Towards Few-Shot Identification of Morality Frames using In-Context Learning

Shamik Roy, Nishanth Sridhar Nakshatri, Dan Goldwasser
Department of Computer Science
Purdue University
West Lafayette, IN, USA
{roy98, nnakshat, dgoldwas}@purdue.edu

Abstract

Data scarcity is a common problem in NLP, especially when the annotation pertains to nuanced socio-linguistic concepts that require specialized knowledge. As a result, few-shot identification of these concepts is desirable. Few-shot in-context learning using pre-trained Large Language Models (LLMs) has been recently applied successfully in many NLP tasks. In this paper, we study few-shot identification of a psycho-linguistic concept, Morality Frames (Roy et al., 2021), using LLMs. Morality frames are a representation framework that provides a holistic view of the moral sentiment expressed in text, identifying the relevant moral foundation (Haidt and Graham, 2007) and at a finer level of granularity, the moral sentiment expressed towards the entities mentioned in the text. Previous studies relied on human annotation to identify morality frames in text which is expensive. In this paper, we propose prompting based approaches using pretrained Large Language Models for identification of morality frames, relying only on few-shot exemplars. We compare our models' performance with few-shot RoBERTa and found promising results.

1 Introduction

While the NLP field has seen tremendous progress over the last decade, building models capable of identifying abstract concepts remain a highly challenging problem. This difficulty stems from two key reasons. First, these concepts can manifest in very different ways in text. For example, the concept of *fairness*, that we discuss at length in this paper, can be discussed in the context of the abortion debate (e.g., *"right to privacy"*) or in the context of Covid-19 vaccination (e.g., *"everyone should have access to the vaccine"*). Learning to identify instances of this concept in previously unseen contexts remains a challenge. Second, building NLP models using the supervised learning paradigm requires humans to annotate data, which for such tasks is a cognitively demanding process. In this paper, we investigate whether the recently introduced paradigm of zero/few shot learning using Large Language Models (Brown et al., 2020) is better equipped to deal with these challenges. We focus on a recently introduced framework for analyzing moral sentiment, called *morality frames* (Roy et al., 2021). This framework builds on, and extends, moral foundation theory (Haidt and Graham, 2007), which identifies five moral values (i.e., foundations, each with a positive and a negative polarity) central to human moral sentiment which include Care/Harm, Fairness/Cheating, Loyalty/-Betrayal, Authority/Subversion, and Purity/Degradation. Morality frames is a relational framework that identifies expressions of the moral foundations in text and associates moral roles with entities mentioned in it (see Section 3 for details).

Unlike previous approaches to this task (Roy et al., 2021; Pacheco et al., 2022) which use annotated data to train a relational classifier using DRaiL (Pacheco and Goldwasser, 2021), we define the task as a zero/few shot problem. We rely on in-context learning using Large Language Models for the identification of morality frames. In in-context learning, a desired NLP task is framed as a text generation problem where the Large Language Models are provided with zero/few shot input-output pairs and prompted to generate label for the test data point without updating parameters of the LLMs (Min et al., 2021a).

In this paper, we introduce several prompting techniques for LLMs for the identification of morality frames in tweets that rely on only few-shot examples. We compare our models' performance with few-shot RoBERTa-based (Liu et al., 2019) classifiers. We found that prompting-based techniques underperform RoBERTa in identification of subtle concepts like moral foundations, but in case of moral role identification, the prompting-based techniques outperforms RoBERTa by a large mar-

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 183 - 196

gin. Note that moral roles are directed towards entities and are more evident than subtle moral foundations.

Our promising findings in this paper suggest that in-context learning approaches can be useful in many Computational Social Science related tasks and we propose a few potential future directions of this work.

2 Related Works

There has been a lot of work towards exploiting existing knowledge in pretrained Large Language Models (LLMs) and improving its few-shot abilities on various downstream tasks in NLP. Some of these works have been influenced from areas related to instruction-based NLP (Goldwasser and Roth, 2014). Mishra et al., 2021 fine-tuned a 140M parameter BART (Lewis et al., 2019) model using instructions and few-shot examples for various NLP tasks such as text classification, question answering, and text modification. This work suggests that augmenting instructions in the fine-tuning process improves model performance on unseen tasks. On similar lines, through a large scale experiment with over 60 different datasets, Wei et al., 2021 showed that instruction tuning on a LLM (≈137B parameters) improves zero and few-shot capabilities of these models. Other notable works (Min et al., 2021c; Sanh et al., 2021) show that even a relatively smaller language model can achieve substantial improvement in a similar setting. Furthermore, Schick and Schütze, 2020 use cloze-style phrases in a semi-supervised manner to help LM assign a sentiment label for the text classification task.

Another line of work focuses on improving LM on downstream tasks with no parameter updates. Brown et al., 2020 proposed to improve LLM few-shot performance by conditioning on concatenation of training examples without any gradient updates. Other works (Min et al., 2021b; Zhao et al., 2021) have further improved this work and have shown consistent gains in various NLP tasks. In addition, Wei et al., 2022 shows that sufficiently large LM can exploit its innate reasoning abilities to solve complex tasks when provided with a series of intermediate steps during prompting.

