ISCAS 中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周

学术论文

提示符可以用于探测预训练语言模型吗？从因果角度理解其潜在的风险

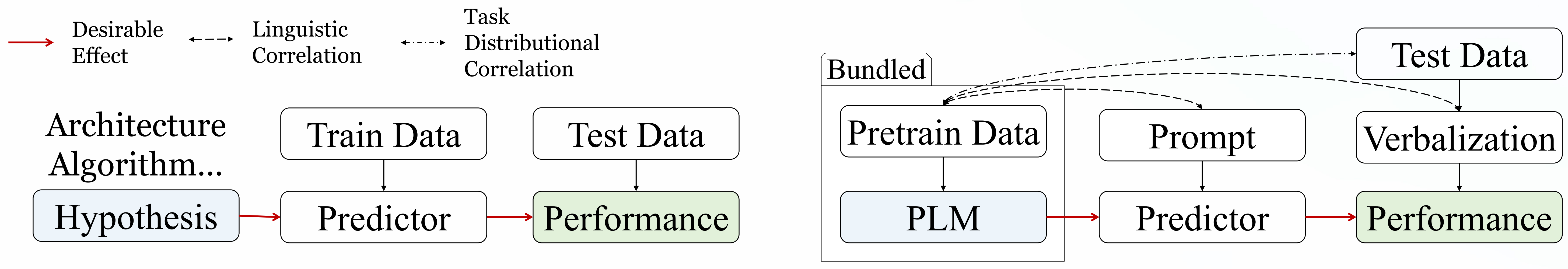
曹博希，林鸿宇, 韩先培, 刘方超, 孙乐
中国科学院软件研究所

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
(Volume 1: Long Papers), pages 5796–5808, Dublin, Ireland.

