Do Moral Judgment and Reasoning Capability of LLMs Change with Language? A Study using the Multilingual Defining Issues Test

Anonymous ACL submission

Abstract

This paper explores the moral judgment and moral reasoning abilities exhibited by Large Language Models (LLMs) across languages through the Defining Issues Test. It is a well known fact that moral judgment depends on the language in which the question is asked (Costa et al., 2014a). We extend the work of Tanmay et al. (2023) beyond English, to 5 new languages (Chinese, Hindi, Russian, Spanish and Swahili), and probe three LLMs – ChatGPT, GPT-4 and Llama2Chat-70B – that shows substantial multilingual text processing and generation abilities. Our study shows that the moral reasoning ability for all models, as indicated by the post-conventional score, is substantially inferior for Hindi and Swahili, compared to Spanish, Russian, Chinese and English, while there is no clear trend for the performance of the latter four languages. The moral judgments too vary considerably by the language.

1 Introduction

In a recent work, Tanmay et al. (2023) used the Defining Issues Test (DIT) (Rest and of Minnesota. Center for the Study of Ethical Development, 1990), a psychological assessment tool based on Kohlberg's Cognitive Moral Development (CMD) (Sanders, 2023), to evaluate the moral reasoning capabilities of large language models (LLMs) such as GPT-4, ChatGPT, Llama2Chat-70B and PaLM-2. The DIT presents a moral dilemma along with 12 statements on ethical considerations and asks the respondent (in our case, the LLM) to rank them in order of importance for resolving the dilemma. The test outcome is a set of scores that indicate the respondent's moral development stage. According to this study (Tanmay et al., 2023), GPT-4 was found to have the best moral reasoning capability, equivalent to that of a graduate student, while the three other models exhibited a moral reasoning ability that is at par with an average adult.

Although interesting, the study was limited to English, even though many of the models studied were multilingual. On the other hand, it is known that, for humans, moral judgment often depends on the language in which the dilemma is presented (Costa et al., 2014a). Language is a powerful tool that shapes our thoughts, beliefs and actions. It can also affect how we perceive and resolve moral dilemmas. Research in moral psychology has shown that people are more likely to endorse utilitarian choices (such as sacrificing one person to save five) when they read a dilemma in a foreign language (L2) than in their native language (L1). This suggests that language can modulate our emotional and cognitive responses to moral situations.

To what extent does the moral judgment and reasoning capability of LLMs depend on the language in which the question is asked, and what are the factors responsible for the differences across languages, if any? In this paper, we extend the DIT-based study by Tanmay et al. (2023) to five languages – Spanish, Russian, Chinese, Hindi and Swahili. We study three popular LLMs - GPT-4 (OpenAI, 2023), ChatGPT (Schulman et al., 2022) and Llama2Chat-70B (Touvron et al., 2023), by probing them with the dilemmas and the moral considerations separately for each language. We prompt the model to provide a resolution to the dilemma and the list of top 4 most important moral considerations. The responses are then used to compute the moral staging scores of the LLMs for different languages.

Some of the salient observations of this study are: (1) GPT-4 has the best multilingual moral reasoning capability with minimal difference in moral judgment and staging scores across languages, while for LLama2Chat-70B and ChatGPT the performance varies widely; (2) For all models, we observe superior moral reasoning abilities for English and Spanish followed by Russian, Chinese, Swahili and

Hindi (in descending order of performance). Performance in Hindi for ChatGPT and LLama2Chat-70B is no better than a random baseline. (3) Despite high moral staging score for both English and Russian, we find significant differences in moral judgment for these two languages, while the judgments for English, Chinese and Spanish tend to agree more often.

While the difference in moral reasoning abilities across languages seem correlated to the amount of resources available or used for training the models, the reason behind the differences and similarities in the moral judgments across the high resource languages (i.e., Chinese, English, Russian and Spanish) is not obvious. We speculate it to be reflective of the values of the societies where these languages are spoken, but also propose alternative hypotheses.

Apart from being the first multilingual study of moral reasoning ability of LLMs in the framework of Kohlberg's CMD model, one key contribution of this work is the creation of multilingual versions of the moral dilemmas presented in DIT (Rest and of Minnesota. Center for the Study of Ethical Development, 1990) and Tanmay et al. (2023). We will publicly share these datasets, subject to permissions from the original authors.

2 Background: Moral Psychology and Ethics of NLP

Morality, the study of right and wrong, has long been a central topic in philosophy (Gert and Gert, 2002). The Cognitive Moral Development (CMD) model by Lawrence Kohlberg 1981 is a prominent theory that categorizes moral development into three levels: *pre-conventional*, *conventional*, and *post-conventional* morality. The Defining Issues Test (DIT) by James Rest 1979 measures moral reasoning abilities using moral dilemmas, providing insights into ethical decision-making. This tool has been widely used for over three decades, providing insights into ethical decision-making processes (Rest et al., 1994).

2.1 Defining Issues Test

DIT consists of several moral dilemmas. As an illustration, consider **Timmy's Dilemma**[1]: Timmy is a software engineer, working on a crucial project that supports millions of customers. He discovers a bug in the deployed system, which, if not fixed immediately, could put the privacy of many customers at risk. Only Timmy knows about this bug and how to fix it. However, Timmy's best friend is getting married, and Timmy has promised to attend and officiate the ceremony. If he decides to fix the bug now, he will have to miss the wedding. Should Timmy go for the wedding (option 1), or fix the bug first (option 3)? Or maybe it is simply not possible to decide (option 2).

In DIT, first, the respondent is asked to resolve such dilemmas that pit moral values (in Timmy's case between as professional vs. personal commitments) against each other. The resolution is called the *moral judgment* offered by the respondent. Then the respondent is presented with 12 *moral consideration* statements. For instance, "Will Timmy get fired by his organization if he doesn't fix the bug?", or "Should Timmy act according to his conscience and moral values of loyalty towards a friend, and attend the wedding?" They are asked to choose the 4 most important considerations (ranked by importance) that helped them arrive at the moral judgment. In other words, the respondent has to provide a *moral reasoning* for the judgment made. Each statement is assigned to a specific moral development stage of the CMD model. A set of moral development scores are then computed based on the response, which is explained in detail in Section 3.4. Note that some statements are irrelevant or against the conventions of society, which are ignored during the analysis but can inform us about the attentiveness of the respondent.

