Statistical Knowledge Assessment for Large Language Models

Qingxiu Dong[1], Jingjing Xu[2], Lingpeng Kong[3], Zhifang Sui[1], Lei Li[4]

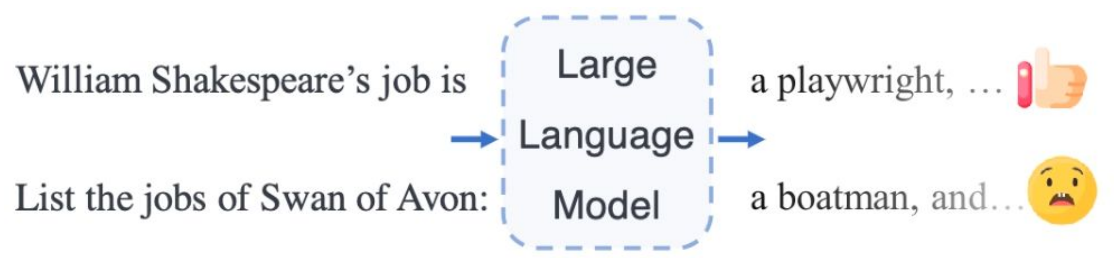
[1]Peking University, [2]Shanghai AI Lab, [3]The University of Hong Kong, [4]Carnegie Mellon University

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Despite the remarkable success of LLMs, critical concerns arise --- LLMs often **generate unreliable answers given varying prompts**.

However, previous studies on model knowledge evaluation primarily assess ***accuracy***, not ***reliability***.



Accurate but Unreliable

Fig1. Accuracy v.s. Reliability

In this work, we evaluate the **reliable knowledge generation ability of LLMs** and present **a statistical approach, KaRR**.

- a vast suite for large-scale knowledge evaluation.
- applied on 20 LLMs, our method effectively assesses their factual knowledge generation reliability.
- KaRR score correlates highly with human evaluation and mitigates evaluation variance and spurious correlation

Graphical Model for Knowledge

To evaluate LLM knowledge reliably, we decompose the knowledge symbols and text forms.

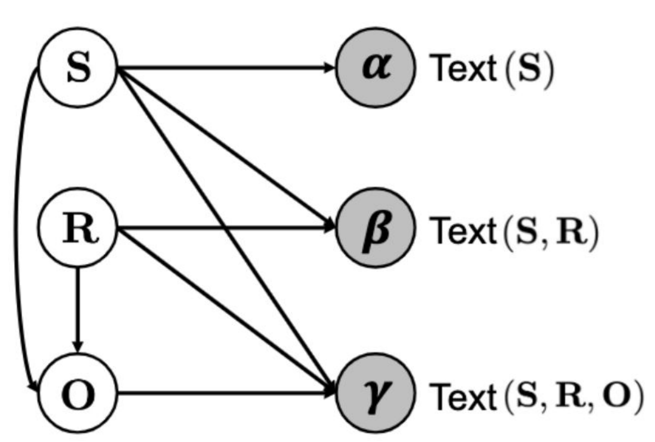


Fig2. Graphical model for knowledge assessment

We establish the connection between symbols and text forms

- Latent variables $\mathbf{S}, \mathbf{R}, \mathbf{O}$ represents the symbolic subject, relation, and object, respectively
- α, β, γ denotes the random variables for the textual forms (textual aliases*)

*"Aliases" are alternative names for entities or relations, defined in Wikidata (https://www.wikidata.org/wiki/Help:Aliases).

Code and data: [dqxiu/KAssess (github.com)](https://github.com/dqxiu/KAssess)

Knowledge Assessment Risk Ratio

Based on the graphical model, we propose a new metric for knowledge assessment, KaRR.

KaRR assesses **the joint impact of subject and relation symbols on the LLMs' ability to generate the object symbol.**

This joint impact comprises two components:
(1) the impact of specifying $\mathcal{r}$ or not on $\mathcal{M}$ generating $\mathcal{O}$ given $\mathcal{S}$.

