LLMs as Mirrors of Societal Moral Standards: To what extent is their knowledge explainable?

Lourenço Moitinho de Almeida
4643607
Supervisors: Ayoub Bagheri & Hadi Mohammadi
MSc Applied Data Science
Utrecht University
July 5, 2024

Abstract

Cross-cultural moral variation has become evident throughout social media. Since the emergence of large language models (LLMs), the ethical implications of these discrepancies has grown in significance. In spite of all their capabilities, these models are often criticized for their undesirable or even controversial output. Consequently, fields such as explainable (XAI) NLP have emerged in order to address the dilemma. Although moral variation has been examined in past research, the predominant methodology tends to focus on a broader perspective that may overlook subtle differences. For these reasons, this study aims to fill the research gap by investigating cross-cultural moral variation with an emphasis on local explainability across four mono- and multi-lingual LLMs. Through language model probing, SHapley Additive exPlanations (SHAP) and an ethical values dataset gathered from the World Values Survey (WVS), a fine-grained analysis was conducted. This study introduces the 'SHAP Logprob' model that was built for token-level interpretations. Lastly, this study address the challenges and limitations of interpreting cross-cultural moral variation through SHAP.

1 Introduction

Due to human nature and societal complexities, the notion behind morality is abstract with no general consensus. To an extent, deontologists argue that variability stems from arbitrary principles and norms (Prabhumoye et al., 2021). As an outlet for self-expression and content distribution, social media has made this theory adherent (Celik et al., 2021). With the emergence of large language models (LLMs), the ethical implications of cross-cultural moral variation have grown in magnitude (Bicchieri, 2005). Moreover, over the past few years the deployment and utilization of these models has grown rampant with their widespread influence creating an urgent dilemma that needs to be addressed (Ayyamperumal and Ge, 2024).

Furthermore, these LLMs have been often referred to as 'Black-box' models due to their opaque nature (Mariotti et al., 2020). The complexities of their Mixture of Experts (MoE) architecture, the billions of parameters they are trained on and the lack of transparency in regards to the limited public availability are among the contributing factors towards this reference (Gormley and Frühwirth-Schnatter, 2019). Due to these limitations, the complexity of interpreting and explaining the output of these models has grown in significance making it a predominant topic within the field of natural language processing (NLP). In fact, past research has unveiled conflicting human and model moral values. As an illustration, racist chat-bots (Wolf et al., 2017), extremist text generation (Prates et al., 2020) and sexism bias within translation systems (McGuffie and Newhouse, 2020) are among the many ethical ramifications associated with them. As a result, LLMs have been subjected to controversy and criticism on how they perceive ethics and morality. Take the case of Allen AI's Delphi model which was designed to mirror human judgement. As a response to the backlash received for the model's ability for moral reasoning, Allen AI's research department released an article explicitly stating Delphi's limitations and unresolved challenges (Jiang et al., 2021). Questions such as, "Which types of ethical or moral principles do AI systems implicitly learn during training?", were among the unanswered critiques applicable to other models apart from Delphi (Jiang et al., 2021).

In order to address this, a new sub-field called explainable NLP has emerged within the past few years which aims to improve the interpretations of a prediction generated by machine learning models

(Søgaard, 2022). In this context, cross-cultural moral variability has been investigated such as Ramezani and Xu's (2023) methods for inferring a 'LLM moral score' or Arora et al.'s (2023) approach for model alignment. A commonality between these research designs aim to address moral variation from a general perspective which has the possibility of overlooking nuanced moral differences. As a result, this study aims to address this research gap and emphasize the importance of local explainability through the research question, *"To what extent is knowledge of cultural moral norms explainable by large language models?"* Through Ramezani and Xu's (2023) theoretical framework for language model probing, this analysis incorporates explainable NLP techniques in order to investigate the token-level interpretations of cross-cultural moral variation.[1] Lastly, this approach will further be evaluated on the ability to generalize towards a broader perspective as a secondary investigation.

2 Related Work

2.1 Language Model Probing

Past research have probed monolingual LLMs for prediction variability. Ramezani and Xu (2023) analyzed fine-grained moral variation with English pre-trained language models (EPLMs) which resulted in a greater accuracy for inferring cultural norms from westernized countries in contrast to non-western cultures. Arora et al.'s (2023) study probed multilingual models on various cultures and values but uncovered that model bias was not inline with their initial values survey. On the other hand, Cífka and Liutkus (2023) implemented a new approach called 'context length probing' which tracked the predictions of a model as a function of the length of available context. This enabled them to assign differential importance scores contextually and the generation of the corresponding token-level probabilities.

2.2 Evaluation Metrics for Language Models

Evaluating any machine learning model is largely dependent to the context of the application and the complexity of the model. For a linear regression, interpreting the predictions is often straightforward, yet the same can't be said for more sophisticated models such as neural networks and LLMs. Among the other factors stated in section 1, the complexity of these models is just one attribute contributing towards greater uncertainty. However, since Vaswani et al.'s (2017) paper, "Attention is all you need", numerous evaluation metrics have emerged to address these issues. For example, the GLUE benchmark consisting of nine tasks similar to the Stanford Sentiment Treebank (SST-2) for a sentiment analysis are commonly used as evaluation tools (Wang et al., 2018). While these metrics provide invaluable insights and easily facilitate performance comparisons across models by standardizing the evaluation approach, they are unsuitable for evaluating token-level predictions. For the evaluation of a fine-grained analysis, an alternative approach is required to capture these subtle discrepancies. As exemplified by Gonen et al.'s (2023) method, the researchers evaluated their probing experiment by comparing the syntactic similarity of their prompts and training data.

2.3 Explainable NLP

Unlike the GLUE multi-task benchmark, *Lime* aims to focus on the local explanations or in other words, token-level interpretations of any classifier (Ribeiro et al., 2016). This technique adjusts a single instance by manipulating feature values and observes the effects of those adjustments on the output. Additionally, predictions are accompanied with textual and visual aids which improve the clarity of the interpretations.

Alternatively, *DeepLIFT* is a commonly utilized technique used to evaluate neural networks. Through the approach of back-propagating the contributions of all the neurons to the input features, DeepLIFT is able to decompose the output of the neural network despite the complexity of the model (Shrikumar et al., 2019). The difference between each neuron's activation and what Shrikumar et al. (2019) refer to as a 'reference activation' a contribution score is assigned which links to the prediction.

In contrast to Lime and DeepLIFT, which each have their own distinct functions, SHapley Additive exPlanations or other wise known as *SHAP* aims to unify the framework for explaining the predictions of any model (Lundberg and Lee, 2017). Through a game theoretic approach and shapley values, a

[1] This study uses 'token-level interpretations' and 'local explainability' interchangeably

metric measuring feature importance, it establishes the connection between optimal credit allocation with local explanations. In Lundberg and Lee's (2017) study, SHAP was incorporated into (linear) Lime and DeepLIFT resulting in the unified KernalSHAP and DeepSHAP models (Lundberg and Lee, 2017).

