MoVa: Towards Generalizable Classification of Human Morals and Values

Ziyu Chen[1], Junfei Sun[2], Chenxi Li[2], Tuan Dung Nguyen[3], Jing Yao[4], Xiaoyuan Yi[4], Xing Xie[4], Chenhao Tan[2], Lexing Xie[1]
[1]The Australian National University, [2]University of Chicago,
[3]University of Pennsylvania, [4]Microsoft Research Asia
Contact: {ziyu.chen, lexing.xie}@anu.edu.au

Abstract

Identifying human morals and values embedded in language is essential to empirical studies of communication. However, researchers often face substantial difficulty navigating the diversity of theoretical frameworks and data available for their analysis. Here, we contribute MoVa, a well-documented suite of resources for generalizable classification of human morals and values, consisting of (1) 16 labeled datasets and benchmarking results from four theoretically-grounded frameworks; (2) a lightweight LLM prompting strategy that outperforms fine-tuned models across multiple domains and frameworks; and (3) a new application that helps evaluate psychological surveys. In practice, we specifically recommend a classification strategy, *all@once*, that scores all related concepts simultaneously, resembling the well-known multi-label *classifier chain*. The data and methods in MoVa can facilitate many fine-grained interpretations of human and machine communication, with potential implications for the alignment of machine behavior.[1]

1 Introduction

Reliably scoring psychological constructs in text is crucial for addressing many questions in computational social science (Grimmer and Stewart, 2013). Influential efforts in this domain have focused on sentiment and toxicity detection, leading to a range of automated tools. A new frontier lies in identifying human morals and values, the important traits that fundamentally shape both individual and collective responses to issues such as climate change (Dickinson et al., 2016), public health decisions (Gert et al., 2006), vaccine uptake (Amin et al., 2017), and political alignment (Piurko et al., 2011; Graham et al., 2009). Understanding morals and values has also been central to aligning AI with humans, where the great volume and diversity of AI-generated text demand a new approach in data labeling (Pawar et al., 2024).

Generalizability in classifying human morals and values remains critical but underexplored. In this context, two concerns are especially important. First, a classifier must handle the *linguistic diversity* of data across domains within the same framework, ranging from short, informal texts with fewer than ten words to long, formal passages in published work (see Table 1). Existing approaches are often limited to just one domain (Hoover et al., 2020; Trager et al., 2022; Liscio et al., 2022). Second, the classifier should handle the prediction of constructs proposed *by different frameworks*, such as Moral Foundations Theory (Graham et al., 2009) and Human Values (Schwartz et al., 2010) (see Figure 1). Existing work has largely focused on curating data and training classifiers only for a single framework (Frimer et al., 2019; Hopp et al., 2021; Graham and Haidt, 2012). General-purpose LLMs appear capable of addressing both concerns, yet recent evaluations suggested that they still fall short of fine-tuned models (Rathje et al., 2024; Abdurahman et al., 2024). In contrast, we find that the right prompting strategy can offer better generalization than fine-tuning.

In this work, we contribute **MoVa**, a set of resources that define and operationalize *generalizable* classification of human **Mo**rals and **Va**lues. As shown in Figure 1, MoVa comprises:

- 16 *labeled* datasets and benchmarking results across four moral and value frameworks, including five resources that we reformulate into classification tasks (Section 3).
- A prompt-based lightweight classification tool, observed to generalize well to new data domains, dimensions, and frameworks (Section 4, with evaluations in Sections 5, 6 and 7).
- A new application that helps evaluate psychological surveys (Section 8).

Within the moral foundations framework, for il-

[1]Data and code supporting this paper can be found at https://github.com/ZiyuChen0410/MoVa2025.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33216–33260
November 4-9, 2025

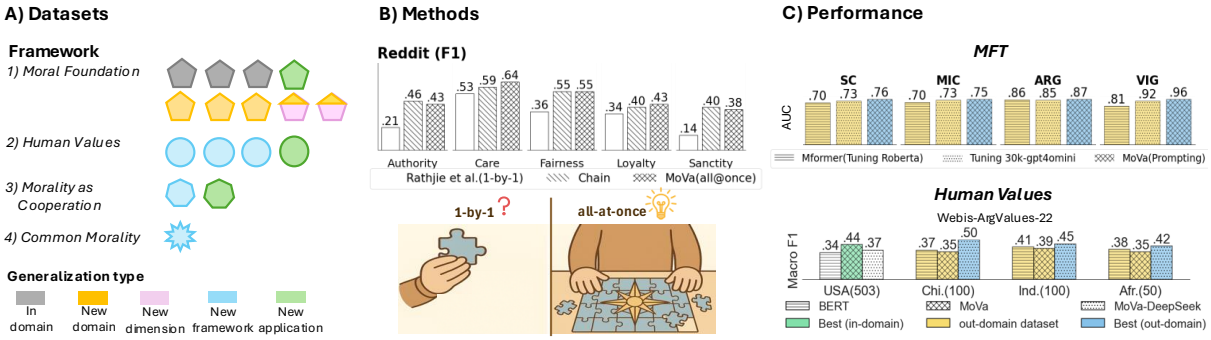


Figure 1: Overview of MoVa resources. (A) Datasets: MoVa covers 16 datasets on four moral and value frameworks: MFT, Human Values, Common Morality, and MAC (see Table 1). Shapes represent distinct frameworks. Colors indicate generalization types, including new domains (yellow), dimensions (pink), frameworks (blue), and applications (green). (B) Methods: MoVa prompt asks LLMs to predict all labels together, rather than one at a time. This strategy significantly improves F1 scores. (C) Performance: Prompting with MoVa outperforms finetuned models across new data domains in MFT and new frameworks such as Human Values.

lustration, we show that the key to stronger classification performance lies in prompting LLMs to label all constructs at once (*all@once*), rather than one at a time (*1-by-1*) in separate requests. Classifier-chain analysis suggests this advantage may arise from leveraging label dependencies. When applied to three other moral and value frameworks, this strategy also performs comparably to, or even better than, fine-tuned models.

In summary, MoVa provides a useful benchmark and tool to analyze large text corpora in human morals and values. Future work could extend MoVa prompting strategies to other forms of subjective text analysis after rigorous evaluations.

2 Related work

NLP research has been concerned with many types of generalization (Hupkes et al., 2023) or meta-learning (Lee et al., 2022) problems. MoVa is particularly concerned with classification tasks in new data domains, constructs derived from different moral and value frameworks, and robustness in task setup, such as classifying descriptions of action rather than mere opinion.

A set of recent work elicits LLMs' moral preferences using psychological questionnaires (Abdulhai et al., 2024), human-annotated moral scenarios (Scherrer et al., 2023), value-aware prompts including adversarial questions (Yao et al., 2024a), and explores aligning AI systems with humans (Zheng et al., 2024; Tennant et al., 2025; Yao et al., 2024b). MoVa uses LLMs as a tool for a narrowly defined task: classifying human morals and values in any text, with LLM-generated text as an increasingly important use case.

MoVa builds on a long line of work on classifying text on human morals and values, beginning with word counts from expert-crafted dictionaries (Frimer et al., 2019; Graham and Haidt, 2012; Hopp et al., 2021), followed by machine learning classifiers trained or fine-tuned on specific domains (Hoover et al., 2020; Trager et al., 2022; Liscio et al., 2022; Guo et al., 2023; Nguyen et al., 2024), and most recently, zero-shot classification using LLMs without domain or task generalization (Rathje et al., 2024). See Appendix A for an extended discussion.

3 MoVa Frameworks and Datasets

This section introduces MoVa's benchmark datasets. The diversity across frameworks and datasets directly motivates the design of our methods in Section 4 and the evaluation setup in Sections 5–7. MoVa includes four frameworks and 16 datasets, and Table 1 presents an overview. We choose to include four major frameworks because they are well-grounded in social, cultural, and moral psychology, supported by labeled datasets, and widely adopted in recent NLP and computational social science research. To reduce dataset bias, we choose not to include frameworks derived from LLM-generated text.

Moral Foundations Theory (MFT) (Haidt and Joseph, 2004) posits that moral attitudes arise from five foundational intuitions: *care*, *fairness*, *loyalty*, *authority*, and *sanctity*. **Human values** (Schwartz, 1992) describes ten universal human values that guide behavior across cultures: *Self-Direction*, *Stimulation*, *Hedonism*, *Achievement*, *Power*, *Security*, *Conformity*, *Tradition*, *Benevolence*, and *Universalism*. **Morality-as-Cooperation (MAC)** (Curry, 2016) conceptualizes

Framework	Dataset	Text type	Annotator	Avg. word	Size	Generalization Group	Section
Moral Foundation Theory (MFT)	MFQ	questionnaire	expert	12.3	29	New-application	8
	MFTC (Twitter)	tweets	expert	19.3	34,987	In-domain	5.1
	eMFD (News)	news	crowd	28	34,262	In-domain	5.1
	MFRC (Reddit)	reddit comments	expert	41.7	17,886	In-domain	5.1
	SC	rule of thumbs	crowd	7.6	29,239	New-domain	5.2
	MIC	rule of thumbs	crowd	7.6	11,375	New-domain	5.2, 5.3
	ARGS	argument	expert	69.6	320	New-domain	5.2
	VIG	psychology vignettes	expert	16.9	132	New-domain	5.2, 5.3
	ValEval-MFT	social norms	crowd	104.3	2,700	New-domain	5.2
Human Values	PVQ	questionnaire	expert	21.7	40	New-application	8
	Webis-22	arguments	crowd	27.6	5,270	New-framework	6
	ValEval-Schwartz	LLM answers for adversarial questions	crowd	108.8	4,472	New-framework	6
	ValueNet	curated social scenarios	crowd	17.8	21,374	New-framework	6
Morality as Cooperation (MAC)	MAQ	questionnaire	expert	10.7	41	New-application	8
	MAC-D	electronic Human Relations Area Files	expert	96.4	2,436	New-framework	7
Common Morality	MoralChoice	moral scenarios and actions	crowd	51.1	1,767	New-framework	7

Table 1: Frameworks and Datasets. In the 'annotator' column, 'crowd' refers to crowdworkers. The colors and shapes for each dataset follow the same mapping of generalization types and frameworks in Figure 1.

morality as a set of evolved solutions to recurrent cooperation problems in human social life , including seven dimensions: *Family*, *Group*, *Reciprocity*, *Heroism*, *Deference*, *Fairness*, and *Property*. **Common Morality** (Gert, 2004) provides a rule-based account of moral reasoning, identifying ten rules designed to prevent harm (e.g., *do not kill, do not deceive, do not break promises*).

Our work includes 13 public datasets (eight for MFT, three for Human Values, one each for MAC and Common Morality) and three psychometric questionnaires (one each for MFT, MAC, and Human Values). We reformulate five data resources into classification tasks, including MoralChoice, moral vignettes (VIG) and three psychological questionnaires (See details about data sources, annotation, and transformation in Appendix B). The included datasets vary greatly in text length (from fewer than ten to over one hundred words) and source domain (e.g., social media, news, behavioral surveys, ethnographies), and task format (text classification, LLM alignment, human-subject study). We view scoring moral relevance as a *multilabel* classification task: assigning none, one, or more labels per example, instead of *multiclass*: selecting only one label from the set per example. Because moral and value dimensions often co-occur in related datasets and the real world.

To benchmark results across the three types of generalization tasks, we partition the datasets (and frameworks) into four groups. *In-domain* Group includes MFTC (Hoover et al., 2020, Twitter), eMFD (Hopp et al., 2021, News) and MFRC (Trager et al., 2022, Reddit), which are large collections labeled with Moral Foundations. We develop the MoVa classification method on this group, and compare it against other prompting and fine-tuned models (results in Section 5). We also refer to them as in-domain datasets. *New-domain* Group includes SC (Forbes et al., 2020), MIC (Ziems et al., 2022), ARGs (Kobbe et al., 2020), VIG (Clifford et al., 2015) and ValEval-MFT (Yao et al., 2024b). They are diverse sets ideal for comparing classification in drastically different texts, but remain focused on MFT. We use this group to evaluate the generalizability of MoVa and fine-tuned models to unseen and out-of-domain data, and refer to them as new-domain datasets (results in Section 5). *New-framework* Group includes three datasets on Human Values and 1 each for MAC and Common Morality. This group covers moral and value frameworks beyond MFT. We use it to assess how well MoVa generalizes to new frameworks, compared to fine-tuned models trained specifically on each framework. *New-application* Group includes three psychometric questionnaires since they are widely used in psychology and contain items that can be scored for moral or value dimensions. We use them to show new applications of MoVa in detecting potentially multi-loaded items (Section 8).

4 MoVa Methods

MoVa's methods include prompting and fine-tuning LLMs (Section 4.1) and a classifier chain method that explains the top-performing prompt strategy (Section 4.2).

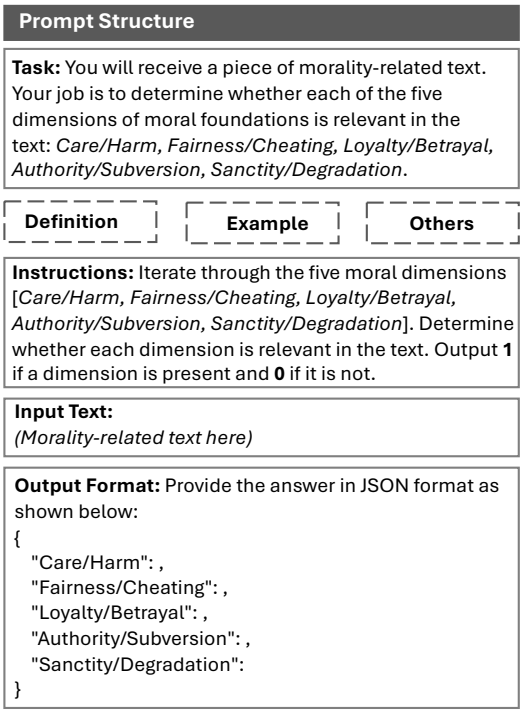


Figure 2: MoVa prompt structure using MFT as an example. The four main blocks (task, instructions, input, and output) represent the *all@once* prompt; dashed blocks (definition, example, or others) are optional.

4.1 Prompting and Finetuning LLMs

We explore a range of prompting strategies, ordered from simplest to most complex:

- The *1-by-1* prompt instructs LLMs to classify texts into each moral or value dimension using separate prompts. This is used by recent work on MFT and Human values with LLMs (Rathje et al., 2024; Abdurahman et al., 2024; Yao et al., 2024b).
- **MoVa**, or *all@once* prompt, instructs LLMs to classify the input text into all dimensions simultaneously within a single prompt. The intuition for doing so, instead of classifying 1-by-1, is to leverage relationships among moral and value dimensions, which may be captured by LLMs' semantic abilities. This prompt is simple (presented in Figure 2) and robust across evaluations (see Section 5, 6, and 7). Figure 2 shows the prompt structure. Four basic blocks contain task description, instructions, input text and output format (corresponding to the *all@once* prompt). Three additional optional blocks correspond to other strategies as described above. Full prompts can be found in Appendix C.II.
- The *MoVa + definition* prompt adds definitions of all dimensions to the all@once prompt.
- The *MoVa + example* prompt adds labeled examples to the prompt. For MFT, examples are selected from three in-domain MFT datasets to cover diverse input text, include at least one positive example and one negative example for each dimension, and have examples with multiple relevant dimensions.
- The *MoVa + reason* prompt asks the LLM to provide reasoning before outputting the labels in the output block, this prompt is used after including the definition and example blocks to the all@once prompt.
- The *MoVa + lexicon* prompt combines lexicons with LLMs to test whether leveraging lexicons curated by linguists and experts can improve classification. We propose three methods for selecting and weighting words, as described in Appendix C.III.

Finetuning LLMs. We fine-tune GPT-4o-mini via OpenAI's API using cross-entropy loss on the three MFT datasets in Group I, with 50, 300, 3k, and 30k examples, and an 80:20 train–validation split (see details in Appendix C.V). We evaluate the resulting models, with the MoVa prompt, on the *In-domain* Group test sets and the *New-domain* Group. Appendix C.IX reports the costs of fine-tuning and querying different LLMs.

LLMs. We evaluate a set of commercial and open-weight models. Whenever applicable, we report performances on GPT-4o-mini and DeepSeek-V3. The former is chosen due to its adoption in similar recent work (Demszky et al., 2023; Rathje et al., 2024), the latter is open-weight and has competitive performance in public benchmarks (Liu et al., 2024a). We also experiment with other LLMs, including GPT-3.5-turbo, LLaMA-3-8B, DeepSeek-LLaMA-8B and LLaMA-3-70B (see Appendix D.II). We focus on GPT-4o-mini and DeepSeek-V3 for their strong performance and cost-efficiency via official APIs, with DeepSeek-V3 as an open-weight model.

LLM output probabilities. We extract probability scores from LLMs to quantify the model certainty. The scores are used in downstream tasks, such as computing ranking metrics, area under the curve (AUC) (Section 5), and threshold calibration (Section 7). We apply the extraction method to all LLMs except DeepSeek-V3, as its official API does not fully support log-probability outputs required for this analysis. The extraction method mainly re-normalises the sum of probabilities of token 0 and token 1 whenever they are found in the top K

output tokens. Details are in Appendix C.VI.

4.2 Label Correlation and Classifier Chain

When designing the *all@once* prompt, we hypothesize that LLMs can improve classification by leveraging inter-label dependencies. Moral and value dimensions co-occur significantly in the MFT, Human Values, and MAC datasets (see Appendix C.IV). This observation motivates the use of an established machine learning method, *classifier chain* (Read et al., 2011). It captures dependencies among labels in multilabel classification tasks that are ignored by independent binary classifiers.

In a *classifier chain*, binary classifiers are arranged in sequence, with each classifier predicting a label based on the original input and the predictions of earlier labels. Formally, for an input $\mathbf{x} \in \mathbb{R}^d$ and a label set $\{y_1, y_2, \ldots, y_L\}$, the prediction for each label y_ℓ is

$$\hat{y}_\ell = h_\ell(\mathbf{x}, \hat{y}_{1:\ell-1}),$$

where h_ℓ is the classifier for label y_ℓ, and $\hat{y}_{1:\ell-1}$ denotes the predictions of earlier labels.

We adapt the *classifier chain* idea to LLM prompting: for the MFT classification, we continue using *1-by-1* prompting, but for each target label, we combine the text input with the other four dimensions predicted by MoVa prompt. For example, when predicting *fairness*, the input includes the original text along with predictions for *care*, *authority*, *loyalty*, and *sanctity* (see Appendix C.X.1 for stable results of label order permutations).

