Bedmar et al., 2013) using strict and partial matching (see §6.2).

A.4.4. Protagonist Evaluation

Fig. 13, 14, 15, 16, 17, and 18 show the detailed results for automatic protagonist evaluation, again with the SemEval-2013 NER-style setup using strict and partial matching (see §6.2).

A.4.5. Demand Evaluation

Fig. 19 shows the results of our manual evaluation of generated and extracted moral demands on Test-150 on a five-point Likert scale for semantic correctness (see §6.2).

Mistral	acc	pre	rec	f1
basic_0shot	0.5208	0.6752	0.6465	0.5181
cot_0shot	0.6921	0.7097	0.7448	0.6849
cot_10shot	0.6157	0.6936	0.7060	0.6151
cot-explain_0shot	0.6441	0.7020	0.7239	0.6422
cot-explain_10shot	0.5666	0.6905	0.6795	0.5662
manual_0shot	0.6620	0.6992	0.7279	0.6578
manual-explain_0shot	0.6586	0.7058	0.7322	0.6556

Table 11: Binary classification results for Mistral.

ensemble	acc	pre	rec	f1
basic_0shot	0.5399	0.6911	0.6645	0.5380
cot_0shot	0.7824	0.7538	0.7863	0.7618
cot_10shot	0.7413	0.7398	0.7819	0.7308
cot-explain_0shot	0.7708	0.7485	0.7863	0.7537
cot-explain_10shot	0.7650	0.7559	0.8000	0.7532
manual_0shot	0.7280	0.7289	0.7691	0.7177
manual-explain_0shot	0.7297	0.7365	0.7777	0.7211

Table 12: Binary classification results for the Ensemble model (majority votes from all models).

A.5. Data Analysis

A.5.1. Agreement Scores for Test-150

In §6.4 we compare human-human, model-model, and human-model agreement using PABAK (as displayed in the main paper) and Fleiss' Kappa as displayed in Fig. 20.

A.5.2. Agreement between Models and Humans

Table 13 displays the results of our agreement evaluation in §6.3, focusing on the most clearly classified cases.

A.5.3. Linguistic Indicators of Moralization

To examine the hypothesis that specific textual features serve as key determinants of both model and human decisions, in §6.5 we conducted an analysis focusing on a selected set of linguistically motivated indicators. Results (in percentage) are displayed in Table 14.

A.5.4. Details on the Deviation Analysis

In §6.6 we investigated possible sources of divergence between model and human decisions. In contrast to the previous analyses – which compared patterns of human and machine agreement

Agreement	Models	Humans	M-H
Moralizations $\geq 80\%$	41.3	22.0	20.7
Moralizations 100%	18.7	12.7	8.7
NVR $\geq 80\%$	24.0	57.3	24.0
NVR 100%	8	48	8

Table 13: Agreement and Disagreement among and between Models and Humans, in percentage.

Figure 9: Moral Value evaluation with MFT category labels in the strict match criteria. Note that the multi-label annotation is allowed for a single moral value span.

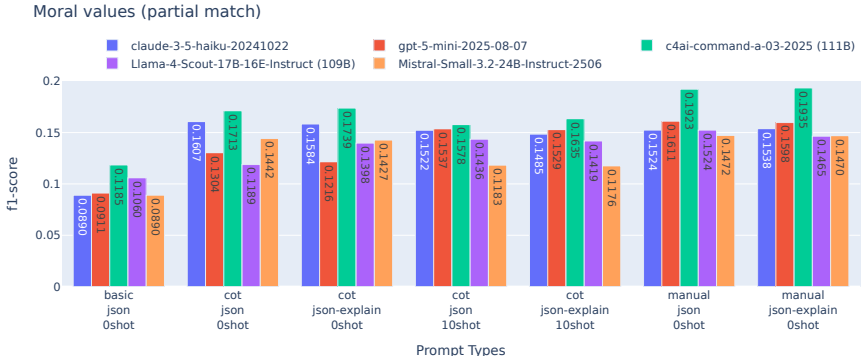

Figure 10: Moral Value evaluation with MFT category labels in the partial match criteria. Again, multi-label annotation is allowed for a single moral value span.

