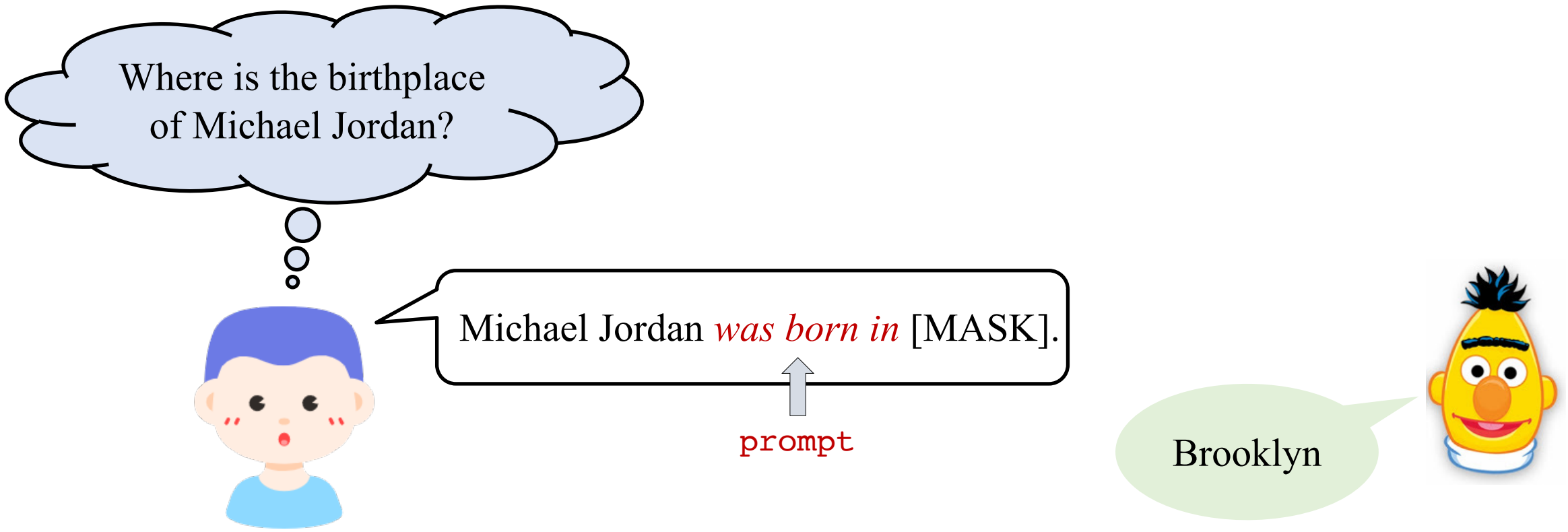
Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, Le Sun
Institute of Software, Chinese Academy of Sciences, Beijing, China

Prompt-based Probing



LAMA- Factual Knowledge

Petroni F, Rocktäschel T, Lewis P, et al. Language models as knowledge bases? In EMNLP 2019.

Evaluating PLMs via Prompt-based Probing

Example
A chicken [MASK] has horns. **A.** never **B.** rarely **C.** sometimes **D.** often **E.** always
A 21 year old person is [MASK] than me in age, If I am a 35 year old person. **A.** younger **B.** older
The size of a airplane is [MASK] than the size of a house . **A.** larger **B.** smaller
It was [MASK] hot, it was really cold . **A.** not **B.** really
What is usually located at hand and used for writing? **A.** pen **B.** spoon **C.** computer
A ferry and a floatplane are both a type of [MASK]. **A.** vehicle **B.** airplane **C.** boat
When did the band where Junior Cony played first form? **A.** 1978 **B.** 1977 **C.** 1980
When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest **A.** second **B.** first **C.** third

oLMpics – Reasoning (Talmor et al. 2020)

Original:
Paul tried to call George on the phone, but he wasn't successful.
Who is he?
Candidate: A. Paul (correct) B. George

Reframed:
A. Paul tried to call George on the phone, but Paul wasn't successful. (Positive sample)
B. Paul tried to call George on the phone, but George wasn't successful. (Negative sample)

CAT – Commonsense (Zhou et al. 2020)