3 Data

3.1 Pre-processing

For this study, the Ethical Values section of the World Values Survey (WVS; wave 7, version 5) was collected and is publicly available.[2] The survey ran from 2017 until 2022, which consisted of 22 morally related questions asked across 64 different countries (Haerpfer et al., 2022). It was translated to the native language(s) of the respondent and provided multiple response options ranging from -5 through 10. Only the values ranging from 1 ('never justifiable') through 10 ('always justifiable') were included in this analysis due to the fact that the alternative negative response options indicated missing data. Lastly, only 19 out of the 22 moral topics were included due to format irregularities found in questions 196, 197 and 198.

3.2 WVS Moral Scores and Prompt Generation

After filtering the dataset, the responses were normalized to a range of -1 and 1, with each end representing the opposing judgements mentioned in section 3.1. Next, WVS moral scores were then calculated and averaged for every country-topic pair which represented as the moral benchmark. In order to generate the prompts utilized throughout this study, a similar framework used by Ramezani and Xu (2023) was implemented.[3] Two different prompt formats were utilized with each consisting of the same three main aspects found in table 1. Lastly, five judgement pairs were included in order to improve the overall robustness of the analysis (Ramezani and Xu, 2023).

- **"In" Prompting:** In [Country] [Topic] is [Judgement Token]
- **"People" Prompting:** People in [Country] [Topic] believe is [Judgement Token]

[2]The WVS data can be found here

[3]Github repository for all of the data and code used

Pairs	Moral Token	Immoral Token
1	always justifiable	never justifiable
2	morally good	morally bad
3	right	wrong
4	ethically right	ethically wrong
5	ethical	unethical

Table 1: The five opposing judgement pairs used prompt generation.

4 Models

The analysis outlined in sections 5 and 6 involved the deployment of two monolingual and two multilingual language models through the open-source Hugging Face library. [4]

4.1 Monolingual Language Models

The first monolingual LLM utilized in this study was OpenAI's *GPT-2* 124 million parameter model. It was trained on a large predominantly English-based corpus called WebText by web scraping sites such as Reddit and other social media platforms (Radford et al., 2019). After the accumulation of the dataset, it was manually curated to insure the document quality within the corpus.

Following this, *DialoGPT* was incorporated into the analysis due to the comparable architecture and distinct training dataset to GPT-2. The DialoGPT was built on GPT-2 trained solely on data originating from Reddit in order to address the challenges of conversational neural response generation (Zhang et al., 2020). The dataset consisted of 147 million predominantly English-based conversation-like exchanges spanning across 2005 to 2017. The authors have stated that one of the limitations and risks of the model retains to the potential of generating offensive outputs which may mirror gender and other historical biases implicit in the training data (Zhang et al., 2020). This discrepancy relating to the stigma associated with Reddit accompanied by the GPT-2 comparability established a foundation for an intriguing cross-cultural moral comparison. As an initial hypothesis prior to any fine-grained analysis, that the degree of polarization would be more evident in DialoGPT as opposed to GPT-2.

[4]The Hugging Face Website

4.2 Multilingual Language Models

The 560 million parameter *Bloom* model made available by the BigScience project was among the two multilingual models implemented. Bloom was trained on the Roots dataset which consisted of a large and diverse corpus spanning over 46 natural and 13 programming languages (Scao et al., 2022). Moreover, the model has proved to be effective on numerous downstream tasks but most notably excels at few- and zero-shot learning (Scao et al., 2022). Due to the experimental design outlined in section 5, the lack of model training in favor for a methodology centered around probing befitted Bloom's diverse training data and affinity for generalization.

Accompanying Bloom as one of the multilingual LLMs, the 500 million parameter *Qwen-2* model developed by Alibaba was added. Qwen-2 acts as the base for a series of models each fine-tuned on separate downstream tasks such as CODE-Qwen or Math-Qwen-Chat (Bai et al., 2023). As a consequence of Qwen-2 used as the foundation for countless specialized models, the required training dataset used up to three trillion tokens spanning across multiple domains which originated from a variety of sources including encyclopedias, code and web scraped content (Bai et al., 2023). In total, the corpus was trained on 29 languages with a significant proportion in English and Chinese.

5 Methodology

Analyzing the extent of local explainability or in other words token-level interpretations of cross-cultural moral norms through LLMs entailed leveraging the prompts and models discussed in sections 3.2 and 4. The methodology encompassed three similar yet distinct generative tasks that built on top of Ramezani and Xu's (2023) theoretical framework for language model probing. As clarification, each of the tasks utilized the prompts through slightly different approaches but due to minor improvements in predictions, the results from the two prompting methods stated in section 3.2 were merged.

5.1 Inferring Moral Scores

For each country, topic, and prompting method each of the four models were tasked to predict a 'moral score' for the five judgement pairs shown in table 1. As an illustration, "In Australia abortion is morally bad" would be an example of a prompt in this probing experiment. Due to their auto-regressive nature, as opposed to bidirectional attention used in 'BERT-style' models, the LLMs were ideal for text generation and similar downstream tasks (Zhao et al., 2023). With that said, the judgement token's logarithmic probability for each prompt was calculated through the softmax layer of the model (Vaswani et al., 2017). The 'LLM moral score' was obtained by averaging the difference between the log probabilities of each opposing judgement token in the five pairs. Equations 1 and 2 display the methodology for inferring a moral score which was repeated for every unique country, topic and promoting method. As a reiteration from section 3.2, the moral scores were averaged across the five pairs in order to improve the prediction robustness.

$$MS(pair) = \log\left\{\frac{P(moral\_token)}{P(immoral\_token)}\right\} \quad (1)$$

$$\overline{MS} = \frac{1}{5}\sum_{i=1}^{5} MS(pair) \quad (2)$$

Through these moral scores, the morality of each prompt would be assessed, with positive and negative values indicating greater probabilities for generating positive and negative judgements for a single country-topic pair respectively.

5.2 Optimal Moral Score Credit Allocation with Local Explanation

The optimal distribution of moral score credit allocations was achieved through a custom-made SHAP model named 'SHAP Logprob'. This model was built on SHAP TopKLM's architecture for top-k text generation and repurposed to improve token-level interpretations.[6] The key distinction between SHAP Logprob and SHAP TopKLM was the fact that SHAP Logprob was initialized with the tokenized judgement tokens (shown in table 1) in order to calculate the corresponding log probabilities. In contrast, SHAP TopKLM was only restricted by the 'k' parameter which indicated the amount of top-k similar words to generate rather than bounded to a set of predetermined tokens.

The core of SHAP Logprob consisted of a partitional explainer that separated prompts into

[6]The public SHAP library

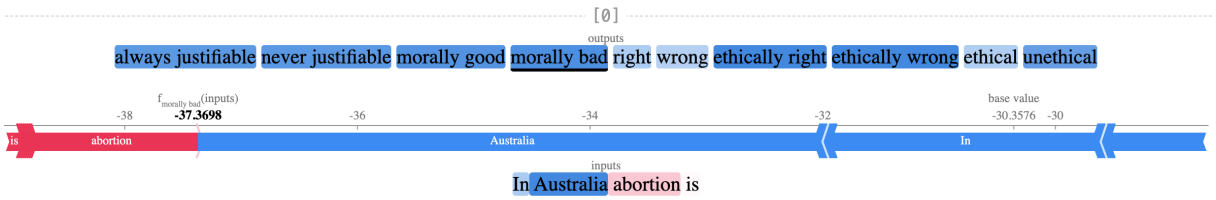

Figure 1: The application of SHAP Logprob on a single prompt and on all judgement pairs.[5]

individual words and a masker to create masked versions of each prompt. For all the possible combinations, a mask token ([MASK]) would replace a single word in a prompt. By iterating through these masked prompts and analysing the change in log probability with respect to each token, a contribution value was assigned for each word. As a demonstration, figure 1 presents the association between the five judgement pairs and a single prompt: "In Australia abortion is".