联系人：曹博希　邮箱：boxi2020@iscas.ac.cn　电话：13051882626

介绍

- ➢基于提示符的探针已经被广泛用于评估预训练模型中的知识和能力。
- ➢现有工作对评测过程中所存在的风险的忽视会误导对模型的理解，甚至产生错误的结论。



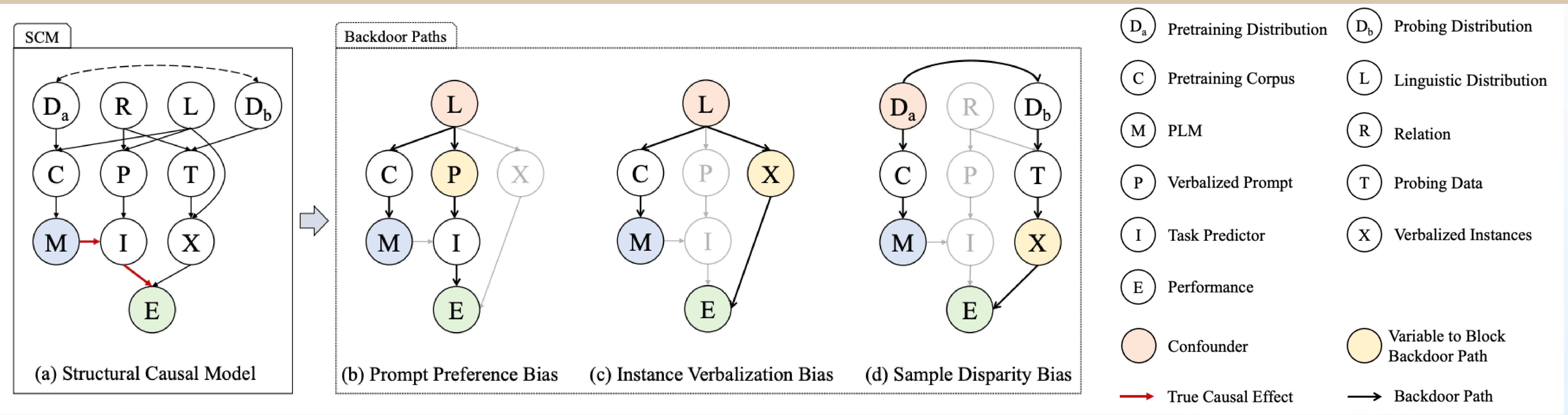
➢ 机器学习中的传统评测范式

I. 评测对象为不同的假设（算法/模型架构），独立于训练测试数据的生成。
II. 数据间关联的影响是透明的、可控的，并且对所有假设均平等。

➢ 基于提示符的探针评测范式

I. 评测对象为不同的预训练模型，均与其特定的预训练数据绑定。
II. 预训练数据、提示符、探针数据间存在的伪相关会给评测结果带来偏差。

结构因果模型

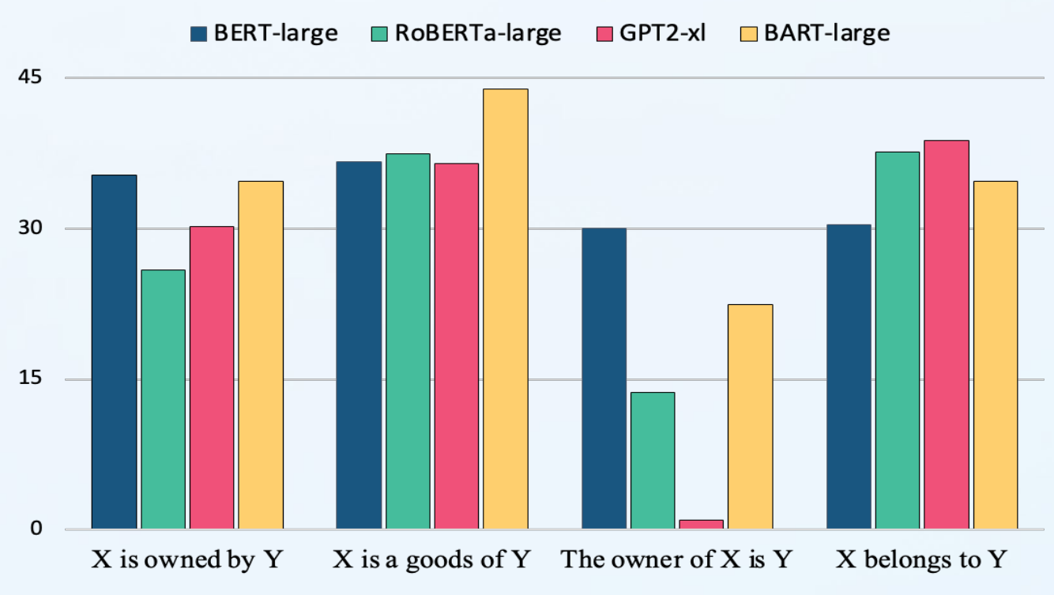


(a) Structural Causal Model　(b) Prompt Preference Bias　(c) Instance Verbalization Bias　(d) Sample Disparity Bias

