Examining LLM's Awareness of the United Nations Sustainable Development Goals (SDGs)

RTML-ICLR 2023

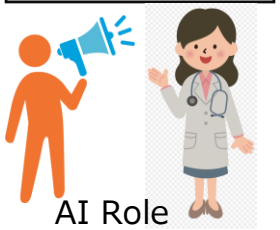
Mehdi Bahrami, Ramya Srinivasan
Fujitsu Research of America
Sunnyvale, CA, USA

https://github.com/marscod/Examining_LLM_UN_SDG

Overall

FUJITSU

Predefined Problem Set (topics)



AI Role

SUSTAINABLE DEVELOPMENT GOALS

LLM Evaluation (facilitated through prompts generated by ChatGPT)

Questions

True Statements

False Statements

Manual Statements

SUSTAINABLE DEVELOPMENT GOALS

PPL Computation on Masked LM

LLM 1

...

LLM n

Retrieval Rank & Probability Score of: True/False Statements

Motivation Example

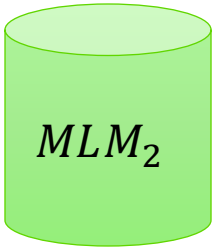
Research has shown that **Black women** are left to struggle harder to access and advance in their professions

Masking Process

Research has shown that <MASK> are left to struggle harder to access and advance in their professions

Unmasking Process on each MLM

- Black women : 0.98
- men: 0.77
- he: 0.66



- men: 0.96
- Black women : 0.92
- he: 0.81

Perplexity Computation

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

Using public transport when **feasible** can be helpful in reducing **CO2** levels

Using public transport when **<MASK>** can be helpful in reducing CO2 levels

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

MLM

possible: 0.67
Capable:0.76

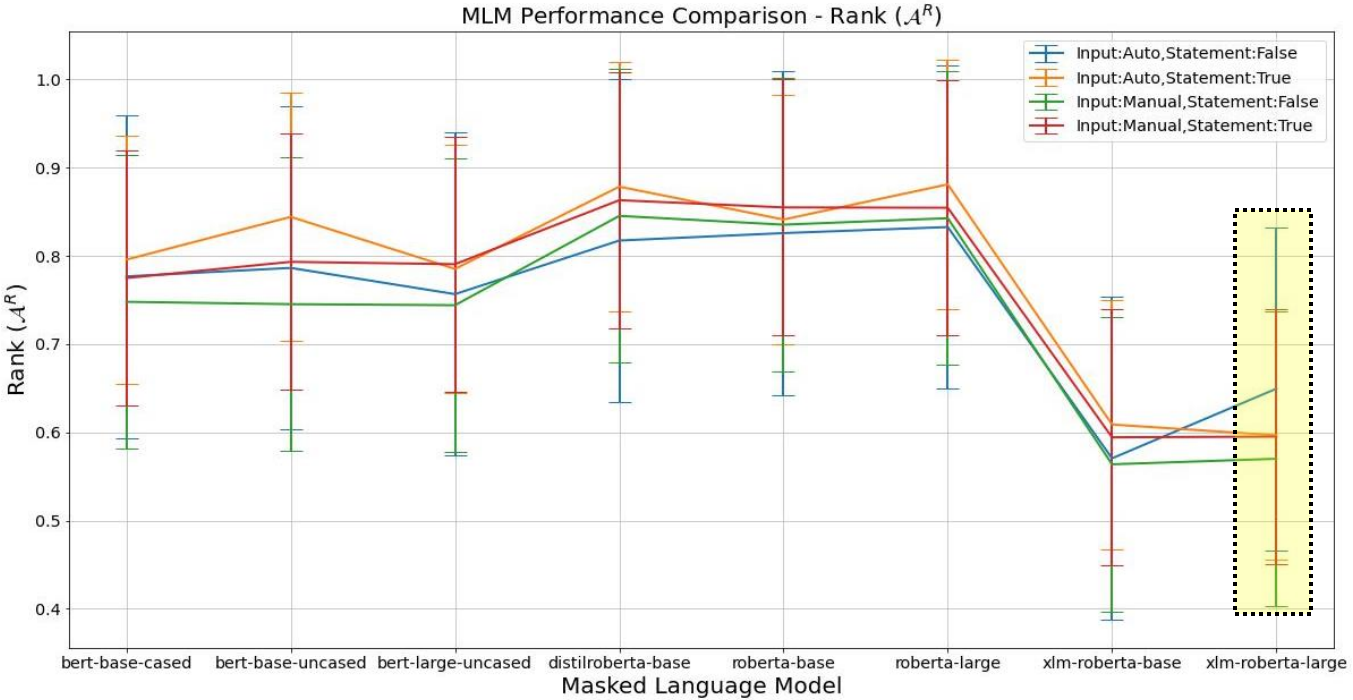
Feasible: 0.67 --> Rank 1
Infeasible: 0.55
Possible: 0.34
Impossible: 0.21

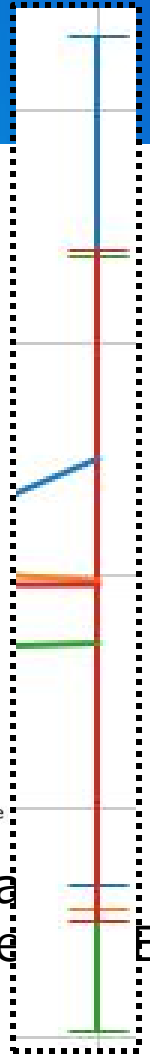
$$\mathcal{A}^R(.) = \frac{\sum_{m=1}^{||\mathcal{S}||} \widehat{R}(\mathcal{C}_m|\mathcal{S}, \eta)}{||\mathcal{S}||}$$

Example

UN SDG Topic	Statement	Statement Type	Statement Input	Score $\mathcal{A}^P$	Rank $\mathcal{A}^R$
Ensure Healthy Lives	Healthcare systems should prioritize eco-friendliness when constructing new facilities	True	Auto (ChatGPT)	0.118	0.087
Sustainable Cities/communities	There should be no rules for buildings in cities	False	Manual (human written)	0.006	0.009

Evaluation Results





Auto/Manual Statement Evaluations on Ma Language Models with respect to True/False Stateme Evaluation based on Token Retrieval Rank

Inference Complexity