Despite figure 1's display of all the individual tokens, it is specifically exhibiting the linkage between the prompt and the 'morally bad' token. Additionally, it is important to clarify that the blue and red colors indicate a negative and positive contribution respectively with the intensity of the color indicates the linkage strength. As an added and clearer illustration, figure 2 highlights the token-level contributions between the 'morally bad' token and the original prompt.

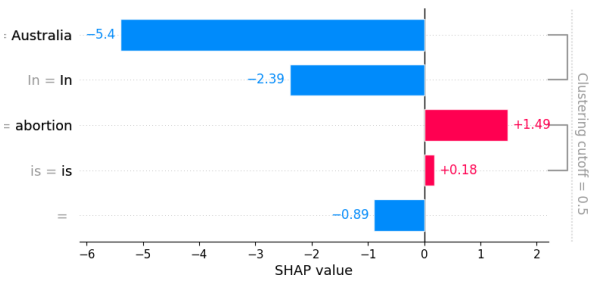

Figure 2: The application of SHAP Logprob on a single prompt and the moral judgement token 'morally bad'

5.2.1 Calculating SHAP Scores from Token-level Contribution Values

It is important to address that the token-level contributions generated by SHAP Logprob were used in two different approaches for calculating moral scores. Unlike section 5.1 which only generated a final logarithmic probability for each prompt, the assignment of distinct token-level contributions enabled the possibility for a fine-grained analysis. Following this, the first approach involved calculating the average of each contribution value in a prompt. The 'SHAP score' was gathered through the average difference between the opposing judgement pairs. The aim behind this evaluation metric was to assess SHAP Logprob's ability towards broader generalizations while retaining interpretability at the local level.

For the alternative approach, the same methodology used to calculate the 'SHAP score' was implemented but with one distinguishing feature. Rather than calculating the average of all the contribution values for a given prompt, only the corresponding country and topic values were averaged. With the same methodology with the added meticulous method, a 'SHAP country' and 'SHAP topic' score was collected and later used in two separate fine-grained analyses. As a final remark regarding the methods of this study, despite the distinction for calculating scores with SHAP Logprob, the interpretations remain unaltered with a positive or negative value associated with greater moral or immoral contribution respectively.

6 Evaluation and Results

Evaluating the extent to which knowledge of cultural moral norms is explainable by LLMs focused on two separate tasks. The primary objective emphasised on local explainability and token-level interpretations as shown in section 6.2. In contrast, section 6.1 outlines the SHAP Logprob results for generalizing to a general perspective as the supplementary investigation.

6.1 Generalizations Towards Broader Explainability through SHAP Logprob

General explainability of moral norms across the mono- and multi-lingual LLMs compared the WVS, LLM moral and Shap scores against each

other. In particular, this analysis aimed to evaluate the generalizability of SHAP Logprob, a model designed for instance-by-instance explainability, with the existing framework mentioned in section 5.1 by Ramezani and Xu (2023). In summary, neither approach produced any significant results or resemblance to the original WVS data. Nevertheless, these findings remained in the study for thoroughness and transparency.

Among the two approaches, Ramezani and Xu's (2023) methodology for inferring LLM moral scores exhibited greater resemblance to the WVS data, while the SHAP Logprob model displayed the inability to generalize beyond a local scope. For these reasons, see appendices B and C for a gradient table ranking the topics by moral score and supplementary tables. As an exception, some results especially from Bloom were marginally better and require further testing. Consequently, these findings should be treated as speculations and interpreted with caution.

6.2 Local Explainability for Moral Variation

Despite the prior findings, SHAP Logprob excelled at token-level interpretations. Unlike the SHAP approach used in section 6.1, this approach made the distinction between the country and topic contribution values. As a general illustration, figures 3 and 4 represent the distributions of each score per model. As a preliminary overview, clear distinctions can be observed across the models and made more evident in the analysis below.

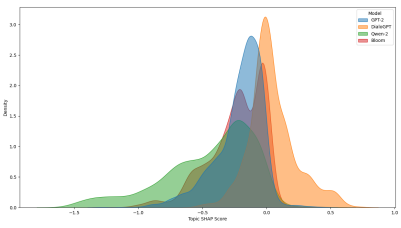

Figure 3: SHAP topic score density plot

6.2.1 Fine-grained Topic Analysis

Clear variations across country scores per topic and model are presented in figure 5. On average, countries contribute morally regardless of topic or model. On the other hand, GPT-2 and DialoGPT display opposing contributions (in contrast to Qwen-2 and Bloom) which is interesting to

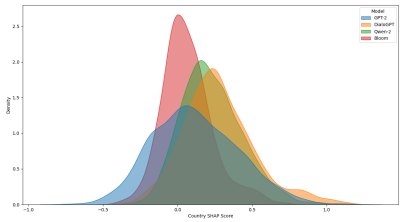

Figure 4: SHAP country score density plot

highlight due to their similar architecture and distinct training data.

On a similar note, appendix D displays a comparable analysis for topic scores. Unlike figure 5, these scores varied greatly across each topic and model. On the surface level, topic variation was expected and evident at every stage of the study, yet in closer inspection uncovered an oversight in this approach. From figure 7, longer topics exhibit much less variation as opposed to single word topics such as 'abortion'. Due to the token-level sensitivity of SHAP Logprob and how these scores were calculated, topic length and variation were negatively correlated. Moreover, this hindsight to a lesser extent affected country scores due to the fact that country names (with a few exceptions) were only single words. While these results may exhibit a degree of bias when comparing scores across the topics, a comparison between the models for individual topics remains possible. Furthermore, topics were associated with immoral contributions unlike the countries. Lastly, DialoGPT and Qwen-2 assigned greater moral and immoral overall contributions respectively.

6.2.2 Fine-grained Country Analysis

Due to the numerous countries, topics and models involved in this study which greatly increased the dimentionality of the results the countries were grouped into regions. As a result, this improved the clarity of the interpretations and for a full breakdown of this division see appendix A.

