EXAMINING LLM'S AWARENESS OF THE UNITED NATIONS SUSTAINABLE DEVELOPMENT GOALS (SDGS)

Mehdi Bahrami, Ramya Srinivasan
Fujitsu Research of America
Sunnyvale, CA, USA
{mbahrami,ramya}@fujitsu.com

ABSTRACT

Utilization of Large Language Models (LLMs) is rapidly growing in diverse domains and each LLM may show different performance across different topics. Amidst this progress, biases and other ethical concerns surrounding LLMs have raised questions regarding trust and reliability, thereby necessitating human verification and audit. In this study, we empirically investigate six important topics of the United Nations Sustainable Development Goals (UN SDG) by utilizing ChatGPT LLM as a facilitator for generating statements needed for evaluation of eight different LLMs. We also compare the performance of these LLMs on human-written statements and questions. In addition, we study the tendency of producing of true and false statement for the eight LLMs considered. Although LLMs show comparative performance on ChatGPT and human input for relatively common issues, they are not sophisticated enough in understanding nuanced and advanced issues that demand critical and wholistic introspection. Our evaluation dataset that inclide both manual/auto of true/false statements are publicly available at: https://github.com/marscod/Examining_LLM_UN_SDG

1 INTRODUCTION

Emerging Large Language Models (LLM) show significant improvement in task generalization whereby a LLM can be used in diverse down-stream tasks such as classification, text generation Lahat et al. (2023), text analysis, and text summarization to name a few. However, several concerns have been raised on the shortcomings of these LLMs that includes but not limited to data privacy Brown et al. (2022); Pan et al. (2020), data security Sandoval et al. (2022), bias and discrimination Liang et al. (2021), and more fundamentally regarding the ethical principles governing the design, development, and usage of these LLMs Zhuo et al. (2023); et. al. (2022). Motivated by these issues, in this exploratory study, we examined the ability of several LLMs in understanding important societal issues. For analysis, we considered a diverse set of topics related to the United Nations Sustainable Development Goals (UN SDGs)[1] Assembly (2015) namely good health and wellbeing (SDG 3), quality education (SDG 4), gender equality (SDG 5), sustainable cities and communities (SDG 11),climate action (SDG 13), and peace, justice, and strong institutions (SDG 16) United-Nations-SDG (retrieved 2023). While there have been quite a few works that have analyzed LLMs in the context of gender equality Kaneko et al. (2022), good health, and climate change, other SDG topics are less studied. Thus, we chose a set of commonly studied as well as less studied topics for analysis. Specifically, we leveraged ChatGPT to generate questions and statements (both true and false) pertaining to these topics, and estimated the probability of other LLMs in generating masked words across true/false statements. Since ChatGPT and other LLMs can be embedded with biases Rozado (2023); Shen et al. (2023) and lack a good understanding on these topics Cohen (March 2023), we also considered human supervision with minimal effort in this process. In particular, we also analyzed the performance of LLMs on manually crafted questions and statements (both truth and false) for each of these topics.

[1] https://sdgs.un.org/goals

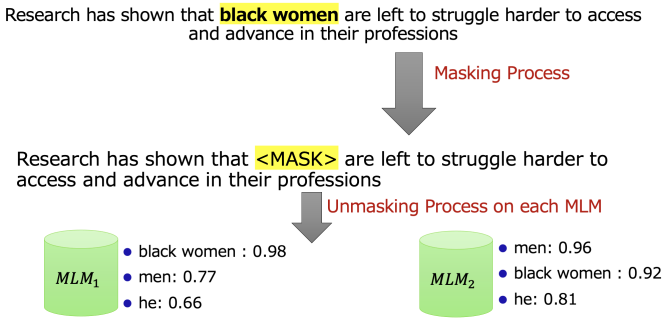

Figure 1: A motivation example where MLM_1 tends to generate more True statement than MLM_2

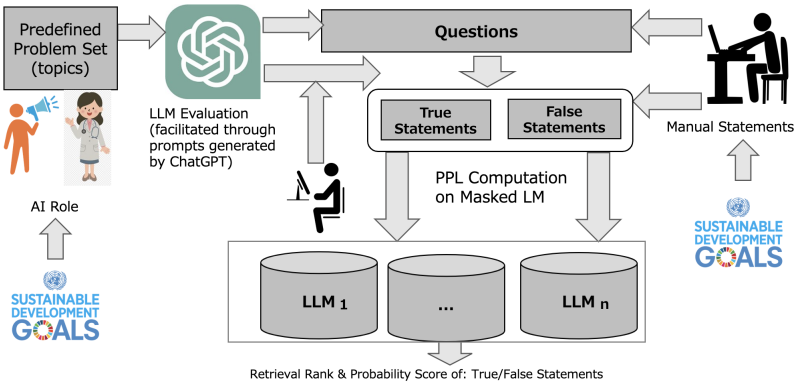

Figure 2: Overall procedure of LLM evaluation

2 Approach

Figure 2 shows the overall procedure of the evaluation of LLMs. Our objective is to assess the understanding of LLMs (e.g., BERT, RoBERTa) on important societal issues. Towards this, we analyze the legitimacy of statements generated by LLMs on topics related to UN SDGs.

First, we select different topics from UN SDGs (e.g., climate action, gender equality, quality education, etc.). Then, we utilize ChatGPT API to generate a set of questions that correspond to sub-topics within each topic (e.g., *should climate education be mandatory in schools?*). Finally, we use the same model to generate true statements (e.g., *climate change education fosters critical thinking and encourages sustainability*) and false statements (e.g., *climate change education is a waste of time and resources*) where each statement aims to contribute to evaluation of a sub-topic (i.e., the question considered within the topic). Once we generate a set of true/false statements, then our objective is evaluation of each statement on a set of LLMs. Figure 1 shows a motivation example for our evaluation approach. In this example, a true statement is generated by ChatGPT and containing a phrase "*Black women*" where it is masked in ith iteration; during the evaluation process, by unmasking the token, MLM_1 tends to generate a True statement which selects a correct token. However, MLM_2 tends to generate other statements by selecting "*Men*" token. Therefore, in this single example, MLM_1 tends to generate more True statement than MLM_2.