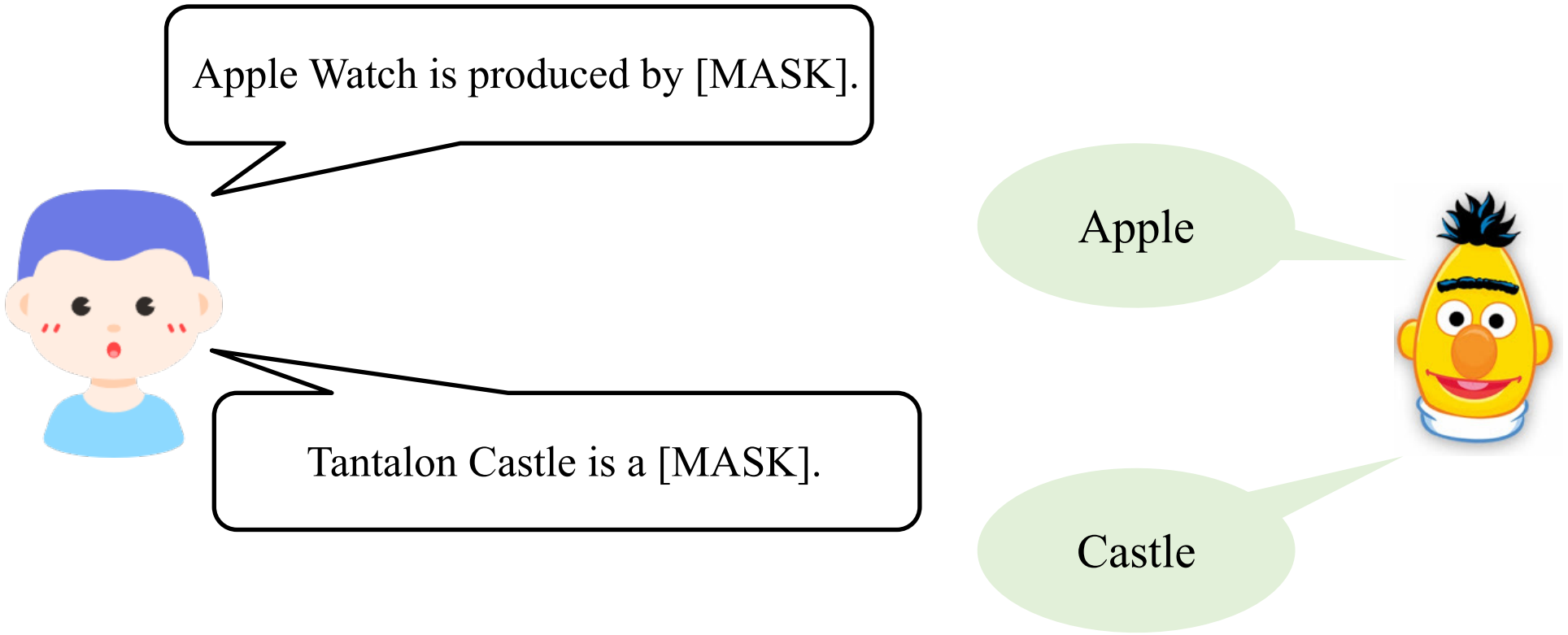
Context	Compl.
the restaurant owner forgot which customer the waitress had ____	*served*
the restaurant owner forgot which waitress the customer had ____	*served*

LM diagnostics – Linguistic (Ettinger et al. 2020)

# Relations: 36	**# Entity Types**: 12	**# Triples**: 49K	**Sources**
medical condition treated	Amantadine has effects on [Y].		
symptoms	Hepatitis has symptoms such as [Y].		
affects binding	Nicotine binds to [Y].		