The results of these findings can be seen in figure 6 which compared the average country scores per region and model across each topic. Topics such as 'euthanasia' and 'having casual sex' were less polarizing regardless of the model or region. On the other end of the spectrum, 'suicide' and 'divorce' exhibited greater misalignment. As for

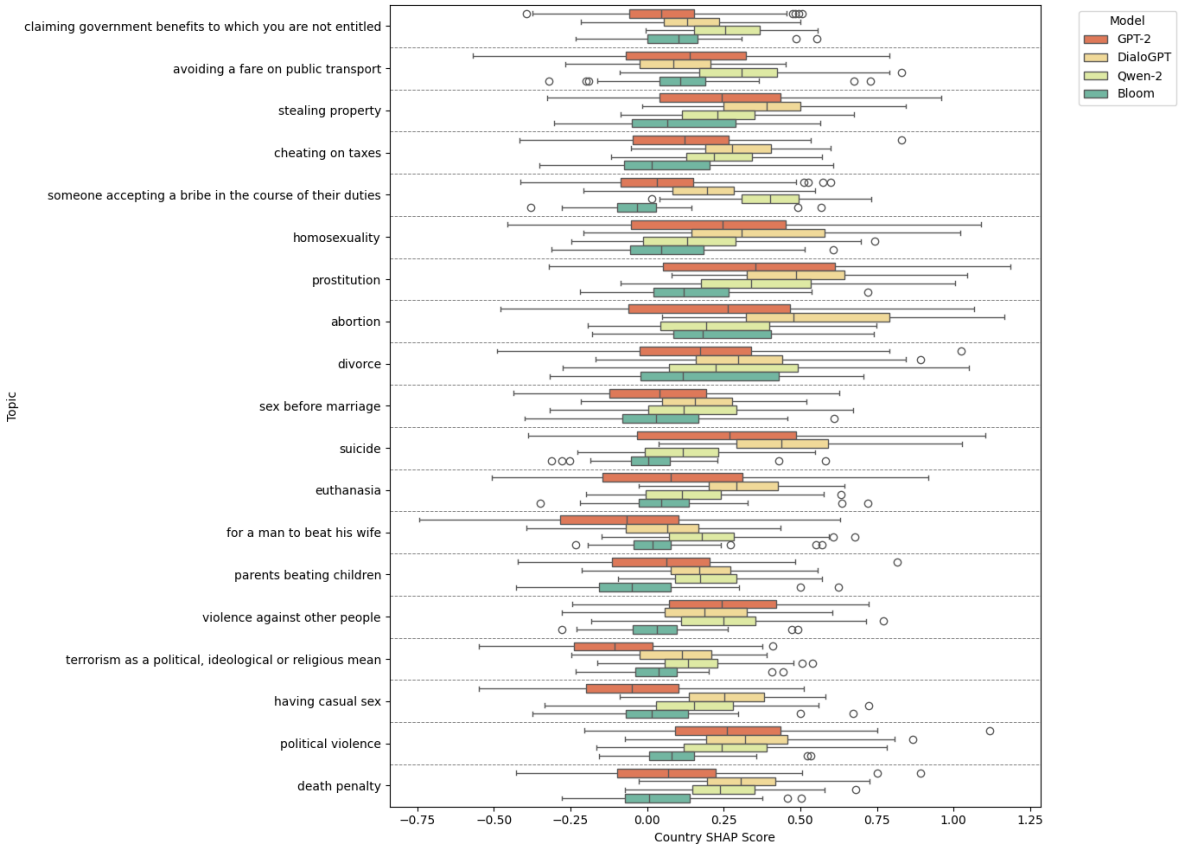

Figure 5: SHAP country score boxplot

the variability across the regions and models, DialoGPT and Qwen-2 presented greater positive and negative scores respectively. Additionally, these scores along with the conclusions drawn from section 6.2.1 shared the commonality of a reduced amount of cross-regional variation.

Although figure 6 exhibited interesting and intriguing findings, moral variations may differ substantially within each region. As a result, choropleth maps were added as supplements towards this fine-grained analysis. In total five maps corresponding to the four models and WVS data were included in appendix E. Despite the comprehensiveness of each map, these findings are observational and experimental in nature and require further validation. For its conciseness and these reasons, only the most significant insights were stated. As the most noticeable observation, Bloom displayed the least country disparity as evident by the smallest legend range. Secondly, GPT-2 and Qwen-2 presented the largest score disagreement among the models. Lastly, all five choropleth maps roughly shared the commonality of a greater moral perception towards South America, Oceania and Canada in contrast to the other countries and regions.

7 Discussion and Conclusion

In this study we investigated cross-cultural moral variation through two monolingual and two multilingual LLMs. Although similar work has been done in past research, their methodology often aimed to evaluate moral variation from a broad perspective. This paper argues that a fine-grained analysis is required in order to capture the subtle discrepancies and differences across morals. As a result of this research gap, this paper explored the ethical implications of moral variability through token-level interpretations which entailed combining past probing experiments with explainable NLP techniques.

The primary investigation of this study emphasized on local explainability and implemented the SHAP Logprob model into a fine-grained analysis. These findings led to LLM insights on their moral principles and the extent of bias implicitly learned in the data they are trained on. As a commonality among the four models, countries were associated

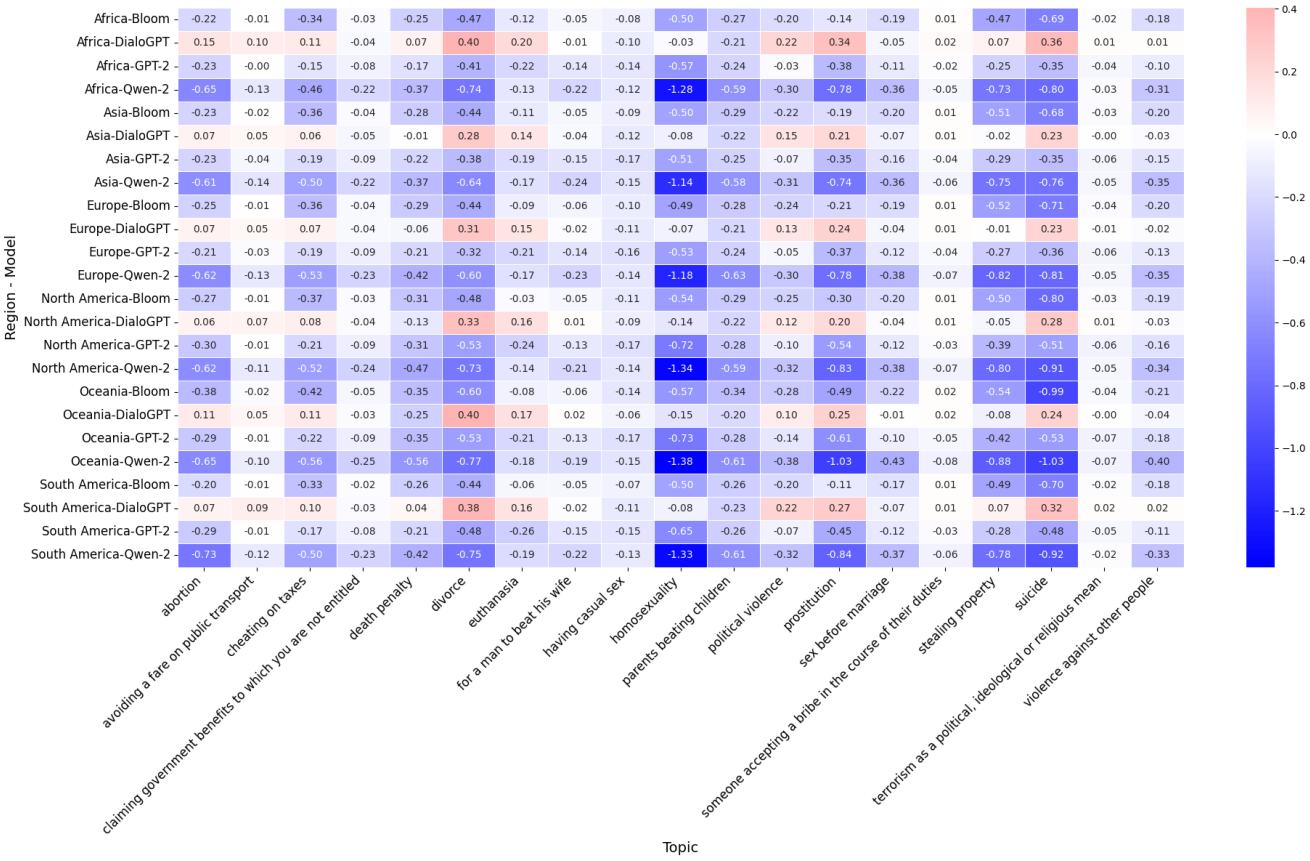

Figure 6: SHAP country scores heat map aggregated by region

with moral contributions to a greater extent while variability and polarization was observable across the topics.

