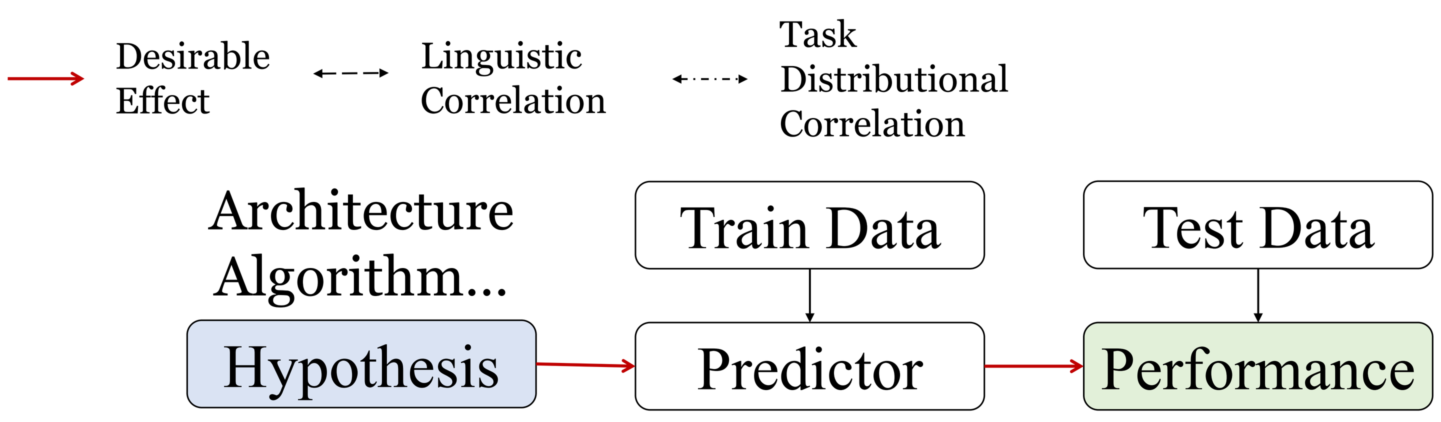
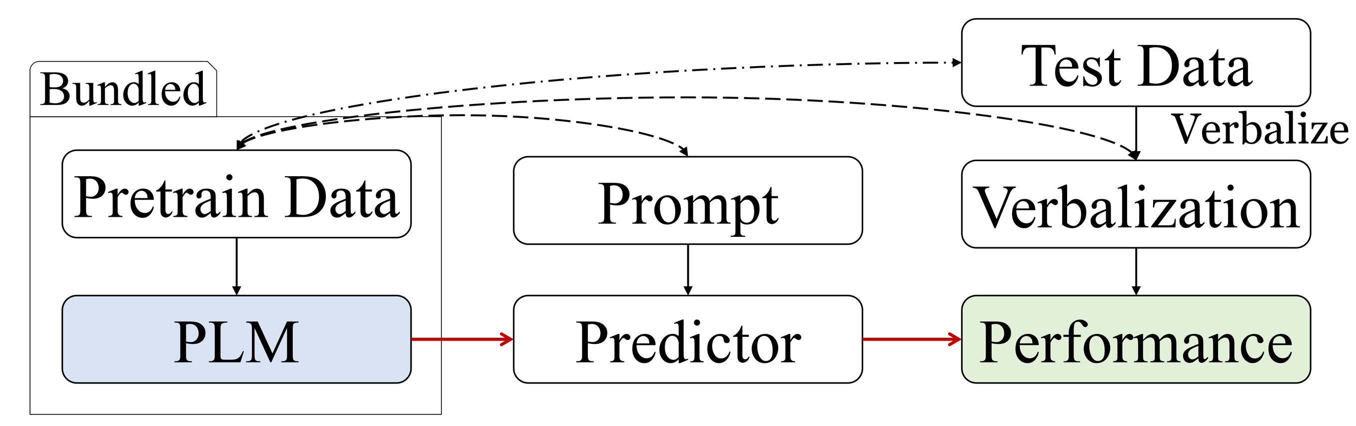
Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risk from a Causal View

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, Le Sun
Institute of Software, Chinese Academy of Sciences, Beijing, China
{boxi2020,hongyu,xianpei,fangchao2017,sunle}@iscas.ac.cn