Bio LAMA – Biological (Sung et al. 2021)

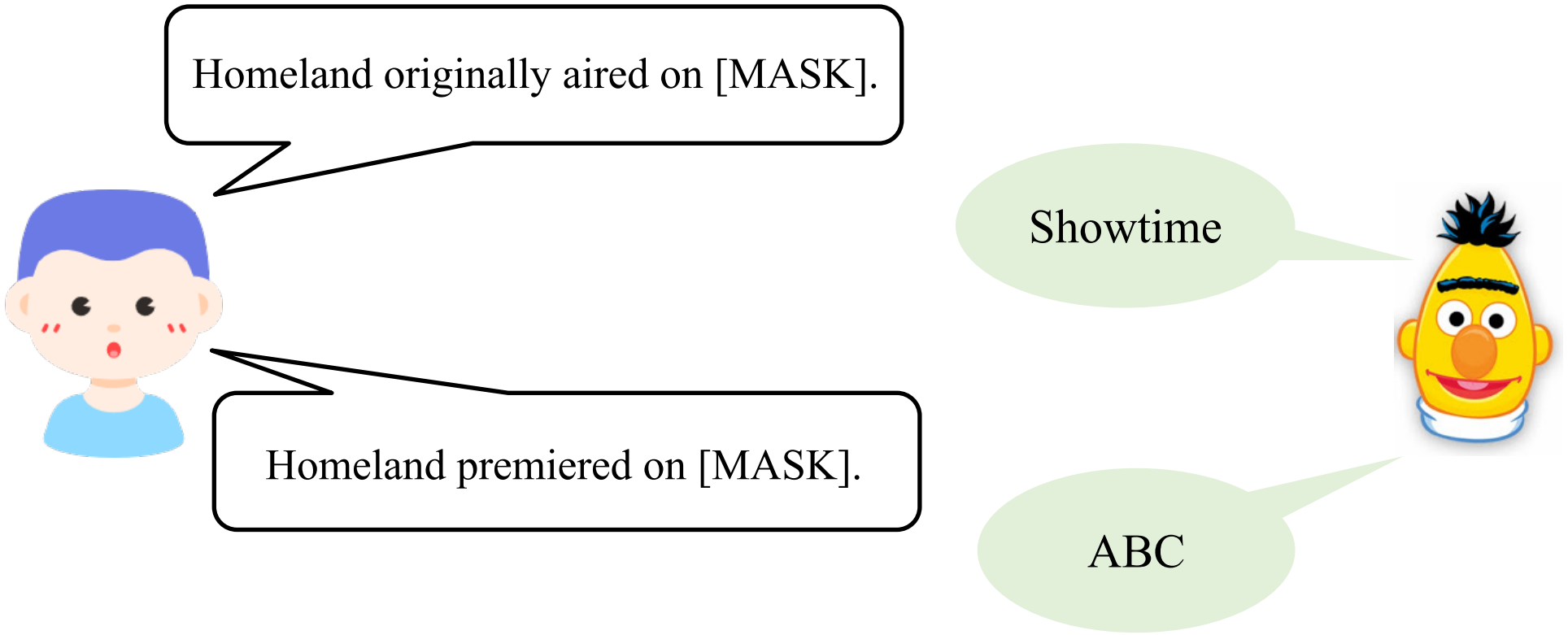
Risks of Prompt-based Probing



- **Performance may be overestimated.**
  I. Predictions rely on surface form shortcuts.

Poerner N, Waltinger U, Schütze H. E-BERT: Efficient-yet-effective entity embeddings for BERT. In findings of EMNLP 2020.