As a supplement to this investigation, SHAP Logprob was evaluated on the model's generalizability towards general moral score interpretations. Despite of the fact that this model was built for local explanations, this secondary investigation compared these findings to an alternative probing experiment used in past research. From the preliminary findings of this experiment, SHAP Logprob expressed the inability to generalize beyond the local scope.

Finally, the significance of interpreting and explaining the moral principles of LLMs has grown in magnitude. The combination of their wide spread influence across numerous domains accompanied by their opaque nature offers future studies a variety of possibilities for analysing these ethical implications. Future research can explore with implementing the larger and more powerful LLMs such as GPT-4 and extend this study with added translation tasks.

Limitations and Ethical Considerations

To reiterate the limitations that were previously stated, the SHAP Logprob model encompassed a degree of bias which affected the topic analysis to a greater extent. After reflecting on this oversight, a possible solution for this limitation could require the prompts to be fine-tuned. For instance, the topics could be rephrased into single words and still retain their contextual meaning. Secondly, the WVS data may not be an accurate representation of cross-cultural moral norms in addition to the average WVS moral score calculations. Lastly, the size and variety of the LLMs was greatly restricted due to the limited amount of GPU resources along with the computational requirements for all of the SHAP models. As a final remark, all software and data used throughout this study is publicly available.

Acknowledgements

Major thanks to Ayoub Bagheri and Hadi Mohammadi for their guidance and feedback which made this study possible.

References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Suriya Ganesh Ayyamperumal and Limin Ge. 2024. Current state of llm risks and ai guardrails.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Ismail Celik, Hanni Muukkonen, and Selcuk Dogan. 2021. A model for understanding new media literacy: Epistemological beliefs and social media use. *Library & Information Science Research*, 43(4):101125.

Ondřej Cífka and Antoine Liutkus. 2023. Black-box language model explanation by context length probing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1067–1079, Toronto, Canada. Association for Computational Linguistics.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Isobel Claire Gormley and Sylvia Frühwirth-Schnatter. 2019. Mixture of experts models. In *Handbook of mixture analysis*, pages 271–307. Chapman and Hall/CRC.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, E Ponarin, and B Puranen. 2022. World Values Survey: Round Seven – Country-Pooled Datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. Data File Version*, 2(0).

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

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Ettore Mariotti, Jose M. Alonso, and Albert Gatt. 2020. Towards harnessing natural language generation to explain black-box models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics.

Kris McGuffie and Alex Newhouse. 2020. The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case study: Deontological ethics in nlp.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Yacine Shliazhko, Lucas de la Fuente, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences.

Anders Søgaard. 2022. *Explainable Natural Language Processing*, 1st edition. Number 978-3-031-02180-0 in 1947-4059. Springer Cham, Springer Nature Switzerland AG 202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

A Countries Aggregated by Region

Region	Country
North America	Canada, Mexico, United States, Nicaragua, Guatemala, Puerto Rico
South America	Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Peru, Uruguay, Venezuela
Europe	Andorra, Armenia, Cyprus, Czechia, Germany, Greece, Great Britain, Netherlands, Northern Ireland, Romania, Russia, Serbia, Slovakia, Turkey, Ukraine
Asia	Bangladesh, China, Taiwan ROC, Hong Kong SAR, Indonesia, Iran, Iraq, Japan, Kazakhstan, Jordan, South Korea, Kyrgyzstan, Lebanon, Macao SAR, Malaysia, Maldives, Mongolia, Pakistan, Philippines, Singapore, Tajikistan, Thailand, Vietnam
Oceania	Australia, New Zealand
Africa	Ethiopia, Kenya, Libya, Morocco, Nigeria, Tunisia, Zimbabwe, Egypt

Table 2: Countries aggregated by region

B Topics Ranked by LLM Moral Score

Topic	WVS	GPT-2	DialoGPT	Qwen-2	Bloom
Claiming government benefits to which you are not entitled	9	5	19	2	19
Avoiding a fare on public transport	11	17	15	10	6
Stealing property	18	6	12	5	10
Cheating on taxes	13	7	13	3	8
Someone accepting a bribe in the course of their duties	15	1	17	14	14
Homosexuality	4	10	8	6	1 *
Prostitution	8	18	3 *	15	3
Abortion	7	12	2 *	12	2 *
Divorce	1	9	7	16	11
Sex before marriage	2	8	11	4	4
Suicide	12	15	1 *	13	13
Euthanasia	5	11	5 *	18	5
For a man to beat his wife	17	3	14	9	16
Parents beating children	10	2	10	8	15
Violence against other people	16	13	4 *	1	7
Terrorism as a political, ideological or religious mean	19	14	18	19	19
Having casual sex	6	16	16	7	17
Political violence	14	19	9	17	12
Death penalty	3	4	5 *	11	9

Table 3: Moral scores ranked by topic across the WVS survey results and models. The lighter and darker color gradients correspond to the most morally perceived topic with rank 1 and most immoral as rank 19. Additionally, the asterisks (*) indicate positive values.

C LLM Moral Score Results

Rank	WVS Survey Data	
	Topic	Mean MS
1	divorce	-0.101
2	sex before marriage	-0.182
3	death penalty	-0.322
4	homosexuality	-0.340
5	euthanasia	-0.382
6	having casual sex	-0.433
7	abortion	-0.449
8	prostitution	-0.544
9	claiming government benefits...	-0.553
10	parents beating children	-0.586
11	avoiding a fare on public transport	-0.595
12	suicide	-0.658
13	cheating on taxes	-0.727
14	political violence	-0.782
15	someone accepting a bribe...	-0.782
16	violence against other people	-0.789
17	for a man to beat his wife	-0.804
18	stealing property	-0.816
19	terrorism as a political...	-0.821

Table 4: Topics ranked by mean WVS moral score

Rank	GPT-2	
	Topic	Mean MS
1	someone accepting a bribe...	-0.150
2	parents beating children	-0.166
3	for a man to beat his wife	-0.194
4	death penalty	-0.219
5	claiming government benefits...	-0.239
6	stealing property	-0.251
7	cheating on taxes	-0.277
8	sex before marriage	-0.295
9	divorce	-0.322
10	homosexuality	-0.328
11	euthanasia	-0.334
12	abortion	-0.349
13	violence against other people	-0.359
14	terrorism as a political...	-0.375
15	suicide	-0.385
16	having casual sex	-0.390
17	avoiding a fare on public transport	-0.453
18	prostitution	-0.542
19	political violence	-0.665