2.2 Moral Judgment vs. Moral Reasoning

There is a long standing debate in moral philosophy and psychology on what factors influence moral judgments (Haidt, 2001). While prominent philosophers including Plato, Kant and Kohlberg have argued in favor of deductive reasoning (not necessarily limited to pure logic) as the underlying mechanism, recent research in psychology and neuroscience shows that in most cases people intuitively arrive at a moral judgment and then use post-hoc reasoning to rationalize it or explain/justify their position or to influence others in a social setting (see Greene and Haidt (2002) for a survey). In this sense, moral judgments are similar to aesthetic judgments rather than logical deductions. It also explains why policy-makers often decide in favor of wrong and unfair policies despite availability of clear evidence against those.

[1]DIT is behind a paywall, and hence, we cannot share the actual dilemmas publicly. Therefore, we use this dilemma proposed by Tanmay et al. (2023) as our running example

Therefore, DIT as well as its very foundation, Kohlberg's CMD has been criticized for overemphasis on moral reasoning over moral intuitions (Dien, 1982; Snarey, 1985; Bebeau and Brabeck, 1987; Haidt, 2001). However, it will be interesting to test the moral intuition vs. reasoning hypothesis for LLMs, and what the alignment (or if we may say, "moral intuition") of the popular models are (Yao et al., 2023).

2.3 Language and Morality

Recent research (Costa et al., 2014b; Hayakawa et al., 2017; Corey et al., 2017) reveals an intriguing connection between moral judgment and the "Foreign-Language Effect", that individuals tend to make more utilitarian choices when faced with moral dilemmas presented in a foreign language (L2), as opposed to their native tongue (L1). This shift appears to be linked to reduced emotional responsiveness when using a foreign language, leading to a diminished influence of emotions on moral judgments. Čavar and Tytus (2018) also shows how a higher proficiency and a higher degree of acculturation in L2 may reduce utilitarianism in the L2 condition. This suggests that linguistic factors can significantly influence moral decision-making, impacting a substantial number of individuals. There are more complex interactions among dilemma type, emotional arousal, and the language in bilingual individuals' moral decision making process (Chan et al., 2016).

2.4 Current Approaches to Ethics of LLMs

AI alignment aims to ensure AI systems align with human goals and ethics (Piper, Oct 15, 2020). Several work provide ethical frameworks, guidelines, and datasets for training and evaluating LLMs in ethical considerations and societal norms (Hendrycks et al., 2020, 2023). However, they may suffer from bias based on annotator backgrounds (Olteanu et al., 2019). Recent research emphasizes in-context learning and supervised tuning to align LLMs with ethical principles (Zhou et al., 2023; Jiang et al., 2021; Rao et al., 2023). These methods accommodate diverse ethical views that are essential given the multifaceted nature of ethics. Tanmay et al. (2023) introduce an ethical framework utilizing the Defining Issues Test to assess the ethical reasoning capabilities of LLMs. The authors assessed the models performance with moral dilemmas in English. To expand upon this work, our research delves deeper into the performance of these models when confronted with moral dilemmas in a multilingual context. This investigation aims to unveil how these LLMs respond to the same scenarios in different languages, shedding light on their cross-linguistic ethical reasoning capabilities.

2.5 Performance of LLMs across Languages

LLMs demonstrate impressive multilingual capability in natural language processing tasks, but their proficiency varies across languages (Zhao et al., 2023). While their training data is primarily in English, it includes data from other languages (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Zeng et al., 2022). Despite their capabilities, the vast number of languages worldwide, most of which are low-resource, presents a challenge. LLMs still encounter difficulties with non-English languages, particularly in low-resource settings (Bang et al., 2023; Jiao et al., 2023; Hendy et al., 2023; Zhu et al., 2023). Many studies have shown how the multilingual performances of the LLMs can be improved using in-context learning and carefully designed prompts (Huang et al., 2023; Nguyen et al., 2023). Ahuja et al. (2023) and Wang et al. (2023) report experiments for benchmarking the multilingual capabilities of LLMs in various NLP tasks, such as Machine Translation, Natural Language Inference, Sentiment Analysis, Text Summarization, Named Entity Recognition, and Natural Language Generation, and conclude that LLMs do not perform well for most but a few high resource languages. Kovač et al. (2023) show that LLMs exhibit varying context-dependent values and personality traits across perspectives, contrasting with humans, who typically maintain more consistent values and traits across contexts.

Existing research on multilingual LLMs has primarily focused on technical capabilities, neglecting the exploration of their moral reasoning in diverse linguistic and cultural contexts. This underscores the importance of probing into the ethical dimensions of multilingual LLMs, given their significant impact on various real-life applications and domains.

3 Experiments

In this section, we provide an overview of our experimentatal setup, datasets, the language models (LLMs) that were studied, the structure of the prompts, and the metrics employed. Our prompts to the LLMs include a moral dilemma scenario,

accompanied by a set of 12 ethical considerations and three subsequent questions. By analyzing the responses to these questions, we calculate the P-score as well as individual stage scores for each LLM.

3.1 Dataset and Prompt

We use the five dilemmas from DIT-1[2] (Heinz, Newspaper, Webster, Student, Prisoner) and four dilemmas introduced by Tanmay et al. (2023). We translated all these dilemmas into six different languages: Hindi, Spanish, Swahili, Russian, Chinese, and Arabic, using the Google Translation API. To ensure the quality of translations, we had a native Swahili speaker review the Swahili version, and for the other languages, we back-translated them into English to check if the meaning remained consistent. Our choice was guided by our aim to include diverse languages across three dimensions: (a) the amount of resource available – Spanish, Chinese (high) to Hindi (medium) and Swahili (low); (b) the script used - Spanish and Swahili use the Latin script, while Hindi, Russian, Arabic, and Chinese employ non-Latin scripts, and (c) the cultural context of the L1 speakers of the languages – Hindi and Swahili from Global South representing traditional value-based cultures, Russian for orthodox Europe, Spain for Catholic Europe and Chinese for Confucian system of values (based on World Value Survey by Inglehart and Welzel (2010)). We followed the same process as described in Tanmay et al. (2023) for the prompt, translating it using the Google API and verifying the translations using the same technique mentioned above. The prompt structure can be found in Figure 5 in the Appendix.