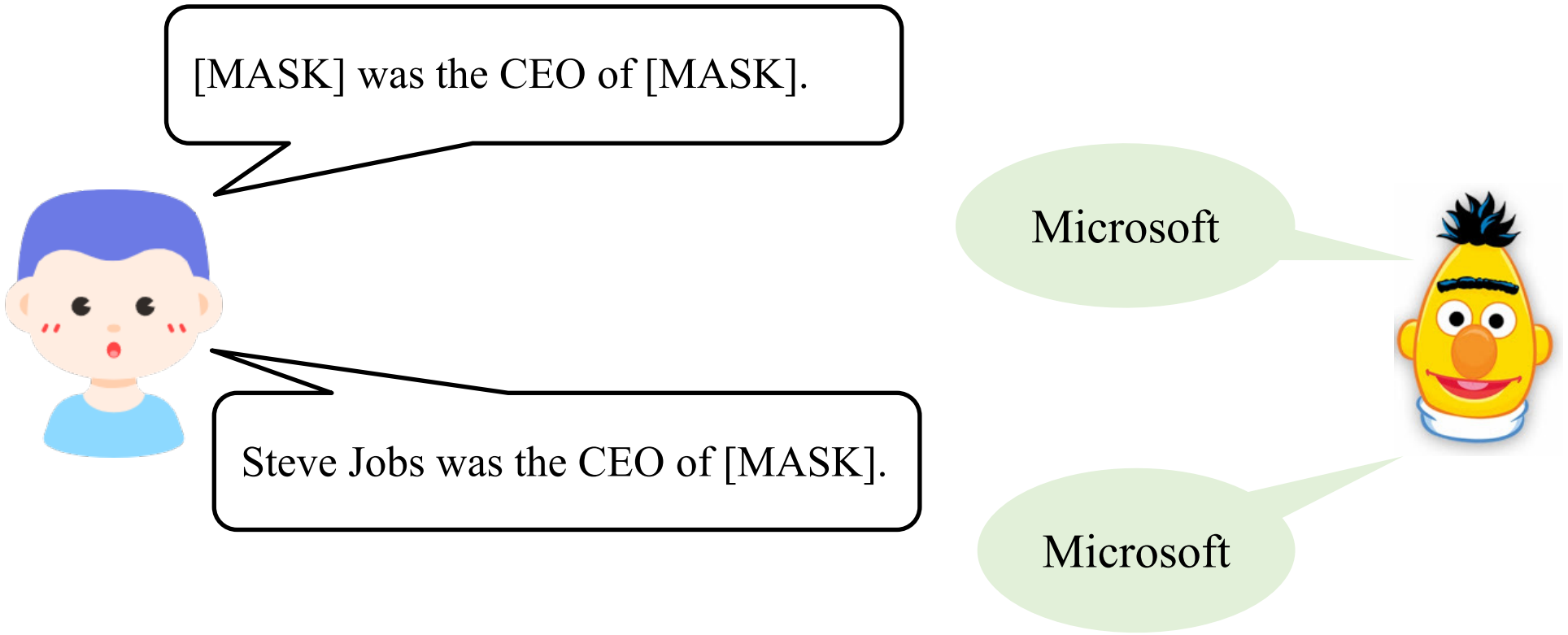
Risks of Prompt-based Probing



- **Predictions are inconsistent.**
  I. Semantically equivalent prompts may result in different predictions.

Yanai Elazar et, al: Measuring and Improving Consistency in Pretrained. In TACL 2021.

Risks of Prompt-based Probing



- **Performance is unreliable.**
  I. PLMs often generate prompt-related but not knowledge related predictions.

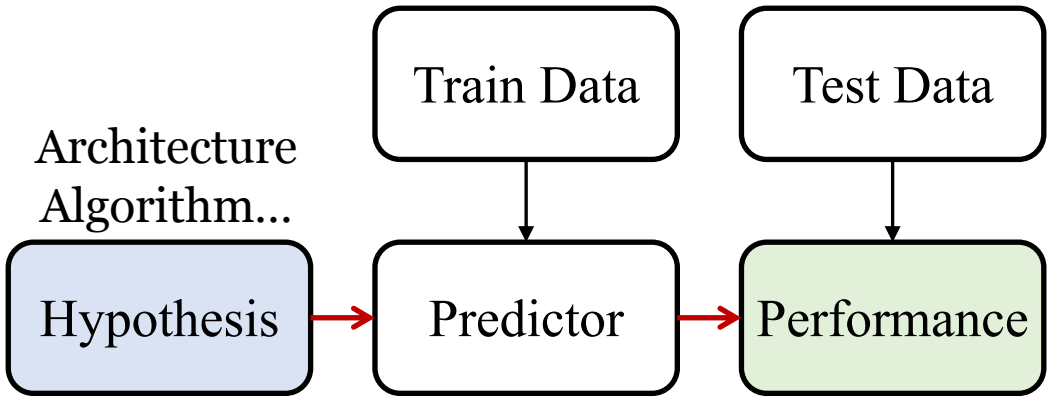
Cao et, al: Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In ACL 2021.