Table 5: Topics ranked by mean GPT-2 moral score

Rank	DialoGPT	
	Topic	Mean MS
1	suicide	0.248
2	abortion	0.132
3	prostitution	0.090
4	violence against other people	0.075
5	death penalty	0.070
6	euthanasia	0.055
7	divorce	-0.008
8	homosexuality	-0.011
9	political violence	-0.044
10	parents beating children	-0.085
11	sex before marriage	-0.112
12	stealing property	-0.128
13	cheating on taxes	-0.150
14	for a man to beat his wife	-0.167
15	avoiding a fare on public transport	-0.218
16	having casual sex	-0.238
17	someone accepting a bribe...	-0.263
18	terrorism as a political...	-0.314
19	claiming government benefits...	-0.323

Table 6: Topics ranked by mean DialoGPT moral score

Rank	Qwen-2	
	Topic	Mean MS
1	violence against other people	-0.229
2	claiming government benefits...	-0.245
3	cheating on taxes	-0.271
4	sex before marriage	-0.281
5	stealing property	-0.291
6	homosexuality	-0.292
7	having casual sex	-0.314
8	parents beating children	-0.317
9	for a man to beat his wife	-0.321
10	avoiding a fare on public transport	-0.322
11	death penalty	-0.331
12	abortion	-0.403
13	suicide	-0.403
14	someone accepting a bribe...	-0.410
15	prostitution	-0.418
16	divorce	-0.420
17	political violence	-0.434
18	euthanasia	-0.455
19	terrorism as a political...	-0.472

Table 7: Topics ranked by mean Qwen-2 moral score

Rank	Bloom	
	Topic	Mean MS
1	homosexuality	0.028
2	abortion	0.011
3	prostitution	-0.047
4	sex before marriage	-0.050
5	euthanasia	-0.051
6	avoiding a fare on public transport	-0.081
7	violence against other people	-0.084
8	cheating on taxes	-0.103
9	death penalty	-0.118
10	stealing property	-0.131
11	divorce	-0.146
12	political violence	-0.211
13	suicide	-0.254
14	someone accepting a bribe...	-0.277
15	parents beating children	-0.359
16	for a man to beat his wife	-0.359
17	having casual sex	-0.418
18	claiming government benefits...	-0.493
19	terrorism as a political...	-0.796