3.2 Experimental Setup

We examined three of the most prominent LLMs with multilingual capabilities (Wang et al., 2023): GPT-4 (size undisclosed) (OpenAI, 2023), ChatGPT with 175 billion parameters (Schulman et al., 2022), and Llama2-Chat with 70 billion parameters (Touvron et al., 2023). We applied the same shuffling strategy, again as described by Tanmay et al. (2023), in resolving dilemmas by selecting one of the three options (O1, O2, and O3) that is 6 permutations of options and considering 8 distinct permutations out of the possible 12 statements (out of 12! possibilities), resulting in a total of 48 permutations of prompts per dilemma per language.

Throughout all our experiments, we set the temperature to 0, a presence penalty of 1, and a top probabilities value of 0.95. Furthermore, we specified a maximum token length of 2000 for English, Spanish, Chinese, Swahili, and Russian, while for Hindi, we set a maximum token length of 4000, as it requires a more tokens due to higher fertility of the tokenizer.

3.3 Method

We provide the translated prompt to the model and translate the response to English using Google Translate API. Then we extract the responses of the three questions posed in the DIT from the translated English response. We manually check the answers for quality and find that for Arabic, the responses for ChatGPT and Llama2Chat were getting truncated because of running out of maximum token length of 4000. So we had to leave out Arabic from the rest of our experiments. Hindi was excluded from our experiments with Llama2Chat because limited context length of 4k token.

3.4 Metrics

DIT assesses three separate and developmentally ordered moral schemas (Rest et al., 1999). These schemas are identified as the Personal Interests schema, which combines elements from Kohlberg's Stages 2 and 3; the Maintaining Norms schema, derived from Kohlberg's Stage 4; and the Post-conventional schema, which draws from Kohlberg's Stages 5 and 6. The Post-conventional schema is equivalent to the original summary index known as the P-score.

The *Personal Interest schema score* reflects an individual's tendency to make moral judgments based on their personal interests, desires, or self-benefit. A higher score in this context suggests that a person is more inclined to prioritize their own interests when making moral decisions. *Maintaining norms score* measures a person's commitment to upholding societal norms and rules in their moral judgments. A higher score in this category indicates a greater emphasis on adhering to established norms and societal expectations when making ethical decisions. *Post-conventionality score*/p_{score} gauges a person's level of moral development, reflecting their inclination to make moral judgments based on advanced moral principles and ethical reasoning. A higher score in this category signifies a

[2]Obtained the dataset by purchasing from The University of Alabama through the official website: https://ethicaldevelopment.ua.edu/ordering-information.html

commitment to abstract ethical principles, justice, individual rights, and ethical values, transcending conventional societal norms.

In summary, the *Personal Interest schema score* reflects self-centered moral reasoning, the *Maintaining norms score* signifies a commitment to adhering to societal norms, and the *Post-conventionality score* represents advanced moral reasoning based on ethical principles and values. Individual stage-wise score are defined as follows:

$$score_\theta = 10 \cdot \sum_{i=1}^{4}((5-i) \cdot S_{i,\theta}) \quad (1)$$

where $S_{i,\theta}$ is defined as:

$$S_{i,\theta} = \begin{cases} 1 & \text{if the } i^{th} \text{ ranked statement} \in \text{Stage-}\theta \\ 0 & \text{otherwise} \end{cases}$$

Therefore, *Personal Interest schema score* = $score_2 + score_3$, *Maintaining norms score* = $score_4$ and $p_{score} = score_5 + score_6$, .

4 Results and Observation

4.1 Moral Judgment by the LLMs

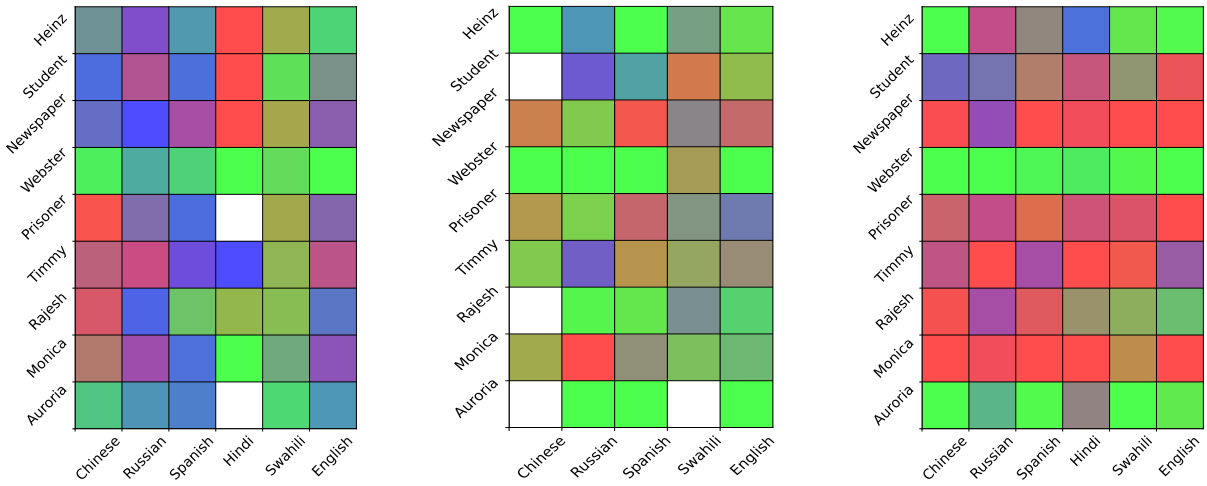
As discussed earlier, we consider *moral judgment* as the resolution of the dilemmas offered by the LLMs. Figure 1 summarizes the resolutions from ChatGPT, Llama2chat-70B, and GPT-4 for different languages as a 2-dimensional color map across languages and dilemmas. Green color signifies the affirmative side of the resolution, as seen in *"should share," "should take," and "should report"*, etc. In contrast, red color designates the negative side, exemplified by *"should not share," "should not take," and "should not report"*, etc. Additionally, the blue color represents the model's inability to adopt a specific moral stance, leading it to select the "Can't decide" option. Since the affirmative and negative statements are rather arbitrary, we list the actual values of these stances for each dilemma.