Two Critical Questions

***What bias** exist in current evaluation criteria via prompt-based probing?*

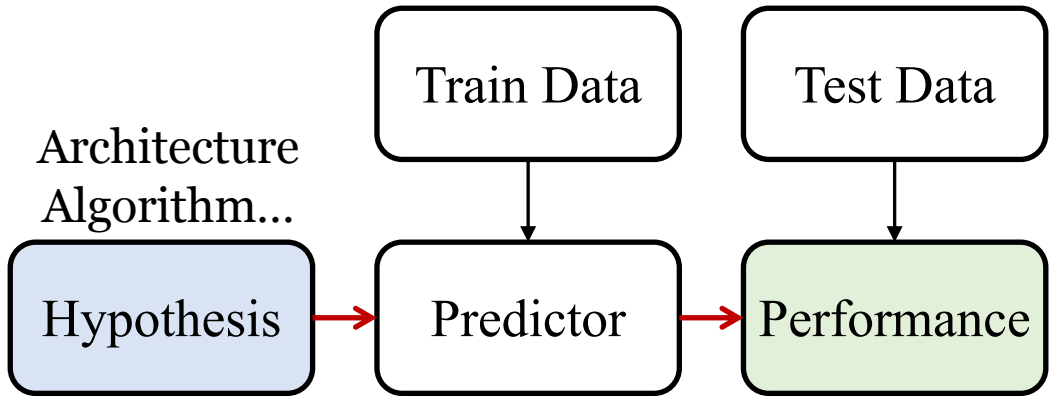
***Where** do these biases come from?*

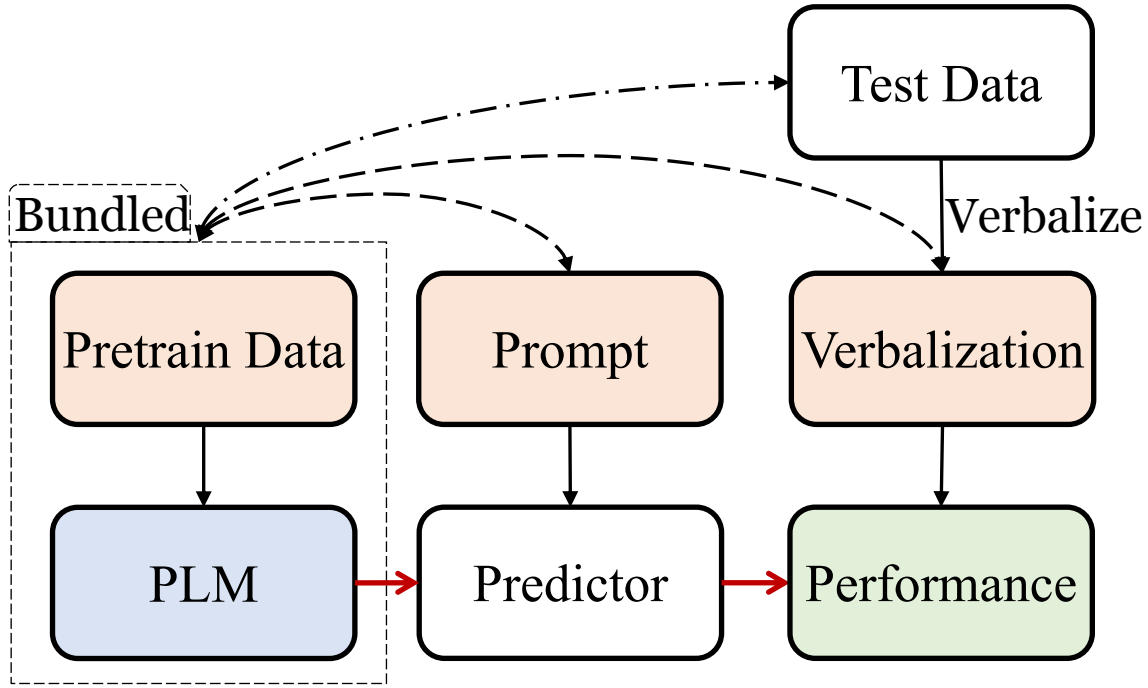
Conventional Evaluation v.s. Prompt-based Probing



- **Conventional evaluation in machine learning:**
  I. Hypotheses are raised independent of train/test data generation.
  II. The impact of correlations is transparent, controllable and equal.