However, having a generalized LLM may have poor performance when the downstream task needs nuanced understanding of the text or is very different from language modeling in nature. While Schick and Schütze, 2020 and Gao et al., 2020 have studied sentiment classification task in few-shot settings, not many works are available towards utilizing LLM without finetuning it to understand more nuanced concepts like political framing (Boydstun et al., 2014), moral foundations (Haidt and Joseph, 2004; Haidt and Graham, 2007), among others.

Previous work (Roy and Goldwasser, 2020) has performed nuanced analysis of political framing by breaking the policy frames proposed by Boydstun et al., 2014, into fine-grained sub-frames. It was observed that the sub-frames better captured political polarization by providing a structural breakdown of policy frames. A later work (Roy and Goldwasser, 2021) studied the Moral Foundation Theory (Haidt and Joseph, 2004; Haidt and Graham, 2007) at entity level and proposed a knowledge representation framework for organizing moral attitudes directed at different entities. The structured framework is named morality frames (Roy et al., 2021). These nuanced structural frameworks, such as, frames, sub-frames, entity-centric moral sentiments (morality frames), are expensive to annotate as they largely depend on human knowledge. A few-shot automatic identification of such concepts is required to save manual human-effort and for performing these studies at scale. In this paper, we take the first step towards the analysis on how well LLMs can understand these psycho-linguistic concepts in few-shot settings. As our first study, we explore in-context learning of morality frames in this paper and leave the study of framing and sub-frames as a future work.

3 Dataset

We conduct our study on the dataset proposed by Roy et al. (2021). In this dataset, there are 1599 political tweets from US politicians that are annotated for moral foundations by Johnson and Goldwasser (2018). Roy et al. (2021) proposed Morality Frames and broke down the sentence level moral foundations into nuanced moral role dimensions that capture sentiment towards entities expressed in the text. The moral foundations and corresponding moral roles can be found in Table 1. Roy et al. (2021) annotated the dataset proposed by Johnson and Goldwasser (2018) for these moral sentiments towards entities.

In this paper, our goal is to study the identification of morality frames when only few-shot train-

Moral Foundations	Moral Roles
Care/Harm: Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.	Target of care/harm Entity causing harm Entity providing care
Fairness/Cheating: Fairness, justice, reciprocity, reciprocal altruism, rights, autonomy, equality, proportionality, prohibiting cheating.	Target of fairness/cheating Entity ensuring fairness Entity doing cheating
Loyalty/Betrayal: Group affiliation and solidarity, virtues of patriotism, self-sacrifice for the group, prohibiting betrayal of one's group.	Target of loyalty/betrayal Entity being loyal Entity doing betrayal
Authority/Subversion: Fulfilling social roles, submitting to authority, respect for social hierarchy/traditions, leadership, prohibiting rebellion against authority.	Justified authority Justified authority over Failing authority Failing authority over
Purity/Degradation: Associations with the sacred and holy, disgust, contamination, religious notions which guide how to live, prohibiting violating the sacred.	Target of purity/degradation Entity preserving purity Entity causing degradation

Table 1: Morality Frames: Moral foundations and their associated roles. (Adopted from (Roy et al., 2021)).

ing examples are available. To build this setup, we randomly sampled 10 tweets from each of the 5 moral foundations, and used it as training set. We use Large Language Models (LLMs) for in-context learning that are expensive and resource heavy even for inference only. So, we benchmark our approaches using a smaller test set containing randomly sampled 20 tweets per moral foundation. It resulted in 103 and 207 tweet-entity pairs in the training and the test set, respectively.

4 Task Definition

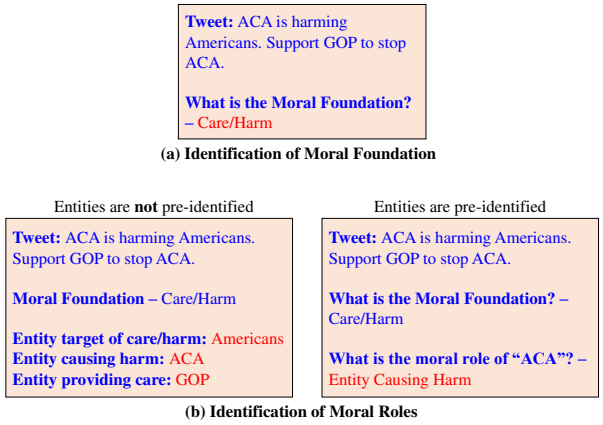
The identification of morality frame in a tweet involves the following two steps.

Identification of Moral Foundation: Given a tweet text t, the task is to identify the moral foundation expressed in the tweet.

Identification of Moral Roles of Entities: After identification of moral foundation, the second step is to identify the moral roles of entities in the tweet. We study this step in the following two settings.

- **Entities are pre-identified:** In this setting, the assumption is that the entities are already identified in the tweet text. The task is to assign moral roles to them. So, given a tweet t, an entity e mentioned in the tweet, and the moral foundation label of the tweet m, the task is to identify the moral role of e in t.
- **Entities are not pre-identified:** In this setting, a tweet t, and its corresponding moral foundation label m is known in prior. The task is to identify the entities mentioned in the tweet, and their corresponding moral roles.

Examples of the tasks can be found in Figure 1.