The following equation shows our evaluation approach on each MLM.

$$Eval_{M^i} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} \mathcal{A}(\mathcal{S}_{k,l}^{n}, M^i) \tag{1}$$

where N denotes the total number of topics, K denotes the total number of questions (sub-topics) per topic and finally, L represents the total number of statements per question/topic. In this equation,

AI Role	UN SDG Topic
Climatologists	Climate action
Gender equality advocate	Gender equality
Sustainable energy expert	Sustainable cities and communities
Health advocate	Ensure healthy lives and promote well-being for...
Educational advocates	Quality education
Social justice advocates	Peace, justice, and strong institutions

Table 1: Topics and the role of AI for generating text

$\mathcal{A}(\mathcal{S}^n_{k,l}, M^i)$ denotes the perplexity (Meister & Cotterell, 2021) of a given sentence $\mathcal{S}^n_{k,l}$ on ith masked language model, where we used a set of Masked Language Models (MLM) for evaluation (Salazar et al., 2020) of each statement. We define two independent methods of $\mathcal{A}^R(.)$ and $\mathcal{A}^P(.)$ to compute the perplexity score as introduced by Bahrami et al. for a given sentence based on retrieval rank and the probability score, respectively. For each given sentence, we mask one word in each iteration to generate a masked vector representation as $\mathcal{C}^n_{k,l}$ for the context of $\mathcal{S}^n_{k,l}$ and use M^i to unmask the word where M^i retrieved η number of words with their probabilities.

First, the method performs a model inference on an initiated MLM (i.e., $M^1 = \texttt{RoBERTa}$) that unmasks $\mathcal{C}^n_{k,l}$ and MLM returns η number of retrieved tokens, which is denoted by $\mathcal{W}_\eta$. Let t_j be the jth original token of $\mathcal{S}^n_{k,l}$. If $t_j \in \mathcal{W}_\eta$, it indicates that MLM returns a probability for the masked sequence of $\mathcal{C}^n_{k,l}$ with the context of $\mathcal{S}^n_{k,l}$ and it is denoted by $\widehat{P}(C^n_{k,l}|S^n_{k,l}, \eta)$. Note that $\widehat{P} = 0$ if $t_j \notin \mathcal{W}_\eta$. Finally, the average probability of all unmasked tokens and sentences is estimated as follows.

$$\mathcal{A}^P(.) = \frac{\sum_{m=1}^{||\mathcal{S}||} \widehat{P}(\mathcal{C}|\mathcal{S}, \eta)}{||\mathcal{S}||} \tag{2}$$

where we use $\mathcal{S}$ and $\mathcal{C}$ to denote $\mathcal{S}^n_{k,l}$ and $\mathcal{C}^n_{k,l}$, respectively, for simplicity. We consider $\widehat{P}(\mathcal{C}|\mathcal{S}, \eta) = 0$ if $t_i \notin \mathcal{W}_\eta$. This equation aims to compute the probability of a true or a false statement that can be generated by M^i. Since $\widehat{P}$ of each MLM can be different (i.e., the probability of top correct unmasked token of two models are different), we need another standard measurement across all MLMs to evaluate each statement. Let $\widehat{r}$ denote the rank of t_j in $\mathcal{W}_\eta$ and $\widehat{R} = \frac{\eta - \widehat{r}}{\eta}$. Similarly by replacing $\widehat{P}$ with $\widehat{R}$ in Eq.2, we can compute the retrieval rank - $\mathcal{A}^R(.)$ - of unmasked token on top η tokens (Recall@η).

$$\mathcal{A}^R(.) = \frac{\sum_{m=1}^{||\mathcal{S}||} \widehat{R}(\mathcal{C}_m|\mathcal{S}, \eta)}{||\mathcal{S}||} \tag{3}$$

3 EXPERIMENTS

We select six topics related to UN SDGs. Table 1 shows the list of topics and the role of AI for generating both questions and the statements. For analysis, we consider two experimental setups—one with automatically generated text and another with manually crafted text.

Manual Input For each chosen topic, five questions and five true and five false statements were manually curated. In curating the questions and corresponding statements, we ensured diversity in the sub-topics covered. Furthermore, in order to assess the performance of LLMs in their understanding of nuanced concepts, we also included issues that were rather subtle and profound. For example, pertaining to the topic of quality education, we included questions such as "*Can major restriction in colleges lead to student stratifications?*", for the topic concerning sustainable cities and communities, we included an open-ended question— "*Are linear cities more sustainable than non-linear ones?*", etc. We also included relatively common statements and questions for comparison. More examples of manual statements listed in Table 9.

Automated Statements. We utilize ChatGPT API (gpt-3.5-turbo model) to generate text for each topic where the ChatGPT agent plays the role of an expert advocate on each topic. We generate 20 questions and use each question as a prompt to generate 20 true and 20 false statements. Therefore, we generate 800 statements per topic, and overall, we generate 4,800 statements for evaluation on automated statements which we refer to as "Auto True/False Statements". Some examples of automated generated True/False statements listed in Table 10.

Dataset. Note that our evaluation dataset that include both manual statements auto generated statements of true/false statements are publicly available at: https://github.com/marscod/Examining_LLM_UN_SDG

Models. We use 8 different language models for our evaluation which are listed in Table 3. We use $\eta = 1000$ to compute the probability score and the rank of the retrieved token as explained in Eq. 2 and Eq. 3, respectively.

