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

Mehdi Bahrami, Ramya Srinivasan
Fujitsu Research of America, Sunnyvale, California

FUJITSU
2023 RTML
ICLR

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

MLM_1
- Black women : 0.98
- men: 0.77
- he: 0.66

MLM_2
- men: 0.96
- Black women : 0.92
- he: 0.81

Examined LLMs

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.

Evaluation Approach

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

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

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

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

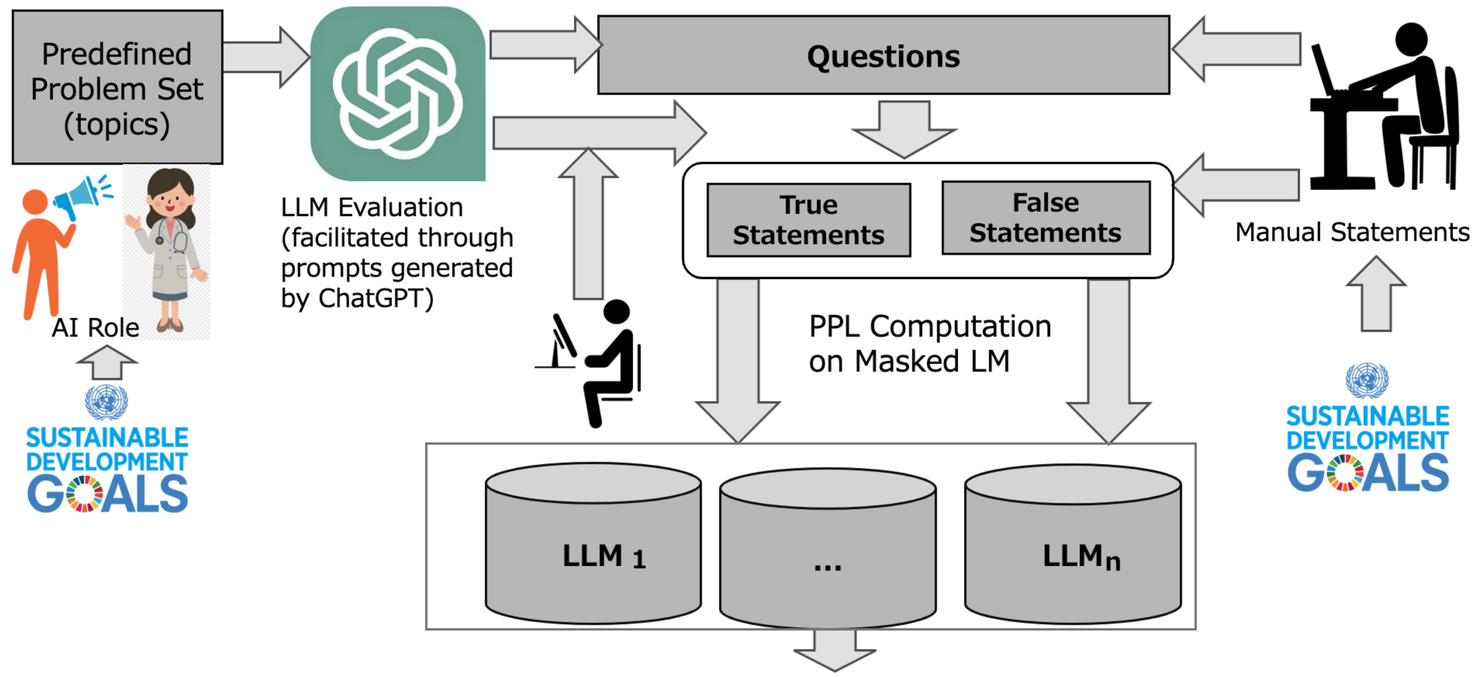
MLM

possible: 0.67
Capable:0.76

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

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

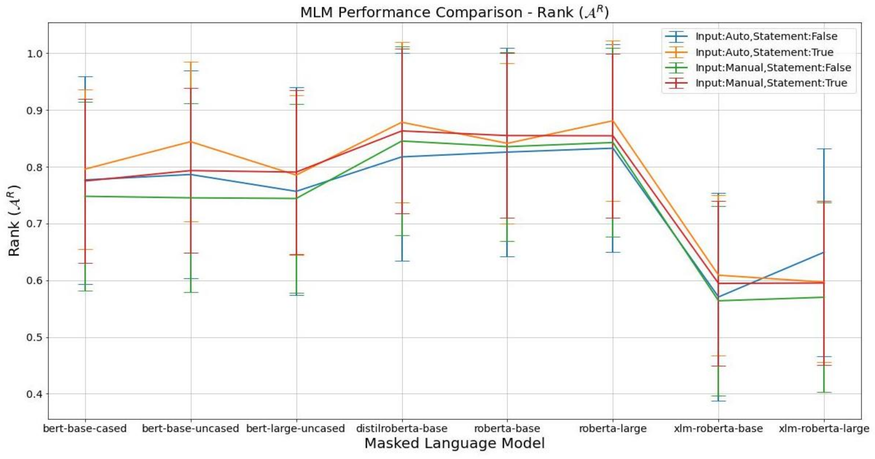
Proposed Approach

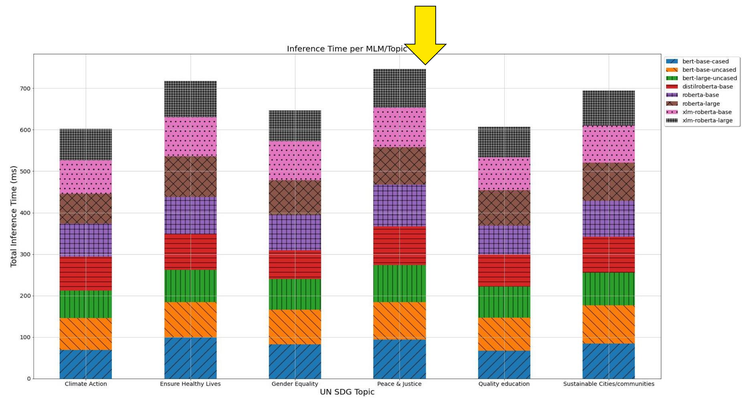


Evaluation Results

UN SDG Topic	Statement	Statement Type	Statement Input	Score	Rank
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

Two examples of Evaluated Statements





Inference Complexity

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

Dataset / GitHub / Paper