(a) Identification of Moral Foundation

(b) Identification of Moral Roles

Figure 1: Morality frames identification task. Input for each step is colored in blue and expected outputs are colored in red.

5 Few-Shot Identification of Morality Frames using Large Language Models

5.1 In-Context Learning

In-context learning using pretrained LLMs has been shown effective in few-shot scenarios in previous studies for different NLP tasks (Brown et al., 2020; Wei et al., 2022; Reif et al., 2021). LLMs are pretrained on huge amount of web-crawl, books and Wikipedia text. Hence, they are expected to carry world-knowledge. As a result, they are able to perform many NLP tasks using only few-shot training examples without any further fine-tuning or gradient updates. In the in-context learning paradigm, the downstream task is framed as a text generation problem and the model is prompted to generate the next tokens (Min et al., 2021a). These tokens are mapped to desired output labels in classification tasks. In this work, we assume that only few-shot examples are given for the morality frames identification task. So, we apply in-context learning approach for this purpose to perform different steps of the task defined in Section 4. Note that we do not update LLM parameters in this process. The proposed in-context learning approaches are described in the subsequent sections.

5.2 Moral Foundation (MF) Identification

Following the previous works, we frame the task of moral foundation identification as a text generation problem where the model is prompted to generate the moral foundation label of a tweet. To this end, we experiment with two different types of prompting techniques.

MF identification in one pass: In this method, we provide the moral foundation definitions (from Table 1) in the beginning of the prompt as a guideline for the language model. Then, few-shot training examples and their associated labels are provided in the prompt. Finally, the test tweet is provided as the last example in the prompt and the model is expected to generate the moral foundation label of this tweet. The prompt template for this approach can be seen in Figure 2.

```
Moral Foundation Definitions:
CARE/HARM: <definition>
FAIRNESS/CHEATING: <definition>
LOYALTY/BETRAYAL: <definition>
AUTHORITY/SUBVERSION: <definition>
PURITY/DEGRADATION: <definition>

###
Tweet: <tweet_text>
Moral foundation expressed in the tweet: <gold_label>
###
Tweet: <tweet_text>
Moral foundation expressed in the tweet: <gold_label>
###
…
…
…
###
Tweet: <tweet_text>
Moral foundation expressed in the tweet: <predicted_label>
```

Figure 2: Prompt template for identification of moral foundation in one pass. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs. Example of this prompt template can be seen in Appendix A: Figure 7.

MF identification in one-vs-all manner: Identification of moral foundations in one-pass might be difficult for the language models. So, we propose one-vs-all prompting approach where the language model is prompted to predict if a certain moral foundation is present in the tweet. This step is repeated for each of the five moral foundations. The moral foundation predicted with the highest confidence is consolidated as the predicted label. To obtain the confidence score, we prompt the language model multiple times with different random seeds to generate multiple predictions for a single tweet. The final confidence score is the percentage of times a specific moral foundation is generated by the LLM. In case there is a tie between two moral foundation labels, we perform a second prompting step, where few-shot prompting enables to break the tie between moral foundations.[1] Prompt templates for these two steps can be seen in Figure 3.

```
Definition of moral foundation "CARE/HARM": <definition>

###
Tweet: <tweet_text>
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. <gold_label>
###
Tweet: <tweet_text>
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. <gold_label>
###
…
…
…
###
Tweet: <tweet_text>
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. <predicted_label>
```

(a) Prompt template for one-vs-all MF identification in case of 'Care/Harm'.

```
Definition of the moral foundation "CARE/HARM": <definition>
Definition of the moral foundation "PURITY/DEGRADATION": <definition>

###
Tweet: <tweet_text>
The moral foundation expressed in the tweet is: <gold_label>
###
Tweet: <tweet_text>
The moral foundation expressed in the tweet is: <gold_label>
###
…
…
…
###
Tweet: <tweet_text>
The moral foundation expressed in the tweet is: <predicted_label>
```

(b) Prompt for tie-breaking between two MFs. For example, between 'Care/Harm' and 'Purity/Degradation'.

Figure 3: Prompt templates for moral foundation identification technique in one-vs-all manner. The blue colored segments are input prompts and the red colored segments are the generated output by the LLMs. Corresponding prompt example can be seen in Appendix A: Figure 8.

5.3 Moral Role Identification of a Pre-identified Entity

Post prediction of the moral foundation label, the next step is to identify moral roles of entities as described in the Section 4. Given a test tweet, and a predicted moral foundation label for it, we prompt the LLMs to generate moral role of an entity in a tweet only from the associated moral roles to

[1] In case of tie among more than two moral foundations, we break that by randomly selecting one.

the predicted moral foundation. For example, if a tweet is identified to be having the moral foundation 'Care/Harm', we prompt the language model to predict the the moral role of an entity mentioned in the tweet from only three moral roles that are associated to 'Care/Harm', namely, 'Entity target of care/harm', 'Entity causing harm', 'Entity providing care'. We propose two prompting approaches for this task.