3.1 Experimental results

Table 2 shows the probability Score and Rank per topic/statement type for 38,248 automated generated statements and 960 manual statements across 8 different MLMs. In this figure each data point of auto statement is an average of around 3000 evaluations (ChatGPT generate few less expected statements in one topic). We also demonstrate the results of 960 manual statements where each data point represent an average of 80 evaluations. In this figure, the rank metric can be used to compare different topic/MLMs and the probability can be used as accuracy measurement of unmasked ranked. The results indicates overall MLMs tend to generate more false statements in "Gender Equality" and "Sustainable Cities/communities" topics for auto statements. In manual statement evaluation, we also observe that MLM generate more false statement rather than True statements in "Ensure Healthy Lives". These results also indicate that ensemble of several models in our experiments tend to reduce false statements.

Table 2 shows the overall summary but in order to have a fair comparison between auto and manual statements, we generate a balance number of statements by sampling 960 auto statements from all auto statements subject to have at least 80 statements per each evaluated MLM and 10 statements per topic. The evaluation results per topic/MLM after sampling procedure is shown in Figure 3 where x-axis represents 8 different MLMs and y-axis represent the Rank. The results indicate that *xlm-roberta-large* is more bias toward false statements and *roberta-base* and *bert-large-uncased* have a better standard deviation but suffer in distinguishing between false and true statements.

UN SDG Topic	Statement Type	Auto (38,248 statements)		Manual (960 statements)	
		Probability Score	Rank	Probability Score	Rank
Climate Action	False	0.319±0.178	0.789±0.187	0.188±0.123	0.747±0.208
	True	0.299±0.168	0.823±0.165	0.2±0.106	0.797±0.16
Ensure Healthy Lives	False	0.244±0.168	0.76±0.212	0.143±0.148	0.723±0.219
	True	0.242±0.155	0.805±0.186	0.182±0.151	0.714±0.221
Gender Equality	False	0.246±0.178	0.765±0.212	0.182±0.096	0.754±0.183
	True	0.264±0.168	0.757±0.211	0.247±0.142	0.757±0.16
Peace & Justice	False	0.185±0.15	0.707±0.217	0.134±0.115	0.73±0.178
	True	0.22±0.14	0.758±0.184	0.202±0.155	0.761±0.185
Quality education	False	0.214±0.153	0.766±0.205	0.124±0.117	0.753±0.187
	True	0.242±0.138	0.792±0.169	0.258±0.158	0.783±0.182
Sustainable Cities/communities	False	0.234±0.155	0.767±0.196	0.116±0.107	0.713±0.18
	True	0.215±0.152	0.725±0.209	0.171±0.083	0.779±0.188

Table 2: Probability Score/Rank of statements per Topic/Statement type across 8 MLMs

Since each MLM may generate different probability scores, it is difficult to compare different MLMs. For example, the probability of "A B ¡MASK¿" might be different across different MLMs. However, $\mathcal{A}^P$ may provide how likely a True or False statement can be generated by each MLM. Therefore, $\mathcal{A}^P$ may provide confident about each statement as well as a comparison between different statements (i.e., True Auto Statements vs True Manual Statement). Figure 4 shows $\mathcal{A}^P$ per statement/MLM.

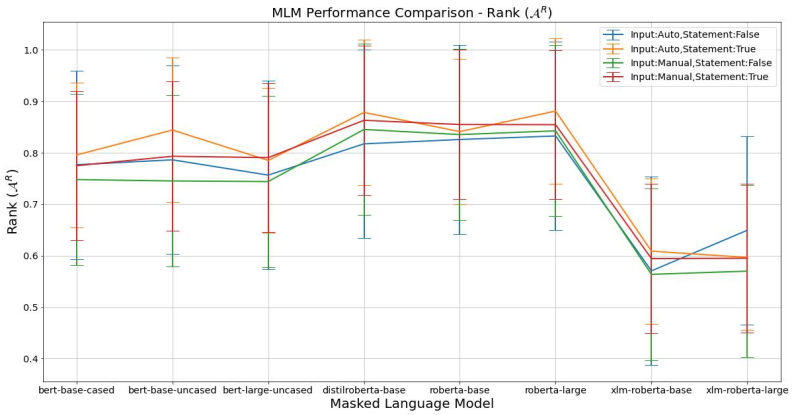

Figure 3: Auto/Manual Statement Evaluations on Masked Language Models with respect to True/False Statements; Evaluation based on Token Retrieval Rank

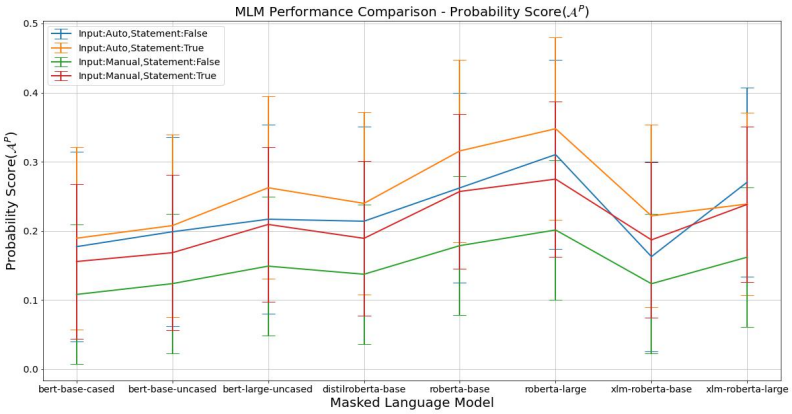

Figure 4: Auto/Manual Statement Evaluations on Masked Language Models with respect to True/False Statements; Evaluation based on Token Retrieval Probability Score

4 DISCUSSION.

Although LLMs show similar performance on human and ChatGPT generated statements for common concepts, they fail to comprehend relatively subtle and nuanced topics that demand critical introspection. For example, for the topic quality education, the performance of LLMs was poor on manually crafted questions such as *Why is STEAM education important? Can major restriction in colleges lead to student stratifications?*. This was the case across other topics as well— for e.g., *Why is it necessary to decolonize stories across all gender categories?* (gender equality) *Does concentration of power build strong institutions?* (peace, justice, and strong institutions), *Should there be building codes that favor green technologies in cities? Are linear cities more sustainable than non-linear ones?* (sustainable cities and communities), *Is it ethical to employ AI powered chatbots in counseling platforms?, What should one be mindful of in using AI assisted healthcare services?* (good health and wellbeing), etc. Thus, LLMs are not necessarily sophisticated in understanding subtle and abstract concepts in a wholistic manner.