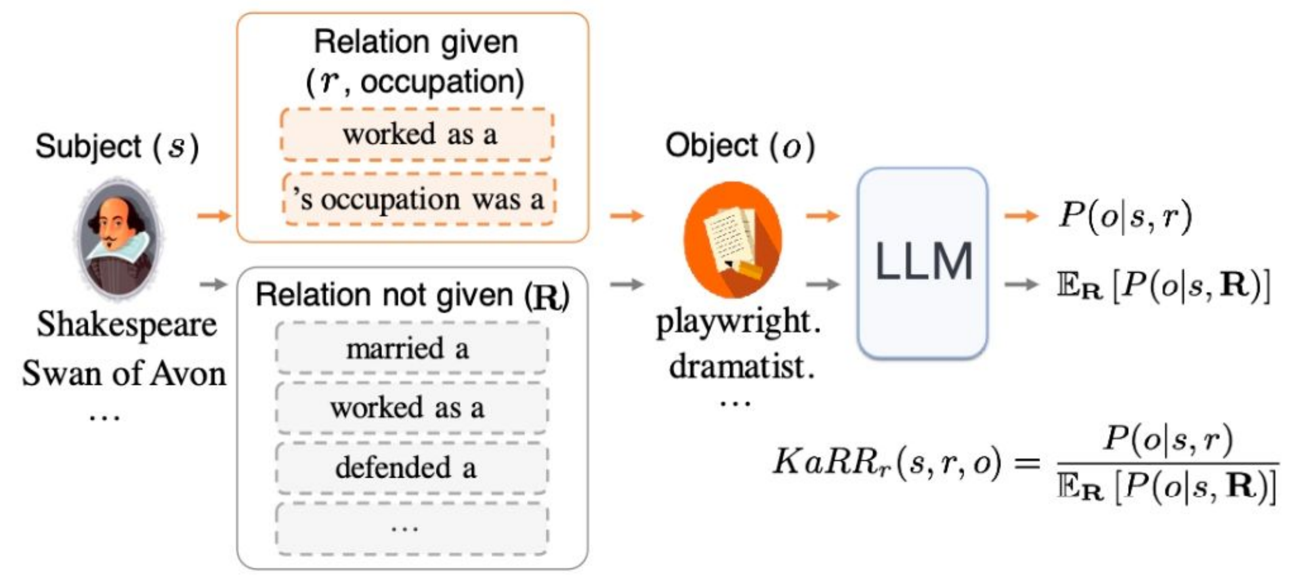


Fig3. Illustration of KaRR_r

(2) the impact of specifying $\mathcal{S}$ or not on $\mathcal{M}$ generating $\mathcal{O}$ given $\mathcal{r}$.