- ➢ 使用结构因果模型建模了基于提示符的探针的流程，用于从理论上分析评测中的偏差。
- ➢ 结构因果模型中存在着三条后门路径，分别对应着一种偏差。

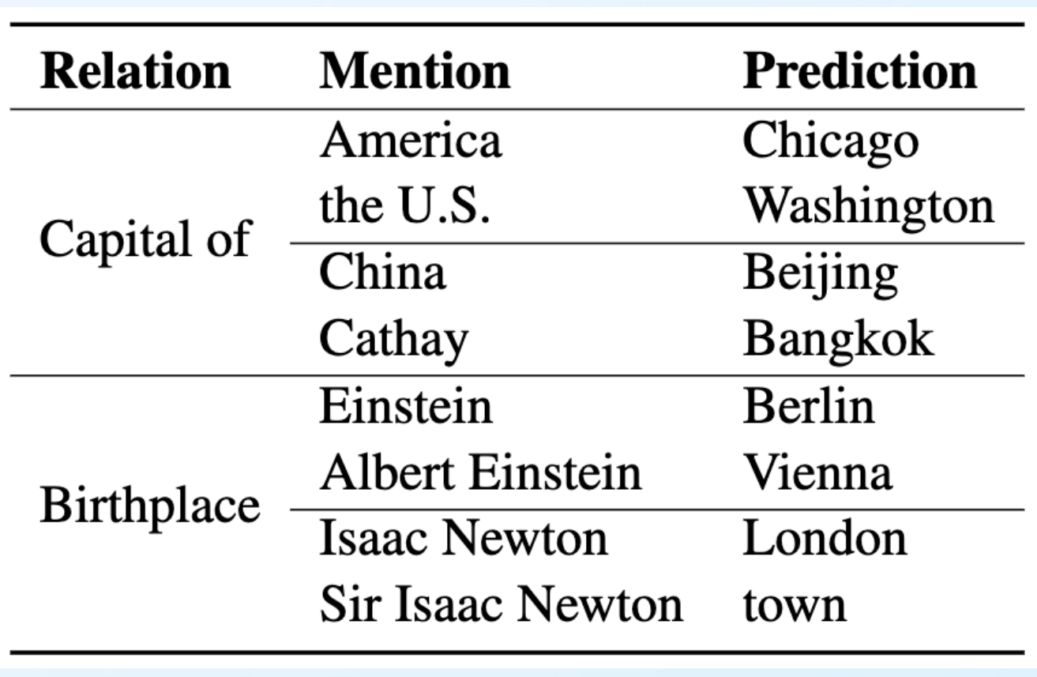
提示符偏好偏差

- ➢ 模型性能会同时受到模型能力和其对特定提示符偏好的影响。

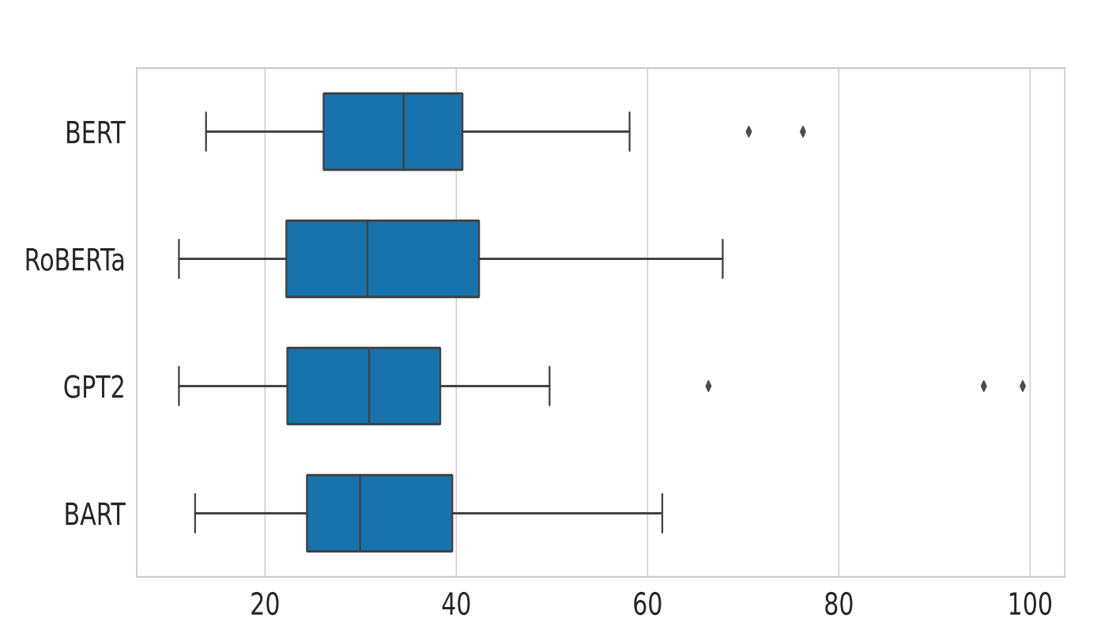


提示符偏好偏差

- ➢ 不同的预训练模型可能会偏好同一个概念的不同自然语言化表达。