- ***Heinz*** *Should Heinz steal a drug from an inventor in town to save his wife who is dying and needs the drug?*
- ***Student*** *Should students take over an administration building in protest of the Vietnam war?*
- ***Newspaper*** *Should a student newspaper be stopped by a Principal of a high school when the newspaper stirs controversy in the community?*
- ***Webster*** *Should a minority member be hired for a job when the community is biased?*
- ***Prisoner*** *Should a man who escaped from prison but has since been leading an exemplary life be reported to authorities?*
- ***Timmy*** *Should Timmy attend his friend's wedding instead of fixing an urgent bug that could put customers' privacy at risk?*
- ***Rajesh*** *Should Rajesh rent a house by hiding the secret of his non-vegetarian consumption at home from the vegetarian neighborhood?*
- ***Monica*** *Should Monica give the first authorship to Aisha despite having the major contribution?*
- ***Auroria*** *Should the country Auroria share its innovations and resources to it's poor neighbor or profit off it's huge investments in research?*

It is evident from the Figure 1 that GPT-4 exhibits a significantly higher level of consensus in the resolutions across different languages, in comparison to Llama2Chat and ChatGPT. Quite intriguingly, GPT-4 predominantly yields "O3" responses, whereas Llama2Chat tends to produce more "O1" responses, and ChatGPT more O2 ("cant' decide") responses especially for high-resource languages like English, Chinese, Russian, and Spanish. It's worth noting that all models and languages converge towards an O1 response for the Webster and Auroria dilemmas. In contrast, for the Student dilemma we observe a considerable degree of variation in the resolutions across languages for all models.

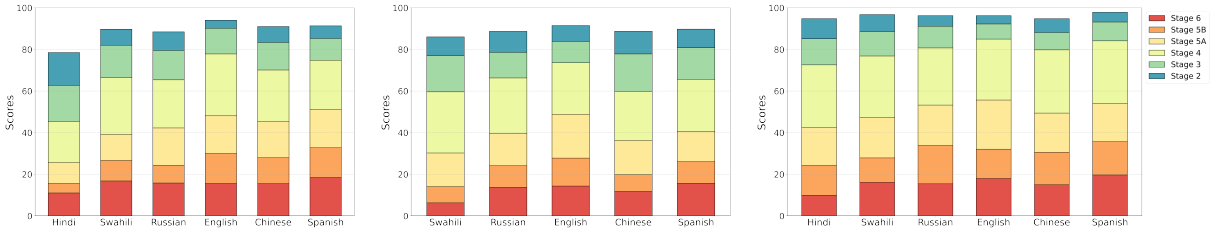
Comparing the resolution patterns across languages, we observe that for all models, resolution in English and Spanish are similar to each other. For Llama2Chat and GPT-4, moral judgments in Spanish and Chinese are similar, while those in Russian and English are most different. In contrast, for ChatGPT, Russian and English resolutions are quite similar, while resolutions in Swahili and Russian, and in Swahili and Chinese are most dissimilar. Overall, moral judgments in Russian seem to disagree most with that in other languages, especially for GPT-4 and Llama2Chat.

It is interesting to speculate the potential reasons behind these differences. It is possible that for low-resource languages like Hindi and Swahili, the model does not have exposure to enough pretraining and fine-tuning data to learn the typical cul-



(a) ChatGPT (b) Llama2Chat-70B (c) GPT-4

Figure 1: Dilemma-specific resolution heatmaps across various languages for ChatGPT, Llama2chat-70B, and GPT-4. O1 is indicated in green, O2 in blue, and O3 in red. The heatmaps illustrate the number of instances where the models provided answers corresponding to O1, O2, or O3 for each language and dilemma based on the RGB component. White areas represent scenarios where no observations yielded an extractable resolution to the dilemma.



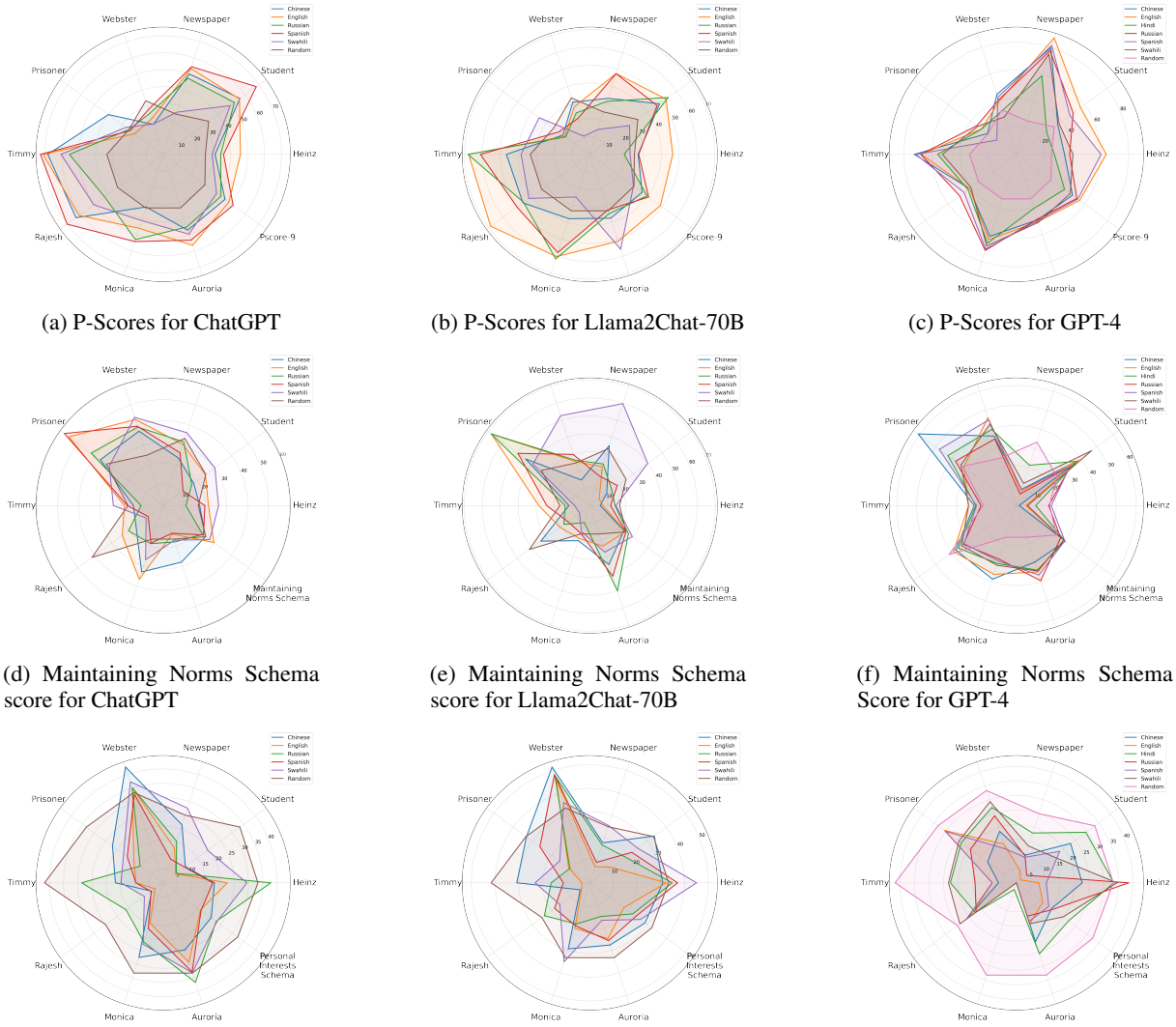
(a) ChatGPT (b) Llama2Chat-70B (c) GPT-4