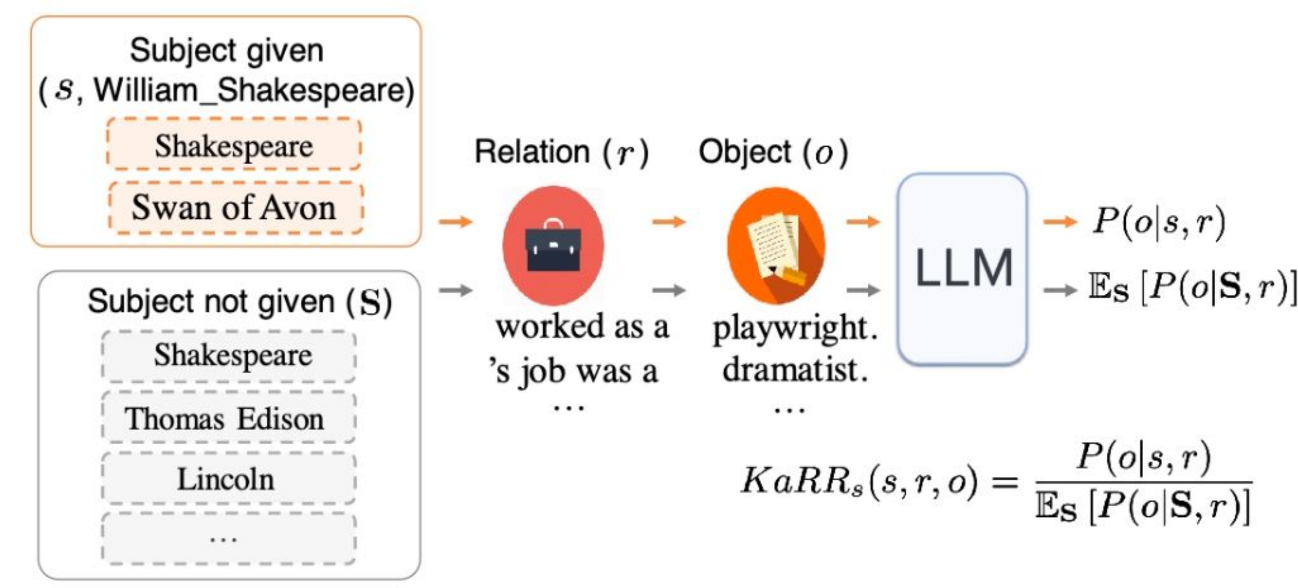


Fig4. Illustration of KaRR_s

Computing KaRR using Graphical Model

- KaRR is formulated based on knowledge symbols, while directly computing the KaRR score using the definition (on symbols) is unfeasible.
- The graphical model for knowledge assessment facilitates the implementation of KaRR by employing model probabilities on the text.

$$\mathrm{KaRR}_r(s,r,o) = \frac{\sum_{k=1}^{|\boldsymbol{\beta}|} P(\beta_k|s,r) \sum_{j=1}^{|\boldsymbol{\gamma}|} P_{\mathcal{M}}(\gamma_j|s,r,\beta_k) P(o|\gamma_j)}{\sum_{i=1}^{|\boldsymbol{\alpha}|} P(\alpha_i|s) \sum_{j=1}^{|\boldsymbol{\gamma}|} P_{\mathcal{M}}(\gamma_j|s,\alpha_i) P(o|s,\alpha_i,\gamma_j)}$$

$$\mathrm{KaRR}_s(s,r,o) = \frac{\sum_{k=1}^{|\boldsymbol{\beta}|} P(\beta_k|s,r) \sum_{j=1}^{|\boldsymbol{\gamma}|} P_{\mathcal{M}}(\gamma_j|s,r,\beta_k) P(o|\gamma_j)}{\sum_{u=1}^{|\boldsymbol{s_u}|} P(s_u|r) \cdot \sum_{k=1}^{|\boldsymbol{\beta}|} P(\beta_k|s_u,r) \sum_{j=1}^{|\boldsymbol{\gamma}|} P_{\mathcal{M}}(\gamma_j|s_u,r,\beta_k) P(o|\gamma_j)}$$

Please refer to the paper for further definitions of the symbols.

Experimental Results

A. Basic Information and Human Evaluation

Method	Subj. Alias	Obj. Alias	Rel. Alias	Rel. Cvg.
LAMA@1	✗	✗	✗	6.83%
LAMA@10	✗	✗	✗	6.83%
ParaRel	✗	✗	✓	6.33%
KaRR	✓	✓	✓	100%

Tab1. Basic information

Method	Recall	Kendall's τ	p-value
LAMA@1	83.25%	0.17	0.10
LAMA@10	65.81%	0.08	0.23
ParaRel	69.15%	0.22	0.02
K-Prompts	78.00 %	0.32	0.03
KaRR	**95.18%**	**0.43**	0.03

Tab2. Results of human evaluation

- Our method has a good coverage of various relations and entity aliases, our assessment suite contains 994,123 entities and 600 relations.
- KaRR exhibits a strong correlation with human assessment.

B. Evaluation Variance and Spurious Correlation

Method	Var (↓)	Std (↓)
LAMA@1	1.90	1.37
LAMA@10	5.14	2.27
ParaRel	0.77	0.94
K-Prompts	2.34	5.47
KaRR	**0.67**	**0.82**

Tab4. Evaluation variance

Method	SP (↓)	ΔP (↓)
LAMA@1	3.81	0.00
LAMA@10	64.29	47.31
ParaRel	2.66	-0.51
K-Prompts	**0.00**	-7.54
KaRR	1.94	**-14.94**

Tab5. Spurious correlation

- Compared to previous methods, KaRR results are **more robust and less influenced by the spurious correlation**.