Introduction





➢**Conventional evaluation in machine learning:**

I. The evaluated hypotheses are raised independent of train/test data generation.

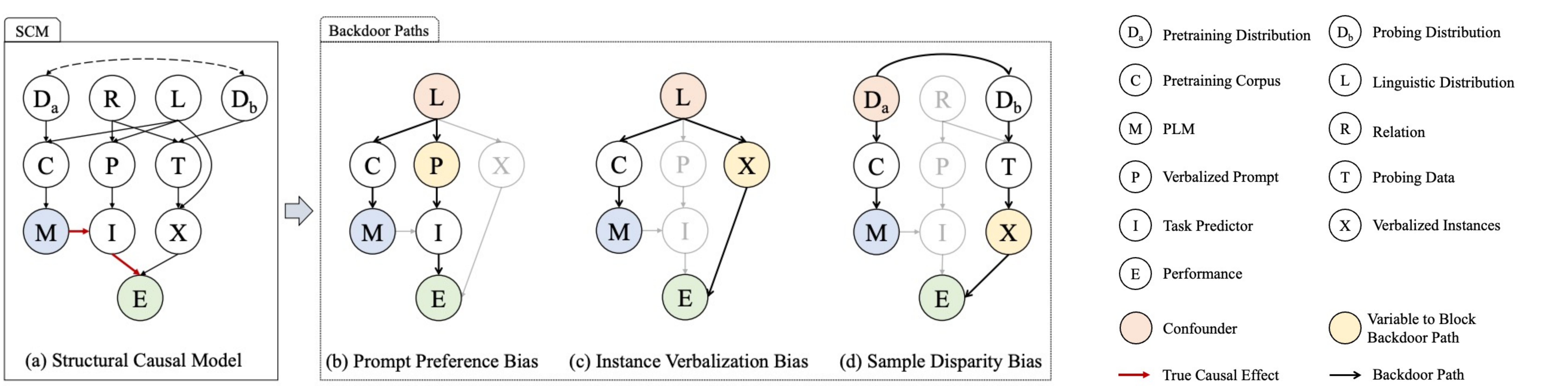
II. The impact of correlations is transparent, controllable and equal for all the hypotheses.

➢**PLM evaluation via prompt-based probing:**

I. Evaluated PLMs are bundled with pretraining corpus.

II. There exist implicit correlations between the pretraining corpus, prompt and probing data which will mislead evaluation.