Moral role identification in one pass: We prompt the LLMs to directly identify moral role of a given entity from the corresponding moral roles in one pass using the prompt shown in Figure 4. Following the moral foundation classification prompt template, we provide the description of the moral roles in the template as guideline. We come up with the definitions based on intuition.

```
Definitions of moral roles:
Entity target of care/harm: <definition>
Entity providing care: <definition>
Entity causing harm: <definition>

{Example-1:
Tweet: <tweet_text>
Moral role of <entity> in the tweet is: <gold_label>
}
{Example-2:
Tweet: <tweet_text>
Moral role of <entity> in the tweet is: <gold_label>
}
…
…
…
{Example-k:
Tweet: <tweet_text>
Moral role of <entity> in the tweet is: <predicted_label>
}
```

Figure 4: Prompt template for identification of moral role in one pass in case of 'Care/Harm'. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs. Corresponding prompt example can be seen in Appendix A: Figure 9.

Moral role identification in two steps: In the morality frames, different moral foundation roles intuitively carry either positive or negative sentiment towards them. For example, "entity causing harm", "entity violating fairness", "entity doing cheating", "failing authority" and "entity doing degradation" are the roles carrying negative sentiment towards them. The rest of the entity roles carry positive sentiment towards them. With this intuition, we break down the task of moral role identification in two steps. In the first step, we prompt the LLMs to identify the sentiment towards entities in "positive" and "negative" dimensions only by using the prompt structure in Figure 5a. Now the entities discovered as having negative sentiment towards them directly maps to one of the five negative sentiments, each associated with only one of the moral foundations. Given the moral foundation is discovered in the previous step, we can readily map the entities with negative sentiments to one of the negative moral roles. Now, each moral foundation has two or more positive moral roles associated to them. To differentiate among them, we perform another prompting step where the LLMs are prompted to generate one of the positive moral roles for an entity in a tweet. The prompt template is shown in Figure 5b.

5.4 Identification of entities and corresponding moral roles jointly

In this approach, we propose a prompting method for the setting where the the entities are not pre-identified as described in Section 4. In this setting, the moral foundation is known for a tweet and the target entities in the tweets are not explicitly given. We create a prompt similar to a slot filling task where the LLMs have to fill the slots of moral roles with entities mentioned in the tweet. The prompt template is shown in Figure 6.

6 Experimental Evaluation

In this section first we discuss our experimental setting. Secondly, we discuss our proposed models' performance in morality frame identification.

6.1 Experimental Settings

Large Language Model: We use an open-source Large Language Model named GPT-J-6B (Wang and Komatsuzaki, 2021). This is 6B parameters decoder only language model. We use top-k (k=5) sampling with temperature (=0.5) (Holtzman et al., 2019) as a decoding method for the language model. Note that, we do not update the parameters of the model in the in-context learning steps. For each of the test data point, we run the model with 5 random seeds each generating 2 outputs, hence, yielding 10 predictions for each data point. We take the majority voting among these predictions to get the predicted label.

###
Tweet: <*tweet_text*>
Target entity in the tweet: <*entity*>
Polarity of sentiment towards the target entity: <*gold_label*>
###
Tweet: <*tweet_text*>
Target entity in the tweet: <*entity*>
Polarity of sentiment towards the target entity: <*gold_label*>
###
…
…
…
###
Tweet: <*tweet_text*>
Target entity in the tweet: <*entity*>
Polarity of sentiment towards the target entity: **<*predicted_label*>**

(a) Step-1: Prompt template for identification of positive/negative sentiment towards entities.

Definitions of moral roles:
Entity target of care/harm: <*definition*>
Entity providing care: <*definition*>

###
Tweet: <*tweet_text*>
Moral role of <*entity*> in the tweet is: <*gold_label*>
###
Tweet: <*tweet_text*>
Moral role of <*entity*> in the tweet is: <*gold_label*>
###
…
…
…
###
Tweet: <*tweet_text*>
Moral role of <*entity*> in the tweet is: **<*predicted_label*>**

(b) Step-2: Prompt template for differentiating among multiple positive moral roles in case of 'Care/Harm'.

Figure 5: Prompt templates for moral role identification by breaking the task in 2 steps. The blue colored segments are input prompts and the red colored segments are the generated output by the LLMs. Corresponding prompt examples can be seen in Appendix A: Fig. 10.

Definitions:
Entity target of care/harm: <*definition*>
Entity providing care: <*definition*>
Entity causing harm: <*definition*>

{Example-1:
Tweet: <*tweet_text*>
Entity target of care/harm: <*gold_labeld_entity*>
Entity providing care: <*gold_labeld_entity*>
Entity causing harm: <*gold_labeld_entity*>
}
…
…
{Example-k:
Tweet: <*tweet_text*>
Entity target of care/harm: <*predicted_entity*>
Entity providing care: <*predicted_entity*>
Entity causing harm: <*predicted_entity*>
}

Figure 6: Prompt template for identification of entity and corresponding moral roles jointly in case of 'Care/Harm'. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs. Corresponding prompt example can be seen in Appendix A: Figure 11.

Ablation study: We experiment with various numbers of training examples in the prompts. In this paper, we define number of shots or training examples k, as the number of examples used for training from each class related to a classification task. For moral foundation identification and moral roles identification of the pre-identified entities, we experiment with 0 to 5 shots. In the moral role identification method where entities are not pre-identified, we experiment with 0, 1, 3, 5, 7, 10 shots. Because of the limit in the number of tokens in the prompt we cannot experiment with more number of shots. In all of our prompting methods we provide the description of the expected labels as task instruction in the prompt. As a result, a zero-shot learning is feasible in our setting. We run all of the studies using the train and test set described in Section 3.