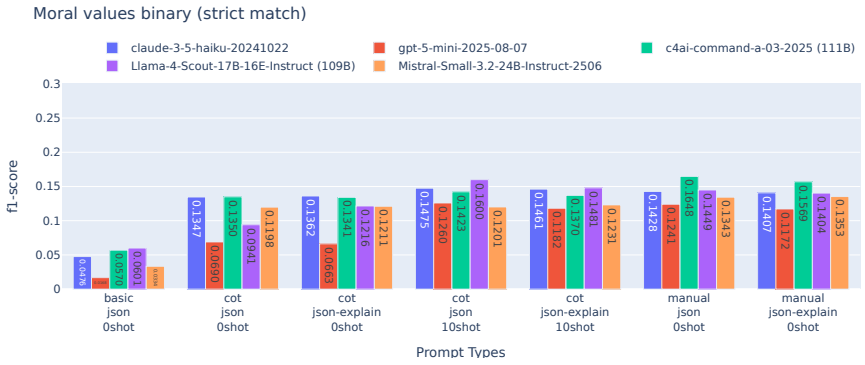

Figure 11: Moral Value evaluation with virtue/vice labels only in the strict match criteria. MFT category difference between gold and prediction is not taken into account. Note that we resolved the multi-label annotation onto multi-class (two-classes virtue or vice) per moral value.

	Moral Phrases	**Explicit Demand**	**Modal Verb**	**Subjunctive**	**Context**
Moralizations (n=31)	28	25	22	7	0
Percentage	90.3	8.6	71	22.6	0
NVR (n=36)	9	0	8	3	11
Percent	25	0	22.2	8.3	30.6

Table 14: Textual features of moralizations and NVIs, numbers in percentage.

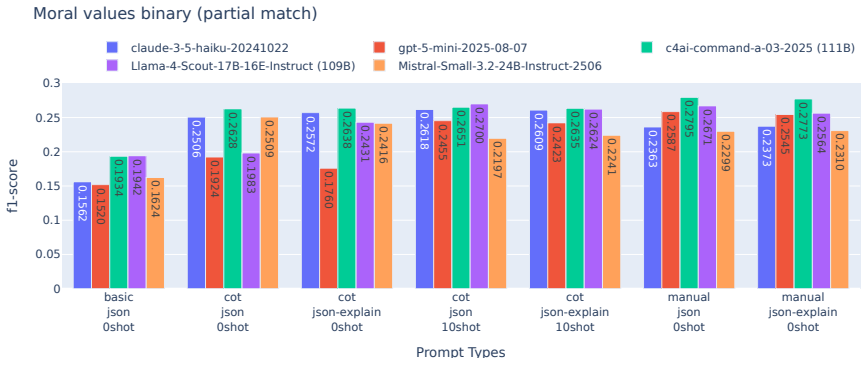

Figure 12: Moral Value evaluation with virtue/vice labels only in the partial match criteria.

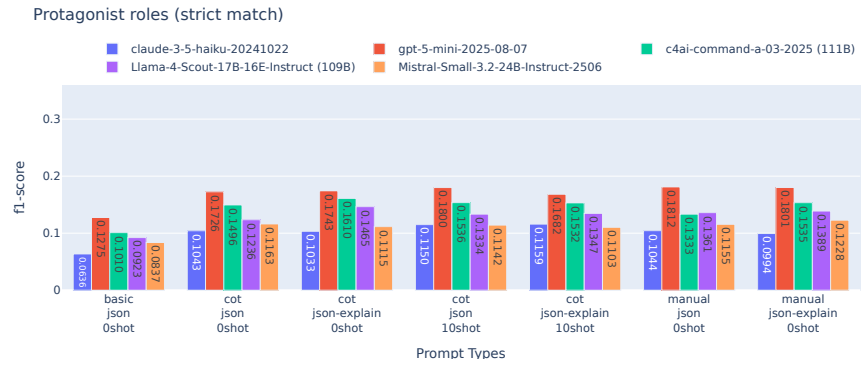

Figure 13: Protagonist evaluation with role labels in the strict match criteria. Multi-label annotation is allowed for a single protagonist span.