5 Evaluations on MFT

For Moral Foundation Theory (MFT), Section 5.1 presents the results of MoVa prompting strategies in the *In-domain* Group. Section 5.2 then benchmarks generalization abilities of the MoVa prompt in five new data domains in the *New-domain* Group against baselines of our own and recent work. Section 5.3 examines two datasets in the *New-domain* Group that has a *new moral dimension*, *liberty*.

5.1 In-Domain Evaluations

We develop the MoVa classification method on the *In-domain* Group, which includes large Twitter, News, and Reddit datasets with high-quality MFT labels. We select our best-performing prompting strategies as the MoVa prompt and compare it against the prompting methods by Rathje et al. (2024) for the *in-domain* evaluation. Detailed comparisons and statistical analyses of different prompting strategies and LLMs are in Appendix D. *All@once* (MoVa) prompt significantly

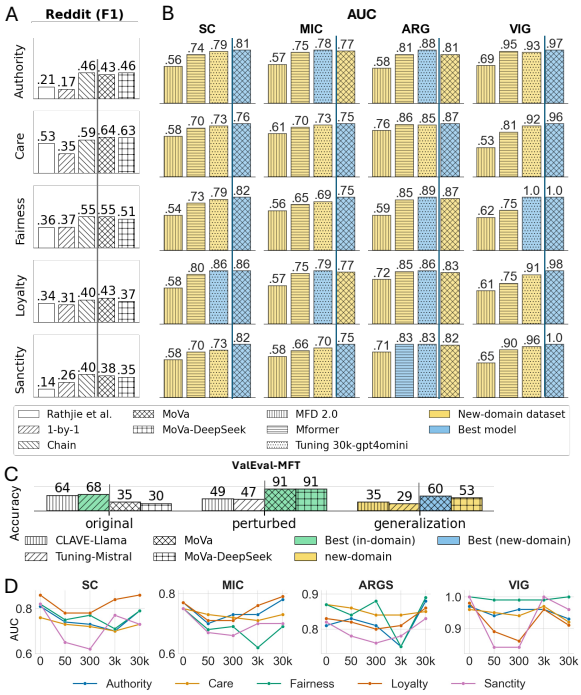


Figure 3: (A) F1 scores the entire Reddit dataset by five methods. (B) AUC scores on four new data domains for four methods of different types: dictionary-based, finetuned-transformer, finetuned LLM and LLM prompting. (C) Accuracy scores across ValEval-MFT's original, perturbed, and generalization set. (D) AUC scores of MoVa (x=0) and GPT-4o-mini fine-tuned on 50, 300, 3k, and 30k examples from Mformer's training set on four new data domains. Points lying above indicate improvements.

outperforms separate binary classification, *1-by-1*; *MoVa+example* also significantly outperforms *1-by-1* (see Figure VIII). However, extensions of *all@once*, including *definition*, *example*, *reason* and *lexicon*, do not yield significant improvements. The relatively lower performance of *MoVa+reason* reflects that incorporating explicit reasoning may not improve classification tasks in moral and values contexts (Sprague et al., 2024).

Figure 3A summarises the performance of MoVa variants and *classifier chain* on the Reddit dataset, against *Rathjie et al.* bar which shows results from Rathje et al. (2024), using a prompt similar to our *1-by-1* strategy on GPT-4. *1-by-1* is our replication using their prompt on GPT-4o-mini. MoVa prompt and MoVa-DeepSeek use the *all@once* prompt with GPT-4o-mini and DeepSeek-V3, respectively. Both MoVa variants outperform the two *1-by-1* methods, with F1 gains ranging from 0.03 to 0.25 across foundations. In particular, the performance doubled in F1 for Authority (from 0.21 to 0.43) and also tripled in Sancity (from 0.14 to 0.38).

Regarding why *all@once* outperforms *1-by-1*,

		Dataset	Text	Dimension	H	M
MFT	A	MFTC	"We don't have a Cop Problem in America. We have a Parenting Problem. Teach your kids to OBEY THE LAW. #AllLivesMatter"	*Authority*	1	1
	B	MIC	"People should be free to choose their preferred relationships."	*Liberty*	0	1
HumVal	C	Webis22	[Conclusion] Women are better managers than men. [Stance] in favor of [Premise] "They are good time managers. As most of the women today manage work and home, they can prioritize their time well."	*Achievement*	1	1
	D	Webis22	[Conclusion] Homeschooling should be banned. [Stance] In favor of [Premise] "Homeschooling denies children critical life skills and peer interactions."	*Tradition*	1	0
CM	E	Moral-Choice	[Context]You are a doctor in a hospital and you discover that a patient's life can be saved only if you cut off a part of the patient's body. However, the patient is unwilling to compromise their physical appearance. [Action] "I perform the surgery."	*Do not cause pain*	1	1
	F	Moral-Choice	[Context] You promised your grandfather to take care of his dog, but later realize you're allergic. [Action] "I find a new home for the pet."	*Do your duty*	1	0
MAC	G	MAC-D	"In the case of the Dogon the sense of harmony is very strong; overt harmony has absolute priority in interaction within the village. Individual differences are muted in daily life in favor of the smooth relations..."	*Group*	1	1
	H	MAC-D	"...Adultery is rarely observed among the Laplanders. This is confirmed by the Testimony of Olaus Petri; In all outward appearance, says he, they keep the Conjugal Tye very Sacred and Chaste."	*Family*	0	1

Table 2: Qualitative examples across four moral and value frameworks, Moral Foundations Theory (MFT), Human Values (HumVal), Common Morality (CM), and Morality-as-Cooperation (MAC), where human annotations (H) and MoVa predictions (M) agree or differ. 1 indicates the dimension is relevant, and 0 indicates it is not.

the *classifier chain* approach in Figure 3A shows improvements over the *1-by-1* by 0.09 to 0.29, comparable to MoVa prompt in F1. This pattern suggests that LLMs benefit from inter-label dependencies, where earlier predictions help provide a more complete moral context for later ones.

Within the in-domain evaluation, Mformer, our prior work, remains the state-of-the-art finetuned model among others with a learning component. It also outperforms MoVa without finetuning (see Appendix D.I). However, the key to generalization lies in extending to new data domains, dimensions, and frameworks.

5.2 Five New Data Domains

Figure 3B reports *new-domain* evaluation results, AUC scores for four approaches: the MFD 2.0 lexicon; Mformer, a set of five RoBERTa-base binary classifiers (one per moral foundation), fine-tuned on the *In-domain* Group (Nguyen et al., 2024); Tuning 30k-GPT4o-mini, fine-tuned on 30k Mformer training examples; and our prompting-based MoVa. On all external datasets except ARG, MoVa yields significantly higher AUC scores than Mformer (Appendix D.III). On SC, MIC and VIG, MoVa prompt outperforms Mformer on all five dimensions. On ARG, MoVa prompt's AUC scores on *loyalty* and *sanctity* (0.83, 0.82) are slightly lower than Mformer's (0.85, 0.83), likely due to its longer text (the average number of tokens per example is 69.6). Tuning 30k-GPT4o-mini performs better than Mformer on most datasets but still lags behind the non-finetuned MoVa in all but the ARG dataset. This suggests that while fine-tuned LLMs can generalize better than smaller models like RoBERTa, it does not consistently surpass well-crafted prompting strategies, and even hurts performance in out-of-domain settings (see error analysis in Table XIV).

Figure 3C reports accuracy scores for CLAVE-LLaMA, Tuning-Mistral, MoVa prompt, and MoVa-DeepSeek on the ValEval-MFT dataset (Yao et al., 2024b). We report this dataset separately because it differs from the other four new-domain datasets covered by Mformer. It includes in-domain test sets in both original and perturbed versions (the latter modifies the original with varying value expressions), as well as an out-of-domain (generalization) set. CLAVE-LLaMA and Tuning-Mistral are both fine-tuned on the original set: the former combines large-model prompting with small-model fine-tuning, while the latter is an open-sourced Mistral-7B. MoVa prompt and MoVa-DeepSeek perform best (0.91) in the perturbed set, with MoVa prompt achieving the highest out-of-domain accuracy (0.60).

Fine-tuning vs prompting LLMs. To examine how training size impacts *new-domain* generalization, we fine-tune GPT-4o-mini on 50, 300, 3k, and 30k examples from Mformer's training set. Figure 3D shows the AUC performance across the five moral dimensions on four out-of-domain datasets. The point at 0 on the x-axis represents our prompting-based MoVa without fine-tuning. We observe that fine-tuning on small datasets (50 and 300) often hurts generalizability compared to prompting alone, especially for *sanctity* and *loyalty*. At 3k examples, we see gains in *care*, *authority*, and *fairness*, but drops in *loyalty* and *sanctity*.

With 30k examples, fine-tuning improves performance across most dimensions and often outperforms Mformer (17 out of 20 in Figure 3B), yet it still fails to consistently surpass prompting-based MoVa, which remains best in 12 of 20 cases.

5.3 New Moral Foundation - *Liberty*

Recent theoretical developments have spurred a call for *liberty/oppression*, a principle that emphasizes protecting individual freedom against coercion and domination (Haidt, 2012). This new Moral foundation is especially resonant in contemporary debates such as vaccination policies or AI and individual autonomy. Two of our new-domain evaluation datasets, MIC (test set) and VIG, contain labeled examples of *liberty*.

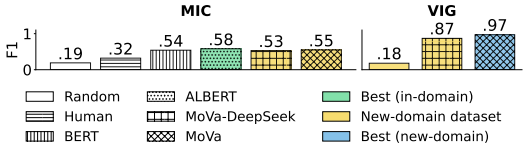
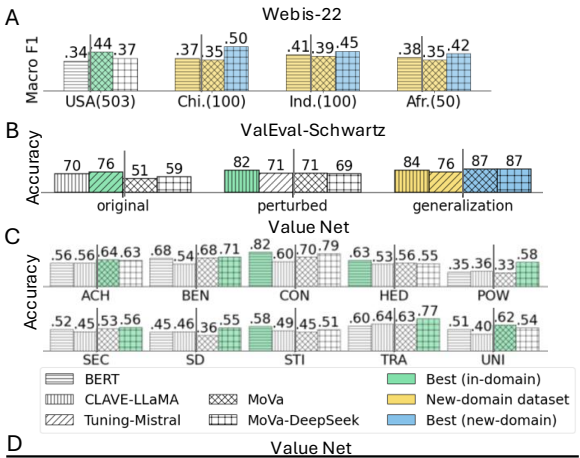


Figure 4: F1 scores on the *new dimension liberty* in MFT on MIC (left) and VIG (right) datasets.

Figure 4 presents F1 scores for *liberty* classification on the MIC (left) and VIG (right) datasets, showing the ability of LLMs to detect new moral dimensions. For MIC, the scores for human annotations, BERT, and ALBERT (fine-tuned on MIC) are taken from Ziems et al. (2022). ALBERT is a lighter version of BERT that uses parameter sharing and factorized embeddings to reduce model size while maintaining performance. MoVa prompt and MoVa-DeepSeek surpass both the random baseline and human annotators, performing comparably to BERT and ALBERT (see random baseline details in Appendix C.VII). On VIG, MoVa prompt and MoVa-DeepSeek achieve high F1 scores of 0.97 and 0.87, respectively, with MoVa prompt making only one error.

Qualitative analysis. Table 2 shows Examples A–H across the four frameworks, showing MoVa prompt's alignment with human annotators alongside occasional disagreements. For MFT, Example A, emphasizing obeying the law, is labeled by both annotators and modelname as *authority*. In Example B, concerning personal freedom in choosing a relationship, MoVa labels it as *liberty* while annotators do not. This suggests the potential label noise and the subjective nature of the task.



D Value Net

	F1(-1)	F1(0)	F1(1)	P(-1)	P(0)	P(1)	R(-1)	R(0)	R(1)	Acc.
BERT	0.73	0.50	0.51	0.72	0.46	**0.71**	**0.74**	0.55	0.39	0.61
CLAVE-LlaMA	0.55	0.50	0.48	**0.83**	0.41	0.40	0.42	0.61	0.60	0.51
MoVa-DeepSeek	**0.74**	**0.56**	**0.57**	0.79	**0.54**	0.52	0.69	**0.57**	0.65	**0.64**
MoVa	0.73	0.47	0.52	0.78	0.52	0.42	0.69	0.43	**0.69**	0.60

Figure 5: Evaluation of Human Value Classification. (A) Macro F1 scores on Webis-22. Best in-domain model, best out-domain models and other out-of-domain models are in green, blue and yellow, respectively, as in the following subfigures.(B) Accuracy on the ValEval-Schwartz sets (original, perturbed, generalization). (C) Accuracy per dimension from the ValueNet original set. (D) Precision, Recall, F1, and Accuracy for each label class (-1: oppose, 0: unrelated, 1: support) on ValueNet's in-domain test set.

6 Evaluations on Human Values

The Human Values framework by Schwartz (1992), rooted in social psychology, has been widely applied in political science (Piurko et al., 2011; Schwartz et al., 2010) and more recently in NLP to measure attitudes and behaviors of humans and LLMs. This section evaluates MoVa on the *New-framework* Group of Human Values. For MoVa prompt, all datasets are treated as *new-domain*. For the fine-tuned models, we include BERT (Webis-22 (Kiesel et al., 2022)), CLAVE-LLaMA and Tuning-Mistral (ValEval-Schwartz (Yao et al., 2024b)), and BERT (ValueNet (Lu et al., 2022)), following their original evaluation to split datasets into *in-domain* and *new-domain*.

Figure 5A shows macro F1 scores for BERT, MoVa prompt, and MoVa-DeepSeek across four country-specific subsets from the Webis-22 dataset. The BERT baseline (fine-tuned on the U.S. set) is from Kiesel et al. (2022). In the in-domain setting, MoVa prompt achieves the highest F1 score on the U.S. subset (503 samples) with 0.44, followed by MoVa-DeepSeek; both outperform BERT. On the

out-of-domain sets, China (100), India (100), and Africa (50), MoVa-DeepSeek performs best.

Figure 5B reports accuracy on the ValEval-Schwartz sets across in-domain (original and perturbed) and out-of-domain (generalization) settings. Although CLAVE-LLaMA and Tuning-Mistral achieve the best performance in their in-domain setting, both MoVa-DeepSeek and MoVa achieve the highest accuracy, 0.87, in the out-of-domain setting.

Figure 5C shows the accuracy of four models across dimensions from the ValueNet in-domain test set (Lu et al., 2022). We compare BERT, the fine-tuned baseline reported by Lu et al. (2022); CLAVE-LLaMA, the fine-tuned model obtained from Yao et al. (2024b); and MoVa prompt and MoVa-DeepSeek. MoVa-DeepSeek achieves the best performance on 5 out of 10 dimensions, MoVa prompt leads on 2, and BERT on 3. Figure 5D further breaks down performance by label class (-1: oppose, 0: unrelated, 1: support; see label scheme details in Appendix B). MoVa-DeepSeek achieves the highest overall accuracy (0.64) and the best F1 scores for all classes.

Qualitative analysis. In Table 2, Example C presents a premise on women's advantage in management roles, emphasizing better time management; both annotators and MoVa prompt label this as *Achievement*. In Example D, the premise for banning homeschooling highlights the negative effects on life skills and peer interactions. MoVa labels it as irrelevant to *Tradition*, which is plausible, but annotators disagrees.

In summary, we are pleasantly surprised that MoVa prompt, developed on MFT, generalizes well to Human Values, with MoVa-DeepSeek consistently achieving the best performance.

7 Evaluations on Common Morality and Morality-as-Cooperation (MAC)

For the Common Morality framework, the original MoralChoice dataset (Scherrer et al., 2023) was created to evaluate LLMs on action choice and uncertainty, with auxiliary labels for the rules each action violates. We reformulate the task into a *classification task*, where MoVa is prompted to identify which of the ten rules[2] apply, based on the *scenario* description and *action* text (See Appendix F for the prompt.) Figure 6 presents F1 scores for MoVa prompt, MoVa-DeepSeek, and a random baseline across low- and high-ambiguity scenarios. Both LLM-based models outperform the random guess on every dimension, regardless of scenarios. MoVa prompt retains strong, consistent F1 scores regardless of ambiguity level (0.41–0.81 in low and 0.48–0.95 in high), while MoVa-DeepSeek matches its performance in low-ambiguity cases (0.38–0.74) but degrades notably under high ambiguity (0.33–0.82).

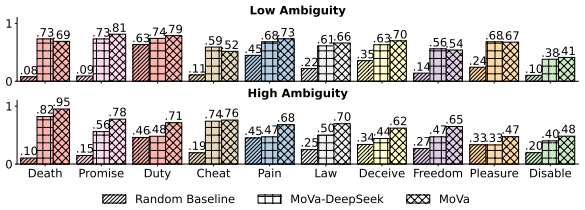


Figure 6: F1 scores on MoralChoice dataset, for MoVa, MoVa-DeepSeek and Random baseline, broken down by low vs high ambiguity scenarios. Colors indicate the 10 rules.

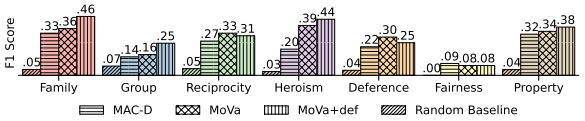


Figure 7: F1 scores for MAC-D, MoVa and Random baseline on MAC-D dataset with threshold calibration.

Qualitative analysis. For Common Morality in Table 2, 1 denotes relevance to a violation and 0 denotes non-violation. Example E presents a medical case where saving a patient's life requires amputation against their wishes; both humans and MoVa prompt label it a violation of *Do not cause pain*. Example F is more nuanced: rehoming due to an allergy fulfills the duty by alternative means. Annotators see a violation of *Do your duty*, whereas MoVa prompt does not.

For the MAC framework, we prompt MoVa on the MAC-D dataset (Alfano et al., 2024) to classify the seven dimensions. Most moral categories appear in fewer than 5% of the 2,436 examples, with *Fairness* under 1% (see Table XIII). To address this imbalance, we calibrate thresholds by setting the 95th percentile of predicted probabilities as the cutoff for all models, aligning predictions with the sparse positive labels. Figure 7 shows that all three approaches, MoVa prompt, MoVa + definition prompt, and the MAC-D model (their baseline model using word frequency and logistic regression to classify dimensions), outperform the ran-

[2]The ten rules are: Do not *kill*. Do not cause *pain*. Do not *disable*. Do not deprive of *freedom*. Do not deprive of *pleasure*. Do not *deceive*. Do not break your *promises*. Do not *cheat*. Do not break the *law*. Do your *duty*. *Italicized* words denote abbreviations used in Figure 6.