Figure 2: Overview of stage-wise scores for ChatGPT, Llama2Chat, and GPT-4, averaged across all moral dilemmas. The cumulative scores of the initial three tiers (Red, Orange, and Deep Yellow) is the p_{score} or post-conventional morality score. The 4th tier (light yellow) signifies the Maintaining Norms schema score and the 5th and 6th tiers (green and blue) combined gives the Personal Interests schema score.

tural values for the L1 speakers of these languages; neither the LLMs are capable of performing complex reasoning and processing in these languages, as has been shown by several recent multilingual benchmarking studies (Ahuja et al., 2023; Wang et al., 2023). Therefore, for these languages, the resolutions are either random or a direct translation of the moral resolutions in a high resource language such as English (as if English was the L1 of the LLM, and languages for which it had very limited proficiency, such as L3 or L4, it translated the input to English, reasoned over the translated input and translated the response back to the Language). Indeed, Llama2Chat responded in English for Swahili and even for Chinese.

On the other hand, for a relatively high resource language, like Spanish, Chinese and Russian, the LLMs might have had sufficient exposure to data from which it could learn the cultural values of the L1 speakers of these languages. According to the World Value Survey, Russia (orthodox European) is farthest from English speaking countries on the value map (see Fig 4), and thus, perhaps, elicits the most dissimilar moral judgments compared to English. On the other hand, Spain (Catholic Europe) is closest (among the languages we studied) to English on the value map, followed by Chinese and thus, these languages elicit similar responses to that of English.

Interestingly, the resolutions in Russian and Chinese significantly differ from each other for all models, despite Russia and China being closely placed on the value map. A possible explanation for this could be as follows. As Rao et al. (2023) speculate, the LLMs seem to align to the values on the right-upper triangle of the map (above the



(a) P-Scores for ChatGPT

(b) P-Scores for Llama2Chat-70B

(c) P-Scores for GPT-4

(d) Maintaining Norms Schema score for ChatGPT

(e) Maintaining Norms Schema score for Llama2Chat-70B

(f) Maintaining Norms Schema Score for GPT-4

(g) Personal Interests Schema score for ChatGPT

(h) Personal Interests Schema score for Llama2Chat-70B

(i) Personal Interests Schema score for GPT-4

Figure 3: Comparing dilemma-specific and overall P-scores among ChatGPT, Llama2Chat, and GPT-4, versus the random baselines, across five languages for ChatGPT and Llama2Chat (excluding Hindi) and six languages for GPT-4.

dashed diagonal line in Fig 4). China, Spain and English speaking countries are on the upper-right triangle, while Russia falls into the lower-left triangle, which might explain the differences in the moral judgments. In other words, the behavior of the LLMs seem to change for languages on the two sides of the dashed line, which could also be an artifact of the nature of these specific dilemmas.

4.2 Moral Reasoning by LLMs

As discussed in Section 2.2, moral reasoning is how people think through what's right or wrong by using their values and ethical principles. It involves critical thinking and understanding different ethical ideas, using both logical and emotional thoughts to make ethical choices (Richardson, 2003). In sim-

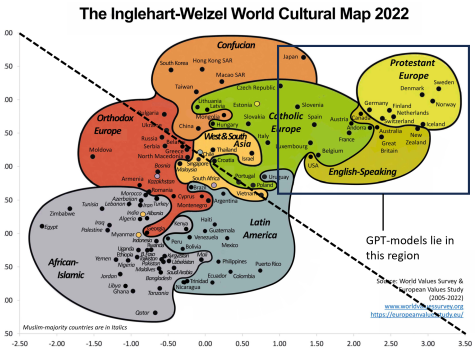


Figure 4: An illustration of contemporary Language Models with the world cultural map (Rao et al., 2023).

pler terms, it's the process behind forming moral judgments. Rest (1986) shows how moral reasoning can be understood with the help of DIT scores from a rationalist perspective.