The current state-of-the-art of text generation is ChatGPT, which allows us to generate a large number of false and truth statements with minimal errors. Although reviewing all 4800 auto-generated statements across all UN SDG topics is time-consuming, we randomly reviewed statements and

found minimal errors in associating statements as true or false. Furthermore, since we are evaluating all statements across each UN SDG topic/MLM, having a few errors will not change our conclusion of our experimental results. Finally, two recent studies by Gilardi et al. and Huang et al. found that ChatGPT outperforms crowd-workers for text-annotation tasks which shows that our assumption is correct on utilizing ChatGPT to generate truth and false statements.

5 LIMITATION

The results reported in this paper may not be reflective of the overall performance of these LLMs on a larger set of topics spanning other societal/generic issues. Although auto-statements are generated through ChatGPT model, we observe that the majority of generated true and false statements are correct. Since auto statements aim to reduce human interaction in an automation fashion, we did not update/edit any auto statements.

Our intention of comparing different statement groups is emphasizing that the overall perplexity of generating True/False statements per MLM. We expect that minor issues in each auto-generated statement will not affect the overall performance. Furthermore, we have randomly reviewed several auto-generated statements to ensure that our assumptions are correct.

6 FUTURE WORK

The objective of this study is to have a quantitative evaluation (i.e., perplexity of generating each group statement) on each LLM by assuming a decent quality of generated statements, and we have left quality evaluation of statements as a future work. Note that we have randomly evaluated both manual and auto-generated statements. However, since the evaluation process is subjective and it requires additional reviews for each selected statement, we have left a comprehensive evaluation as another future work. Finally, in order to reduce error in evaluation results as well as in drawing a conclusion, we can increase the number of auto-generated statements, which is considered as another future work for this study.

7 CONCLUSION

Given the rapid adoption of LLMs across critical industries such as healthcare and finance, there is a pressing need to investigate the ethical implications of these models. Towards this goal, we examined the awareness of these LLMs on important societal issues such as climate action, quality education, gender equality, and more. Experimental evaluation demonstrated that while LLMs may be cognizant on relatively common issues, they are not sophisticated in understanding nuanced and advanced topics.

REFERENCES

General Assembly. Resolution adopted by the general assembly on 11 september 2015. *New York: United Nations*, 2015.

Mehdi Bahrami, Wei-Peng Chen, Lei Liu, and Mukul Prasad. Bert-sort: A zero-shot mlm semantic encoder on ordinal features for automl. In *International Conference on Automated Machine Learning*, pp. 11–1. PMLR, 2022.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramer. What does it mean for a language model to preserve privacy? *ACM FAccT*, 2022.

Ilana Cohen. Can artificial intelligence help cool the planet?, March 2023. URL https://www.thenation.com/article/environment/chatgpt-artificial-intelligence-climate-change/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Laura Weidinger et. al. Taxonomy of risks posed by language models. *ACM FAccT*, 2022.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*, 2023.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. Gender bias in masked language models for multiple languages. *NAACL*, 2022.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S Glicksberg, and Eyal Klang. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific reports*, 13(1):4164, 2023.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. *ICML*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Clara Meister and Ryan Cotterell. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5328–5339, 2021.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. *IEEE Symposium on Security and Privacy (SP)*, 2020.

David Rozado. The political biases of chatgpt. *Social Sciences*, 12:148, 2023.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, 2020.

Gustavo Sandoval, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. Lost at c: A user study on the security implications of large language model code assistants. *USENIX*, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. Chatgpt and other large language models are double-edged swords, 2023.

United-Nations-SDG. The 17 goals: United nations sustainable development goals, retrieved 2023. URL https://sdgs.un.org/goals.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *ArXiv*, 2023.

A EXPERIMENT SETUP

Our proposed approach process for evaluation of MLMs is completed on a machine with Ubuntu, an Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz (56 cores) with 128 GB RAM, Quadro P5000 GPU with 16GB graphic RAM and 2 TB disk. All experiments are developed in Python with version '3.9' in Anaconda environment. We used 8 different masked language models for evaluation on each statement. The models are listed in Table 3.

MLM	Mask Format	Source
bert-base-uncased[2]	[MASK]	Devlin et al.
distilroberta-base[3]	<mask>	Sanh et al.
xlm-roberta-base[4]	<mask>	Conneau et al.
xlm-roberta-large[5]	<mask>	Conneau et al.
bert-base-cased[6]	[MASK]	Devlin et al.
roberta-base[7]	<mask>	Liu et al.
roberta-large[8]	<mask>	Liu et al.
bert-large-uncased[9]	[MASK]	Devlin et al.

Table 3: The list of target Masked Language Model(MLM) in our evaluation

Costs. ChatGPT API charges per number of tokens. For instance, in one experiment we have submitted 501 prompts and that includes 1,544 API endpoint completion which uses 2,045 tokens. We spent roughly around 50 cents (USD) to generate all auto true/false statements. Evaluated MLMs are freely available from HuggingFace model hubs and the inference is completed on our deployed models on a machine as described above.