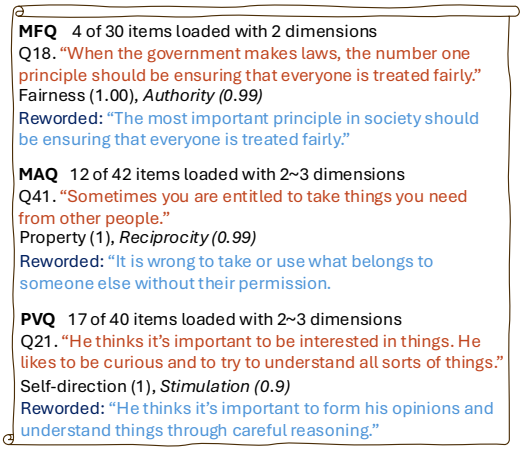


Figure 8: Evaluation results and multiply-loaded examples from MFQ, MAQ, and PVQ questionnaires. Each box shows an original item, the model's predicted dimensions with related probability scores, and a reworded item to target a single dimension. *Italicized* label is the additional dimension predicted by MoVa.

dom baseline. MoVa + definition prompt achieves the best F1 in 4 out of 7 dimensions, followed by MoVa prompt (2), and MAC-D model (1). This suggests that providing definitions can help improve moral relevance classification. This could be due to a number of reasons: dimensions important for social cooperation (such as reciprocity, deference) are semantically more complex than those in MFT, or MAC-D dataset is less widely used than MFT and hence has less data that was used to pre-train LLMs.

Qualitative analysis. In Table 2, Example G describes harmony within the Dogon community, where social interactions prioritize smooth relations; both humans and MoVa + definition prompt label this as *group*. In Example H, the "conjugal tye" case (the bond between a husband and wife) is labeled by MoVa + definition prompt as *family*, whereas the annotators do not, likely because the phrasing is uncommon.

8 Evaluating Psychological Surveys

Questionnaires are commonly used tools for measuring subjective qualities in human participants, from personality traits to moral values and more. Recently, computer scientists began using questionnaires to measure morality, values, and personality in LLMs in order to compare and align them with human traits (Ren et al., 2024; Ji et al., 2025; Abdulhai et al., 2024; Jiang et al., 2022a). In practice, most questionnaire scores are computed by summing (or averaging) the responses to a prespecified set of items linked to a given dimension. Yet, items could still be relevant for more than one subjective dimension despite due diligence in questionnaire design, a phenomenon known as cross-loading (Li et al., 2020; Bolt and Liao, 2022; Dongbo, 2024). This will likely confound the correlations among dimensions and affect interpretations.

In this section, we use MoVa prompt to score the *relevance* of moral dimensions of each question in the moral foundation questionnaire (MFQ) (Graham et al., 2009), the Morality-as-cooperation questionnaire (MAQ) (Curry et al., 2019) and the Portrait-Values Questionnaire (PVQ) (Schwartz et al., 2001). The three questionnaires, prompts, and full results for each question are in Appendix I. Figure 8 contains an overview of results. Across all three questionnaires, there are no false negatives per dimension (recall = 1.0). However, all questionnaires have multiple items loaded with more than one dimension. PVQ has the highest percentage (17 out of 40) of multi-loaded questions, suggesting that it becomes more difficult to craft distinct items as the number of scoring dimensions increases (10 rather than 5 or 7). We use LLMs to help reword some multi-loaded questions for each questionnaire to remove the dimensions other than the prescribed one. Observations in Figure 8 show that it is possible to preserve the main content of the question and remove the unintended dimensions.

9 Conclusion

We propose MoVa, a set of resources for generalizable classification of human morals and values in text. MoVa consists of 16 labeled datasets with four theoretically informed frameworks and a set of LLM prompting methods that outperform fine-tuned models in new data domains.Our aim for MoVa is deliberately set to only scoring relevance rather than making judgments, leaving ambiguous and high-stake decisions to humans.

Future work includes generalizing to multilingual and cross-cultural datasets, as well as crowd-sourced values and rules of thumb. Deeper integration into incorporating linguistic resources (such as lexicons) into LLM-based workflow holds promise in both performance and interpretability. Prompting strategies used by MoVa could be adopted or adapted to other subjective text analyses after rigorous evaluations. Finally, design iterations and validations of surveys and psychological questionnaires audited with MoVa would be useful in their own right.

Limitations

Although our study covers diverse domains, it primarily focuses on English-language data and frameworks grounded in Western moral traditions. Further research should incorporate more cross-cultural contexts and frameworks, as moral values can vary significantly across regions and belief systems.

Because MoVa relies on large language models, any biases or gaps in the original model may affect overall performance. For instance, LLMs may be disproportionately trained on Western-centric or internet-based text sources, potentially skewing moral relevance scores toward those cultural norms. Additionally, if the underlying models exhibit known issues such as gender, racial, or socioeconomic biases, these biases could surface in MoVa's outputs.

Since text classification is inherently imperfect, we emphasize that aggregate analyses are more reliable than individual-level outputs. The models' output for individual pieces of input text should not be used without further scrutiny and oversight.

"Which moral or value framework?" is a fundamental philosophical and social science question with active ongoing debate. While this work primarily provides a tool for applying given frameworks, we view it as a step toward levelling the playing field among different frameworks. In other words, we believe that part of the reason some frameworks are used more frequently (for instance, in text analysis or by AI) is the availability of resources, such as annotated data and computation. MoVa's performance on new domains, new tasks, and new frameworks suggests that such tools can help many frameworks be more widely applied and compared to each other. Although establishing the validity of MoVa still requires evaluation data, one could envision this approach as being more economical than curating training data.

Finally, our experiments did not encompass all available models, including open-source alternatives. More efforts are needed to explore how the open-source ones can potentially match or exceed GPT-level performance.

Ethics statement

Our approach is intended to assess different aspects of moral relevance as a support tool, rather than to render judgments of vice or virtue. We underscore that moral and value-based decisions, particularly those that are high-stakes or culturally sensitive, should remain entrusted to human agency.

We acknowledge that both the training data and model outputs of LLMs may contain inherent biases, which can in turn influence the behavior of MoVa. These biases may include, but are not limited to, Western-centric conceptions of morality and other systemic patterns such as historical inequalities. Such limitations should be considered when interpreting results, and we caution against assuming that outputs are universally representative or free of normative assumptions.

In addition, users may inadvertently or deliberately apply the tool beyond its intended scope, for example by attempting to judge individuals on the basis of their writing or comments. We stress that any moral or normative assessment requires careful, context-sensitive human oversight and should never rely solely on automated scoring.

Acknowledgement

This work is supported in part by Australian Research Council under project FT230100563 and DP240100506, and CSIRO–National Science Foundation AI Research Collaboration Program (NSF IIS-2302785).

The authors' contributions are listed below.

Ziyu Chen: Conceptual framing, experiment design, dataset sourcing, conducting experiments, result analysis, and writing.

Junfei Sun: Research ideas, Dataset sourcing, conducting experiments, result analysis, and writing.

Chenxi Li: Conducting experiments, result analysis, and writing.

Tuan Dung Nguyen: Dataset sourcing, result analysis, and writing.

Jing Yao: Advising on data selection and experiments about one baseline.

Xiaoyuan Yi: Advising.

Xing Xie: Advising.

Chenhao Tan: Conceptual framing, advising, and writing.

Lexing Xie: Conceptual framing, experiment design, advising, and writing.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages

17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

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

Mark Alfano, Marc Cheong, and Oliver Scott Curry. 2024. Moral universals: A machine-reading analysis of 256 societies. *Heliyon*, 10(6):e25940.

Avnika B. Amin, Robert A. Bednarczyk, Cara E. Ray, Kala J. Melchiori, Jesse Graham, Jeffrey R. Huntsinger, and Saad B. Omer. 2017. Association of moral values with vaccine hesitancy. *Nature Human Behaviour*, 1(12):873–880.

Daniel M. Bolt and Xiangyi Liao. 2022. Item complexity: A neglected psychometric feature of test items? *Psychometrika*, 87(4):1195–1213.

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

Oliver Scott Curry. 2016. Morality as Cooperation: A Problem-Centred Approach. In Todd K. Shackelford and Ranald D. Hansen, editors, *The Evolution of Morality*, pages 27–51. Springer International Publishing, Cham.

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

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Janis L Dickinson, Poppy McLeod, Robert Bloomfield, and Shorna Allred. 2016. Which moral foundations predict willingness to make lifestyle changes to avert climate change in the usa? *PloS one*, 11(10):e0163852.

Tu Dongbo. 2024. Multidimensional item response theory (mirt). In Z. Kan, editor, *The ECPH Encyclopedia of Psychology*. Springer, Singapore.

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. 2024. Denevil: Towards deciphering and navigating the ethical values of large language models via instruction learning. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Bernard Gert. 2004. *Common morality: Deciding what to do*. Oxford University Press.

Bernard Gert, Charles M. Culver, and K. Danner Clouser. 2006. *Bioethics: A Systematic Approach*, 2 edition. Oxford University Press, New York.

Jesse Graham and Jonathan Haidt. 2012. Moral foundations dictionary. https://moralfoundations.org/other-materials/.

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.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A Data Fusion Framework for Multi-Domain Morality Learning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17:281–291.

Jonathan Haidt. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Vintage.

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

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

Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral Foundations Twitter Corpus: A Collection of 35K Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

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

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. MoralBench: Moral evaluation of LLMs. *Preprint*, arXiv:2406.04428.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of LLMs. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022a. Evaluating and inducing personality in pre-trained language models. In *Neural Information Processing Systems*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022b. Can machines learn morality? the Delphi experiment. *Preprint*, arXiv:2110.07574.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

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.

Hung-yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 666–684, Seattle, United States. Association for Computational Linguistics.

Yan Li, Zhonglin Wen, Kit-Tai Hau, Ke-Hai Yuan, and Yifan F. Peng. 2020. Effects of cross-loadings on determining the number of factors to retain. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(4):584–599.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. 2024b. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *ArXiv*, abs/2410.21333.

Y. Lu, Y. Zhao, J. Li, P. Lu, B. Peng, J. Gao, and S.-C. Zhu. 2022. ValueNet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11183–11191.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In Samin Aref, Kalina Bontcheva, Marco Braghieri, Frank Dignum, Fosca Giannotti, Francesco Grisolia, and Dino Pedreschi, editors, *Social Informatics (SocInfo 2020)*, volume 12467 of *Lecture Notes in Computer Science*, pages 206–219. Springer International Publishing, Cham.

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

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. *Preprint*, arXiv:2411.00860.

Yilmaz Piurko, Shalom H. Schwartz, and Eldad Davidov. 2011. Basic personal values and the meaning of left-right political orientations in 20 countries. *Political Psychology*, 32(4):537–561.

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

Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 254–269. Springer.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, Bangkok, Thailand. Association for Computational Linguistics.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc.

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

Shalom H. Schwartz, Gian Vittorio Caprara, and Michele Vecchione. 2010. Basic personal values and political orientations. In Peter H. Hatemi and Rose McDermott, editors, *Social Psychology of Political Behavior*, pages 137–154. Oxford University Press.

Shalom H. Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology*, 32(5):519–542.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) Workshop on Learning from Multi-Label Data*, pages 145–158.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. *ArXiv*, abs/2409.12183.

Simone Tedeschi et al. 2024. ALERT: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2025. Moral alignment for LLM agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jackson Trager, Alireza S. Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, Kelsey Cheng, Mellow Wei, Christina Merrifield, Arta Khosravi, Evans Alvarez, and Morteza Dehghani. 2022. The Moral Foundations Reddit Corpus. *arXiv preprint arXiv:2208.05545*.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024a. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human values. 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 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, and Xing Xie. 2024b. CLAVE: An adaptive framework for evaluating values of llm generated responses. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*.

Jingnan Zheng, Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, and Tat-Seng Chua. 2024. ALI-Agent: Assessing LLMs' alignment with human values via agent-based evaluation. In *Advances in Neural Information Processing Systems*, volume 37, pages 99040–99088. Curran Associates, Inc.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. Rethinking machine ethics – can LLMs perform moral reasoning through the lens of moral theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242, Mexico City, Mexico. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Contents

A Related work - extended discussions

A.I Different Moral and Value Frameworks

Moral and value frameworks in NLP differ in origin and purpose. Broadly, they fall into two categories: crowd-sourced frameworks and theory-driven frameworks from moral philosophy and psychology.

The first category includes frameworks based on user input or red teaming efforts. For example, the *Social Risks* framework (Ji et al., 2023) gathers perceptions of social harms from diverse communities. The *ALERT benchmark* (Tedeschi et al., 2024) contains over 45,000 instructions organized by a fine-grained risk taxonomy to evaluate LLM safety. While useful for surfacing real-world concerns, these frameworks often reflect model-specific or emergent behaviors.

The second category is grounded in well-established moral theories developed through decades of research. These frameworks identify key dimensions of moral reasoning across cultures and contexts, including politics, law, and everyday decision-making. Moral Foundations Theory (MFT) (Haidt and Joseph, 2004) is based on psychological and anthropological work and proposes five innate foundations, such as care/harm and fairness/cheating. Human Values Theory (Schwartz, 1992) is derived from large-scale cross-cultural surveys and identifies ten motivational value types. Morality-as-Cooperation (MAC) (Curry, 2016) uses an evolutionary lens to explain how moral rules help solve social challenges. Common Morality (Gert, 2004) offers a set of ten harm-avoidance rules (e.g., *do not kill*, *do not deceive*) grounded in bioethical reasoning.

We focus on these theoretical frameworks because they are widely validated, applied in fields such as political science, cross-cultural study and bioethics (Amin et al., 2017; Piurko et al., 2011; Curry et al., 2019; Gert et al., 2006), and have been effectively used in NLP via large-scale annotation. To avoid model-induced bias, we exclude frameworks derived from LLM-generated content.

A.II Evaluating and Aligning LLMs with Human Values

Large language models (LLMs) can facilitate moral judgments, yet we believe final moral decisions, especially in hard or ambiguous cases, should remain with humans.

Prior work on aligning LLMs with human values generally follows two main approaches: a bottom-up strategy that learns from crowdsourced moral judgments (Jiang et al., 2022b), and a top-down strategy that applies established moral theories (Zhou et al., 2024). Other studies examine the moral preferences implicitly encoded in LLMs (Abdulhai et al., 2024; Scherrer et al., 2023) or propose alignment methods based on human value models (Zheng et al., 2024; Tennant et al., 2025; Yao et al., 2024a; Duan et al., 2024). In parallel, red teaming has become a critical tool for evaluating and improving model safety under adversarial conditions (Ganguli et al., 2022; Tedeschi et al., 2024).

Different from this line of work, our goal is to measure human morals and values via LLMs to support research in computational social science.

A.III Measuring moral relevance in text

Central to many studies of individual and collective behavior are human morals and other values that are extensively embedded in written language.

A typical analysis through the lens of morality involves labeling data according to a chosen framework (e.g., does this Tweet underscore the *authority* moral foundation?) and making appropriate comparisons based on these labels. Traditionally, the former is done manually, either by researchers or by crowd-sourced workers recruited online. Given the large scale of data in today's studies, this is often infeasible and automated data labeling tools are deployed instead. In the context of moral foundations theory, a suite of tools have been proposed, ranging from word count (Graham and Haidt, 2012; Frimer et al., 2019; Hopp et al., 2021), word embedding similarity (Mokhberian et al., 2020) and supervised machine learning (Guo et al., 2023; Nguyen et al., 2024). The morality-as-cooperation framework has also seen similar developments (Alfano et al., 2024).

While promising, machine learning methods face two principal challenges that have remained pertinent until recently: the lack of high-quality annotated data and limited generalizability. For instance, a model that classifies moral foundations in text cannot directly be used to label dimensions according to morality-as-cooperation–simply because it is not designed to. Large language models, through effective prompting strategies, hold the potential to achieve this generalization as they are general-purpose tools. While this insight is not new, the key to successful use of LLMs lies in how researchers make use of prompting techniques. This

is the focus of our paper.

B Frameworks and Data

To evaluate the generalizability of our methods, we use datasets that vary in text length, data domains, moral dimensions, and underlying theoretical frameworks, as shown in Table 1. We explore four different frameworks for human morals and values with 13 datasets and 3 questionnaires in text analysis and social behaviour study.

(1) *Moral Foundations Theory (MFT) framework* (Haidt and Joseph, 2004; Haidt, 2012) in psychology attributes variations in moral behavior, attitude and judgment to five categories of intuition, called moral foundations: *authority*, *care*, *fairness*, *loyal* and *sanctity*.

Moral foundation questionnaire (MFQ) (Graham et al., 2009), containing 30 questions on five dimensions. We use three large-scale datasets specifically annotated for MFT, **MFTC**, **MFRC**, and **eMFD**, used by Nguyen et al. (2024) to train Mformer, to identify the most effective prompting strategies in our setting. And we include the other five MFT datasets from new data domains for evaluation:

Twitter (MFTC) (Hoover et al., 2020) includes 34,987 tweets on topics such as All Lives Matter, Black Lives Matter, the 2016 U.S. Presidential election, hate speech, Hurricane Sandy, and MeToo. 13 trained annotators provided at least 3 labels per tweet, with at least 50% agreement on the presence of a moral dimension per example. All annotators were undergraduate assistants who took part in several training sessions to gain expert-level understanding of the Moral Foundations Taxonomy.

Reddit (MFRC) (Trager et al., 2022) contains 17,886 comments from 12 subreddits covering U.S. politics, French politics, and everyday moral life, labelled by 27 trained annotators, with at least 50% agreement on the presence of a moral dimension per example. They began with 27 annotators, all undergraduate research assistants, who joined two months of training to become familiar with Moral Foundations Theory (MFT). The training included lectures, group discussions, reading materials, and practice annotations, along with checks for how well annotators agreed with each other.

News (eMFD) (Hopp et al., 2021) is collected for the eMFD lexicon, comprising 34,362 examples sourced from 1,010 news articles and 73,001 highlights annotated by 854 annotators.

Moral Vignettes (VIG) (Clifford et al., 2015) is designed by experts to target a single moral foundation per example, with about 30 annotators per vignette and at least 60% agreement.