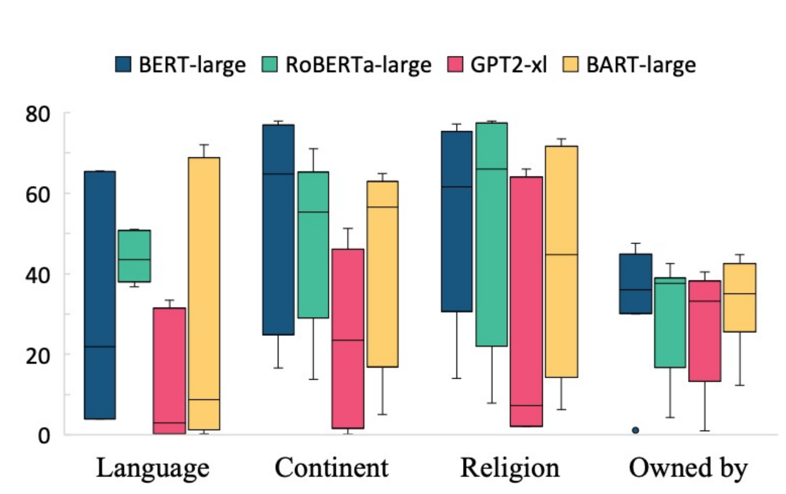
Structural Causal Model

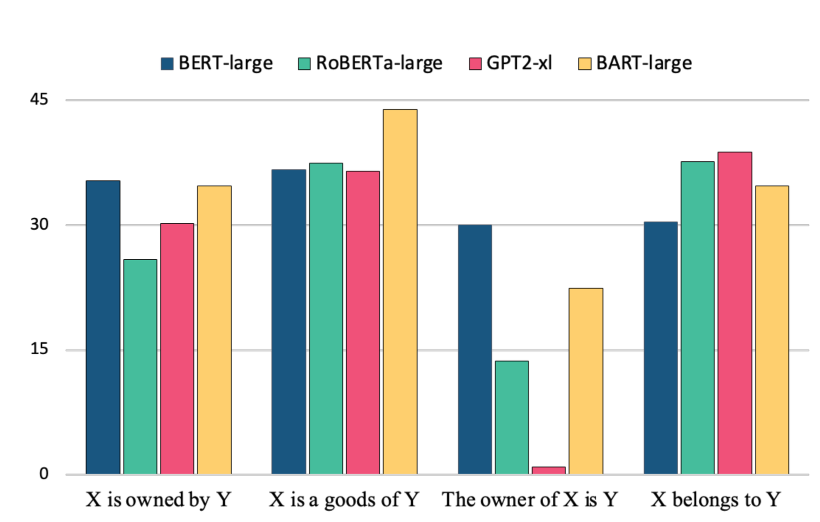


➢There are three backdoor paths in the structural causal model, and each backdoor path corresponds to one bias.

Prompt Preference Bias

➢The model performance will be affected by both the task ability of PLM and preference fitness of a prompt.

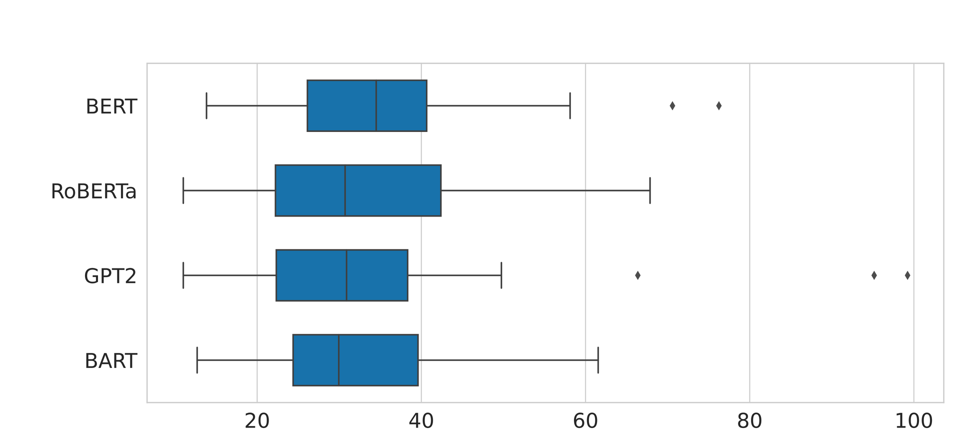




Instance Verbalization Bias

➢Different PLMs may prefer different verbalizations due to mention coverage, expression overlap, etc.

Relation	Mention	Prediction
Capital of	America	Chicago
	the U.S.	Washington
	China	Beijing
	Cathay	Bangkok
Birthplace	Einstein	Berlin
	Albert Einstein	Vienna
	Isaac Newton	London
	Sir Isaac Newton	town



Sample Disparity Bias

➢The performance difference between different PLMs may due to the sample disparity of their pretraining corpus, rather than the ability divergence.

γ%	BERT-base	BERT-large	GPT2-base	GPT2-medium
0%	30.54	33.08	15.22	22.11
20%	35.77	39.56	22.02	28.21
40%	38.68	39.75	24.32	30.29
60%	38.72	40.68	25.42	31.16
80%	39.79	41.48	25.65	31.88
100%	40.15	42.51	26.82	33.12
None	37.13	39.08	16.88	22.60

Bias Elimination

➢Causal intervention can significantly improve the evaluation consistency.

☐ Propose to reduce bias via backdoor adjustment.

$$\mathcal{P}(E|do(M=m),R=r) = \sum_{p\in P}\sum_{x\in X}\mathcal{P}(p,x)\mathcal{P}(E|m,r,p,x).$$

Model	Original	Random	+Intervention
BERT-base	56.4	45.4	**86.5**
BERT-large	100.0	78.1	**100.0**
RoBERTa-base	75.7	44.0	**77.8**
RoBERTa-large	56.1	42.2	**86.5**
GPT2-medium	63.5	40.7	**98.2**
GPT2-xl	74.2	35.7	**77.8**
BART-base	63.4	61.6	**98.2**
BART-large	97.7	61.3	**100.0**
Overall Rank	25.5	5.5	**68.5**

Conclusion

➢ A causal analysis framework is proposed to effectively identify, interpret and eliminate evaluation biases with a theoretical guarantee.

➢ Our conclusions echo that we need to rethink the criteria for identifying better PLMs.