C. KaRR Scores on 20 LLMs

Model	Size	KaRR Score	Model	Size	KaRR Score
GPT	0.12B	9.57	GLM	10B	5.59
XLNet	0.12B	5.86	Dolly	12B	15.60
T5-large	0.74B	3.22	LLaMA	13B	13.86
Phi-1.5	1.3B	10.58	Alpaca	13B	8.24
GPT2-XL	1.56B	12.27	Vicuna	13B	19.50
GPT-NEO	2.65B	13.44	WizardLM	13B	16.90
T5-3B	3B	9.52	Moss	16B	11.20
Falcon	7B	7.97	LLaMA	65B	14.56
BLOOM	7B	7.72	LLaMA2	65B	19.71
LLaMA	7B	12.37	OPT	175B	23.06

Tab6. Evaluation results on 20 LLMs

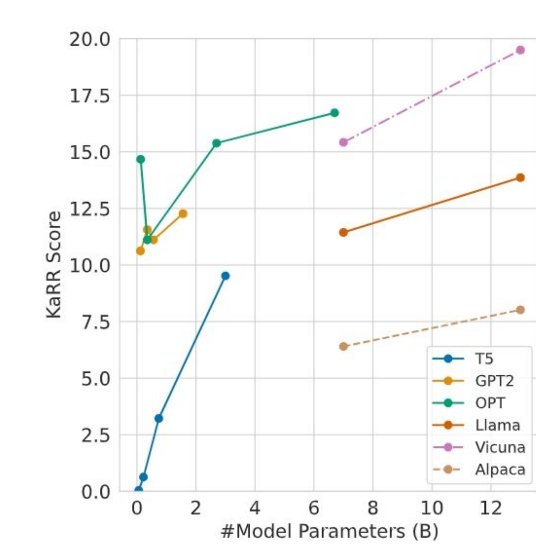


Fig5. Scaling results

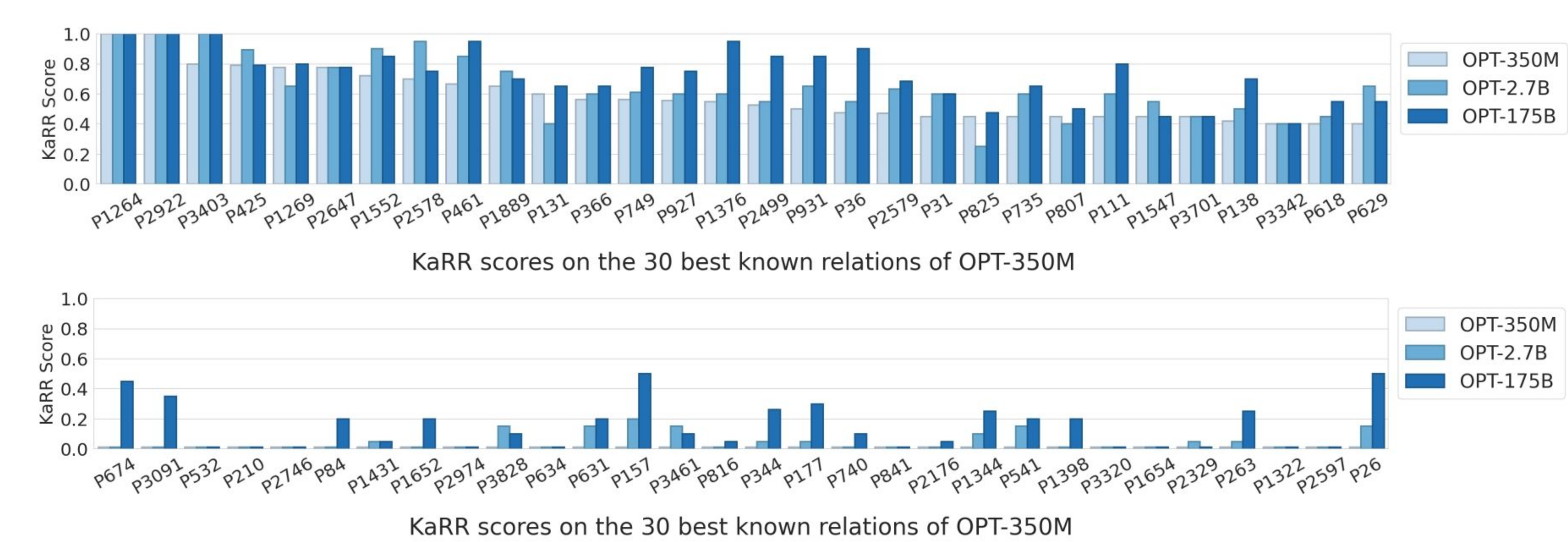


Fig6. KaRR scores on different relations when scaling up OPT

- small and medium-sized LLMs struggle with generating correct facts consistently.
- instruction tuning could influence knowledge consistency and correctness.
- scaling law: larger models generally hold more factual knowledge.