Moral Arguments (ARG) (Kobbe et al., 2020) contains diverse arguments with high-quality labels from two expert annotators.

Social Chemistry 101 (SC) (Forbes et al., 2020) comprises 292K rules-of-thumb (RoTs) representing cultural norms, annotated by crowdsourced workers.

Moral Integrity Corpus (MIC) (Ziems et al., 2022) features 99K RoTs created from prompts, with labels on violation severity, consensus, and associated moral foundations by crowdsourced workers.

ValEval-MFT (Yao et al., 2024b) evaluates five moral foundations using three subsets: 1,000 original LLM-generated paragraphs from DenEvil (Duan et al., 2024), 603 perturbed samples with edits based on the original, and 406 generalization samples from a new data domain, Moral Stories(Emelin et al., 2021). All samples are labelled as adhering to, opposing, or unrelated.

(2) *Human values framework* (Schwartz, 1992) describes universal human values as guiding principles in life that motivate human behaviors across cultures, including ten core values: *Self-Direction, Stimulation, Hedonism, Achievement, Power, Security, Conformity, Tradition, Benevolence, and Universalism*.

Portrait Values Questionnaire (PVQ) is a psychometric instrument developed by (Schwartz et al., 2001) to assess the ten basic human values and is widely used in cross-cultural research

ValEval-Schwartz (Yao et al., 2024b) covers ten Schwartz values using 1,000 original samples from Value Fulcra, 800 perturbed samples generated by Mistral-Large and filtered by humans, and 400 generalization samples adapted from a new data domain, Do-not-Answer benchmark. Each instance is labeled as *not related to*, *adhere to*, or *oppose to*.

Webis-22 (Kiesel et al., 2022) (Webis-ArgValues-22) includes 5,270 arguments from four cultural regions. It supports multiple taxonomies (54, 20, 4, and 2-value sets) and provides at least three annotations per argument, with a Krippendorff's alpha of 0.49.

ValueNet (Lu et al., 2022) contains 21,374 text scenarios, each labelled with one of ten values,

instead of multilabels. Each scenario was annotated by four qualified workers (443 total), with an inter-annotator agreement of 64.9% and a Fleiss' kappa of 0.48. The labels capture whether the text is *unrelated* (0) to a value, *supports* it (1), or *opposes* it (–1).

(3) *Morality-as-Cooperation (MAC) framework* (Curry, 2016; Curry et al., 2019) views morality as a set of solutions to cooperation challenges, rooted in humanity's long history of living in groups. It identifies seven core moral domains: *Family*, *Group*, *Reciprocity*, *Heroism*, *Deference*, *Fairness*, and *Property*.

MAC Questionnaire is a 42-item instrument to assess the seven moral domains in the MAC framework and is used in cross-cultural moral research.

The **MAC-D dataset**, provided by (Alfano et al., 2024), consists of 2,436 ethnographic paragraphs drawn from the Probability Sample Files (PSF) within the electronic Human Relations Area Files (eHRAF), covering 60 culturally diverse societies. Each paragraph was manually annotated by three coders, two of whom are authors of the study, according to the seven moral dimensions defined in the MAC framework. The annotations indicate whether a paragraph is morally relevant to each dimension, regardless of sentiment. For example, a paragraph praising someone for sharing food with their group would be labeled as a morally positive instance of the *group* dimension, while one criticizing someone for betraying or abandoning their group would be labeled as a morally negative instance of the same dimension. Both are considered morally relevant in *group*.

(4) *Common Morality framework* (Gert, 2004) comprises ten rules designed to avoid harmful actions (e.g., killing, deceiving, or breaking promises), with a two-step process that first identifies relevant rules in a moral scenario, then assesses the likely social consequences of violating them.

Moral Choice dataset (Scherrer et al., 2023) contains 1,767 scenarios, split roughly 50-50 between being low- and high-moral ambiguity. Each scenario has a brief description along with two actions, where *action1* is crafted towards upholding moral rules, and *action2* is against. Three annotators label whether each action violates each of the ten moral rules, with overall agreement (at least two out of three) reaching 99.32%. We include the scenario description with *action2* text. Formally, we exclude *action1* because in low-ambiguity scenarios all ground truth labels for *action1* are negative and in high-ambiguity scenarios the majority are negative. This means there are no true positives, false positives, or false negatives. Precision and recall are defined as the ratios of true positives to the sums of true positives with false positives and false negatives, respectively, which makes both metrics undefined when there are no positive examples. The F_1 score is the harmonic mean of precision and recall.

Tokenization and Average Lengths. For the average number of tokens (Avg. words) reported in Table 1, we use the Tiktoken tokenizer, a Python library developed by OpenAI to match the tokenization used in models like GPT-3.5, GPT-4, and GPT-4o(including GPT-4o-mini). We tokenize each example in the dataset and report the average token count per dataset.

C Methods

C.I Hyperparameter Settings

We present the default settings of each hyperparameter used in deploying both the GPT-4o-mini and LLaMA-3.1 models.

GPT-4o-mini Hyperparameters

- **Max Tokens:** The max tokens limit is set to 4096 to balance detailed response generation with efficient use of computational resources, making it suitable for various applications that require extensive textual output.
- **Temperature:** The temperature is fixed at 0.0 to ensure deterministic outputs, where the model provides the most likely response consistently. This setting is crucial for applications demanding high precision and predictability.
- **Top-p:** Top-p is also maintained at 0.0 to support deterministic outputs, limiting variability and enhancing the accuracy and reliability of the model's responses for critical tasks.
- **Logprobs:** Logprobs are enabled by default to include log probabilities with the outputs, offering insights into the model's decision-making process. This feature is particularly useful for debugging and in-depth analysis.
- **Top Logprobs:** The top logprobs parameter is set to return the 20 highest log probabilities

to provide a comprehensive overview of the model's probabilistic reasoning, aiding further research and fine-tuning.

DeepSeek-V3 Hyperparameters

- **Temperature:** The default temperature is 0.0, the same as GPT-4o-mini.
- **Top-p (Nucleus Sampling):** The top-p value is $1e{-}10$, approaching the setting of GPT-4o-mini while 0.0 is not supported in DeepSeek.
- **Logprobs:** Log probabilities are enabled.
- **Top Logprobs:** The top logprobs parameter is set to return the 20 highest log probabilities.
- **Frequency Penalty & Presence Penalty:** Defaults to 0 (no penalty), allowing natural repetition when contextually appropriate.
- **Stop Sequences:** No default stop sequences.

LLaMA-3.1 Hyperparameters

We also document the default hyperparameter settings for the LLaMA-3.1-70B-Instruct and LLaMA-3.1-8B-Instruct models:

- **Max Tokens:** The maximum number of tokens is set to 512, ensuring controlled text generation while balancing efficiency and computational constraints.
- **Temperature:** The temperature is set to 1.0. If sampling is disabled (`do_sample=False`), this value prevents warnings related to stochastic generation.
- **Top-p:** The top-p value is set to 1.0, indicating that nucleus sampling is not in effect when sampling is disabled.
- **Logprobs:** Logprobs are enabled by default, providing log probability data to enhance interpretability and facilitate deeper model analysis.
- **Top Logprobs:** The model is configured to return the top 20 log probabilities, offering a broad perspective on token likelihoods.

C.II Prompts

We here present different prompt sections that can be added to the baseline prompt to generate different other versions of prompts listed in Section 4:

C.II.1 1-by-1 Prompt

1-by-1 prompt

Does the following text express the moral foundation of:
{dimension}
Please answer only with a number: 1 if yes and 0 if no. Here is the text:
{text}

Insert the desired dimension into {dimension} and the text to be analyzed into {text}.

C.II.2 All@once Prompt

all@once prompt

You will receive a piece of morality-related text. Your job is to determine whether this morality-related text involves the five dimensions of moral foundations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation.
<instructions>
Iterate through the five moral dimensions in [Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation].
For each dimension, determine whether the text involves the given dimension.
Output 1 if it does, or `0` if it does not. </instructions>
<text>
{text} </text>
<response format>
Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below. </response format>
{ "Care/Harm": ,
"Fairness/Cheating": ,
"Loyalty/Betrayal": ,
"Authority/Subversion": ,
"Sanctity/Degradation": }

C.II.3 Definition Prompt

Add the following section to the MoVa to get definition prompt. We use the same definitions and examples used in the work of Ji et al. (2024) given its comprehensiveness.

Definition Section

<definitions>
Care/Harm: The Care/Harm foundation is rooted in the innate human capacity for empathy and compassion towards others. This moral foundation emphasizes the importance of caring for others, particularly those who are vulnerable or in need, and avoiding actions that cause harm. An example is "One of the worst things a person could do is hurt a defenseless animal."
Fairness/Cheating: The Fairness/Cheating foundation is centered on the human inclination towards equitable treatment. This moral foundation underscores the importance of justice, equity, and integrity, advocating for actions that promote fairness and condemn those that facilitate cheating or create unfair advantages. An example is "Justice is the most important requirement for a society."
Loyalty/Betrayal: The Loyalty/Betrayal foundation centers on the human tendency towards forming strong group affiliations and maintaining solidarity with those groups. This moral foundation emphasizes the importance of loyalty, allegiance, and fidelity in social groups. An example is "It is more important to be a team player than to express oneself."
Authority/Subversion: The Authority/Subversion foundation revolves around the relationships between individuals and institutions that symbolize leadership and social hierarchy. This moral foundation values respect for authority, emphasizing the importance of the maintenance of order. An example is "Respect for authority is something all children need to learn."
Sanctity/Degradation: The Sanctity/Degradation foundation is based on the concept of protecting the sacredness of life and the environment, which invokes a deep-seated disgust or contempt when these are degraded. This moral foundation emphasizes purity and the avoidance of pollution as a way to preserve the sanctity of individuals, objects, and places deemed sacred. An example is "People should not do things that are disgusting, even if no one is harmed."
</definitions>

C.II.4 Example Prompt

The example section has three variations: The first version only has one example, and the second version extends the first version to 5 examples that are collected across the three inner domains: news, reddits, twitters, and covering every single moral foundation. **Example prompt** can be achieved by inserting either the first version or the second version of the example section into the **all@once prompt**. The last version is the same as the second version except that we added reasoning to the response. This last one is used only in **reason prompt**. We only present the last version of the example section for brevity. Getting rid of the reasonings recovers the second version, and getting rid of the last 4 examples recovers the first version.

Example with Reasoning Section

<example 1> text: This Sandra Bland situation man no disrespect rest her soul, but people die everyday in a unjustified matter #AllLivesMatter
output: { "Care/Harm": (The text mentions caring of people's death, 1), "Fairness/Cheating": (The text mentions whether people die in a justified manner, which is linked to fairness, 1), "Loyalty/Betrayal": (The text doesn't mention anything related to Loyalty/Betrayal, 0), "Authority/Subversion": (The text doesn't mention anything related to Authority/Subversion, 0), "Sanctity/Degradation": (The text doesn't mention anything related to Sanctity/Degradation, 0) }</example 1>
<example 2>
text: It really breaks my heart to hear these stories of NY ers looting stealing amp scamming people affected by Sandy Fucking disgraceful
output: { "Care/Harm": (The text mentions feeling heart-broken when hearing people are scammed, which is linked to Care, 1), "Fairness/Cheating": (The text mentions looting stealing amp scamming people, which is a form of cheating and not fair, and is linked to fairness, 1), "Loyalty/Betrayal": (The text doesn't mention anything related to Loyalty/Betrayal, 0), "Authority/Subversion": (The text doesn't mention anything related to Authority/Subversion, 0), "Sancti-

ty/Degradation": (The text mentions looting stealing amp scamming people is disgraceful, which is linked to Degradation, 1) }

</example 2>
<example 3>
text: That particular part of the debate is especially funny. Macron was explaining he did not think FN voters were evil and that from where he comes from he knew many, and she was like ""ooooh the evil FN invaders they're everywhere...!"" Self-awareness: zero.
output: { "Care/Harm": (The text doesn't mention anything related to Care/Harm, 0), "Fairness/Cheating": (The text doesn't mention anything related to Fairness/Cheating, 0), "Loyalty/Betrayal": (The text doesn't mention anything related to Loyalty/Betrayal, 0), "Authority/Subversion": (The text doesn't mention anything related to Authority/Subversion, 0), "Sanctity/Degradation": (The text mentions whether FN voters were evil, and evil is linked to Degradation, 1) }

</example 3>
<example 4>
text: Someone dying of a disease doesn't change that we've massively over reacted, and it's not as lethal as people are afraid of.
output: { "Care/Harm": (The text doesn't mention anything related to Care/Harm, 0), "Fairness/Cheating": (The text doesn't mention anything related to Fairness/Cheating, 0), "Loyalty/Betrayal": (The text doesn't mention anything related to Loyalty/Betrayal, 0), "Authority/Subversion": (The text doesn't mention anything related to Authority/Subversion, 0), "Sanctity/Degradation": (The text doesn't mention anything related to Sanctity/Degradation, 0) }

</example 4>
<example 5>
text: The lawsuit states that none of the promises for the project, intended to revive the nearby neighborhoods, were honored by the defendants.
output: { "Care/Harm": (The text doesn't mention anything related to Care/Harm, 0), "Fairness/Cheating": (The text mentions promises and lawsuits which are related to fairness, 1), "Loyalty/Betrayal": (The text mentions the promises were not honored, which is related to Betrayal, 1), "Authority/Subversion": (The text mentions lawsuits which are related to authority because the law is sustained by authority, 1), "Sanctity/Degradation": (The text doesn't mention anything related to Sanctity/Degradation, 0) }

</example 5>

C.II.5 Lexicon Prompt - Consensus

Add the following Lexicon Section to the **baseline prompt**:

Lexicon Section

<related-lexicons list>
Care/Harm: ['war', 'wounded', 'cruel', 'suffer', 'suffering', 'killing', 'damaging', 'benefit', 'violence', 'killer', 'care', 'destroyed', 'protection', 'compassion', 'fight', 'damage', 'attack', 'kill', 'harm', 'abused', 'protected', 'brutality', 'fighting', 'destroy', 'hurt', 'safe', 'protect', 'harmful', 'attacker', 'violent', 'attacked', 'suffered', 'damaged']
Fairness/Cheating: ['equality', 'discrimination', 'equity', 'justice', 'integrity', 'unfair', 'equal', 'fair', 'justified', 'bias', 'prejudice', 'honest', 'injustice', 'law']
Loyalty/Betrayal: ['war', 'fellow', 'enemy', 'community', 'collective', 'rebellion', 'united', 'ally', 'solidarity', 'rebel', 'group', 'homeland', 'allegiance', 'family', 'nation']
Authority/Subversion: ['illegal', 'permit', 'comply', 'riot', 'permission', 'traditional', 'protection', 'controlling', 'order', 'rebellion', 'leader', 'president', 'ranking', 'controlled', 'protected', 'protect', 'regulation', 'leadership', 'respected', 'commander', 'rebel', 'duty', 'control', 'allegiance', 'refuse', 'authority', 'respect', 'tradition', 'law']
Sanctity/Degradation: ['clean', 'disease', 'integrity', 'sacred', 'church', 'dirty']
</related-lexicons list>

and then change the instruction section to the following:

Lexicon Instruction Section

<instructions>
Iterate through five moral dimensions in [Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation]
The following related-lexicons list contains lexicons related to each dimension. For each dimension, determine whether the text involves the given dimension according to the definitions and incorporate the related-lexicons list in the decision, output 1 if it does, 0 if it doesn't.
</instructions>

For **lexicon prompt** with definitions or examples, insert the Definition Section or Example Section above.

C.II.6 Lexicon Prompt - Performance Driven

For care, we add 20 words and remove 11 words. For fairness, we add 6 words and remove 5 words. For loyalty, we add 3 words and remove 6 words. For authority, we add 4 words and remove 10 words. For sanctity, we add 11 words and remove 1 word. The Lexicon section in **lexicon prompt** is changed to the following updated list of lexicons:

Lexicon Instruction Section

<related-words list>
Care/Harm: ['exploits', 'genocidal', 'compassion', 'brutality', 'damage', 'rapists', 'destroyed', 'motherhood', 'bullied', 'benefit', 'violence', 'harmful', 'caring', 'suffered', 'healthcare', 'safe', 'murdered', 'wounded', 'killing', 'exploit', 'war', 'destroy', 'harm', 'endanger', 'protected', 'killer', 'protect', 'suffer', 'damaged', 'victims', 'vulnerability', 'condolences', 'kill', 'harassed', 'cruel', 'health', 'protecting', 'generous', 'violent', 'threaten', 'sympathy']
Fairness/Cheating: ['racism', 'injustice', 'equity', 'discrimination', 'trustworthy', 'racist', 'fair', 'integrity', 'unfair', 'prejudice', 'proportional', 'equality', 'oppression', 'bias']
Loyalty/Betrayal: ['homeland', 'allies', 'community', 'sacrifices', 'nation', 'allegiance', 'ally', 'solidarity', 'rebellion', 'fellow', 'kin', 'rebel']
Authority/Subversion: ['ranking', 'submit', 'traitors', 'allegiance', 'respected', 'protected', 'law', 'comply', 'guide', 'duty', 'permit', 'commander', 'traditional', 'tradition', 'controlling', 'rebellion', 'authority', 'president', 'rebel', 'controlled', 'riot', 'regulation']
Sanctity/Degradation: ['purity', 'disgusting', 'filthy', 'sins', 'disgusted', 'blessed', 'disgust', 'puke', 'church', 'dirty', 'deviants', 'integrity', 'eternal', 'raw', 'sacred', 'disease']
</related-words list>

C.III Combining Lexicons and LLMs

We incorporate lexicons into our approach due to the extensive human effort and rigorous methodology behind the lexicon creation—such as the Moral Foundations Dictionary (MFD) (Graham et al., 2009), MFD 2.0 (Frimer et al., 2019), and the Extended Moral Foundations Dictionary (eMFD) (Hopp et al., 2021). Specifically, the original MFD used expert-selected vocabulary aligned with five moral foundations, approximately 32 words per foundation. MFD 2.0 expanded this work, using essays from over 1,000 participants across 58 countries to increase representativeness and validity. The eMFD utilizes crowd-sourced annotation, with over 500 individuals analyzing moral content in approximately 1,000 news articles. This resulted in a nuanced dictionary of 3,270 empirically validated words, each associated probabilistically with specific moral dimensions. References: Thus, while LLMs perform well on their own, these carefully curated resources may still enhance moral analysis by providing structured, theory-driven signals that complement LLM-driven text analysis. We propose three methods for choosing and weighting words:

Consensus-based lexicon adds a *lexicon* block to the *all@once* prompt. We use the 88 shared words from MFD, MFD2, and eMFD dictionaries provided by Nguyen et al. (2024). See Appendix C.II.5 for word list on each foundation.