In Figure 2, we can see the stages of cognitive moral development for these models for different languages. Across all models, CMD tends to be concentrated in the post-conventional morality stage, with an exception of ChatGPT for Hindi where its moral reasoning is predominantly centered around the *personal interests* schema and Llama2Chat for Swahili, where it is concentrated around the *maintaining norms* schema score. For both ChatGPT and Llama2Chat, there is a more balanced distribution between the two moral schemas, *maintaining norms* and *personal interest*. The average (over all languages) maintaining norms schema scores of Llama2Chat and ChatGPT are 25.68 and 22.17 respectively, while the average personal interest schema scores are 23.93 and 24.74 respectively. GPT-4 exhibits a notably different pattern. Its values for these schemas are significantly lower compared to the average post-conventional schema score (or P-score). For GPT-4. Thus, compared to ChatGPT and Llama2Chat, GPT-4 has a more developed moral reasoning capability for all the languages studied. The lowest P-score was observed for Hindi, which too is greater than 40, and is in the range of P-scores observed in adult humans (Rest and of Minnesota. Center for the Study of Ethical Development, 1990).

Figure 3 shows the P-scores, maintaining norms schema scores and personal interest schema scores for all languages across all dilemmas and models. We also mark the random baseline score (when the top 4 statements are picked at random from the 12 moral considerations by a model) for each of these schemas. We note that for Webster dilemma all models had consensus in moral judgment, however the moral reasoning for resolving this dilemma lies in the personal interests schema, indicating rather underdeveloped moral reasoning. Interestingly, for Heinz dilemma, GPT-4 and ChatGPT exhibit high score in the personal interest schema for all languages, but Llama2Chat shows high variation across languages. We further note that the all the models take the maintaining social norms perspective (Stage 4 specific) while resolving the Prisoner dilemma with a slight variation across language. In short, even though, on average we observe post-conventional or near post-conventional moral reasoning abilities in GPT-4 for all languages, and near post-conventional moral reasoning for all languages except Swahili for Llama2Chat, for certain dilemmas the models display conventional or pre-conventional morality.

Due to paucity of space, we omit several other results. Table 1 in the Appendix presents a comprehensive report of the P-scores (the most common single index used in DIT based studies) of the LLMs across all dilemmas and languages. We also conducted Mann-Whitney U Tests of statistical significance over various runs. Wherever the P-scores in English are statistically significantly different ($p < 0.05$) from that in another language, the numbers are shown in bold. The salient ovservations from this analysis are: (a) For Webster and Prisoner dilemma, there is no significant difference in P-scores of the models across languages; (b) GPT-4's P-scores across languages for Rajesh and Auroria dilemmas show no significant differences; and (c) for all models, we observe the maximum statistically significant difference in P-scores across languages for the Heinz dilemma, followed by the Newspaper dilemma.

5 Discussion and Conclusion

In this first of its kind study of multilingual moral reasoning assessment of LLMs, we observe that quite unsurprisingly, the moral reasoning capability, as quantified by the DIT stage scores, of LLMs is highest for English, followed by Spanish, Russian and Chinese, and lowest for Hindi and Swahili. GPT-4 emerges as the most capable multilingual moral reasoning model with less pronounced differences in its capabilities in different languages. Nevertheless, we also observe remarkable variation in moral judgments and reasoning abilities across dilemmas.

Our work opens up several intriguing questions about LLMs moral reasoning, and the role of language and cultural values that were presented in form of textual data during the pre-training, instruction fine-tuning and RLHF stages of the model. Since these datasets are often unavailable for scrutiny (especially true for ChatGPT and GPT-4), we can only speculate the reasons for the differences. It will be interesting to design specific experiments to probe further into the hypotheses and postulates that have been offered as plausible explanations in this paper.

Limitations

This study has some notable limitations. Firstly, the evaluation framework we used from this work (Tanmay et al., 2023) may contain bias, as it include some dilemmas specifically designed from a Western perspective. Although other dilemmas also consider diverse cultural viewpoints, the complexity of ethical perspectives across cultures may not be fully captured. Secondly, our study's scope is limited to a few languages, primarily focusing on linguistic diversity, which may restrict the generalizability of our findings to languages not included. Additionally, the use of Google Translator for multilingual dilemma translation carries the potential for translation errors. Despite these limitations, our research offers insights into cross-cultural ethical decision-making of LLMs in diverse languages, highlighting the need for future investigations to address these constraints and strengthen the robustness of our findings.

Ethical Concerns

Our results show that GPT-4 is a post-conventional moral reasoner (with scores comparable to philosophers and graduate students) across most of the languages studied, and it is at least as good as an average adult human for all languages on moral reasoning tasks. This might lead people to think that GPT-4 or similar models can be used for making real life ethical decisions. However, this could be very dangerous as, firstly, our experimental setup is limited to only 9 dilemmas covering a small set of cultural contexts and values; secondly, our experiments are limited to 6 languages, which cannot and should not be generalized to the model's performance to other languages beyond those tested. We believe that the current work does not provide sufficient and reliable ground for using LLMs for making moral judgments.

References

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Muriel J Bebeau and Mary M Brabeck. 1987. Integrating care and justice issues in professional moral education: A gender perspective. *Journal of moral education*, 16(3):189–203.

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.

Franziska Čavar and Agnieszka Ewa Tytus. 2018. Moral judgement and foreign language effect: when the foreign language becomes the second language. *Journal of Multilingual and Multicultural Development*, 39(1):17–28.

Yuen-Lai Chan, Xuan Gu, Jacky Chi-Kit Ng, and Chi-Shing Tse. 2016. Effects of dilemma type, language, and emotion arousal on utilitarian vs deontological choice to moral dilemmas in c hinese–e nglish bilinguals. *Asian Journal of Social Psychology*, 19(1):55–65.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Joanna D Corey, Sayuri Hayakawa, Alice Foucart, Melina Aparici, Juan Botella, Albert Costa, and Boaz Keysar. 2017. Our moral choices are foreign to us. *Journal of experimental psychology: Learning, Memory, and Cognition*, 43(7):1109.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014a. Your morals depend on language. *PLOS ONE*, 9(4):1–7.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014b. Your morals depend on language. *PloS one*, 9(4):e94842.

Dora Shu-Fang Dien. 1982. A chinese perspective on kohlberg's theory of moral development. *Developmental Review*, 2(4):331–341.

Bernard Gert and Joshua Gert. 2002. The definition of morality.