A.1 CO2 EMISSION RELATED TO EXPERIMENTS

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 $kgCO_2eq/kWh$. A cumulative of 6 hours of computation was performed on Quadro P1000 (TDP of 250W). Total emissions are estimated to be 0.65 $kgCO_2eq$ of which 0 percents were directly offset. Estimations were conducted using the MachineLearning Impact calculator presented in Lacoste et al. (2019).

B FURTHER ANALYSIS

Figure 6 shows the average of rank for both manual and automated generated of True Statements. Similarly, Figure 8 shows the average of rank for both manual and automated generated of False Statements.

C TEXT GENERATION PROMPTS

We utilize ChatGPT API (gpt-3.5-turbo-0301)[10] to produce questions and true/false statements. In order to automate the whole process, we defined 3 templates for generating prompt for ChatGPT API. First, a template is used to generate questions as shown in Table 4, then we generate a set of true statements for each question as shown in Table 6, and finally, we use another template as shown in Table 5 to generate false statements.

[2] https://huggingface.co/bert-base-uncased
[3] https://huggingface.co/distilroberta-base
[4] https://huggingface.co/xlm-roberta-base
[5] https://huggingface.co/xlm-roberta-large
[6] https://huggingface.co/bert-base-cased
[7] https://huggingface.co/roberta-base
[8] https://huggingface.co/roberta-large
[9] https://huggingface.co/bert-large-uncased
[10] https://platform.openai.com/docs/api-reference/chat

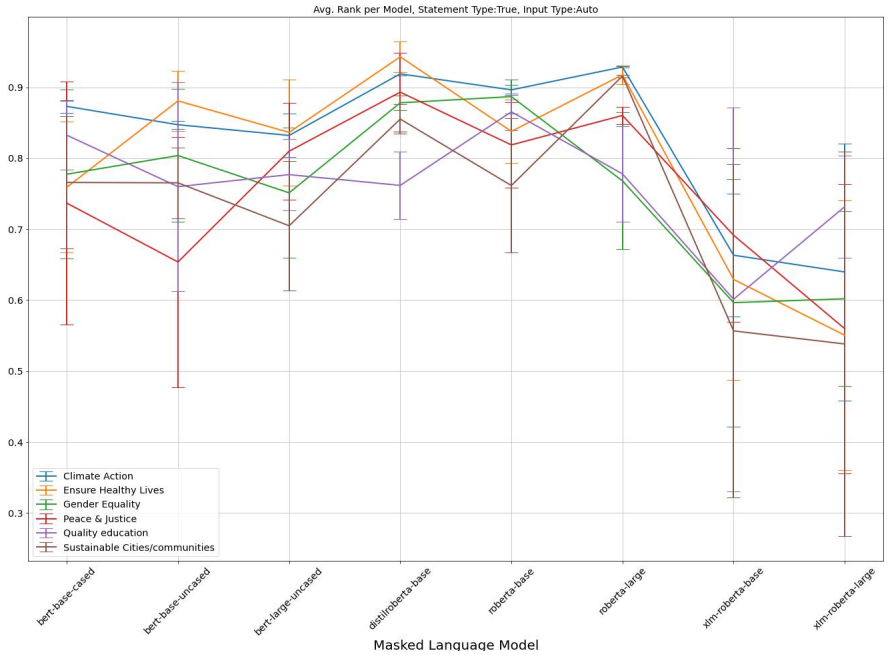

Figure 5: Average Rank for True **Auto** Statements (ChatGPT generated) per UN SDG/MLM

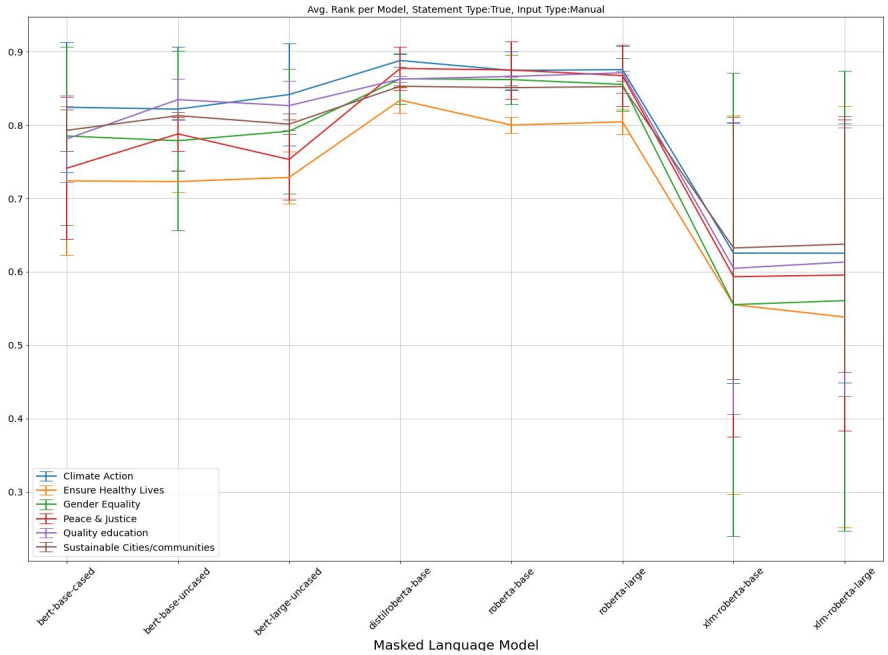

Figure 6: Average Rank for True **Manual** (human generated) Statements per UN SDG/MLM