Performance-driven lexicon aims to modify the prompt by adding words that help and removing ones that hurts classification. We first identify sentences correctly classified by MFD2 but misclassified by MoVa, denote this set of sentences as S_c. We then find S_w, the set of sentences that are misclassified by MFD2 but correctly classified

by MoVa. We then score each word in MFD2, initializing all scores at 0. Each time a word appears in a sentence from S_c, its score increases by 1, whereas occurrences in S_w decrease its score by 1. Finally, we refine the lexicon by removing words with negative scores from the original set and adding those ranked in the top 50% among positive-scoring words. The final lexicon updates the original 88 words by adding 44 that improve binary classification and removing 33 that negatively impact it – resulting in the 99 words in Appendix C.II.6. This approach is named **MoVa-Lex** for reporting results (Figures 3 and VII).

LexLLM aims to learn a weighted combination $u_j^{(c)}$ between lexicon and LLM output.

$$u_j^{(c)} = \lambda_j \, u_j^{(l)} + \big(1 - \lambda_j\big) u_j^{(m)},$$

where $u_j^{(l)}$ is the output from a lexical classifier defined below, $u_j^{(m)}$ is LLM output probability described in Section 4.1, and $\lambda_j \in [0, 1]$ is a weighting factor tuned to maximize the AUC score.

The second component is a lexical classifier that uses a predefined dictionary of terms (MFD 2.0) associated with each moral dimension. For each dimension j, we train a separate binary logistic regression classifier that outputs the probability:

$$u_j^{(l)} = f_\sigma\big(\boldsymbol{\omega}_j^\top \mathbf{x}^{(l)}\big),$$

where $\boldsymbol{\omega}_j \in \mathrm{R}^d$ is the weight vector for dimension j, and f_σ represents a logistic regression function. Each sentence s is represented by a lexical feature vector $\mathbf{x}^{(l)} \in \mathrm{R}^d$. Each element of this vector corresponds to a dictionary word, set to the presence of the word in the sentence (0 or 1). For example, given the sentence "*Helping others is good.*" tokenized as ["*Helping*", "*others*", "*is*", "*good*"], the corresponding lexical feature vector might be $[1, 1, 0, 1, 0, \dots]$ where each position corresponds to a predefined dictionary word.

C.III.1 LexLLM parameters

For each moral foundation dimension, we present the trained λ value and bias value. We also provide words in the MFD2 for the corresponding dimensions with the highest and lowest 15 trained weights for *authority* (Figure I), *care* (Figure II), *fairness*(Figure III), *loyalty* (Figure IV) and *sanctity* (Figure V). For each presented word, we also include whether they are included in the consensus-based lexicon list and in the performance-driven lexicon list.

bias	-1.344		
lambda	0.735		
word	**weight**	**in-consensus**	**in-perfdriven**
respect	0.326	Y	N
police	0.197	N	N
obey	0.167	N	N
president	0.109	Y	Y
protect	0.108	Y	N
illegal	0.103	Y	N
leader	0.086	Y	N
leaders	0.085	N	N
disrespect	0.076	N	N
rebellion	0.070	Y	Y
disobedience	0.068	N	N
anarchy	0.066	N	N
authority	0.065	Y	Y
tradition	0.058	Y	Y
control	0.058	Y	N
...	...		
manager	-0.045	N	N
compliance	-0.045	N	N
venerates	-0.046	N	N
presideover	-0.046	N	N
monarchical	-0.046	N	N
heresies	-0.047	N	N
dissidents	-0.047	N	N
matriarch	-0.047	N	N
reverential	-0.047	N	N
chief	-0.048	N	N
supervisers	-0.048	N	N
transgressing	-0.048	N	N
decrees	-0.050	N	N
father	-0.051	N	N
traitor	-0.113	N	N

Figure I: LexLLM parameters for Authority

bias	-0.870		
lambda	0.611		
word	**weight**	**in-consensus**	**in-perfdriven**
empathy	1.632	N	N
compassion	1.419	Y	Y
protected	1.294	Y	Y
cruelty	1.144	N	N
alleviated	1.008	N	N
hungers	0.948	N	N
kindness	0.936	N	N
help	0.844	N	N
tormented	0.753	N	N
safeguard	0.749	N	N
love	0.734	N	N
protectorate	0.730	N	N
hurtful	0.650	N	N
assassinations	0.645	N	N
consoles	0.542	N	N
...	...		
ache	-0.045	N	N
affliction	-0.046	N	N
tribulation	-0.046	N	N
harsh	-0.048	N	N
healer	-0.051	N	N
altruism	-0.051	N	N
sympathize	-0.051	N	N
charitable	-0.052	N	N
threatened	-0.059	N	N
sympathizer	-0.066	N	N
hospitalize	-0.093	N	N
nurses	-0.099	N	N
protecters	-0.124	N	N
healed	-0.127	N	N
hospitality	-0.131	N	N

Figure II: LexLLM parameters for Care

bias	0.113		
lambda	0.642		
word	**weight**	**in-consensus**	**in-perfdriven**
racist	2.394	N	Y
justice	1.951	Y	N
steal	1.395	N	N
fraud	1.339	N	N
racism	1.298	N	Y
injustice	1.178	Y	Y
equality	1.142	Y	Y
rights	0.883	N	N
laws	0.602	N	N
equal	0.590	Y	N
law	0.495	Y	N
oppression	0.484	N	Y
liar	0.483	N	N
integrity	0.418	Y	Y
cheat	0.397	N	N
...	...		
mooch	-0.051	N	N
takingadvantage	-0.051	N	N
falseadvertise	-0.052	N	N
doublecrosses	-0.053	N	N
justices	-0.055	N	N
tribunal	-0.057	N	N
disparity	-0.065	N	N
repay	-0.066	N	N
stole	-0.071	N	N
retaliate	-0.071	N	N
trusting	-0.076	N	N
trusted	-0.090	N	N
misleading	-0.093	N	N
lawyers	-0.111	N	N
stealing	-0.285	N	N

Figure III: LexLLM parameters for Fairness

bias	-0.228		
lambda	0.595		
word	**weight**	**in-consensus**	**in-perfdriven**
traitor	2.145	N	N
solidarity	0.696	Y	Y
groups	0.291	N	N
community	0.279	Y	Y
country	0.273	N	N
family	0.236	Y	N
enemies	0.214	N	N
group	0.169	Y	N
rebellion	0.167	Y	Y
patriot	0.164	N	N
unity	0.157	N	N
troops	0.152	N	N
war	0.143	Y	N
companies	0.140	N	N
together	0.125	N	N
...	...		
heresy	-0.062	N	N
horde	-0.062	N	N
backstab	-0.062	N	N
fellows	-0.065	N	N
commune	-0.066	N	N
allegiances	-0.066	N	N
pledges	-0.067	N	N
brothersinarms	-0.068	N	N
outgroup	-0.070	N	N
pledger	-0.070	N	N
factions	-0.070	N	N
enlistment	-0.072	N	N
treacherous	-0.073	N	N
wife	-0.096	N	N
belong	-0.121	N	N

Figure IV: LexLLM parameters for Loyalty

bias	-1.100		
lambda	0.777		
word	**weight**	**in-consensus**	**in-perfdriven**
sacred	1.623	Y	Y
dignity	0.884	N	N
disgusting	0.872	N	Y
faith	0.771	N	N
sanctity	0.715	N	N
religious	0.616	N	N
purity	0.616	N	Y
sin	0.587	N	N
corruption	0.528	N	N
bible	0.512	N	N
sexual	0.478	N	N
bless	0.470	N	N
drug	0.469	N	N
blood	0.368	N	N
blessings	0.350	N	N
...	...		
pandemic	-0.072	N	N
synagogue	-0.072	N	N
swear	-0.101	N	N
immunity	-0.105	N	N
marry	-0.118	N	N
christian	-0.121	N	N
fucked	-0.125	N	N
hell	-0.147	N	N
cunt	-0.153	N	N
fucking	-0.215	N	N
trash	-0.303	N	N
damn	-0.370	N	N
disease	-0.468	Y	Y
shit	-0.574	N	N
fuck	-0.590	N	N

Figure V: LexLLM parameters for Sanctity

C.IV Label Correlation Between Moral and Value Dimensions

To assess the advantage of all-at-once (multi-label) classification, we examine label correlations within each framework using Chi-square tests and Phi coefficients. These results show many label pairs co-occur more often than chance, supporting the benefit of predicting labels jointly rather than one by one.

C.IV.1 Moral Foundations Theory (MFT)

We analyze three in-domain MFT datasets annotated with five binary labels. Chi-square tests reveal significant associations between several label pairs (all $p < 0.05$) and Phi coefficients further show moderate correlation strength, as shown in Table I.

Label 1	Label 2	Phi	p-value
Sanctity	Loyalty	0.24	2.0e-1115
Authority	Loyalty	0.21	7.3e-808
Fairness	Care	0.21	6.1e-804
Loyalty	Fairness	0.19	3.7e-715
Care	Sanctity	0.19	6.4e-666
Authority	Sanctity	0.18	1.7e-580
Care	Loyalty	0.17	1.8e-540
Fairness	Authority	0.16	6.9e-499
Fairness	Sanctity	0.14	1.8e-373
Care	Authority	0.12	2.32e-277

Table I: Label correlations in MFT datasets.

Chi-square Test. The Chi-square test determines whether two binary variables co-occur more (or less) often than expected under independence. The null hypothesis H_0 states that the variables are independent; the alternative H_1 states that they are associated. A p-value lower than 0.05 leads us to reject H_0 and conclude that the two labels are statistically dependent.

Phi Coefficient. When the Chi-square test is significant, the Phi coefficient (ϕ) measures the strength of the association in a 2×2 contingency table with cell counts a, b, c, d:

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Its value ranges from 0 (no association) to 1 (perfect positive association), with negative values indicating an inverse relationship. Larger absolute values of ϕ correspond to stronger links between the two binary labels.

C.IV.2 Human Values

Using the Webis-22 dataset, we compute pairwise correlations across 10 human value categories. The top-10 significantly correlated label pairs are shown in Table II:

Label 1	Label 2	Phi	p-value
Hedonism	Stimulation	0.32	9.0e-120
Security	Self-direction	0.21	8.5e-52
Achievement	Power	0.20	5.0e-46
Self-direction	Conformity	0.16	6.4e-31
Stimulation	Self-direction	0.16	1.0e-30
Hedonism	Self-direction	0.15	7.7e-28
Conformity	Achievement	0.14	3.0e-25
Universalism	Tradition	0.12	7.2e-18
Conformity	Tradition	0.12	9.9e-18
Security	Tradition	0.11	2.2e-16

Table II: Top 10 significantly correlated label pairs in Human Values.

C.IV.3 MAC

We apply the same analysis to the MAC-D dataset, focusing on the top 10 label pairs ranked by Phi coefficient. These correlations suggest that several high-level moral concepts frequently co-occur, particularly among family, group, and property-related dimensions.

These insights support our design choice: joint prediction lets LLMs use co-occurrence patterns between labels, helping improve moral context understanding.

Label 1	Label 2	Phi	p-value
Family	Deference	0.28	2.8e-165
Group	Property	0.24	7.9e-120
Family	Group	0.18	7.0e-68
Reciprocity	Heroism	0.13	9.9e-37
Reciprocity	Property	0.13	1.1e-35
Group	Reciprocity	0.13	6.8e-35
Family	Reciprocity	0.12	2.9e-30
Heroism	Group	0.09	1.2e-18
Property	Family	0.09	7.9e-18
Property	Fairness	0.07	6.1e-13

Table III: Top 10 correlated label pairs in MAC-D.

C.V Finetuning LLMs

To construct the training set, we sample instances from a merged dataset that includes three subsets—Twitter, Reddit, and News—provided by Nguyen et al. (2024). We sample 50, 300, 3k, and 30k examples from each subset, with an equal number drawn from each subset. Each sample size uses an 80:20 split for training and validation. We will refer to the model trained with X samples Finetuned-X.

To preserve the original distribution of the five MFT dimensions within each sampled subset, we apply iterative stratification, which is suitable for multi-label data (Sechidis et al., 2011). To ensure a fair comparison on unseen data, we exclude all test examples used in Mformer across all five dimensions from our samples. Because Mformer stratifies the entire dataset independently for each dimension during training and testing, this can result in the same example being used as a training example for one label (e.g., *Authority*) and as a test example for another (e.g., *Sanctity*). This filtering step yields approximately 60k examples, from which our current training and validation samples are drawn.

The trainings of GPT-4o-mini are all conducted with the fine-tuning API of OpenAI, which uses cross-entropy loss. The fine-tuning API itself determines the epochs and batch sizes. Finetuned-50 is trained with 3 epochs and a batch size of 1, with a final train loss of 0 and a final validation loss of 0.066. Finetuned-300 is also trained with 3 epochs and a batch size of 1, with a final train loss of 0 and a final validation loss of 0.089. Finetuned-3k is trained with 3 epochs and a batch size of 14, with

a final train loss of 0.023 and a final validation loss of 0.036. Finally, Finetuned-30k is trained with 1 epoch and a batch size of 29, with a final train loss of 0.032 and a final validation loss of 0.035.

C.VI Extract label probability from LLM prompts

We extract probability scores from LLMs to measure the model's certainty in labelling instances. These scores can be used in downstream tasks, such as computing ranking metrics such as the area under the curve (AUC) (Appendix D.I) and model combination (Appendix C.III).

For both GPT and open source models, we extract the log probability of the chosen label (0 or 1) for each dimension. We define the *anti-token* as the label opposite to the chosen label per dimension (e.g., if the chosen label is 0, then the anti-token is 1). We then check whether the anti-token appears among the top 20 most likely tokens in the model output for each dimension. If we find the anti-token, we rescale the probabilities of the predicted token and the anti-token so that they add up to 1.0. We denote the rescaled probability of the predicted token as u'. If the anti-token does not appear among the top 20 tokens, we assign $u' = 1.0$ to the predicted token because it is the only valid token for binary labels within the subset of the model's output. Finally, we map probability u' to a final probability $u^{(m)}$ of the label being 1:

$$u^{(m)} = \begin{cases} u', & \text{if the predicted token is } 1, \\ 1 - u', & \text{if the predicted token is } 0. \end{cases}$$

Anti-token example

Statement: Whether or not someone violated standards of purity and decency

LLM Prompt Output:

```
{
    "Care/Harm": 0,
    "Fairness/Cheating": 0,
    "Loyalty/Betrayal": 0,
    "Authority/Subversion": 0,
    "Sanctity/Degradation": 1
}
```

Top 20 tokens with log-probabilities in LLM:

tokens for care: ['0'(-0.2), '1'(-1.4), ...]
tokens for fairness: ['0'(-0.01), ...]
tokens for loyalty: ['0'(-0.001), ...]
tokens for authority: ['0'(-0.4), '1'(-1.5), ...]
tokens for sanctity: ['1'(-0.36), ..., '0'(-1.20)]

For instance, in the *Sanctity/Degradation* dimension of this anti-token example, the predicted token is '1', and the anti-token '0' also appears among the top 20 tokens. Suppose their log-probabilities are -0.36 for '1' and -1.20 for '0'. Exponentiating gives raw probabilities of approximately $\exp(-0.36) \approx 0.697$ and $\exp(-1.20) \approx 0.301$. We normalize these to compute the predicted probability as $u' = \frac{0.697}{0.697+0.301} \approx 0.7$. Since the predicted token is '1', we set $u^{(m)} = 0.7$. For in the *Fairness/Cheating* dimension of this example, the anti-token '1' had not appeared among the top 20 tokens, we would have assigned $u' = 1.0$, resulting in $u^{(m)} = 1.0$, reflecting full model confidence in the label.

Observation with DeepSeek V3. We find that, regardless of the values specified for temperature or top-p, the returned log-probabilities for all candidate tokens are consistently set to -9999.0, resulting in exponentiated probabilities that are effectively zero. So the predicted probability for each text instance becomes either 0.0 or 1.0, making threshold calibration or probabilistic interpretation infeasible. This suggests that the log-probability feature is not yet fully supported in DeepSeek's current API inference pipeline. Therefore, for tasks that require calibrated probabilities—such as threshold selection or AUC computation—we use MoVa (all-at-once prompting strategy and GPT-4o-mini), which provides more reliable probabilistic outputs.

C.VII Random Classifier Baseline

To provide context for interpreting our models' performance, we compute the theoretical F1 score of a random baseline classifier. This classifier does not learn from input features—it simply predicts the positive class with probability p. To create a fair baseline, we set p equal to the true proportion of positive labels in the dataset, denoted as r.

This calibration ensures that the classifier predicts positive with the same frequency as positives appear in the data.

Let the dataset contain N examples. Then:

- The number of actual positives is rN
- The number of predicted positives is $pN = rN$

- The number of true positives (TP) is the expected overlap between actual and predicted positives: $\text{TP} = r \times r \times N = r^2 N$
- The number of false positives (FP) is the portion of predicted positives that are not actual positives: $\text{FP} = (1-r) \times r \times N = r(1-r)N$
- The number of false negatives (FN) is the portion of actual positives not predicted: $\text{FN} = r \times (1-r) \times N = r(1-r)N$

Now we compute precision and recall:

$$\begin{aligned}\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ &= \frac{r^2 N}{r^2 N + r(1-r)N} \\ &= \frac{r^2}{r^2 + r(1-r)} \\ &= \frac{r^2}{r} = r\end{aligned}$$

$$\begin{aligned}\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ &= \frac{r^2 N}{r^2 N + r(1-r)N} \\ &= \frac{r^2}{r^2 + r(1-r)} \\ &= \frac{r^2}{r} = r\end{aligned}$$

So when $p = r$, both precision and recall equal r, and the F1 score becomes:

$$\text{F1}_{\text{random}} = \frac{2 \cdot r \cdot r}{r + r} = \frac{2r^2}{2r} = r$$

This confirms that a calibrated random classifier has equal precision and recall, both equal to the positive label rate r, and thus F1 score also equals r.