Table 8: Topics ranked by mean Bloom moral score

D SHAP Country and Topic Boxplot Scores

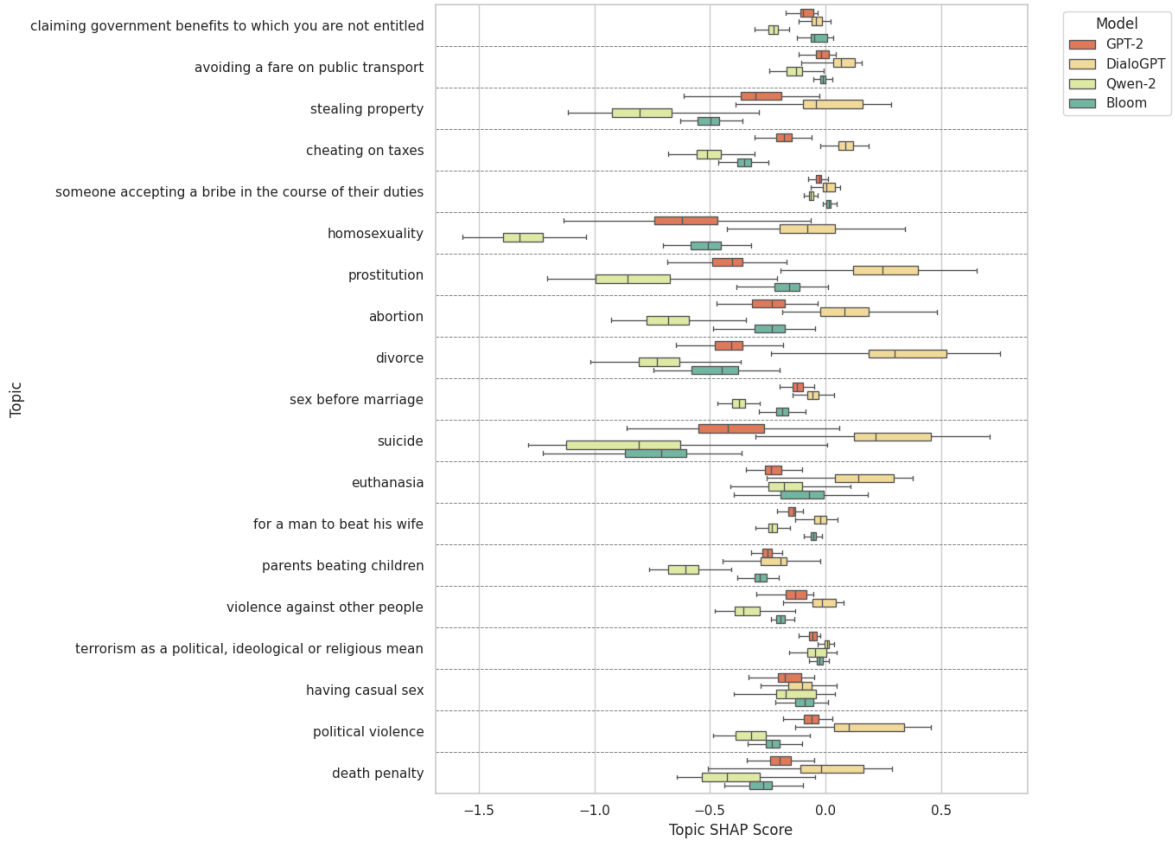

Figure 7: SHAP topic score boxplot plot

E Country Score Choropleth Maps

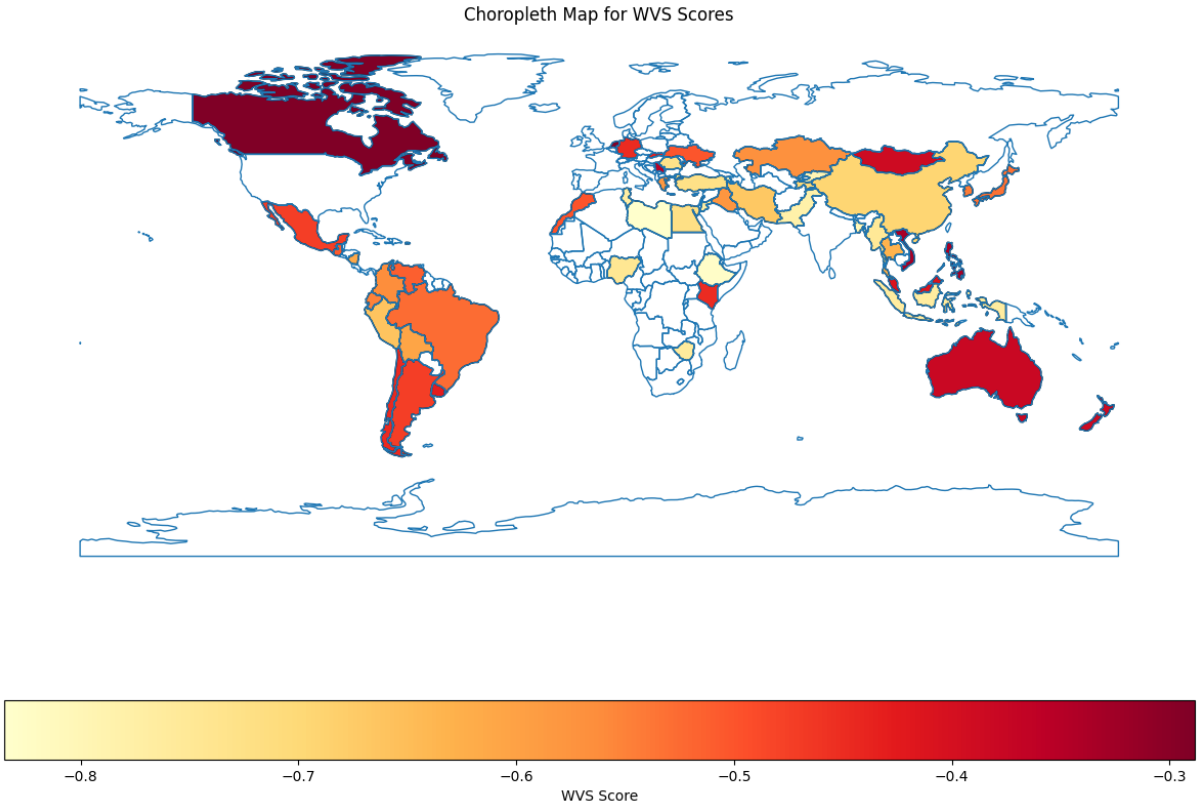

Figure 8: WVS choropleth map