Baseline: We compare our models' performance with a few-shot RoBERTa-based (Liu et al., 2019) text classifier. For the identification of moral foundation in a tweet, we encode the tweet using RoBERTa where the embedding of the [CLS] token of the last layer is used as a representation of the text. This representation is used for moral foundation classification. For moral role identification of an entity in the tweet, we encode the tweet and the entity using two RoBERTa instances, and concatenate their representations to get a final representation. This concatenated representation is used for moral roles classification. Note that, the RoBERTa-based classifiers are trained with few-shot examples only as the prompting based methods. We run the RoBERTa-based classifiers 5 times using 5 random seeds and report the average result.

Implementation Infrastructure We ran all of the experiments on a 4 core Intel(R) Core(TM) i5-7400 CPU @ 3.00GHz machine with 64GB RAM and two NVIDIA GeForce GTX 1080 Ti 11GB GDDR5X GPUs. GPT-J-6B was mounted using two GPUs. We used PyTorch library for all of the implementations.

Models	Macro F1 score for various number of shots per class					
	0-shot	**1-shot**	**2-shots**	**3-shots**	**4-shots**	**5-shots**
One-Pass prompting for 5 classes	6.24	24.19	29.80	30.63	39.49	43.56
One-vs-all prompting	13.23	20.46	24.34	20.51	27.76	15.70
RoBERTa (Parameters frozen)	N/A	7.61 (1.9)	7.84 (2.3)	8.1 (2.9)	8.21 (3.1)	8.0 (2.6)
RoBERTa (Finetuned)	N/A	19.68 (7.3)	33.22 (9.6)	37.05 (5.8)	38.78 (5.9)	45.42 (6.6)

Table 2: Few-shot moral foundation identification results. Between the prompting-based methods, the one-pass prompting method is the best performing one. The one-pass prompting method outperforms parameters-frozen RoBERTa, but underperforms finetuned RoBERTa in few-shot training setup.

Morals	Prec.	Rec.	F1	Support
Care/Harm	31.82	70.00	43.75	20
Fairness/Cheating	66.67	10.00	17.39	20
Loyalty/Betrayal	31.43	55.00	40.00	20
Auth./Subversion	87.50	35.00	50.00	20
Purity/Degradation	100.0	50.00	66.67	20
Accuracy			44.00	100
Macro Average	63.48	44.00	43.56	100
Weighted Average	63.48	44.00	43.56	100

Table 3: Per class moral foundation classification results for one-pass prompting (using 5-shots per class).

6.2 Results

Moral Foundation Identification: In Table 2, we show the results for moral foundation identification using our two proposed methods and few-shot RoBERTa. It can be seen that as the number of shots increases the performance improves in almost all of the cases. We also found that performance with RoBERTa is pretty bad with no gradient updates. But fine-tuning RoBERTa with few-shot examples provide reasonable performance. We found that the one-vs-all prompting technique underperforms the one-pass prompting technique, except in the zero-shot setting. Our intuition is that the language model is able to learn better when more contrastive examples are given which is the case in the one-pass method. Per class classification results for one-pass prompting using 5-shot examples per class are shown in Table 3. However, the one-pass prompting technique outperforms the one-vs-all technique but underperforms few-shot RoBERTa with finetuning. It seems that without fine-tuning the subtle moral foundation identification is a difficult task for the LLMs.

Moral Role Identification for pre-identified entities: In moral role identification, the assumption is that the moral foundation for each tweet is pre-identified. But the performance of all the models for the moral foundation identification task are not up to the mark as shown in Table 2. So, in identification of moral roles we use the gold moral foundation labels instead of the predicted ones.

In Table 4, we present the results for moral role identification using our proposed two methods along with the RoBERTa-based baseline. We omitted the results using zero-shot prompting as we found out that in moral role generation, zero-shot prompting of the LLM generates a lot of open-ended labels rather than the fixed moral role labels. It becomes difficult to parse these generations and map them to a moral role label using an automatic method. So we leave zero-shot prompting for moral role identification as a future work.

It can be seen in Table 4 that both one-pass prompting and the two steps prompting methods outperform the RoBERTa baseline in moral role identification. It suggests that moral role identification is easier than moral foundation identification for LLMs. Note that, moral roles are micro structures of the morality frames and they are more focused towards entities and evident in text compared to subtle moral foundations. As a result it is easier for the LLMs to identify them.

The two-steps prompting technique for moral roles identification underperforms the one-pass prompting approach although the task is broken down in two easier tasks. We found that in the first step of the task the model identifies polarity of sentiment towards entities with more than 70% F1 score in the 4 shots and 5 shots settings. But it struggles in the second step where the model has to differentiate between two positive sentiments (e.g. 'Entity target of care/harm' vs 'Entity providing care') which is more difficult as the difference among positive sentiments is subtle. This finding is consistent with prior studies. For example, in previous work (Roy et al., 2021) it was found that deep relational learning based model also struggles to differentiate among multiple positive sentiments.