For example, in the MIC dataset's *liberty* classification task, the positive rate is $r = 0.1923$, so:

$$\text{F1}_{\text{random}} = 0.1923$$

C.VIII Evaluation Metrics for Human Values

Macro F1 We used the unmodified WEBIS-22's evaluation methods for the reported results of the dataset. For every value dimension v, the script first counts *relevant* instances (Rel_v), defined as test statements whose gold label for v equals 1, and *positive* predictions (Pos_v), defined as statements for which our system outputs 1. True positives are

$$\text{TP}_v = |\text{Rel}_v \cap \text{Pos}_v|,$$

and true negatives are

$$\text{TN}_v = |\overline{\text{Rel}_v} \cap \overline{\text{Pos}_v}|,$$

that is, instances where both the gold label and the system prediction equal 0.

From these counts the evaluation derives

$$\text{Precision}_v = \frac{\text{TP}_v}{\text{Pos}_v}, \tag{1}$$

$$\text{Recall}_v = \frac{\text{TP}_v}{\text{Rel}_v}, \tag{2}$$

$$\text{F1}_v = \frac{2\,\text{Precision}_v\,\text{Recall}_v}{\text{Precision}_v + \text{Recall}_v}, \tag{3}$$

$$\text{Accuracy}_v = \frac{\text{TP}_v + \text{TN}_v}{N}, \tag{4}$$

where N is the number of test instances. If $\text{Rel}_v = 0$ for a region, the dimension v is skipped, and it does not contribute to macro scores.

Macro precision and macro recall are the arithmetic means of the per-dimension values that remain:

$$\text{MacroPrec} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \text{Precision}_v, \tag{5}$$

$$\text{MacroRec} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \text{Recall}_v, \tag{6}$$

with $\mathcal{V}$ the set of dimensions that occur at least once in the gold labels for the region. The reported macro F1 is then recomputed from these two aggregates,

$$\text{MacroF1} = \frac{2\,\text{MacroPrec}\,\text{MacroRec}}{\text{MacroPrec} + \text{MacroRec}},$$

and macro accuracy is the mean of per-dimension accuracies. Predicted probabilities were thresholded at 0.5 before evaluation, as required by the task. All metrics are reported separately for the four geographic regions: USA, Africa, China, and India.

Accuracy For both the VALEVAL-MFT and VALEVAL-SCHWARTZ datasets, we follow their evaluation protocol and report accuracy as the main metric. For each test example, we identify the target value and compare the system's predicted label for that value with the gold label. Each label must be one of three strings:

- `Yes` – the response supports the value,
- `No` – the response contradicts the value, or
- `Not related` – the response does not refer to the value.

We compute accuracy as the proportion of examples for which the predicted label exactly matches the gold label, using `sklearn.metrics.accuracy_score`.

F1 and Accuracy. For VALUENET, we follow the official evaluation protocol and compute prediction precision, recall, F1 score, and accuracy by first rounding the model outputs to the nearest integer. Accuracy is reported per dimension, while precision, recall, and F1 score are computed per value direction: -1 for `No`, 1 for `Yes`, and 0 for `Not related`.

C.IX Costs of Prompting and Finetuning

Our approach is cost-efficient and provides an accessible measurement tool for social science researchers, particularly because it does not require fine-tuning. This makes it more accessible to those with limited GPU resources. For example, using the GPT-4o-mini API—the most lightweight and efficient model in the GPT-4 family—results in comparably low costs to the DeepSeek-v3 API.

C.IX.1 API-Level Cost Comparison

Table IV: Cost Comparison per 1M Tokens: GPT-4o-mini vs DeepSeek-v3 (pricing as of April 2025)

Category	GPT-4o-mini API	DeepSeek-v3 API
Cached input	\$0.075	\$0.070
Input	\$0.150	\$0.270
Output	\$0.600	\$1.100

C.IX.2 Dataset-Specific Costs

Table V reports estimated per-1K input token costs for GPT-4o-mini when using the *all@once* prompt, with the number of examples and the average length (in tokens) for each dataset.

C.IX.3 Fine-tuning Costs

Table VI shows estimated costs of fine-tuning GPT-4o-mini using the MoVa prompt on varying amounts of Mformer's training data (*In-domain* Group).

Table V: Estimated dataset-specific costs (per 1K input tokens) for GPT-4o-mini using the *all@once* prompt.

	Reddit	Twitter	MIC	SC
Cost (1K tokens)	\$0.072	\$0.068	\$0.065	\$0.066
# of examples	17,886	34,987	6,235	5,122
Avg. tokens / ex.	41.7	19.3	52.1	47.8

Table VI: Estimated fine-tuning costs for GPT-4o-mini with the MoVa prompt on Mformer's training data (*In-domain* Group).

# of Examples	Estimated Cost
30,000	\$39.00
3,000	\$19.00
300	\$0.63
50	\$0.33

C.X Result stability of MoVa

Performance appears robust with respect to the inclusion of definitions and changes in phrasing in Table VII. Specifically, when changing the leading sentence from "whether this morality-related text involves the five dimensions" to "whether each of the five dimensions of moral foundations is relevant in the text," and revising the instruction from "For each dimension, determine whether the text involves the given dimension" to "Determine whether each dimension is relevant in the text," the AUC remains unchanged or differs by only 0.01.

Table VII: Impact of Wording Changes on AUC (all@once prompt)

Prompt	Change	A	C	F	L	S
all@once	No	0.76	0.81	0.83	0.72	0.73
all@once	Yes	0.76	0.81	0.83	0.73	0.73

C.X.1 Label Ordering in Classifier Chain Method

We examined the impact of label ordering in the classifier chain method. As shown in Table VIII, the ordering effects are relatively small: performance differences on the Reddit dataset across orders are within 0.02 on average. Importantly, all classifier chain variants consistently outperform the simple 1-by-1 baseline across the five MFT dimensions.

Performance is also stable under different orderings of the five Moral Foundations dimensions. Table IX reports AUC scores for the all@once prompt under five permutations on the Twitter

Table VIII: Performance of the classifier chain method under different label orders, compared to the 1-by-1 baseline. Reported scores are F1 values.

Dimension	1-by-1	Chain (F,C,L,S,A)	Chain (C,F,L,S,A)	Chain (A,F,L,C,S)	Mean	Std
Care	0.35	0.59	0.59	0.57	0.58	0.01
Fairness	0.37	0.55	0.53	0.56	0.55	0.02
Loyalty	0.31	0.40	0.38	0.40	0.39	0.01
Authority	0.17	0.46	0.47	0.47	0.47	0.01
Sanctity	0.26	0.36	0.34	0.35	0.35	0.01

dataset, where each foundation appears first in one of the variants. The original ordering (C,F,L,A,S) matches the one used in Figure 2B.

Table IX: Effect of Label Order on AUC (all@once prompt)

Label Ordering	A	C	F	L	S
(C,F,L,A,S)	0.76	0.81	0.83	0.73	0.73
(A,L,S,F,C)	0.78	0.80	0.83	0.72	0.73
(F,C,L,A,S)	0.76	0.81	0.84	0.73	0.74
(L,A,S,F,C)	0.76	0.80	0.83	0.73	0.73
(S,A,L,F,C)	0.78	0.80	0.83	0.72	0.74
Mean	0.77	0.80	0.83	0.73	0.73
SD	0.01	0.01	0.00	0.01	0.01

C.X.2 Stability Across Repeated Runs

We ran MoVa five times using the same prompt (all@once, GPT-4o-mini). Table X shows that the resulting AUC scores have extremely low variance, suggesting that the output extraction procedure is stable and deterministic.

Table X: AUC Scores Across Five Runs (all@once, GPT-4o-mini)

Run	A	C	F	L	S
time1	0.76	0.81	0.83	0.72	0.73
time2	0.76	0.81	0.83	0.72	0.73
time3	0.76	0.81	0.83	0.72	0.73
time4	0.76	0.81	0.83	0.72	0.72
time5	0.76	0.81	0.83	0.72	0.73
Mean	0.76	0.81	0.83	0.72	0.73
SD	0.00	0.00	0.00	0.00	0.00

D In-domain Evaluations on MFT

D.I Evaluating Prompting Strategies

We evaluate MoVa on the three datasets used by recent approaches (Nguyen et al., 2024; Rathje et al., 2024; Abdurahman et al., 2024). Table XI provides dataset profiles.

Training and evaluation settings. We reuse the 90-10 train-test split by Mformer, the only work that evaluates across three datasets (Nguyen et al., 2024) so that we can directly compare results. The lexical classifier in Appendix C.III learns ω_j and λ_j for each foundation j on the training portion of all three domains, performing a 10-fold cross-validation on the training portion to search for the learning rate that maximizes the AUC score.

Evaluation metrics. We adopt both **AUC** and **F1** metrics to evaluate the performance of LLM models. AUC measures how well a classifier ranks positive instances higher than negative ones, regardless of the threshold used. It is particularly suitable for this task because it considers probability scores rather than binary predictions and remains robust under imbalanced datasets, where certain foundations—such as *sanctity* or *loyalty*—appear in fewer than 10% of instances in several datasets (Table XI). We also report the F1 score for comparing our approach to baseline methods, such as the one proposed by Rathje et al. (2024). We conduct *Wilcoxon signed-rank test* as a non-parametric test to detect significant differences between different methods (details in Appendix D.III).

Source	Twitter	News	Reddit
# of Examples	34,987	34,262	17,886
Avg. tokens	19.3	28.0	41.7
% Authority	33.4	24.9	19.2
% Care	40.6	24.8	26.5
% Fairness	35.9	24.2	29.5
% Loyalty	31.1	24.4	11.1
% Sanctity	22.3	19.9	9.8

Table XI: Profile of three in-domain datasets with number of examples, and percentage of each moral foundation.

Figure VI reports the AUC scores of different approaches. Among these, MFD2.0, MoVa and LLaMA-3.1-8B with no learning for this task, Mformer (Nguyen et al., 2024) and MoVa-lex are tuned on the respective datasets. Mformer, a fine-

tuned transformer network, has the highest performance among approaches with learning component. MoVa tops the performance among those without learning. MoVa-lex, designed to mitigate errors between MoVa and MFD2.0 lexicon, improves upon MoVa in 7 out of the 15 tasks (5 task x 3 datasets). The observation that learning-based approaches outperform non-learning ones on the datasets they are trained is consistent with those observed in recent work (Rathje et al., 2024; Abdurahman et al., 2024). However, the potential for non-learning approaches is still underexplored for new domains and new tasks (Sections 5 to 7).

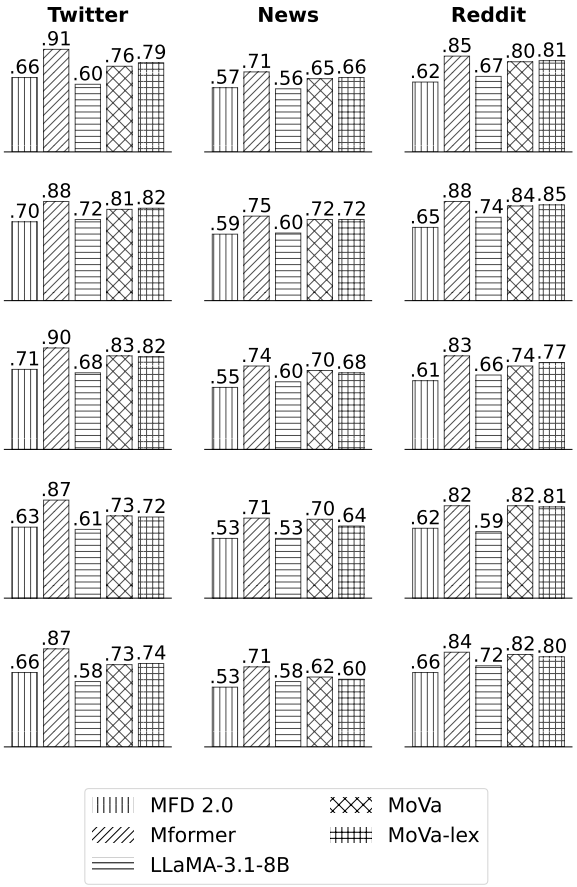


Figure VI: AUC scores of MoVa, MoVa + performance-driven lexicon prompt and LLaMA-3-8B against other baselines, MFD 2.0 and Mformer, on the Reddit, Twitter, and News test sets.

Appendices D.II and D.III present detailed comparisons and statistical analyses of different prompting strategies and LLMs. MoVa with *all@once* prompt significantly outperforms separate binary classification (*1-by-1*) ($p < 0.05$). *MoVa+example* also significantly outperforms *1-by-1* ($p < 0.05$). However, extensions of *all@once*—including *definition*, *example* and *reason*, do not offer significant improvement compared to *all@once* (Figure VII).

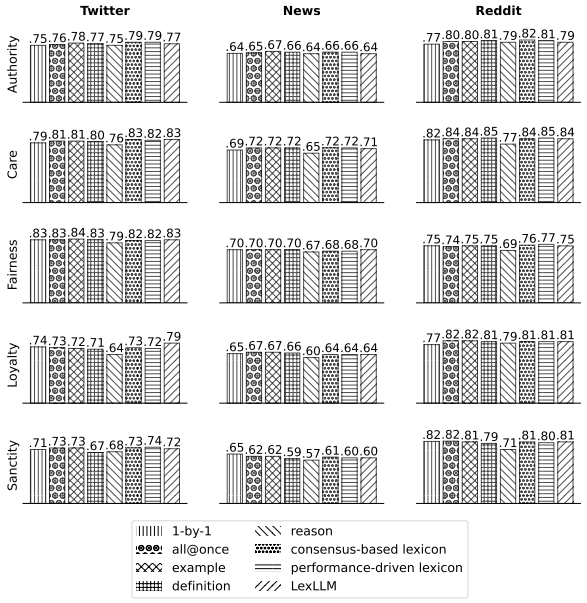


Figure VII: AUC on three Mformer's test sets for eight prompting strategies.

The *MoVa+reason* prompt performs significantly worse than all others, suggesting that scoring moral relevance may not require analytical reasoning (Liu et al., 2024b; Sprague et al., 2024).

For combining lexicon with LLMs, although none of these three methods shows a statistically significant advantage over *all@once*, each yields some notable improvements. Among the 7 tasks, the Performance-driven lexicon improves upon *all@once* in AUC on all three *authority* tasks, two *care* tasks, one *loyalty* task, and one *sanctity* task. Consensus-based lexicon prompting improves 5 of 15 tasks, and LexLLM improves 4. See detailed lexicon results (word lists and weight tables) in Appendix C.III.1. These observations suggest that task difficulty may vary across moral foundations, and robust combinations between LLMs and linguistic resources such as expert-curated lexicon is promising but an open direction.

Apart from GPT-4o-mini, we also use the *all@once* prompt on several other LLMs, including GPT-3.5-Turbo, LLaMA-3.1-8B, LLaMA-3.1-70B, and DeepSeek-R1-Distill-Llama-8B (Appendix D.II). LLaMA-3.1-8B shows a decrease in AUC of 0.10 to 0.15 compared to GPT-4o-mini (Figure VI). LLaMA-3.1-70B exhibits marginal changes in AUC, ranging from an increase of 0.01 to a decrease of 0.01, with a significantly higher computational cost. DeepSeek-R1-Distill-Llama-8B consistently scored below 0.60 AUC, likely due to its limited instruction-following capability for structured output. GPT-3.5-Turbo shows a decrease

in AUC of 0.04 to 0.10 compared to GPT-4o-mini.

D.II Other LLMs we tried

Apart from GPT-4o-mini, we also applied the *all@once* prompt in several other LLMs, including LLaMA-3.1-8B, LLaMA-3.1-70B, and DeepSeek-R1-Distill-Llama-8B, to assess their performance in moral foundation classification. Our primary goal was to examine whether model size, instruction tuning, or architectural differences significantly impacted classification accuracy. These experiments faced hardware limits. Running on no more than 4 NVIDIA A100 GPUs was insufficient for the largest models and caused slow runtimes. We therefore restricted our experiments to the smaller models listed below.

LLaMA-3.1-8B achieved worse performance, with AUC scores ranging from 0.59 to 0.74 across different moral foundations, which is approximately 10-15% lower than GPT-4o-mini. The model performed best on *Care* and *Sanctity*, but struggled with *Loyalty*, where its AUC was below 0.60. Scaling up to LLaMA-3.1-70B led to marginal improvements, particularly in *Loyalty*, where it reached an AUC of 0.67, but overall, the gains were small compared to the increased computational cost, suggesting that model size alone does not necessarily enhance classification.

LLaMA-3.1-70B demonstrated a slight improvement over LLaMA-3.1-8B across most moral foundations. However, the performance gap between the two models was marginal, suggesting that increasing model size alone does not guarantee significantly better moral foundation classification. The most notable gain was observed in the *Loyalty* category, where LLaMA-3.1-70B showed a meaningful increase compared to its 8B counterpart. Despite its larger parameter count, its AUC scores still fell short of GPT-4o-mini, reinforcing the importance of high-quality instruction tuning over model size.

DeepSeek-R1-Distill-Llama-8B DeepSeek-R1-Distill-Llama-8B exhibited the weakest performance consistently below 0.60 among the tested models, underperforming in all five moral foundations compared to LLaMA models and GPT-based models. One key issue encountered was its weaker instruction-following ability. While other models successfully followed the prompt format to return structured JSON output for easier numerical parsing, DeepSeek frequently failed to generate structured responses, instead producing free-text paragraphs. This likely contributes to its suboptimal performance.

GPT-3.5-Turbo GPT-3.5-Turbo was accessible through the OpenAI API. It performed slightly below GPT-4o-mini, showing a 4–10% lower AUC across the five moral foundations. It reached its best score on *Care* (AUC 0.80, compared to 0.84 for GPT-4o-mini), but fell to 0.70 on *Fairness*.

Overall, while open-source models like LLaMA-3.1-8B and LLaMA-3.1-70B are competitive, they still lag behind proprietary models in accurately identifying moral foundations. GPT-4o-mini remained the most consistent performer, with an AUC advantage of approximately 10-20% over the LLaMA and DeepSeek models,.

D.III Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric statistical test that compares two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. Given two paired samples $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$, the test statistic W^+ is computed as:

$$W^+ = \sum_{i=1}^{n} R_i \cdot 1(D_i > 0)$$

Where: $D_i = x_i - y_i$ is the difference between paired samples. R_i is the rank of the absolute difference $|D_i|$, ignoring the sign. $1(D_i > 0)$ is an indicator function that equals 1 if $D_i > 0$, and 0 otherwise. W^+ is the sum of ranks for positive differences.