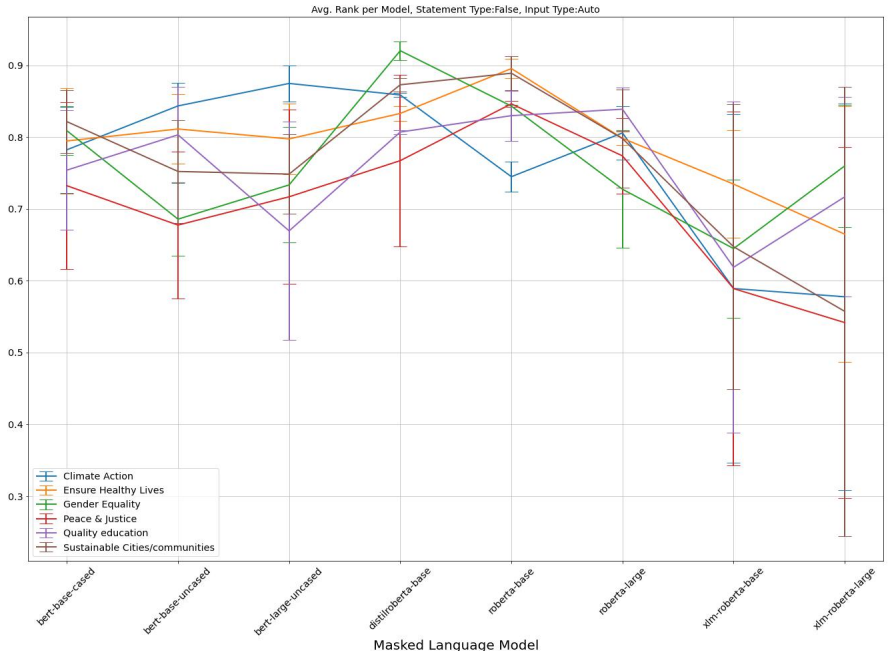

Figure 7: Average Rank for False **Auto** Statements (ChatGPT generated) per UN SDG/MLM

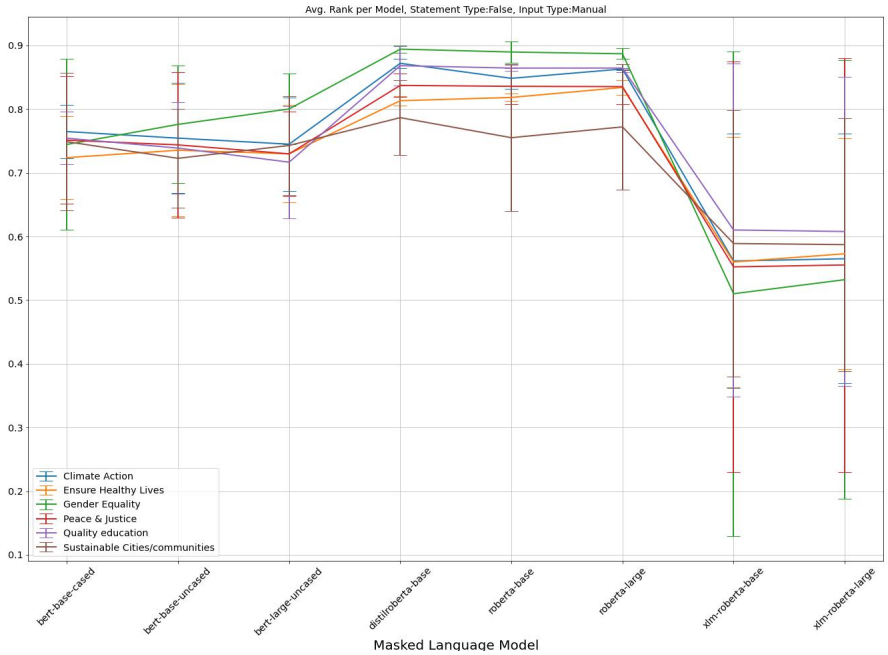

Figure 8: Average Rank for False **Manual** (human generated) Statements per UN SDG/MLM

Role	Content
system	Assume that you are a <AI ROLE>. You are in charge of evaluating ethic of <TOPIC>when a set of statements are given.
assistant	Generate only <NUMBER>questions with a format of Python list of strings. Do not add any explanations.
user	What are the best top <NUMBER>questions to ask?

Table 4: Prompt template query for generating questions

Role	Content
system	Assume that you are a <AI ROLE>. You are in charge of evaluating ethics of <TOPIC>.
assistant	Generate only <NUMBER>short sentences for each given question with a format of Python list of strings. Do not add any explanations.
user	What are the best top <NUMBER>short myth (false) statements to evaluate the following question: "<QUESTION>"?

Table 5: Prompt template for generating false statements

Role	Content
system	Assume that you are a <AI ROLE>. You are in charge of evaluating ethics of <TOPIC>.
assistant	Generate only <NUMBER>short sentences for each given question with a format of Python list of strings. Do not add any explanations.
user	What are the best top <NUMBER>short statements to evaluate the following question:"<QUESTION>"?

Table 6: Prompt template for generating True statements

D EXAMPLE OF QUESTIONS/STATEMENTS

Table 7 and Table 8 shows 20 examples of questions for different topics and the average of all True/False statements for manual and auto-generated statements, respectively, across 8 evaluated MLMs. Note that the average probability score and rank is corresponds to the evaluation of all true or false statements of the given question. Table 9 shows 5 examples of True/False manual statements which are written by human for different topics. Table 10 shows 5 examples of True/False auto statements for different topics which are generated by ChatGPT.

E INFERENCE TIME

In order to avoid noise, we masked only non-stop words, therefore each statement, based on its length and the number of masked elements, may require different inference time. We divide the total inference time by the number of masked elements to compute the actual cost of inference time per element.