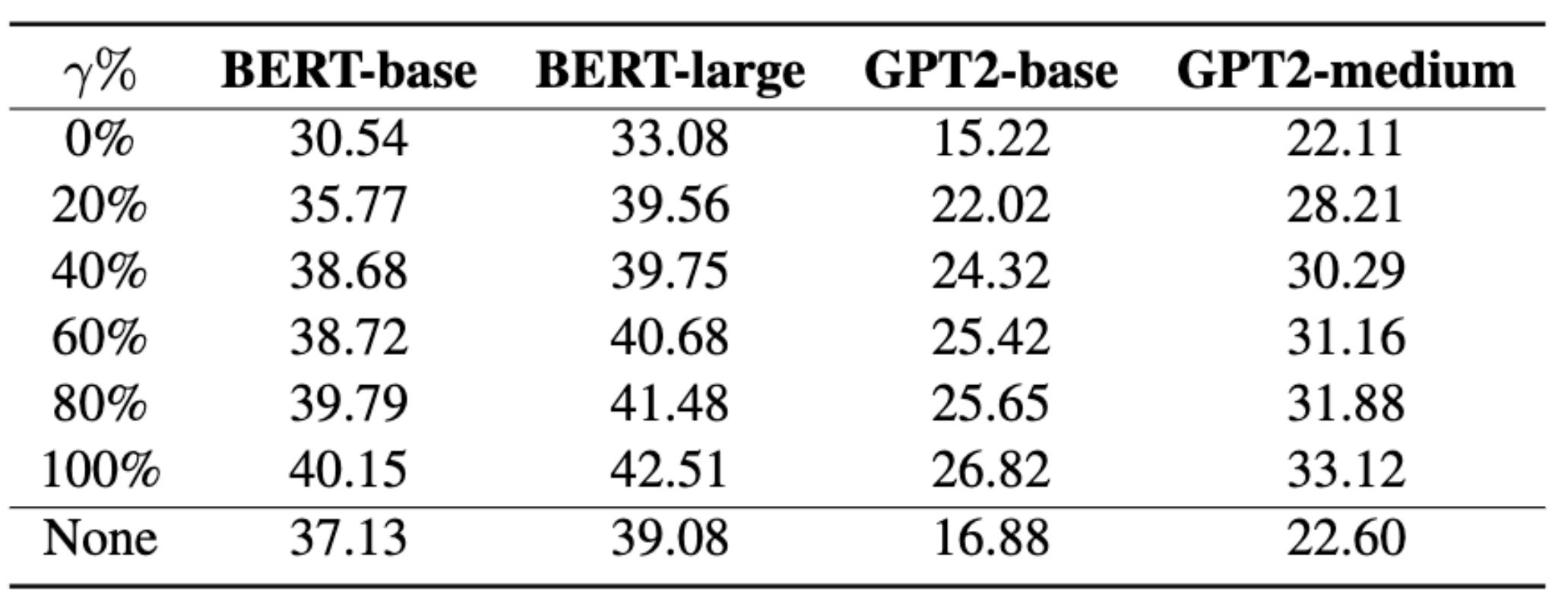


Relation	Mention	Prediction
Capital of	America	Chicago
	the U.S.	Washington
	China	Beijing
	Cathay	Bangkok
Birthplace	Einstein	Berlin
	Albert Einstein	Vienna
	Isaac Newton	London
	Sir Isaac Newton	town