– this analysis required a reference point. Therefore, model predictions were compared with the majority vote of the human annotators for Test-150 to identify typical deviations. It is important to emphasize that, in the context of inherently subjective annotation tasks such as moralization classification, deviations from the reference are not regarded as errors but as variance (Falk and Lapesa, 2025; Chochlakis et al., 2025). For ease of general comprehensibility, however, we refer to these cases as false positives (FPs) – instances labeled by humans as NVIs but predicted as moralizations by a model, and false negatives (FNs) – instances labeled by humans as moralizations but predicted as NVIs.

Evaluation Setup. In a bottom-up process, potential causes for divergences between model and human decisions were first explored on the entire test set (excluding Test-150) and then grouped into categories for FPs and FNs separately. Based on these observations, an annotation manual was developed, including detailed category definitions, examples, and annotation guidelines. Two research assistants with a background in linguistics conducted parallel manual annotations of all false positives and false negatives (across all five models and seven configurations) within Test-150. Annotations were performed blind, meaning the annotators were not informed about which model or configuration produced which prediction.[12] We measured inter-annotator agreement (IAA) using Cohen's Kappa and achieved an agreement of 0.72% for the false-positive categories and 0.69% for the false-negative categories. Remaining disagreements were resolved by an expert annotator (one of the authors).

Results for False Positives. For model deviations classified as false positives – that is, instances labeled by humans as NVIs but predicted by a model as moralizations – three main categories emerged:

1. **Use of moral vocabulary in neutral contexts**: Many models label passages as moralizing whenever words such as duty, moral, responsibility, right, or wrong appear, even when used descriptively or within quotations. In our dataset, this often occurs in historical reporting, where frequent negative moral terms such as *horrible epochs, time of plague, Thirty Years' War* appear, without constituting moralization as a discursive strategy in our sense.

[12] Annotators were, in fact, not told that the data originated from model predictions at all.

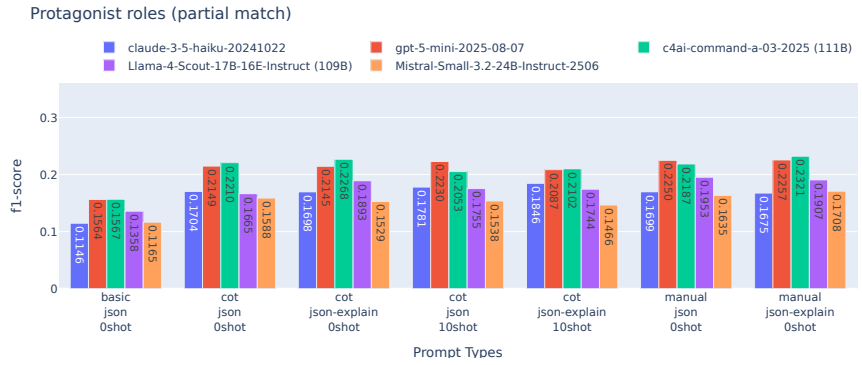

Figure 14: Protagonist evaluation with role labels in the partial match criteria. Especially in the partial match, we relaxed the criteria further for noun phrases, so that the start position difference (whether a preceding determinat is present in the span or not) does not affect much.

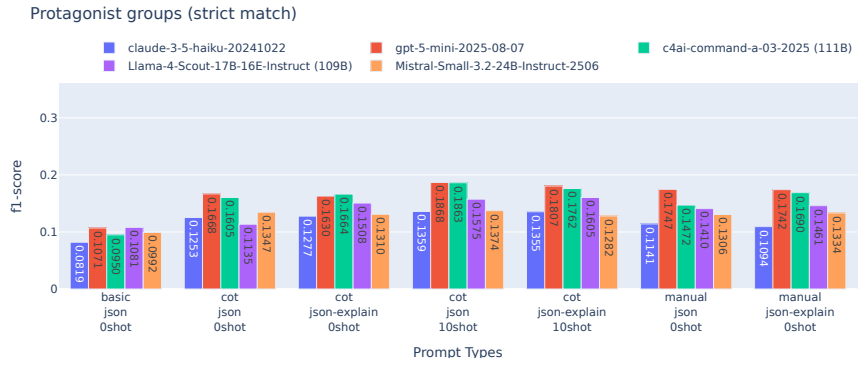

Figure 15: Protagonist evaluation with group labels in the strict match criteria. Note that the group labels are mutually exclusive, which means the original multi-class evaluation is applied without adjustment for multi-label setup.