Our objective of this ablation study is answering the following research question *Is there any correlation between inference time and type of statements (i.e., True/False or Auto/Manual statement)?* In order to answer this question, we plot inference time based on each statement group and MLMs. Figure 9 shows the result of this evaluation. Interestingly the results show that False statements in both Auto and Manual are required more computation time except "xlm-roberta-base" for Auto statements. Although different reasons may affect inference time, these results indicate that False statements are more complex than True statements.

Similarly, we compute inference time per UN SDG Topic that includes both True and False statements. The results are shown in Figure 10. This analysis indicate that "Ensure Healthy Life" and "Peace and Justice" are top complex statements in compared to other topics.

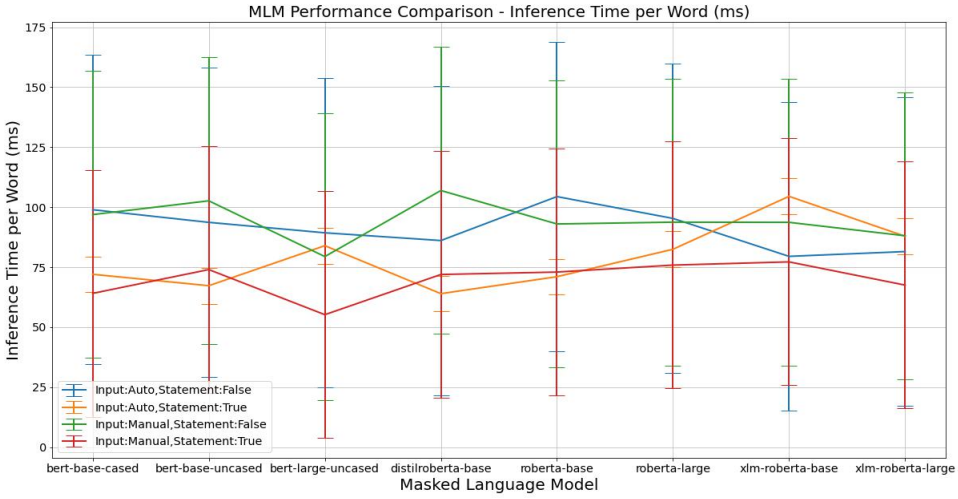

Figure 9: Average Inference Time (ms) per masked token w.r.t. Statements Type, UN SDG Topic and MLM

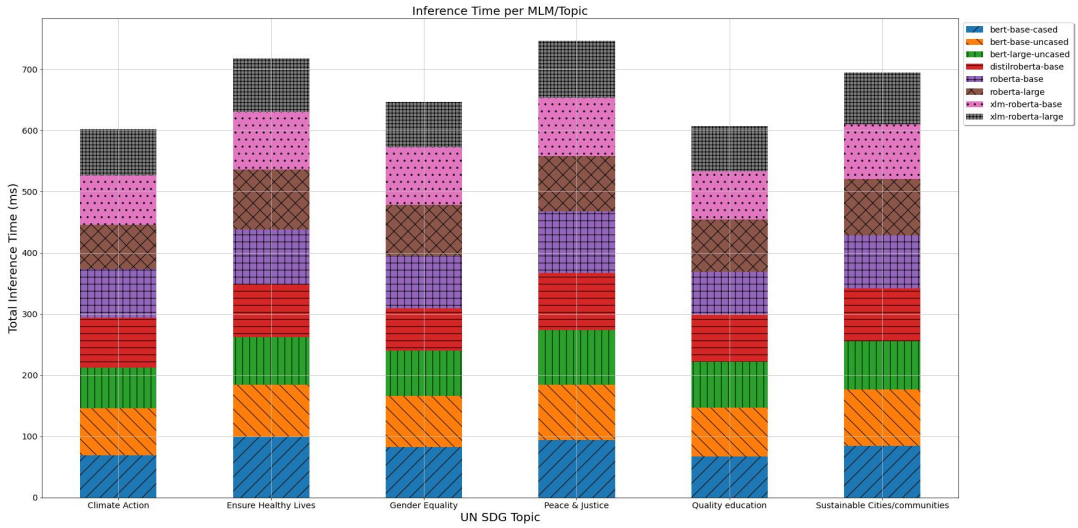

Figure 10: Total inference time (ms) per UN SDG Topic across both True and False statements for 8 evaluated MLMs

Topic	Question	Statement Type	Score($\mathcal{A}^P$) Avg	Score($\mathcal{A}^P$) std	Rank ($\mathcal{A}^R$) Avg	Rank ($\mathcal{A}^R$) std
Quality education	Should English fluency be mandatory for school admissions?	True	0.083	0.061	0.620	0.267
	How can access to education be enhanced in rural areas?	True	0.602	0.062	0.957	0.050
Gender Equality	Is gender inequality aggravated by a person's age?	True	0.225	0.121	0.704	0.184
Peace & Justice	Is equality same as justice?	True	0.143	0.117	0.883	0.078
Sustainable Cities/communities	What are some ways of making cities sustainable?	False	0.321	0.162	0.709	0.009
Gender Equality	How can the inclusion of "preferred pronouns" in conference badges enhance participant experience?	True	0.160	0.070	0.823	0.230
Quality education	How can access to education be enhanced in rural areas?	False	0.281	0.069	0.752	0.050
Sustainable Cities/communities	Can a city without pedestrian and bike friendly sidewalks be called sustainable?	True	0.279	0.084	0.810	0.120
Climate Action	How can the percentage of vegan consumers be increased?	False	0.232	0.085	0.843	0.160
Ensure Healthy Lives	How does anger and anxiety affect health?	False	0.108	0.104	0.571	0.102
Peace & Justice	Does empathy enhance peace?	True	0.281	0.035	0.854	0.003
Ensure Healthy Lives	Why is clean environment necessary for wellbeing?	False	0.050	0.039	0.984	0.010
Gender Equality	How can the representation of women be increased in STEM fields?	False	0.046	0.020	0.722	0.114
Quality education	Why is STEAM education important?	True	0.154	0.024	0.526	0.069
Climate Action	Will an outer-space human colony help alleviate climate crisis?	True	0.293	0.091	0.847	0.164
Sustainable Cities/communities	Can deurbanization enhance sustainability?	True	0.107	0.053	0.828	0.002
	is limiting pollution levels necessary for sustainability?	False	0.103	0.049	0.571	0.215
Gender Equality	is gender inequality more prevalent in non-Western societies?	True	0.342	0.023	0.924	0.151
Climate Action	What policies should governments enforce to reduce coal usage?	False	0.134	0.083	0.519	0.333
	How can the impact of storms and floods be reduced?	True	0.195	0.034	0.720	0.301