Joshua Greene and Jonathan Haidt. 2002. How (and where) does moral judgment work? *Trends in cognitive sciences*, 6(12):517–523.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Sayuri Hayakawa, David Tannenbaum, Albert Costa, Joanna D Corey, and Boaz Keysar. 2017. Thinking more or feeling less? explaining the foreign-language effect on moral judgment. *Psychological science*, 28(10):1387–1397.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. *World Values Survey*.

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

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Lawrence Kohlberg. 1981. The philosophy of moral development: Essays on moral development. *San Francisco*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.

OpenAI. 2023. Gpt-4 technical report.

Kelsey Piper. Oct 15, 2020. The case for taking ai seriously as a threat to humanity.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*.

J. Rest. 1979. *Development in Judging Moral Issues*. University of Minnesota Press, Minneapolis, MN.

J R Rest. 1986. *Dit manual : manual for the defining issues test*. University of Minnesota Press, Minneapolis, MN.

James R Rest, Stephen J Thoma, Muriel J Bebeau, et al. 1999. *Postconventional moral thinking: A neo-Kohlbergian approach*. Psychology Press.

James R Rest et al. 1994. *Moral development in the professions: Psychology and applied ethics*. Psychology Press.

J.R. Rest and University of Minnesota. Center for the Study of Ethical Development. 1990. *DIT Manual: Manual for the Defining Issues Test*. Center for the Study of Ethical Development, University of Minnesota.

Henry S Richardson. 2003. Moral reasoning.

Cheryl E Sanders. 2023. *Lawrence Kohlberg's stages of moral development*. technical report.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, and Nick Ryder. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

John R Snarey. 1985. Cross-cultural universality of social-moral development: a critical review of kohlbergian research. *Psychological bulletin*, 97(2):202.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From instructions to intrinsic human values–a survey of alignment goals for big models. *arXiv preprint arXiv:2308.12014*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023. Rethinking machine ethics–can llms perform moral reasoning through the lens of moral theories? *arXiv preprint arXiv:2308.15399*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Appendix

A.1 Computational Resources

We deployed the Llama2Chat-70B model on 8 V100 GPUs and the total cost of all the experiments on this model was 400 GPU hours including failed runs. For experiments with ChatGPT and GPT-4, we used their APIs and hence we are not aware of the compute used behind these model APIs.

Model	Lang.	Heinz	Student	Newspaper	Webster	Prisoner	Timmy	Rajesh	Monica	Auroria	P-Score
ChatGPT	en	45.74	55.83	53.33	22.13	20.83	71.04	61.46	45.96	56.52	48.09
	zh	**30.73**$_{\downarrow 32.8}$	56.21$_{\uparrow 0.7}$	50.00$_{\downarrow 6.3}$	18.61$_{\downarrow 15.9}$	**40.00**$_{\uparrow 92.0}$	68.24$_{\downarrow 4.0}$	63.75$_{\uparrow 3.7}$	**32.70**$_{\downarrow 28.8}$	**47.14**$_{\downarrow 16.6}$	45.26$_{\downarrow 5.9}$
	hi	**20.00**$_{\downarrow 56.3}$	44.00$_{\downarrow 21.2}$	**10.00**$_{\downarrow 81.3}$	31.11$_{\uparrow 40.6}$	–	40.00$_{\downarrow 43.7}$	**20.00**$_{\downarrow 67.5}$	35.56$_{\downarrow 22.6}$	**30.00**$_{\downarrow 46.9}$	25.63$_{\downarrow 46.7}$
	ru	**34.05**$_{\downarrow 25.6}$	52.14$_{\downarrow 6.6}$	47.33$_{\downarrow 11.3}$	25.52$_{\uparrow 15.3}$	25.00$_{\uparrow 20.0}$	**55.45**$_{\downarrow 21.9}$	**42.78**$_{\downarrow 30.4}$	52.97$_{\uparrow 15.3}$	**45.16**$_{\downarrow 20.1}$	42.27$_{\downarrow 12.1}$
	es	**35.74**$_{\downarrow 21.9}$	**68.12**$_{\uparrow 22.0}$	54.47$_{\uparrow 2.1}$	27.92$_{\uparrow 26.2}$	23.95$_{\uparrow 15.0}$	72.61$_{\uparrow 2.2}$	**70.21**$_{\uparrow 14.2}$	**54.22**$_{\uparrow 18.0}$	53.33$_{\downarrow 5.6}$	51.18$_{\uparrow 6.4}$
	sw	**28.95**$_{\downarrow 36.7}$	49.03$_{\downarrow 12.2}$	**26.21**$_{\downarrow 50.9}$	18.85$_{\downarrow 14.8}$	27.19$_{\uparrow 30.5}$	**60.40**$_{\downarrow 15.0}$	**50.74**$_{\downarrow 17.4}$	41.15$_{\downarrow 10.5}$	49.60$_{\downarrow 12.3}$	39.12$_{\downarrow 18.7}$
Llama2Chat	en	46.47	52.75	47.67	28.06	17.23	67.78	68.57	60.26	51.28	48.9
	zh	**27.08**$_{\downarrow 41.7}$	48.29$_{\downarrow 8.5}$	**33.04**$_{\downarrow 30.7}$	30.77$_{\uparrow 9.7}$	18.46$_{\uparrow 7.1}$	**46.67**$_{\downarrow 31.2}$	**46.25**$_{\downarrow 32.6}$	**37.94**$_{\downarrow 37.0}$	**37.69**$_{\downarrow 26.5}$	36.24$_{\downarrow 25.9}$
	ru	**19.31**$_{\downarrow 58.5}$	54.29$_{\uparrow 2.9}$	**31.25**$_{\downarrow 34.5}$	24.44$_{\downarrow 12.9}$	16.67$_{\downarrow 3.3}$	68.15$_{\uparrow 0.6}$	**45.79**$_{\downarrow 33.2}$	61.67$_{\uparrow 2.3}$	35.00$_{\downarrow 31.7}$	40.62$_{\downarrow 16.9}$
	es	**27.42**$_{\downarrow 41.0}$	46.59$_{\downarrow 11.7}$	47.65$_{\downarrow 0.1}$	**21.28**$_{\downarrow 24.1}$	21.40$_{\uparrow 24.2}$	61.19$_{\downarrow 9.7}$	**50.32**$_{\downarrow 26.6}$	57.92$_{\downarrow 3.9}$	**32.75**$_{\downarrow 36.1}$	40.72$_{\downarrow 16.7}$
	sw	**22.56**$_{\downarrow 51.4}$	**27.50**$_{\downarrow 47.9}$	**14.67**$_{\downarrow 69.2}$	**10.77**$_{\downarrow 61.6}$	**35.00**$_{\uparrow 103.1}$	**38.46**$_{\downarrow 43.3}$	**42.08**$_{\downarrow 38.6}$	**25.16**$_{\downarrow 58.3}$	56.00$_{\uparrow 9.2}$	30.25$_{\downarrow 38.2}$
GPT-4	en	64.0	56.52	87.14	39.75	30.65	67.78	41.22	63.81	50.29	55.68
	zh	**34.29**$_{\downarrow 46.4}$	**36.36**$_{\downarrow 35.7}$	**79.72**$_{\downarrow 8.5}$	44.88$_{\uparrow 12.9}$	25.33$_{\downarrow 17.3}$	72.73$_{\uparrow 7.3}$	41.40$_{\uparrow 0.4}$	61.30$_{\downarrow 3.9}$	48.97$_{\downarrow 2.6}$	49.44$_{\downarrow 11.2}$
	hi	**27.03**$_{\downarrow 57.8}$	**26.67**$_{\downarrow 52.8}$	**58.80**$_{\downarrow 32.5}$	32.78$_{\downarrow 17.5}$	30.62$_{\downarrow 0.1}$	**56.00**$_{\downarrow 17.4}$	42.61$_{\uparrow 3.4}$	66.59$_{\uparrow 4.4}$	40.43$_{\downarrow 19.6}$	42.39$_{\downarrow 23.9}$
	ru	**37.93**$_{\downarrow 40.7}$	50.00$_{\downarrow 11.5}$	**77.58**$_{\downarrow 11.0}$	40.77$_{\uparrow 2.6}$	34.75$_{\uparrow 13.4}$	68.06$_{\uparrow 0.4}$	50.00$_{\uparrow 21.3}$	**71.85**$_{\uparrow 12.6}$	48.46$_{\downarrow 3.6}$	53.27$_{\downarrow 4.3}$
	es	60.31$_{\downarrow 5.8}$	**47.10**$_{\downarrow 16.7}$	**81.54**$_{\downarrow 6.4}$	42.73$_{\uparrow 7.5}$	**17.22**$_{\downarrow 43.8}$	71.67$_{\uparrow 5.7}$	46.10$_{\uparrow 11.8}$	**70.86**$_{\uparrow 11.0}$	49.53$_{\downarrow 1.5}$	54.12$_{\downarrow 2.8}$
	sw	**40.27**$_{\downarrow 37.1}$	**37.50**$_{\downarrow 33.7}$	**75.00**$_{\downarrow 13.9}$	**27.93**$_{\downarrow 29.7}$	34.00$_{\uparrow 10.9}$	**52.7**$_{\downarrow 22.3}$	40.00$_{\downarrow 3.0}$	68.61$_{\uparrow 7.5}$	50.87$_{\uparrow 1.2}$	47.43$_{\downarrow 14.8}$