Neutral uses of moral vocabulary also appear in passages that weigh different moral perspectives without articulating a demand or persuasive stance, which by our definition likewise do not qualify as moralizing.

2. **Lack of context**: In some cases, the surrounding context is missing, making it impossible to determine whether moralization is present.

3. **Borderline cases**: Since moralization detection is a subjective task, certain instances can reasonably be interpreted either way. These ambiguous cases are those where moralization could plausibly be argued even though the majority human label is NVI.

Across all five models and seven configurations, Test-150 contains 729 false positives.[13] The distribution of categories (neutral / context / borderline) is visualized in Fig. 22.

[13]Accordingly, these findings are based on a relatively small dataset, and the scalability of the results should be validated in future work.

Findings. Models such as LLaMA, Cohere, and GPT – which also achieved the best performance in the binary evaluation (see §3) – exhibit the highest proportion of borderline cases, suggesting that these systems may be more nuanced than raw metrics indicate. By contrast, Mistral and Claude, which performed weakest overall, struggle most with identifying neutral texts as non-moralizing, consistent with their lower precision scores. Missing context affects all models to a similar extent (except for LLaMA, which shows slightly greater robustness). This context problem, already discussed above as a potential signal for NVIs, here reappears as a factor influencing moralization decisions as well. Missing context thus represents a general limitation for moralization analysis and should be addressed in future work through targeted dataset expansion or context reconstruction. When mapping these categories (neutral, context, borderline) back to the 12 cases where models and humans strongly diverge ($\geq$ 80% of models vs. $\geq$ 80% of humans giving opposing labels), 10 out of 12 are classified as borderline. This indicates that the cases in which humans and models diverge most

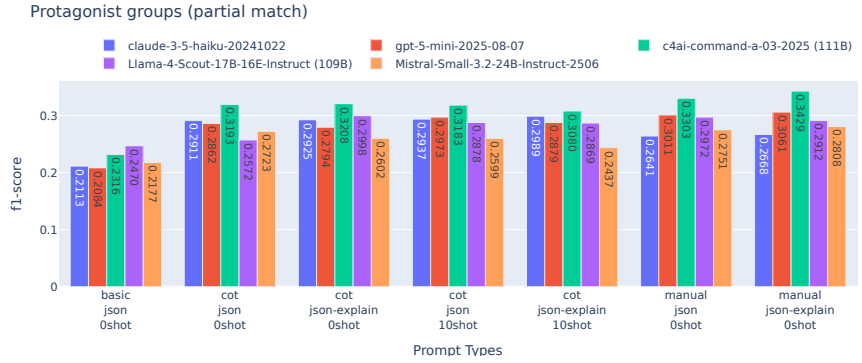

Figure 16: Protagonist evaluation with group labels in the partial match criteria. We applied the relaxed rule for noun phrases described above.

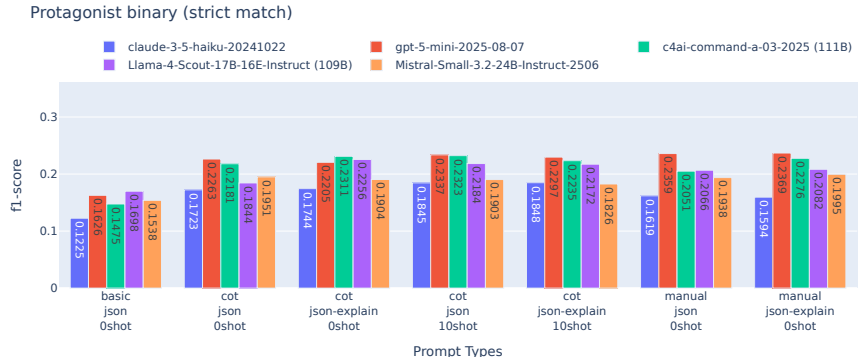

Figure 17: Protagonist evaluation without labels in the strict match criteria. We focus on the span start and end positions only, no label differences (neither role labels nor group labels) are taken into account.

are precisely those that are highly open to interpretation – a plausible and consistent finding that warrants further validation with additional data.