Table 7: 20 examples of manual questions for different topics and the average of all False/True statements for the given question across 8 evaluated MLMs.

Topic	Question	Statement Type	Score ($\mathcal{A}^P$) Avg	Score ($\mathcal{A}^P$) std	Rank ($\mathcal{A}^R$) Avg	Rank ($\mathcal{A}^R$) std
Gender Equality	Should paternity leave be offered as a standard benefit to all working fathers?	True	0.049	0.059	0.696	0.099
Peace & Justice	Is the organization committed to promoting environmental sustainability and addressing climate change, and if so, how does it integrate this into its work?	True	0.202	0.076	0.861	0.012
Sustainable Cities/communities	What is the city/community's policy for waste reduction and management?	False	0.138	0.090	0.705	0.135
Peace & Justice	How does the organization ensure that its work is culturally sensitive and respectful of local customs and traditions?	True	0.262	0.073	0.883	0.093
Sustainable Cities/communities	How is the city/community investing in energy-efficient street lighting?	False	0.239	0.059	0.835	0.037
Climate Action	How can we address the issue of climate refugees and displaced persons?	True	0.103	0.098	0.795	0.202
Peace & Justice	Does the organization support the autonomy and self-determination of communities in its work, or does it impose its own solutions and ideas from outside?	False	0.198	0.119	0.930	0.071
Gender Equality	Are you against gender-based discrimination or bias in any form?	False	0.354	0.266	0.912	0.113
Sustainable Cities/communities	Is there a water conservation policy in place in the city/community?	False	0.511	0.085	0.950	0.080
	What are the measures taken towards energy-efficient buildings?	False	0.197	0.089	0.583	0.166
Quality education	Are the assessment and evaluation methods used fair and valid in measuring student learning?	True	0.117	0.111	0.755	0.253
Ensure Healthy Lives	What steps can we take to reduce the prevalence of non-communicable diseases like heart disease and diabetes?	False	0.412	0.252	0.816	0.024
Quality education	Is there support for students who face economic or family-related challenges?	True	0.241	0.160	0.646	0.197
Gender Equality	Do you believe that men should be involved in conversations about gender equality?	True	0.213	0.003	0.631	0.047
Climate Action	Should carbon pricing be implemented to incentivize individuals and companies to reduce emissions?	True	0.436	0.164	0.966	0.053
	Should individuals be held financially responsible for their carbon footprint?	False	0.359	0.278	0.840	0.273
Ensure Healthy Lives	How can we support individuals who live in rural or remote areas to ensure that they have access to quality healthcare?	False	0.376	0.088	0.968	0.046
	How can we ensure that all individuals have access to safe and effective contraception and family planning services?	True	0.490	0.332	0.954	0.064
	What policies can we put in place to ensure that healthcare is affordable and accessible for everyone?	True	0.399	0.271	0.875	0.109
Sustainable Cities/communities	How is the city/community promoting sustainable tourism?	True	0.162	0.082	0.764	0.255

Table 8: 20 examples of automated generated questions for different topics and the average of all False/True statements for the given question across 8 evaluated MLMs.

Note that the following false statements may contain biases, discrimination, and prejudice and true statement may not necessarily reflect the truth.

Topic	Statement	Statement Type	Score ($\mathcal{A}^P$)		Rank ($\mathcal{A}^R$)	
			Avg	std	Avg	std
Sustainable Cities/communities	there should be no rules for buildings in cities	False	0.006	0.009	0.567	0.123
Quality education	Quality education does not entail gender parity	False	0.118	0.074	0.738	0.206
	English fluency should not be a mandatory criterion for admission	True	0.083	0.061	0.620	0.267
Gender Equality	Research has shown that Black women are left to struggle harder to access and advance in their professions, with occupational underrepresentation and wage disparities to show for it.	True	0.162	0.047	0.735	0.084
Ensure Healthy Lives	Healthy lives can be ensured for children by strengthening primary healthcare systems to reach every child	True	0.192	0.098	0.848	0.119

Table 9: 5 examples of True/False manual statements (human written statements) and the average score/rank over 8 evaluated MLMs

Topic	Statement	Statement Type	Score ($\mathcal{A}^P$)		Rank ($\mathcal{A}^R$)	
			Avg	std	Avg	std
Quality education	The policy should encourage students to report any form of discrimination, harassment or bullying experienced	True	0.289	0.097	0.754	0.094
Ensure Healthy Lives	Reliable access to contraception can reduce maternal mortality	True	0.388	0.172	0.710	0.239
Peace & Justice	Our way of doing things is superior to local customs	False	0.257	0.081	0.852	0.105
Ensure Healthy Lives	Condoms are the only effective form of contraception	False	0.281	0.115	0.749	0.230
	Healthcare systems should prioritize eco-friendliness when constructing new facilities	True	0.118	0.087	0.662	0.151

Table 10: 5 examples of True/False automated statements (ChatGPT generated statements) and the average score/rank over 8 evaluated MLMs