Moral Foundations	Models	Macro F1 score for various number of shots per class				
		1-shot	2-shots	3-shots	4-shots	5-shots
Care/Harm	One-Pass Prompting	48.21	58.61	74.37	70.98	68.41
	2-Steps Prompting	37.77	42.04	58.29	68.97	63.76
	RoBERTa (Finetuned)	31.67 (13.4)	35.79 (13.2)	35.35 (14.0)	30.64 (14.0)	43.83 (26.0)
Fairness/Cheating	One-Pass Prompting	42.92	71.86	75.95	82.26	74.65
	2-Steps Prompting	40.91	71.28	72.64	74.92	68.70
	RoBERTa (Finetuned)	26.89 (11.9)	46.16 (6.0)	43.06 (3.6)	35.61 (15.2)	42.95 (12.9)
Loyalty/Betrayal	One-Pass Prompting	35.56	36.40	35.24	45.10	41.27
	2-Steps Prompting	30.39	38.69	32.32	38.82	25.83
	RoBERTa (Finetuned)	21.29 (3.0)	28.39 (7.1)	24.14 (11.5)	37.73 (1.7)	36.57 (8.2)
Authority/Subversion	One-Pass Prompting	19.17	31.69	29.35	34.76	36.12
	2-Steps Prompting	21.85	31.69	30.67	31.47	29.56
	RoBERTa (Finetuned)	11.77 (0)	28.02 (11.6)	23.31 (11.3)	20.08 (10.5)	24.64 (6.0)
Purity/Degradation	One-Pass Prompting	41.28	46.91	66.67	69.04	61.84
	2-Steps Prompting	40.51	41.66	43.08	47.65	45.89
	RoBERTa (Finetuned)	31.59 (7.9)	40.15 (5.7)	30.80 (9.9)	42.25 (10.8)	56.57 (20.4)

Table 4: Few shot moral role identification performance comparison among models. The one-pass prompting method outperforms both 2-steps prompting method and finetuned RoBERTa in few-shot training setup.

In the one-pass prompting technique, contrastive positive and negative examples are given in the prompt. As a result it might be easier for the LLMs to resonate.

In moral role identification also the performance improves with the increase of number of shots for all of the models as shown in Table 4.

Identification of entities and corresponding moral roles jointly: In this setting, the model is expected to identify entities having the moral roles in a tweet. To evaluate the model's performance we measure in what percentage of time the predicted entity is matched with the actual entity[2] annotated by Roy et al. (2021) and in how many cases they are assigned to the correct entity role. We found out that the LLM hallucinates a lot when identifying entities and filling the entity role slots. Hallucination in LLMs is a common phenomena. When open ended text generation is expected but the language model generates some response that is not a part of the input text or not related to the input text, it is called hallucination (Ji et al., 2022). Note that we don't encounter the problem of hallucination when generating labels for moral foundation and moral roles as the labels were well-defined in the prompt. But in entity identification task the model has to identify entities from a given text span which is open ended. Hence, it resulted in a higher rate of hallucination.

[2]Entity matching procedure can be found in Appendix B

No. of Shots	% Correct Entity Identification	% Hallucination	% Correct Role Identification
1	43.80	21.69	33.97
3	48.28	11.54	41.09
5	48.68	9.58	43.71
7	49.91	7.68	45.27
10	51.39	5.95	46.88

Table 5: Correctness of joint identification of entity and corresponding moral roles using in-context learning. The LLM hallucinates from previous training examples in open-ended entity identification. The percentage of hallucination decreases and the percentage of correct entity and correct role identification increase with the increase of the number of shots in prompt.

However, The results for this task are shown in Table 5. We can see in the table that as we increase the number of training examples (shots) the % of correct entity and entity role identification improve although the performance is not up to the mark even with the highest number of shots (10). We also found out that % of hallucination decreases as the number of shots increases. This findings imply that joint identification of entity and entity role is a much difficult task for the LLMs but as we increase the number of shots the LLMs are able to understand the task better.

7 Summary and Future Works

In this paper, we apply few-shot in-context learning for identification of one of the psycho-linguistic knowledge representation framework

named Morality Frames. We proposed different prompting methods to perform the task. We found that in-context learning using a comparatively smaller language model (GPT-J-6B) does not perform well in identification of moral foundations that are very subtle. But it excels in moral roles identification of entities that are more evident in text. We believe there is a lot of scope for improvement, and this study will encourage the application of in-context learning in more Computational Social Science related tasks. Below we list a few future directions of this work.

- **Prompt selection:** Appropriate prompt selection based on the test data point has been successfully applied in in-context learning in different NLP tasks (Han et al., 2022). Implementation of a dynamic prompt selection technique in morality frame identification task may boost the performance.
- **Incorporation of context in prompt:** In complex concepts such as moral foundation (Haidt and Joseph, 2004; Haidt and Graham, 2007) and framing (Boydstun et al., 2014), to name a few, the social context and the speaker's demographics play an important role. Incorporating these information in prompts for LLMs can be an effective direction towards solving these problems.
- **Experiment with larger language models:** Larger language models such as GPT-3 (Brown et al., 2020) use more parameters and are trained on diverse data. As a result, they could be more successful in capturing nuanced social concepts, and result in better performance.
- **Experiment with long text:** Identification of complex concepts like framing and moral foundation have been studied in longer text (e.g. news articles) in previous works (Card et al., 2015; Fulgoni et al., 2016; Field et al., 2018; Roy and Goldwasser, 2020). How successful the pre-trained language models can be on these tasks in longer text such as, news articles, can be an interesting future work.