Results for False Negatives. The second part of the analysis concerns false negatives, i.e., instances identified by human annotators as moralizations but not recognized as such by the models. The annotation parameters (deviation categories) were derived from the main defining features of moralizations (see §1). The underlying hypothesis is that the models were prompted to first identify the key components of a moralization (moral values, demands, and their interrelation) and then make a classification decision based on this. We therefore assume that divergences from human judgments occur when at least one of these components is not recognized by the model. The categories used are (1) unrecognized moral word(s) or phrase(s); (2) unrecognized demand; (3) both unrecognized moral values and unrecognized demand; and (4) unrecognized connection between moral values and demand. The last category applies when a model correctly identifies both a moral value and a demand but still labels the instance as a NVI.[14]

Findings. The analysis of false negatives (n = 85)[15] shows that the detection of demands poses a particular challenge even for the best-performing models (Cohere, LLaMA, GPT). A close inspection of the affected instances and their gold annotations reveals that 43 of 51 unrecognized demands are implicit, supporting our hypothesis that implicit moral demands are generally more difficult for language models to detect than explicit ones. Many of these cases involve descriptions of a negative (and thus implicitly undesirable) state of affairs without an explicit call for change.

Failure to detect moral vocabulary was a major source of divergence only for LLaMA and Cohere.[16] Manual inspection revealed that this typically occurs when identifying moral terms requires world knowledge or sociocultural context.

[14] Although in rare cases moral values and demands may co-occur without an argumentative connection, we assume – in line with the human majority label and for analytical simplicity – that this is not the case in the present data.

[15] Given the limited data and low number of false negatives for some models/configurations, these findings should be interpreted with caution.

[16] Cohere failed to detect both moral vocabulary and demands in 6 of its 15 false negatives.

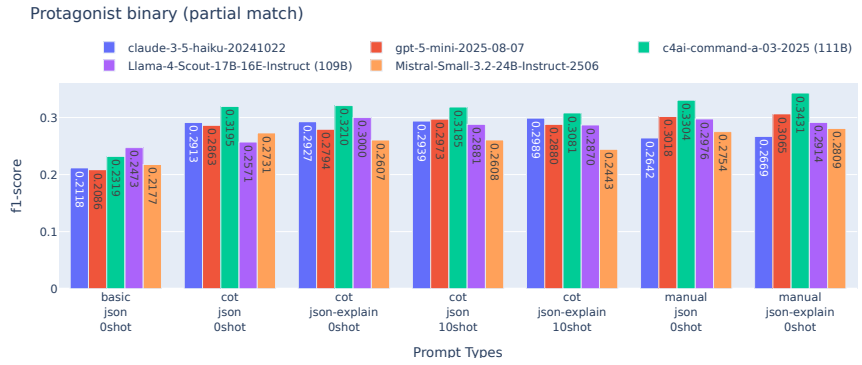

Figure 18: Protagonist evaluation without labels in the partial match criteria. Again, no label differences (neither role labels nor group labels) are taken into account.

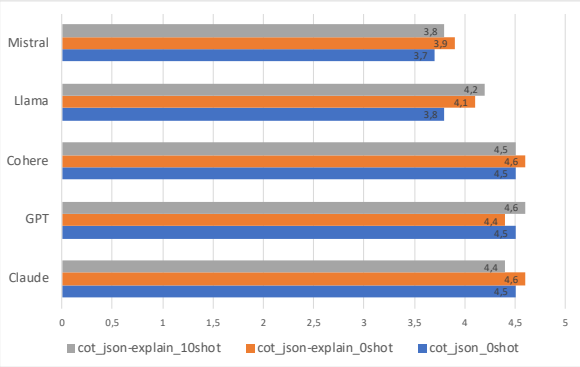

Figure 19: Manual evaluation of demands on a five-point Likert scale for semantic correctness, i.e., how accurately the demand conveyed the intended moral claim.