采样差异偏差

- ➢ 不同模型性能差异可能来源于其预训练语料的采样差异，而非仅仅模型能力的差别。



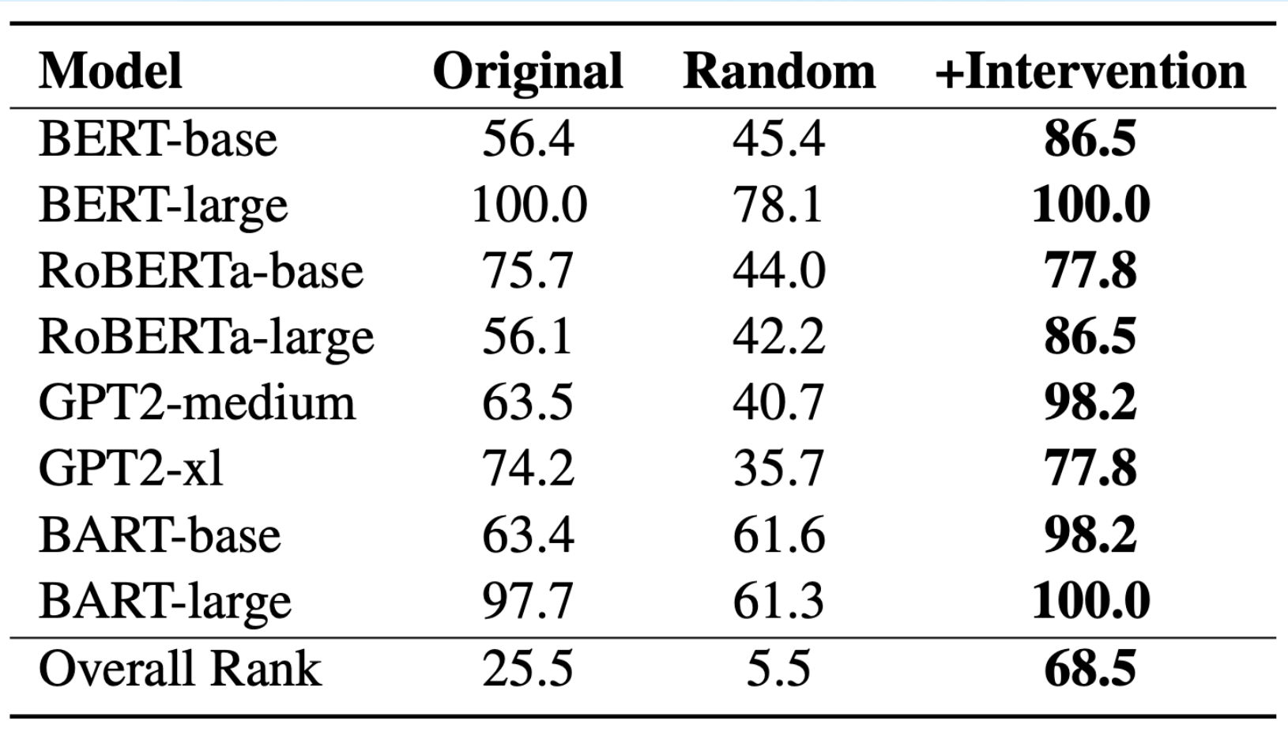
γ%	BERT-base	BERT-large	GPT2-base	GPT2-medium
0%	30.54	33.08	15.22	22.11
20%	35.77	39.56	22.02	28.21
40%	38.68	39.75	24.32	30.29
60%	38.72	40.68	25.42	31.16
80%	39.79	41.48	25.65	31.88
100%	40.15	42.51	26.82	33.12
None	37.13	39.08	16.88	22.60

偏差消减

- ➢ 因果干预能够显著提升评测结果的稳定性。

- 通过后门准则来减小评测偏差。

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



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**

结论

- ➢ 提出了一个因果分析框架，能够**从理论层面**识别、解释和减小基于提示符的探针中的偏差。
- ➢ 基于结构因果模型的分析框架能够**原则性地扩展**到其他的评测场景中。
- ➢ 本文为设计更好的探针框架，可靠的评估范式，以及推动偏差分析**从经验化到理论到发展**提供宝贵的参考价值。同时提醒研究者们应该重新思考如何更好地评测语言模型。