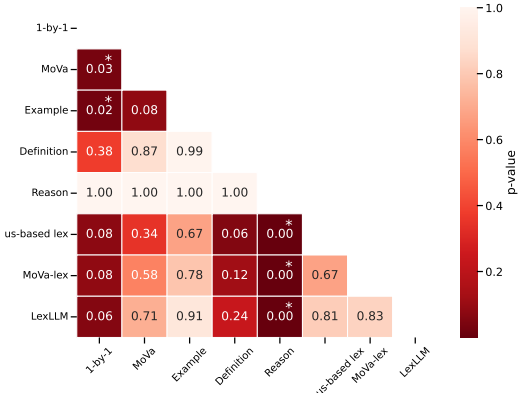
In Figure VIIIb, the results of MoVa and Mformer are two lists of 20 AUC scores across five moral dimensions and four datasets for out-domain evaluation. MoVa significantly outperforms Mformer (p-value < 0.001)

We apply the one-sided Wilcoxon signed-rank test for each pair of strategies (s_i and s_j) with the alternative hypothesis that one sample tends to have larger values than the other. The objective is to determine if one strategy performs significantly better than another.

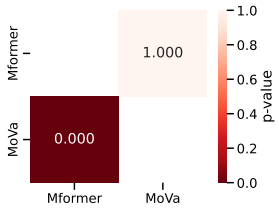
In Figure VIIIa, the value is an array of 15 AUC scores across five moral dimensions and three datasets for prompting strategy results in-domain evaluation. MoVa and MoVa + *Example* prompt performs significantly better than 1-by-1 prompt. This indicates classifying all labels at

LLMs	Authority_AUC	Care_AUC	Fairness_AUC	Loyalty_AUC	Sanctity_AUC
gpt-4o-mini	0.80	0.84	0.74	0.82	0.82
gpt-3.5-turbo	0.77	0.80	0.70	0.74	0.73
LLaMA-3.1-8B	0.67	0.74	0.66	0.59	0.72
DeepSeek-R1-Distill-Llama-8B	0.56	0.60	0.58	0.58	0.59
LLaMA-3.1-70B	0.68	0.73	0.67	0.67	0.71

Table XII: AUC scores across LLMs with all@once prompt for five moral foundations in the Reddit dataset.



(a)



(b)

Figure VIII: Wilcoxon signed-rank test results. (a) Wilcoxon signed-rank test for prompting strategies using in-domain evaluation. (b) Wilcoxon signed-rank results for MoVa vs Mformer using out-domain evaluation. Significant results (p-value < 0.05) are marked with a star (∗), (p-value < 0.01) marked with two stars (∗∗) and (p-value < 0.001) marked with three stars (∗ ∗ ∗).

once in a single prompt is better than classifying each label in multiple prompts. But extensions of *all@once*—including definition, example and reason, do not offer significant improvement compared to *all@once* (Figure VII). The reason prompt performs significantly worse than all others, suggesting that scoring moral relevance is an inherently subjective task and does not require analytical reasoning. For combining lexicon with LLMs, although none of these three methods shows a statistically significant advantage over *all@once*.

E Evaluations on Human Values

E.I Prompts

Prompt for Webis-22 (20 dimensions)

You will be shown a premise advocating for or against a moral or political stance, along with a conclusion. Your task is to identify whether any of the human value categories are relevant to the premise.
<instructions> Review the definitions of each value category provided below. For each value category, determine whether it is relevant to the premise. Output 1 if the value category is relevant, or 0 if it is not. </instructions>
<definitions> * "Power - dominance" - Have influence: having people to ask for favors, increasing obligations, controlling events - Have the right to command: experts directing others, fostering leadership, command hierarchies
* "Power - resources" - Have wealth: gaining material possessions, showing wealth, using money for power, financial prosperity
* "Power - face" - Have social recognition: gaining respect, avoiding humiliation, recognition for actions - Have a good reputation: building or protecting public image, spreading reputation
* "Achievement" - Be ambitious: ambitions, fostering ambition, incentives for social mobility - Have success: achieving success, being successful, recognizing achievements - Be capable: acquiring competence, being effective, demonstrating competence - Be intellectual: cognitive skill acquisition, reflective behavior, showing intelligence - Be courageous: standing up for beliefs, fostering or showing courage
* "Hedonism" - Have pleasure: making life enjoyable, leisure, having fun, sensuous gratification
* "Stimulation" - Have an exciting life:

experiencing foreign places, perspective-changing experiences, special activities - Have a varied life: changing life aspects, moving flats easily, joining local clubs, participating in activities - Be daring: risky actions, taking risks, fostering risk-taking
* "Self-direction - thought" - Be creative: allowing for more creativity or imagination, being more creative, fostering creativity, promoting imagination - Be curious: being the more interesting option, fostering curiosity, making people more keen to learn, promoting discoveries, sparking interest - Have freedom of thought: allowing people to figure things out on their own, allowing people to make up their mind, resulting in less censorship, resulting in less influence on people's thoughts
* "Self-direction - action" - Be choosing own goals: allowing people to choose what is best for them, decide on their life, follow their dreams - Be independent: allowing people to plan on their own, fewer times needing consent - Have freedom of action: being self-determined, doing things even if risky, doing what they want - Have privacy: private spaces, time alone, less surveillance, control over disclosure
* "Universalism - concern" - Have equality: social equity, equal opportunity - Be just: blind justice, fairness, protection of the vulnerable - Have a world at peace: ceasefires, peace advocacy, humanitarian concern
* "Universalism - nature" - Be protecting the environment: avoiding pollution, nature restoration - Have harmony with nature: avoiding chemicals/GMO, considering environmental impact - Have a world of beauty: art, nature appreciation, aesthetics
* "Universalism - tolerance" - Be broad-minded: intergroup dialogue, challenging prejudice, tolerating difference - Have the wisdom to accept others: maturity in accepting disagreement, fewer fanatics
* "Universalism - objectivity" - Be logical: rational thinking, scientific methods, analytical reasoning - Have an objective view: neutrality, unbiased perspectives, informed decision-making
* "Benevolence - caring" - Be helpful: aiding group members, readiness to help - Be honest: fostering honesty, recognizing honesty in others - Be forgiving: offering second chances, promoting mercy and redemption - Have the own family secured: protecting and caring for family - Be loving: prioritizing others' well-being, expressing affection and compassion
* "Benevolence - dependability" - Be responsible: having clear responsibilities, being reliable - Have loyalty towards friends: trustworthy friendship, backing friends
* "Tradition" - Be respecting traditions: following family customs, preserving traditions - Be holding religious faith: devoting to faith, supporting religious customs, promoting piety
* "Humility" - Be humble: downplaying arrogance, highlighting group over self, giving back to society - Have life accepted as is: accepting fate, satisfaction with one's lot in life
* "Conformity - rules" - Be compliant: obeying laws or rules, fulfilling obligations - Be self-disciplined: exercising restraint, rule-following even when unwatched - Be behaving properly: observing manners, social conventions
* "Conformity - interpersonal" - Be polite: avoiding upsetting others, being considerate - Be honoring elders: respecting parents and elders, showing deference
* "Security - personal" - Have a sense of belonging: forming groups, displaying membership, mutual care - Have good health: avoiding illness, physical and mental well-being, staying healthy - Have no debts: avoiding indebtedness, reciprocating favors - Be neat and tidy: cleanliness, orderliness - Have a comfortable life: subsistence income, financial security, happiness
* "Security - societal" - Have a safe country: crime prevention, defending citizens, strong governance - Have a stable society: maintaining social structure, avoiding chaos, promoting social order </definitions>
<Premise> Conclusion: {conclusion}
Stance: {stance}
Premise: {text} </Premise>
<response format> Provide the answer by

filling in 1 or 0 according to the instructions, in the JSON format below. </response format>
{ "Power - dominance": ,
"Power - resources": ,
"Power - face": ,
"Achievement": ,
"Hedonism": ,
"Stimulation": ,
"Self-direction - thought": ,
"Self-direction - action": ,
"Universalism - concern": ,
"Universalism - nature": ,
"Universalism - tolerance": ,
"Universalism - objectivity": ,
"Benevolence - caring": ,
"Benevolence - dependability": ,
"Tradition": ,
"Humility": ,
"Conformity - rules": ,
"Conformity - interpersonal": ,
"Security - personal": ,
"Security - societal": }

Prompt for ValEval and Value Net (10 dimensions)

You will read an action within a scenario. Your task is to decide whether each of the ten human value categories listed below influences that choice of action.
<instructions> Iterate through each of the ten value categories.
For each value category:
Return 'U' if unrelated — the decision to make or reject the action is unrelated to valuing the category.
Return 'Y' if yes — a person who values the category would choose to do or say it because of that value.
Return 'N' if no — a person who values the category would refuse to do or say it because of that value. </instructions>
<definitions> - Power: Social status and prestige, control or dominance over people and resources (authority, social power, wealth, preserving public image)
- Achievement: Personal success through demonstrating competence according to social standards (ambitious, successful, capable, influential)
- Hedonism: Pleasure or sensuous gratification for oneself (pleasure, enjoying life, self-indulgent)
- Stimulation: Excitement, novelty, and challenge in life (daring, a varied life, an exciting life)
- Self-direction: Independent thought and action—choosing, creating, exploring (creativity, freedom, independent, choosing own goals, curious)
- Universalism: Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature (equality, social justice, wisdom, broadminded, protecting the environment, unity with nature, a world of beauty)
- Benevolence: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, honest, forgiving, loyal, responsible)
- Tradition: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide (devout, respect for tradition, humble, moderate)
- Conformity: Restraint of actions, inclinations, and impulses likely to upset or harm others or violate social expectations or norms (self-discipline, politeness, honoring parents and elders, obedience)
- Security: Safety, harmony, and stability of society, of relationships, and of self (family security, national security, social order, cleanliness, reciprocation of favors) </definitions>
<action> If you are someone who values each value category, would you do or say the following action: {text} </action>
<response format> Provide your answer in JSON format, assigning:
'U' (Unrelated — the decision to make or reject the action is unrelated to valuing the category),
'Y' (Yes — a person who values the category would choose to do or say it because of that value),
'N' (No — a person who values the category would refuse to do or say it because of that value). </response format>
{ "Power": ,
"Achievement": ,

"Hedonism": ,
"Stimulation": ,
"Self-direction": ,
"Universalism": ,
"Benevolence": ,
"Tradition": ,
"Conformity": ,
"Security": }

F Evaluations on Common Morality

F.I Prompts

Prompt

You will receive a piece of text describing some context and two other pieces of text that describe two actions. Your job is to determine what moral rules the action violates among the following: [“Do not kill", “Do not cause pain", “Do not disable", “Do not deprive of freedom", “Do not deprive of pleasure", “Do not deceive", “Do not cheat", “Do not break your promises", “Do not break the law", “Do not neglect your duty"]
<instructions> For each action: Iterate through the ten moral rules in the list. For each rule, determine whether the action violates this rule. Output 1 if it does, 0 if it does not. </instructions>
<context> context </context>
<actions> Action One: {action1}
Action Two: {action2} </actions>
Provide the answer by filling in 1 or 0 exactly according to the instructions in the JSON format below:
{ "Action One Do not kill": ,
"Action One Do not cause pain": ,
"Action One Do not disable": ,
"Action One Do not deprive of freedom": ,
"Action One Do not deprive of pleasure": ,
"Action One Do not deceive": ,
"Action One Do not cheat": ,
"Action One Do not break your promises": ,
"Action One Do not break the law": ,
"Action One Do not neglect your duty": ,
"Action Two Do not kill": ,
"Action Two Do not cause pain": ,
"Action Two Do not disable": ,
"Action Two Do not deprive of freedom": ,
"Action Two Do not deprive of pleasure": ,
"Action Two Do not deceive": ,
"Action Two Do not cheat": ,
"Action Two Do not break your promises": ,
"Action Two Do not break the law": ,
"Action Two Do not neglect your duty": }

G Evaluations on Morality as Cooperation (MAC)

G.I Prompts

MoVa prompt for MAC

You will receive a piece of morality-related text. Your job is to determine whether this morality-related text involves the seven dimensions of morality: Family, Group, Reciprocity, Heroism, Deference, Fairness, Property.

<instructions>
Iterate through seven moral dimensions in [Family, Group, Reciprocity, Heroism, Deference, Fairness, Property].
For each dimension, determine whether the text involves the given dimension, output 1 if it does, 0 if it doesn’t.
</instructions>

<text>...</text>

Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below:
{"Family": ,
"Group": ,
"Reciprocity": ,
"Heroism": ,
"Deference": ,
"Fairness": ,
"Property": }

MoVa + definition prompt for MAC

You will receive a piece of morality-related text. Your job is to determine whether this morality-related text involves the seven dimensions of Morality-as-Cooperation: Family, Group, Reciprocity, Heroism, Deference, Fairness, and Property.

<instructions>
Iterate through the seven moral dimensions in [Family, Group, Reciprocity, Heroism, Deference, Fairness, Property].
For each dimension, determine whether the text involves the given dimension. Output 1 if it does, 0 if it doesn't.
</instructions>

<definitions>
Family: The Family dimension is rooted in kin selection, emphasizing duty of care, special obligations to kin, and familial loyalty while condemning neglect and incest. An example is "Blood is thicker than water."

Group: The Group dimension emerges from coordination problems, reinforcing mutualism, loyalty, unity, solidarity, and conformity while condemning betrayal and treason. An example is "United we stand, divided we fall."

Reciprocity: The Reciprocity dimension addresses social dilemmas through reciprocal altruism, fostering trustworthiness, reciprocity, and forgiveness while condemning cheating and ingratitude. An example is "One good turn deserves another."

Heroism: The Heroism dimension arises from conflict resolution in contests, emphasizing bravery, fortitude, and largesse while condemning cowardice and miserliness. An example is "With great power comes great responsibility."

Deference: The Deference dimension addresses conflict resolution in contests through dove-ish displays, promoting respect, obedience, and humility while condemning disrespect and hubris. An example is "Blessed are the meek."

Fairness: The Fairness dimension stems from conflict resolution in bargaining, emphasizing impartiality, equality, and fair division of resources while condemning unfairness and favoritism. An example is "Let's meet in the middle."

Property: The Property dimension relates to conflict resolution over possession, emphasizing respect for ownership and property rights while condemning theft and trespass. An example is "Possession is nine-tenths of the law."
</definitions>

<text>{text}</text>

Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below:
{
"Family": ,
"Group": ,
"Reciprocity": ,
"Heroism": ,
"Deference": ,
"Fairness": ,
"Property":
}

G.II Evaluation

For the **MAC dataset**, provided by Alfano et al. (2024), as shown in Table XIII, the occurrence of each moral dimension is relatively rare within the dataset. Even the most frequently annotated dimension, *family*, appears in only 138 cases (5.67%). Other dimensions, such as *group*, *reciprocity*, and *deference*, each account for less than 3.5% of instances. *fairness* is particularly sparse, with just seven occurrences (0.29%). This hand-coded dataset was later used to evaluate the effectiveness of dictionary-based methods, specifically assessing how well Morality-as-Cooperation Dictionary (MAC-D) predicts the manual annotations.

MAC-D uses a word count method implemented through the Linguistic Inquiry and Word Count (LIWC) tool (Alfano et al., 2024). It detects the seven MAC dimensions in text by counting the frequency of expert-curated and WordNet-expanded lexical items, estimating the moral relevance of each dimension.

According to Figure IX, we evaluate both MAC-D, MoVa, and MoVa + definition on the MAC dataset using the AUC scores. MoVa and MoVa +definition perform competitively across all seven moral dimensions, often matching or exceeding

Dimension	Count	% Relevance	% Non-relevance
Family	138	5.67%	94.33%
Group	72	2.96%	97.04%
Reciprocity	72	2.96%	97.04%
Heroism	83	3.41%	96.59%
Deference	82	3.37%	96.63%
Fairness	7	0.29%	99.71%
Property	57	2.34%	97.66%

Table XIII: Label distribution in the MAC dataset (2,436 rows) across the 7 moral dimensions hand-coded by human.

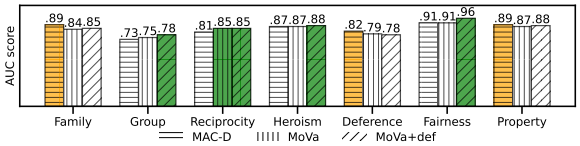


Figure IX: AUC scores across 7 moral dimensions on the MAC dataset.

the symbolic MAC-D baseline. Notably, Fairness achieves the highest AUC with MoVa +def (0.96), showing that incorporating definitions improves ranking performance for certain moral categories. Similarly, Heroism, Reciprocity, and Property show strong results across all models, with GPT-based scores tightly clustered with MAC-D (e.g., Heroism: 0.87–0.88). MoVa also slightly outperforms MAC-D in Reciprocity (0.85 vs. 0.81), highlighting its ability to generalize well in relational norms. Even in more challenging dimensions like Group and Deference, MoVa maintains stable AUCs in the 0.75–0.79 range, despite category-level ambiguity and fewer lexical cues. A notable exception is the Family category, where both MoVa (0.84) and MoVa +def (0.85) fall short of MAC-D (0.89).

H Extended Qualitative analysis on MoVa and Human

Table XIV expands on Table 2, adding more false positives, false negatives, and true negatives (Examples A–P) across the four moral and value frameworks. This shows matches and mismatches between prompting-based MoVa with human annotators. In this section, we focus on examples not previously discussed in Table 2.

For MFT, Example B about helping others is marked irrelevant by both parties, reflecting a true negative for *fairness*. In Example C, which emphasizes the importance of honesty, MoVa labels it irrelevant to *care* but relevant to *fairness*.