Failure to detect the connection between moral value and demand emerged as a key source of error particularly for LLaMA, Mistral, and Claude.[17] Apart from the 14 LLaMA cases falling into this category, only eight additional instances across models met this criterion – a finding that supports the assumption that co-occurrence of moral vocabulary and demands within a short span usually indicates moralization. Example 20 illustrates a case in which models (Claude, Mistral, LLaMA) correctly detect moral components but fail to recognize that they serve as argumentative support for a demand.

Additional, more subtle sources of error emerged from manual inspection: Figurative language or irony can obscure moral content, leading models to miss moralizations; complex argument structures, particularly when the link between moral value and demand spans multiple sentences, pose difficulties; and negated demands, where a text argues that a certain moral action is not required (e.g., "no need to make amends"), are often overlooked, even though such negations are constitutive of moralization by our definition.

[17] Given the small number of affected instances, caution is warranted when interpreting these results, especially for Claude and Mistral.

A.6. Examples

For reasons of space, only a limited number of examples could be included in the paper itself. However, throughout the chapters, references were made to additional examples, which are listed below, organized according to the respective chapters. All translations by the authors of this paper.

A.6.1. Illustrating Examples for §4

Example 1 (Wikipedia Discussions) – Moral demands for neutral description in Wikipedia articles: *In my view, a neutral description of reality always means describing what was formative for both the past and the present in a balanced way.*

Example 2 (Wikipedia Discussions) – Moral demands for neutral description in Wikipedia articles: *I don't think it's the task of an encyclopedia to express such judgments. If you criticize that something isn't written neutrally, then point out the passages and explain what exactly isn't neutral. Just because you feel the article sounds too positive doesn't mean it is. . . maybe your own attitude toward the topic isn't neutral enough.*

Example 3 (Letters to the editors) – Moralization where the actual addressee is a diffuse public audience: *Beckmann's angry speech is justified. It is outrageous how arrogantly our politicians ignore the needs of our children. From a political perspective, children do not pay off—but for society they certainly do. Yet recognizing that requires a certain degree of foresight.*

Example 4 (Parliamentary debates) – Moralization that serves to legitimize and morally reinforce desirable actions: *Anyone who wants to fight right-wing extremism must fundamentally revise and abolish xenophobic provisions and laws. When people from Vietnam who have lived here for more than ten years, or people from crisis regions like Sri Lanka, are deported, this must be addressed too—just like the ghettoization of*

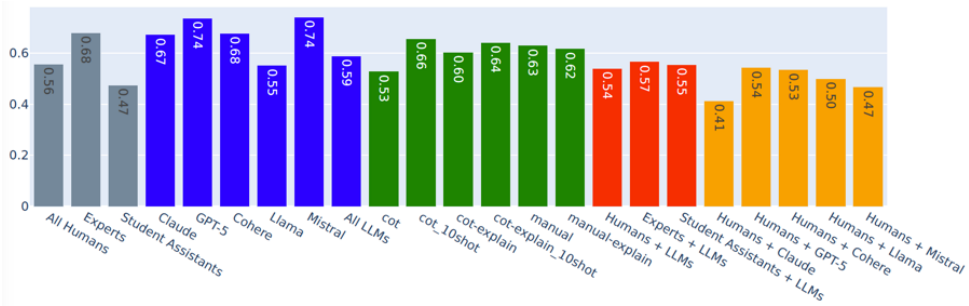

Figure 20: Fleiss' Kappa scores for different comparisons of agreement between and within human annotators and models.

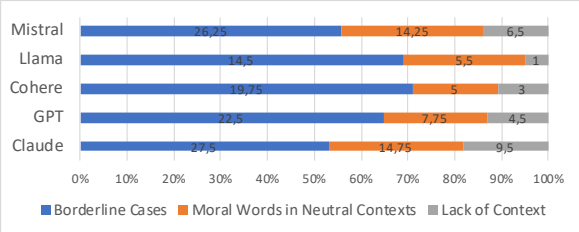

Figure 21: Error causes for false positives, mean values (in total) per model across all prompt configurations.