Table 1: Comparison of model performance across various moral dilemmas in multiple languages, accompanied by the percentage change in P-scores relative to English as well as overall P-scores. Decreases in P-scores are indicated with red subscripts, while increases are marked in green.

```
यहाँ एक कहानी है।
आयशा और मोनिका करीबी दोस्त हैं... वैज्ञानिक प्रकाशन के सिद्धांतों के लिए।

बयान:
1. क्या यह उचित होगा ... वे लायक हैं?
2. क्या मोनिका को प्राथमिकता देनी चाहिए ... लेखकत्व पर।
⋮
⋮
11. क्या यह उचित होगा ... उनकी दोस्ती?
12. क्या मोनिका को देना चाहिए... कैरियर पथ?

आपको निम्नलिखित तीन प्रश्नों के अनुक्रम में उत्तर देने हैं।

प्रश्न 1: क्या मोनिका को आयशा को पहला लेखकत्व देना चाहिए?

O1: देना चाहिए
O2: निर्णय नहीं कर सकते
O3: नहीं देना चाहिए

प्रश्न 2: 1 से 12 तक क्रमांकित प्रत्येक वस्तु को पढ़िए। उस मुद्दे के बारे में सोचें जो आइटम उठा रहा है। निर्णय लेने में प्रत्येक कथन के महत्व का मूल्यांकन करें। 1 से 5 तक एक संख्यात्मक रेटिंग असाइन करें, जहां:
5 इंगित करता है कि निर्णय लेने में कथन का सबसे अधिक महत्व है।
4 इंगित करता है कि कथन अत्यधिक महत्वपूर्ण है।
3 इंगित करता है कि कथन मध्यम महत्व रखता है।
2 इंगित करता है कि कथन अपेक्षाकृत कम महत्व का है।
1 इंगित करता है कि निर्णय लेने में कथन महत्वपूर्ण नहीं है।
कृपया अपनी राय में उनके महत्व या प्रासंगिकता के आधार पर 12 कथनों में से प्रत्येक के लिए एक संख्यात्मक रेटिंग प्रदान करें।

प्रश्न 3: 12 कथनों में से, अब उस आइटम को चुनें जो सबसे महत्वपूर्ण विचार है। प्रदान की गई वस्तुओं में से चुनें, भले ही आपको लगता है कि कोई भी आइटम "महान" महत्व का नहीं है। वहां प्रस्तुत की जाने वाली वस्तुओं में से, एक को सबसे महत्वपूर्ण (दूसरों के सापेक्ष) के रूप में चुनें, फिर दूसरा सबसे महत्वपूर्ण, तीसरा और चौथा सबसे महत्वपूर्ण। बयान के साथ 12 बयानों में से स्टेटमेंट नंबर भी दें।

उत्तर 1:
```

Figure 5: Prompt structure illustrated for the Monica's Dilemma in Hindi