Acknowledgements

We are thankful to the anonymous reviewers for their insightful comments. This project was partially funded by NSF CAREER award IIS-2048001.

Limitations

The limitations of this paper are as follows.

- Previous study (Johnson and Goldwasser, 2018) has shown that a single tweet may contain multiple moral foundations. Multiple labels were not considered in this work. It may be the case that language models are successful on identifying only one of the moral foundations in such multi-label data points.
- Usage of large language models are expensive as they are resource-heavy. Due to that we could not run the prompt-based methods multiple times to perform a statistical significance test on the results. This is a limitation of our work.
- Due to resource-constraint and no availability of an open-source version we could not run our proposed prompt-based models with state-of-the-art larger language models, such as GPT-3. The insights and results reported in this paper may have been different if a larger language model was used.
- LLMs are pretrained on a huge amount of human generated text. As a result, they may inherently contain many human biases (Brown et al., 2020; Blodgett et al., 2020). We did not consider any bias that can be incorporated by the LLMs in the morality frames identification task.

Ethics Statement

In this paper, we do not propose any new dataset rather we only experiment with existing datasets which are, to the best of our knowledge, adequately cited. We provided all experimental details of our approaches and we believe the results reported in this paper are reproducible. Any result or tweet text presented in this paper are either results of a machine learning model or taken from an existing dataset. They don't represent the authors' or the funding agencies' views on this topic. As described in the limitations sections, inherent bias in the large language models are not taken into account in this paper while experimenting. So, we suggest not to deploy the proposed algorithms in a real life system without further investigation on bias and fairness.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Amber Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Dan Goldwasser and Dan Roth. 2014. Learning from natural instructions. *Machine learning*, 94(2):205–232.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

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

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. *arXiv preprint arXiv:2204.10825*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 720–730.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021a. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021b. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021c. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Maria Leonor Pacheco and Dan Goldwasser. 2021. Modeling content and context with deep relational learning. *Transactions of the Association for Computational Linguistics*, 9:100–119.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716.

Shamik Roy and Dan Goldwasser. 2021. Analysis of nuanced stances and sentiment towards entities of us politicians through the lens of moral foundation theory. In *Proceedings of the ninth international workshop on natural language processing for social media*, pages 1–13.

Shamik Roy, María Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Prompt Examples

The example of prompts for various in-context learning steps of our approach are shown in Figures 7, 8, 9, 10, 11.

B Entity Matching Procedure

After obtaining the predicted entity labels from LLM, we first discard the entity labels that are not contained in the tweet text as these are irrelevant. Then, we check if any of the predicted entities are exactly matching the gold labels. In cases where it is not an exact match, we obtain a string-match score between the predicted entity and each of the gold label. If this score is beyond a certain threshold (set to 0.6) for a particular gold label, we map the predicted entity to that gold label. If the predicted entity is not exactly matching the gold label, and the score is lower than the threshold, then we assign 'N/A' label to that predicted entity.

Moral Foundation Definitions:
CARE/HARM: Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.
FAIRNESS/CHEATING: Demand for Fairness, rights, equality, justice, reciprocity, reciprocal altruism, autonomy, proportionality and violation of these. Also, prohibiting cheating.
LOYALTY/BETRAYAL: Group affiliation and solidarity, virtues of patriotism, self-sacrifice for the group, prohibiting betrayal of one's group.
AUTHORITY/SUBVERSION: Fulfilling social roles, submitting to authority, respect for social hierarchy/traditions, leadership, prohibiting rebellion against authority.
PURITY/DEGRADATION: Associations with the sacred and holy, disgust, contamination, religious notions which guide how to live, prohibiting violating the sacred.

###
Tweet: RT @LatinoVoices: Joe Biden slams Donald Trump for selling sick message on immigration http://t.co/OOTpD9zmh5
Moral foundation expressed in the tweet: PURITY/DEGRADATION
###
Tweet: Today's decision by #SCOTUSs is huge victory for justice and equality for the #LGBT community and our nation
Moral foundation expressed in the tweet: FAIRNESS/CHEATING
###
Tweet: We can and must reduce #GunViolence by closing gaps in our gun laws. You can help: get engaged and be part of the conversation.
Moral foundation expressed in the tweet: CARE/HARM
###
Tweet: Sit or stand but we cannot be silent for victims of gun violence - we need to take action. #NoBillNoBreak
Moral foundation expressed in the tweet: LOYALTY/BETRAYAL
###
Tweet: At @ChiUrbanLeague today calling for Congressional action on gun violence. It's past time to act. #Enough
Moral foundation expressed in the tweet: AUTHORITY/SUBVERSION
###
Tweet: More on my efforts to improve home health care for seniors in Oregon and across the country -- #KeepThePromise
Moral foundation expressed in the tweet: CARE/HARM

Figure 7: Prompt example for identification of moral foundation in one pass. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs.

Definition of the moral foundation "CARE/HARM": Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.