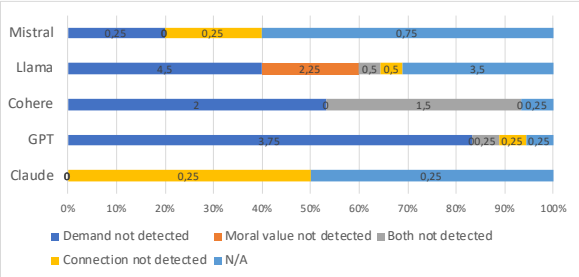

Figure 22: Error causes for false negatives, mean values (in total) per model across all prompt configurations.

refugees. This, in my view, plays directly into the hands of the arsonists, because inhibitions are being lowered.

Example 5 (Parliamentary debates) – Moralization that emphasize positive social outcomes of moral action: *That is why it is not enough to simply stand up and say: We reject all forms of violence with disgust and indignation. We must not only show that we do not tolerate racism and violence, but also convey the democratic values by which we want to convince our children and the youth in both East and West. Therefore, I believe that under no circumstances should we restrict civil rights, such as the right to assembly. That would amount to a capitulation of the rule of law to right-wing extremists, ladies and gentlemen.*

Example 6 (Parliamentary debates) – Moralization where the beneficiary stays implicit: *We need stricter laws to prevent racially motivated violence.*

Example 7 (Letters to the editors) – Moralization where generic expressions are used to evoke shared societal concerns: *I would like to arrive at a definition that grants all people equal humanitarian protection—regardless of whether they are persecuted by the state or by non-state actors.*

A.6.2. Illustrating Examples for §6.3

Example 8 (Letters to the editor) – Moralization with total agreement between all models and humans: *Dear people, if you want to make a difference, then you should be consistent and vote out the responsible, established parties, boycott the products of the polluters and warmongers, and switch to fair trade. Immediately and rigorously align your own life ecologically and economically, and fight for a just world. Only then will anything improve.*

Example 9 (Nonfiction books) – NVI with total agreement between all models and humans: *On the eve of the First World War, faith in progress was accompanied by fear of the abyss. Nietzsche had become the herald of the apocalyptic spirit of the age, who in Ecce Homo spoke of wars "such as have never yet existed on earth." It was into such a war — in which "the nineteenth century [...] was shot to pieces" (AH1/SW9, 133) and the ideal of eternal peace destroyed — that Ernst Juenger was to go.*

Example 10 (Letters to the editor) – 33 out of 35 models predict a moralization, 4 out of 5 humans annotate a NFR: *The penal order procedure may be suitable for tax evaders. For those who make up the bulk of the accused, however, the opportunity for an oral hearing is necessary. With the proliferation of written procedures, the judiciary has become alienated from the citizens. Yet without a simplification of legal language, even the strengthening of oral hearings will not achieve much. The mechanically recited indictment has degenerated into an undignified ritual.*

A.6.3. Illustrating Examples for §6.6

Example 11 (Nonfiction Books) – Use of moral vocabulary in neutral contexts: *Beneath the smoking ruins lay around seventy million dead. In breathtaking recklessness, the politicians had unleashed the dogs of war and triggered a frenzy of self-destruction. History has known terrible epochs such as the time of the plague or the Thirty Years' War, but never before had there been massacres of such magnitude as in the Thirty Years'*

War of 1914 to 1945 (setting aside the brief pause in between).

Example 12 (Letters to the editor) – Missing Context: *The explicit goal of the award ceremony was, not least, to bring the forgotten fate of this endangered people before the world public and the conscience of humanity.*

Example 13 (Parliamentary Debates) – Borderline Cases: *Empirical evidence is sometimes more valuable than the occasional utopian idea, no matter how often it has been written down. That is why we are well advised to build on what Ms. Kotting-Uhl has introduced here. It is forestry from which the concept of sustainability originally stems — developed not so much out of a love of trees and forests, but from a purely pragmatic pursuit of profit. This again shows that ecology and economy can go very well together.*