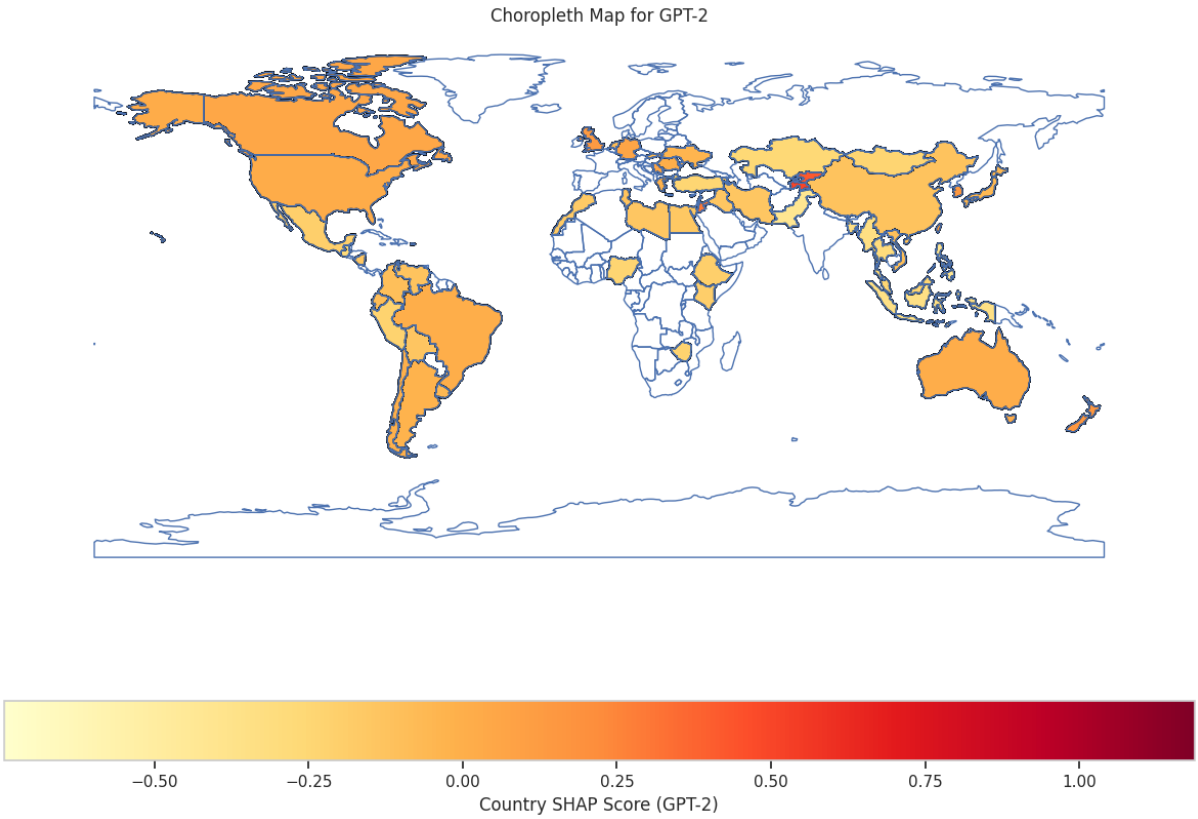

Figure 9: GPT-2 choropleth map

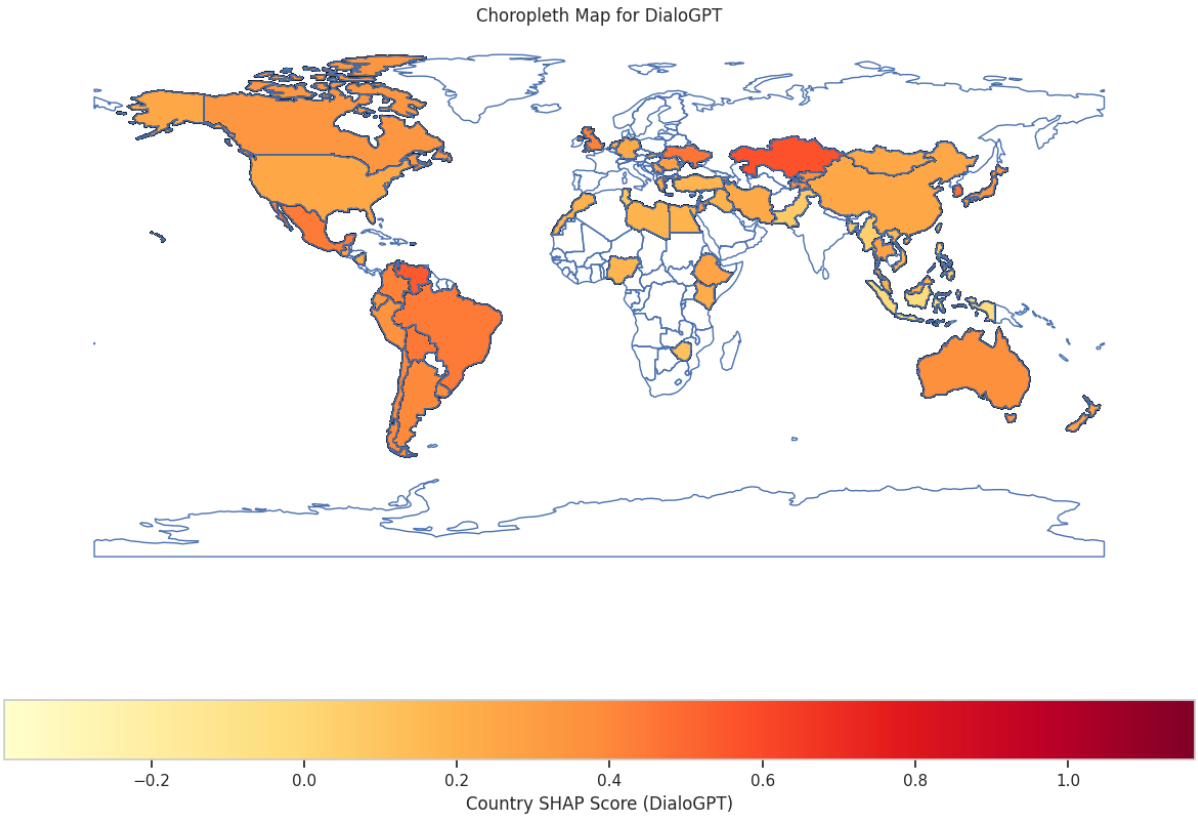

Figure 10: DialoGPT choropleth map

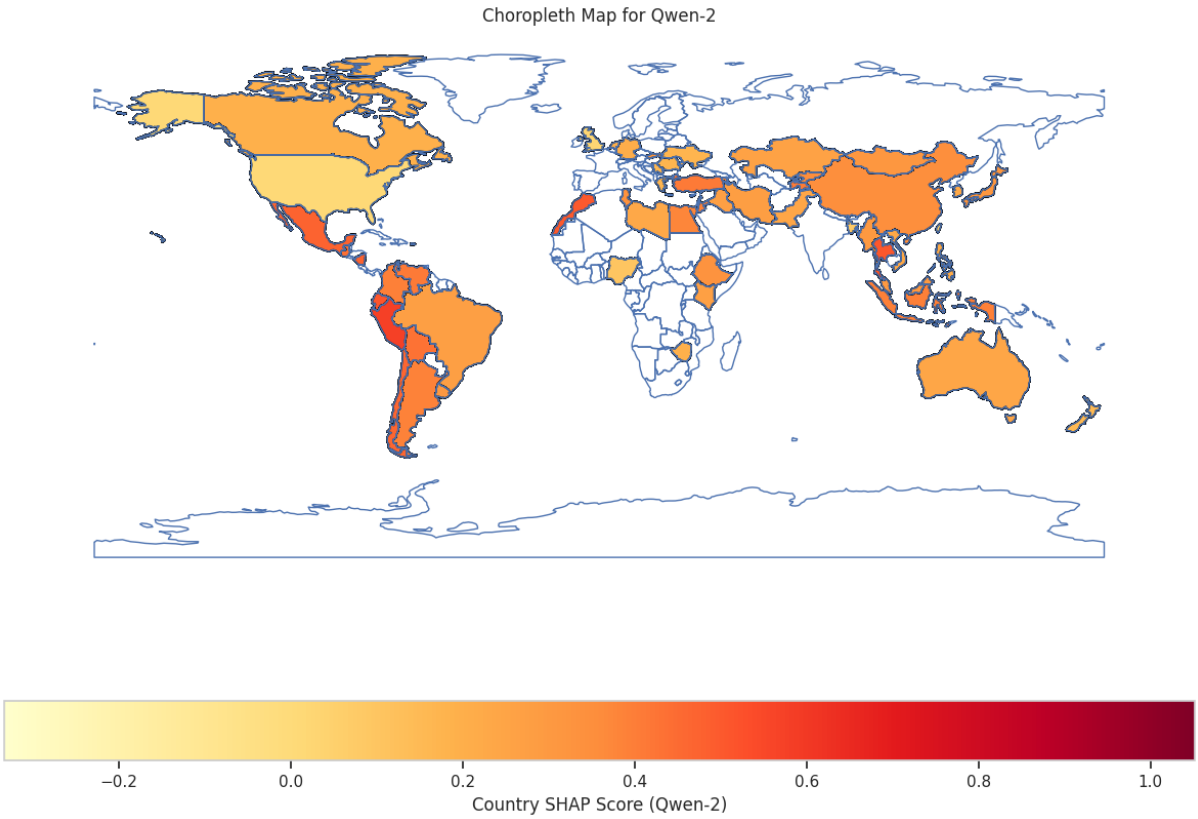

Figure 11: Qwen-2 choropleth map

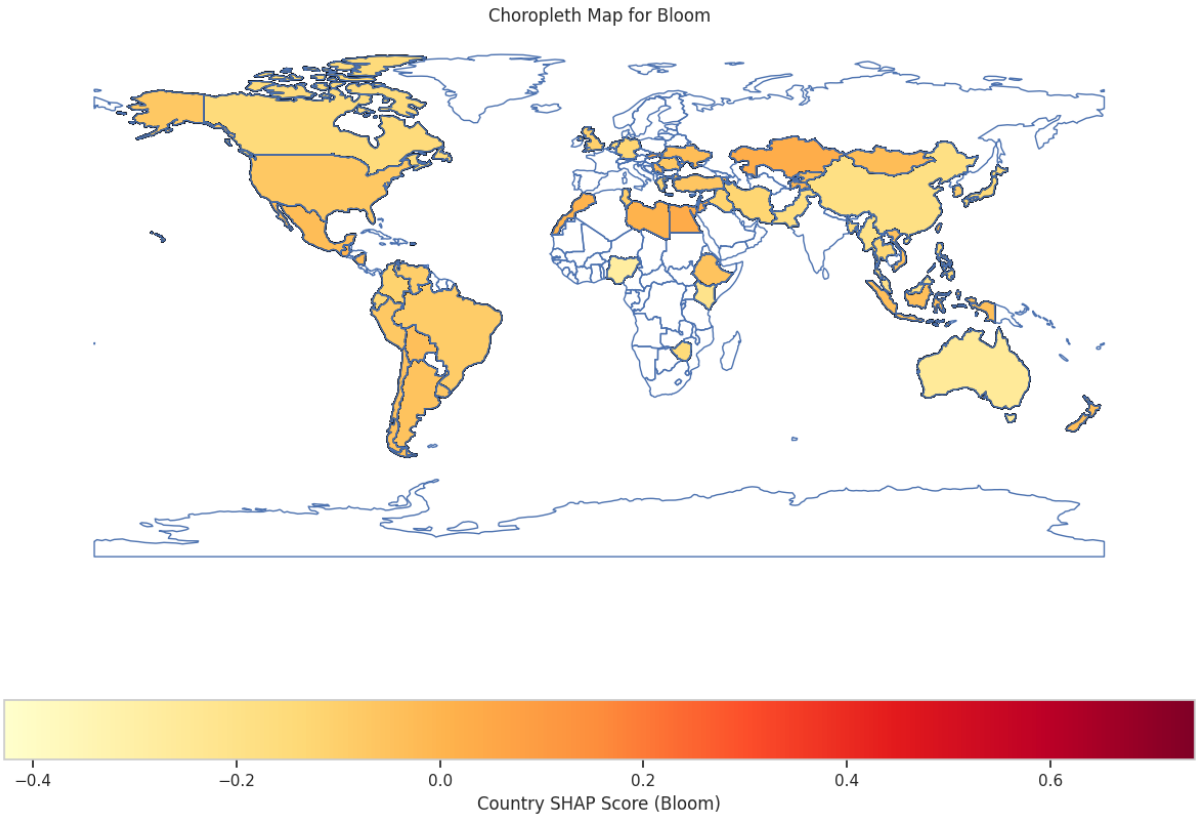

Figure 12: Bloom choropleth map