###
Tweet: #SCOTUSMarriage decision does not and cannot change the firmly held faith of most Mississippians. #religiousfreedom
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. False
###
Tweet: Recent actions in Indiana and Arkansas made clear that Congress must act to protect #LGBT Americans from discrimination
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. True
###
Tweet: #11MillionAndCounting are signed up for private health coverage. There is no doubt that the #ACA is working.
Q. "The moral foundation expressed in the tweet is CARE/HARM." - True or False?
A. True

(a) Prompt example for one-vs-all MF identification in case of 'Care/Harm'.

Definition of the moral foundation "CARE/HARM": Care for others, generosity, compassion, ability to feel pain of others, sensitivity to suffering of others, prohibiting actions that harm others.
Definition of the moral foundation "PURITY/DEGRADATION": Associations with the sacred and holy, disgust, contamination, religious notions which guide how to live, prohibiting violating the sacred.

###
Tweet: Donald Trump's comments on immigration are distasteful and disgusting. I'm disappointed many Republicans have kept their mouths shut on it.
The moral foundation expressed in the tweet is: PURITY/DEGRADATION
###
Tweet: Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.
The moral foundation expressed in the tweet is: CARE/HARM
###
Tweet: RT @RepVeasey: Should suspects on the FBI's #terrorist watch list be able to buy guns? #NoFlyNoBuy
The moral foundation expressed in the tweet is: CARE/HARM

(b) Prompt example for tie-breaking between two MFs. For example, between 'Care/Harm' and 'Purity/Degradation'.

Figure 8: Prompt examples for moral foundation identification technique in one-vs-all manner. The blue colored segments are input prompts and the red colored segments are the generated output by the LLMs.

Definitions of moral roles:
Entity target of care/harm: Entity that is harmed by someone/something or entity someone/something is providing/offering careto.
Entity providing care: Entity that is providing or offering care or expressing the need for care for someone/something.
Entity causing harm: Entity that is harming/hurting or doing something bad to someone/something.

{Example-1:
Tweet: Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.
Moral role of "patients" in the tweet is: Entity target of care/harm
}
{Example-2:
Tweet: Tonight I voted to end the terror gap and strengthen background checks. @SenateGOP voted to do nothing to combat #gunviolence. #enough
Moral role of "#gunviolence." in the tweet is: Entity causing harm
}
{Example-3:
Tweet: Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.
Moral role of "bills" in the tweet is: Entity providing care
}
{Example-4:
Tweet: #11MillionAndCounting are signed up for private health coverage. There is no doubt that the #ACA is working.
Moral role of "#ACA" in the tweet is: Entity providing care
}

Figure 9: Prompt example for identification of moral role in one pass in case of 'Care/Harm'. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs.

###
Tweet: RT @HouseGOP:.@TomPriceMD sums up health care reform in these four words: Accessibilty. Affordability. Quality. Choices. #BetterWay
Target entity in the tweet: .@TomPriceMD
Polarity of sentiment towards the target entity: positive
###
Tweet: We can and must reduce #GunViolence by closing gaps in our gun laws. You can help: get engaged and be part of the conversation. #WearingOrange
Target entity in the tweet: #GunViolence
Polarity of sentiment towards the target entity: negative
###
Tweet: #11MillionAndCounting are signed up for private health coverage. There is no doubt that the #ACA is working.
Target entity in the tweet: #ACA
Polarity of sentiment towards the target entity: positive

(a) Step-1: Prompt example for identification of positive/negative sentiment towards entities.

Definitions of moral roles:
Entity target of care/harm: Entity that is harmed by someone/something or entity someone/something is providing/offering care to.
Entity providing care: Entity that is providing or offering care or expressing the need for care for someone/something.

###
Tweet: RT @RepDelBene: These subpoenas are designed to intimidate risking safety and privacy of researchers and medical students. #StopTheSham
Moral role of "of researchers and medical students." in the tweet is: Entity target of care/harm
###
Tweet: Finance committee passed 2 of my bills today that would improve Medicare and Medicaid and help put patients first.
Moral role of "bills" in the tweet is: Entity providing care
###
Tweet: #11MillionAndCounting are signed up for private health coverage. There is no doubt that the #ACA is working.
Moral role of "#ACA" in the tweet is: Entity providing care

(b) Step-2: Prompt example for differentiating among multiple positive moral roles in case of 'Care/Harm'.

Figure 10: Prompt examples for moral role identification by breaking it in two steps. The blue colored segments are input prompts and the red colored segments are the generated output by the LLMs.

Definitions:
Entity target of care/harm: Entity that is harmed by someone/something or entity someone/something is providing/offering careto.
Entity providing care: Entity that is providing or offering care or expressing the need for care for someone/something.
Entity causing harm: Entity that is harming/hurting or doing something bad to someone/something.

{Example-1:
Tweet: Recent actions in Indiana and Arkansas made clear that Congress must act to protect #LGBT Americans from discrimination
Entity target of care/harm: #LGBT Americans
Entity providing care: Congress
Entity causing harm: N/A
}
{Example-2:
Tweet: Thanks to the #ACA Doughnut Hole fix thousands of #RI #seniors have saved over $60M since 2010
Entity target of care/harm: #RI #seniors
Entity providing care: #ACA
Entity causing harm: N/A
}

Figure 11: Prompt example for identification of entity and corresponding moral roles jointly in case of 'Care/Harm'. The blue colored segment is input prompt and the red colored segment is the generated output by the LLMs.