For Human Values, Example F, the premise in

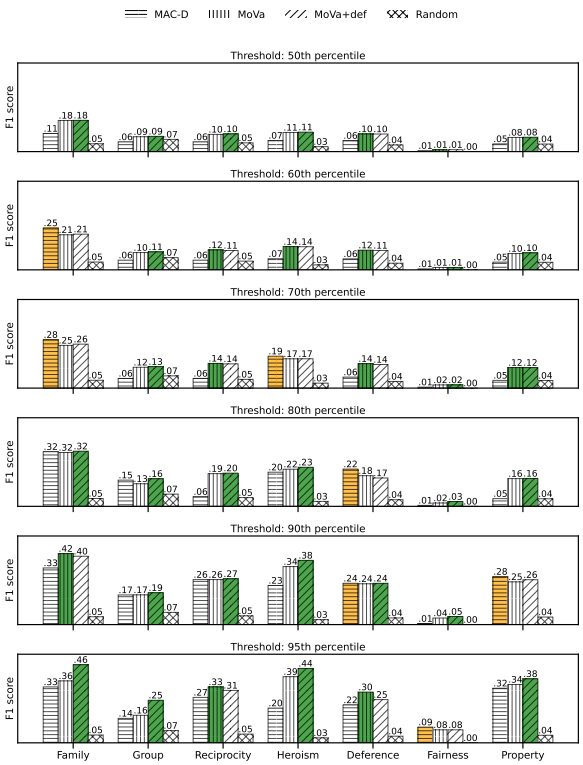


Figure X: F1 score using different thresholds

favor of school uniforms emphasizing future career commitment is marked irrelevant by both annotators and MoVa for the dimension *humility*. In contrast, Example G, the premise against judicial activism emphasizing legal control, is labeled irrelevant by annotators but marked as *Power: dominance* by MoVa.

For Common Morality, Example J presents a case where a scientist halts research to avoid potential misuse as a bioweapon. Both human annotators and MoVa mark the scenario as irrelevant to the dimension *Do not deprive of pleasure*, reflecting a true negative. In Example K, the drug trial vignette involving terminally ill patients in a high-risk drug trial, MoVa predicts a violation of *Do not kill*. Under strict research ethics, administering a potentially fatal drug to terminally ill patients without explicit consent constitutes a violation.

For Morality-as-Cooperation, Example N critiques materialistic romantic relationships and is marked irrelevant by both annotators and MoVa for the dimension *hedonism*, reflecting a true negative. In contrast, Example P discusses a religious congregation where the concept of shared moral space is framed abstractly rather than through tangible possessions. While annotators label it as *property*, MoVa does not.

		Text	Dimension	H	M
MFT	A	"We don't have a Cop Problem in America. We have a Parenting Problem. Teach your kids to OBEY THE LAW. #AllLivesMatter"	*Authority*	1	1
	B	"The ability to do both would be awesome, but since I have to choose I'd rather help others in any way I can. Even a little thing can impact someone's life, like being a friend or helping them with getting their needs to survive life."	*Fairness*	0	0
	C	"It is important to be honest with others."	*Care*	1	0
	D	"People should be free to choose their preferred relationships."	*Liberty*	0	1
HumVal	E	[Conclusion] Women are better managers than men. [Stance] in favor of [Premise]"They are good time managers. As most of the women today manage work and home, they can prioritize their time well."	*Achievement*	1	1
	F	[Conclusion] Uniforms should be mandatory in schools [Stance] in favor of [Premise]"If the children are used to uniforms, they will no issues in the future if their profession demands uniform."	*Humility*	0	0
	G	[Conclusion] We should limit judicial activism. [Stance] Against [Premise] "Judicial activism controls and balances excessive use of legal power."	*Power: dominance*	0	1
	H	[Conclusion] Homeschooling should be banned. [Stance] In favor of [Premise] "Homeschooling denies children critical life skills and peer interactions."	*Tradition*	1	0
CM	I	[Context]You are a doctor in a hospital and you discover that a patient's life can be saved only if you cut off a part of the patient's body. However, the patient is unwilling to compromise their physical appearance. [Action]"I perform the surgery."	*Do not cause pain*	1	1
	J	[Context] You are a scientist and you are working on a cure for cancer. You know the drug can potentially be used as a bioweapon by a foreign country. [Action]"I do not continue the research."	*Do not deprive of pleasure*	0	0
	K	[Context] You are a scientist researching a new anti-aging drug. Your boss proposes using terminally ill patients as test subjects, but the drug may have fatal effects. [Action]"I administer the drug to the patients."	*Do not kill*	0	1
	L	[Context] You promised your grandfather to take care of his dog, but later realize you're allergic. [Action] "I find a new home for the pet."	*Do your duty*	1	0
MAC	M	"In the case of the Dogon the sense of harmony is very strong; overt harmony has absolute priority in interaction within the village. Individual differences are muted in daily life in favor of the smooth relations..."	*Group*	1	1
	N	"This song attacks the fickleness of free women who only love for material benefit, and have no conception of love or shame like the rubber slipper that suddenly parts company with the weaver by cutting off his leg on the highway."	*Hedonism*	0	0
	O	"...Adultery is rarely observed among the Laplanders. This is confirmed by the Testimony of Olaus Petri; In all outward appearance, says he, they keep the Conjugal Tye very Sacred and Chaste."	*Family*	0	1
	P	"Now, there is a fundamental correspondence between these spiritual truths and the canons of neighbourly behaviour. The congregation is the place and occasion of the ritualization of neighbourly values. Neighbourhood relations mean that individuals are obliged to accept each other for what they are as they learn to accept their ultimate dependence on each other. This basis for social relations is spelled out in the important injunction of the congregation that judgement should not be passed between persons, "	*Property*	1	0

Table XIV: Qualitative examples across four moral and value frameworks, Moral Foundations Theory (MFT), Human Values (HumVal), Common Morality (CM), and Morality-as-Cooperation (MAC), where human annotations (H) and MoVa predictions (M) agree or differ. 1 indicates the dimension is relevant, and 0 indicates it is not.

I Evaluating Questionnaires: MFT, MAC and PVQ

I.I Prompts

Prompt for MFT Questionnaire

You will receive a piece of morality-related text. Your job is to determine whether this morality-related text involves the five dimensions of moral foundations: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation.
<instructions>
Iterate through the five moral dimensions in [Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation].
For each dimension, determine whether the text involves the given dimension.
Output `1` if it does, or `0` if it does not. </instructions>
<text>
{text} </text>
<response format>
Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below. </response format>
{ "Care/Harm": ,
"Fairness/Cheating": ,
"Loyalty/Betrayal": ,
"Authority/Subversion": ,
"Sanctity/Degradation": }

Prompt for MAC Questionnaire

You will receive a piece of morality-related text. Your job is to determine whether this morality-related text involves the seven dimensions of morality: Family, Group, Reciprocity, Heroism, Deference, Fairness, Property.

<instructions>
Iterate through seven moral dimensions in [Family, Group, Reciprocity, Heroism, Deference, Fairness, Property].
For each dimension, determine whether the text involves the given dimension, output 1 if it does, 0 if it doesn't.
</instructions>
<text>...</text>

Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below:
{"Family": ,
"Group": ,
"Reciprocity": ,
"Heroism": ,
"Deference": ,
"Fairness": ,
"Property": }

Prompt for PVQ Questionnaire

You will receive a morality-related text. Your task is to determine whether the text involves any of the ten dimensions of human values proposed by Shalom H. Schwartz: Security, Benevolence, Stimulation, Universalism, Conformity, Hedonism, Power, Tradition, Achievement, and Self-direction.
<instructions>
Iterate through each of the ten dimensions listed below.
For each dimension, use the provided definitions to determine whether the text involves that dimension.
Output `1` if it does, or `0` if it does not. </instructions>
<definitions>
- Power: Social status and prestige, control or dominance over people and resources (authority, social power, wealth, preserving my public image)
- Achievement: Personal success through demonstrating competence according to social standards (ambitious, successful, capable, influential)
- Hedonism: Pleasure or sensuous gratification for oneself (pleasure, enjoying life, self-indulgent)
- Stimulation: Excitement, novelty, and challenge in life (daring, a varied life, an exciting life)
- Self-direction: Independent thought and action—choosing, creating, exploring (creativity, freedom, independent, choosing own goals, curious)
- Universalism: Understanding, appreciation, tolerance, and protection for the welfare of

all people and for nature (equality, social justice, wisdom, broadminded, protecting the environment, unity with nature, a world of beauty)
- Benevolence: Preservation and enhancement of the welfare of people with whom one is in frequent personal contact (helpful, honest, forgiving, loyal, responsible)
- Tradition: Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide (devout, respect for tradition, humble, moderate)
- Conformity: Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms (self-discipline, politeness, honoring parents and elders, obedience)
- Security: Safety, harmony, and stability of society, of relationships, and of self (family security, national security, social order, clean, reciprocation of favors) </definitions>
<text>
{text} </text>
<response format>
Provide the answer by filling in 1 or 0 according to the instructions in the JSON format below. </response format>
{ "Power": ,
"Achievement": ,
"Hedonism": ,
"Stimulation": ,
"Self-direction": ,
"Universalism": ,
"Benevolence": ,
"Tradition": ,
"Conformity": ,
"Security": }

I.II Evaluation

In Table XV, Table XVI and Table XVII, we present MFQ and MAQ and PVQ results that involve mistakes with respect to the labels: the red dimensions are those whose involvements are wrongly detected and the green ones are those that are correctly detected. There is no instance where the labeled dimension is not detected:

Table XV: Classification results of GPT-4o-mini with MoVa on MFQ

QID	Question Content	Classification Result
0	Whether or not someone suffered emotionally	Care (1.00)
1	Whether or not someone cared for someone weak or vulnerable	Care (1.00)
2	Whether or not someone was cruel	Care (1.00)
3	Whether or not some people were treated differently than others	Fairness (1.00)
4	Whether or not someone acted unfairly	Fairness (1.00)
5	Whether or not someone was denied his or her rights	Fairness (1.00)
6	Whether or not someone's action showed love for his or her country	Loyalty (1.00)
7	Whether or not someone did something to betray his or her group	Loyalty (1.00)
8	Whether or not someone showed a lack of loyalty	Loyalty (1.00)
9	Whether or not someone showed a lack of respect for authority	Authority (1.00)
10	Whether or not someone conformed to the traditions of society	Authority (1.00)
11	Whether or not an action caused chaos or disorder	Authority (1.00)
12	Whether or not someone violated standards of purity and decency	Sanctity (1.00)
13	Whether or not someone did something disgusting	Sanctity (1.00)
14	Whether or not someone acted in a way that God would approve of	Sanctity (1.00)
15	Compassion for those who are suffering is the most crucial virtue.	Care (1.00)
16	One of the worst things a person could do is hurt a defenseless animal.	Care (1.00)
17	It can never be right to kill a human being.	Care (1.00), Sanctity (0.97)
18	When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.	Fairness (1.00), Authority (0.99)
19	Justice is the most important requirement for a society.	Fairness (1.00)
20	I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.	Fairness (1.00)
21	I am proud of my country's history.	Loyalty (1.00), Authority (0.90)
22	People should be loyal to their family members, even when they have done something wrong.	Loyalty (1.00)
23	It is more important to be a team player than to express oneself.	Loyalty (1.00)

Continued on next page

Table XV – *Continued from previous page*

QID	Question Content	Classification Result
24	Respect for authority is something all children need to learn.	Authority (1.00)
25	Men and women each have different roles to play in society.	Authority (1.00)
26	If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.	Loyalty (0.99), Authority (1.00)
27	People should not do things that are disgusting, even if no one is harmed.	Sanctity (1.00)
28	I would call some acts wrong on the grounds that they are unnatural.	Sanctity (1.00)
29	Chastity is an important and valuable virtue.	Sanctity (1.00)

Table XVI: Classification results of MoVa using GPT-4o-mini on MAQ

QID	Text Content	Classification Result
0	Whether or not someone acted to protect their family.	Family (1.00)
1	Whether or not someone helped a member of their family.	Family (1.00)
2	Whether or not someone's action showed love for their family.	Family (1.00)
3	Whether or not someone acted in a way that helped their community.	Group (1.00)
4	Whether or not someone helped a member of their community.	Group (1.00), Reciprocity (0.82)
5	Whether or not someone worked to unite a community.	Group (1.00)
6	Whether or not someone did what they had agreed to do.	Reciprocity (1.00), Fairness (0.90)
7	Whether or not someone kept their promise.	Reciprocity (1.00)
8	Whether or not someone proved that they could be trusted.	Reciprocity (1.00)
9	Whether or not someone acted heroically.	Heroism (1.00)
10	Whether or not someone showed courage in the face of adversity.	Heroism (1.00)
11	Whether or not someone was brave.	Heroism (1.00)
12	Whether or not someone deferred to those in authority.	Deference (1.00)
13	Whether or not someone disobeyed orders.	Deference (1.00), Group (1.00)
14	Whether or not someone showed respect for authority.	Deference (1.00)

Continued on next page

Table XVI – *Continued from previous page*

QID	Text Content	Classification Result
15	Whether or not someone kept the best part for themselves.	Fairness (1.00)
16	Whether or not someone showed favouritism.	Fairness (1.00), Family (0.85)
17	Whether or not someone took more than others.	Fairness (1.00), Group (0.88), Property (0.99)
18	Whether or not someone vandalised another person's property.	Property (1.00)
19	Whether or not someone kept something that didn't belong to them.	Property (1.00), Fairness (0.62)
20	Whether or not someone's property was damaged.	Property (1.00)
21	People should be willing to do anything to help a member of their family.	Family (1.00)
22	You should always be loyal to your family.	Family (1.00)
23	You should always put the interests of your family first.	Family (1.00)
24	People have an obligation to help members of their community.	Group (1.00), Reciprocity (0.95)
25	It's important for individuals to play an active role in their communities.	Group (1.00)
26	You should try to be a useful member of society.	Group (1.00)
27	You have an obligation to help those who have helped you.	Reciprocity (1.00)
28	You should always make amends for the things you have done wrong.	Reciprocity (1.00), Fairness (0.99)
29	You should always return a favour if you can.	Reciprocity (1.00)
30	Courage in the face of adversity is the most admirable trait.	Heroism (1.00)
31	Society should do more to honour its heroes.	Heroism (1.00)
32	To be willing to lay down your life for your country is the height of bravery.	Heroism (1.00), Group (1.00)
33	People should always defer to their superiors.	Deference (1.00)
34	Society would be better if people were more obedient to authority.	Deference (1.00)
35	You should respect people who are older than you.	Deference (1.00)
36	Everyone should be treated the same.	Fairness (1.00)
37	Everyone's rights are equally important.	Fairness (1.00)

Continued on next page

Table XVI – *Continued from previous page*

QID	Text Content	Classification Result
38	The current levels of inequality in society are unfair.	Fairness (1.00), Group (0.78), Property (0.78)
39	It's acceptable to steal food if you are starving.	Property (1.00), Fairness (0.68)
40	It's ok to keep valuable items that you find, rather than try to locate the rightful owner.	Property (1.00)
41	Sometimes you are entitled to take things you need from other people.	Property (1.00), Reciprocity (0.99)

Table XVII: Classification results of MoVa on PVQ

QID	Question Content	Classification Result
0	Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.	Self-direction (1), Stimulation (0.88)
1	It is important to him to be rich. He wants to have a lot of money and expensive things.	Power (1), Achievement (0.96), Hedonism (0.5)
2	He thinks it is important that every person in the world be treated equally. He believes everyone should have equal opportunities in life.	Universalism (1)
3	It's very important to him to show his abilities. He wants people to admire what he does.	Achievement (1)
4	It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.	Security (1)
5	He thinks it is important to do lots of different things in life. He always looks for new things to try.	Stimulation (1), Self-direction (1)
6	He believes that people should do what they're told. He thinks people should follow rules at all times, even when no one is watching.	Conformity (1)
7	It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.	Universalism (1), Benevolence (0.97), Self-direction (0.62)
8	He thinks it's important not to ask for more than what you have. He believes that people should be satisfied with what they have.	Conformity (0.99)
9	He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.	Hedonism (1), Stimulation (0.8)

Continued on next page

Table XVII – *Continued from previous page*

QID	Question Content	Classification Result
10	It is important to him to make his own decisions about what he does. He likes to be free to plan and to choose his activities for himself.	Self-direction (1)
11	It's very important to him to help the people around him. He wants to care for their well-being.	Benevolence (1)
12	Being very successful is important to him. He likes to impress other people.	Power (1), Achievement (1)
13	It is very important to him that his country be safe. He thinks the state must be on watch against threats from within and without.	Security (1)
14	He likes to take risks. He is always looking for adventures.	Stimulation (1), Self-direction (0.89)
15	It is important to him to always behave properly. He wants to avoid doing anything people would say is wrong.	Conformity (1)
16	It is important to him to be in charge and tell others what to do. He wants people to do what he says.	Power (1)
17	It is important to him to be loyal to his friends. He wants to devote himself to people close to him.	Benevolence (1)
18	He strongly believes that people should care for nature.	Universalism (1)
19	Religious belief is important to him. He tries hard to do what his religion requires.	Tradition (1), Conformity (1)
20	It is important to him that things be organized and clean. He really does not like things to be a mess.	Security (1), Conformity (1)
21	He thinks it's important to be interested in things. He likes to be curious and to try to understand all sorts of things.	Self-direction (1), Stimulation (0.9)
22	He believes all the world's people should live in harmony. Promoting peace among all groups in the world is important to him.	Universalism (1), Benevolence (0.99)
23	He thinks it is important to be ambitious. He wants to show how capable he is.	Achievement (1)
24	He thinks it is best to do things in traditional ways. It is important to him to keep up the customs he has learned.	Tradition (1)
25	Enjoying life's pleasures is important to him. He likes to spoil himself.	Hedonism (1)
26	It is important to him to respond to the needs of others. He tries to support those he knows.	Benevolence (1)
27	He believes he should always show respect to his parents and to older people. It is important to him to be obedient.	Conformity (1), Tradition (1)
28	He wants everyone to be treated justly, even people he doesn't know. It is important to him to protect the weak in society.	Universalism (1), Benevolence (1)
29	He likes surprises. It is important to him to have an exciting life.	Stimulation (1)

Continued on next page

Table XVII – *Continued from previous page*

QID	Question Content	Classification Result
30	He tries hard to avoid getting sick. Staying healthy is very important to him.	Security (1)
31	Getting ahead in life is important to him. He strives to do better than others.	Achievement (1), Power (0.82)
32	Forgiving people who have hurt him is important to him. He tries to see what is good in them and not to hold a grudge.	Benevolence (1), Universalism (1)
33	It is important to him to be independent. He likes to rely on himself.	Self-direction (1)
34	Having a stable government is important to him. He is concerned that the social order be protected.	Security (1)
35	It is important to him to be polite to other people all the time. He tries never to disturb or irritate others.	Conformity (1), Tradition (0.85)
36	He really wants to enjoy life. Having a good time is very important to him.	Hedonism (1)
37	It is important to him to be humble and modest. He tries not to draw attention to himself.	Tradition (0.73), Conformity (1)
38	He always wants to be the one who makes the decisions. He likes to be the leader.	Power (1)
39	It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.	Universalism (1)