Conventional Evaluation v.s. Prompt-based Probing

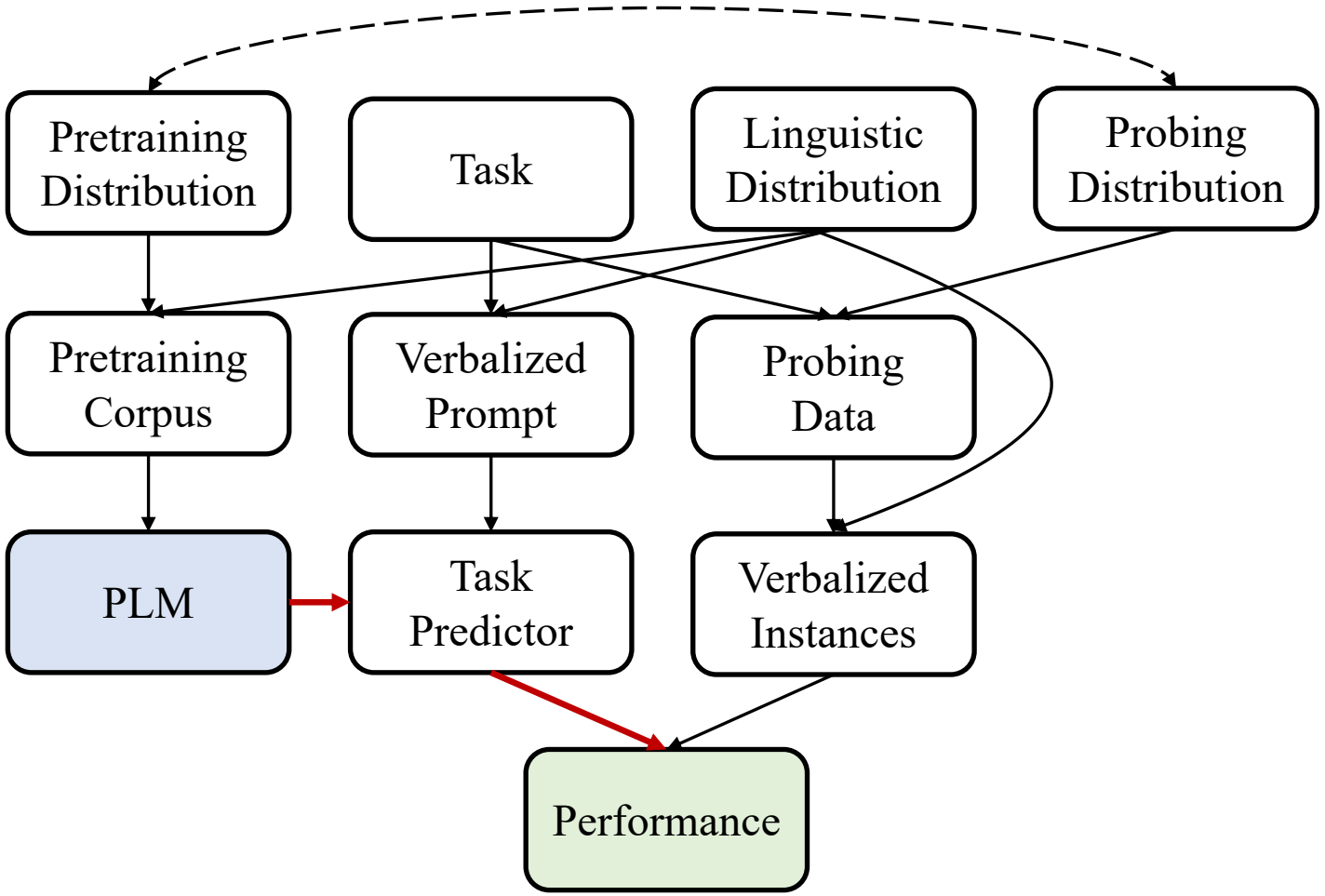




- **Conventional evaluation in machine learning:**
  I. Hypotheses are raised <u>independent of</u> train/test data generation.
  II. The impact of correlations is <u>transparent, controllable and equal.</u>

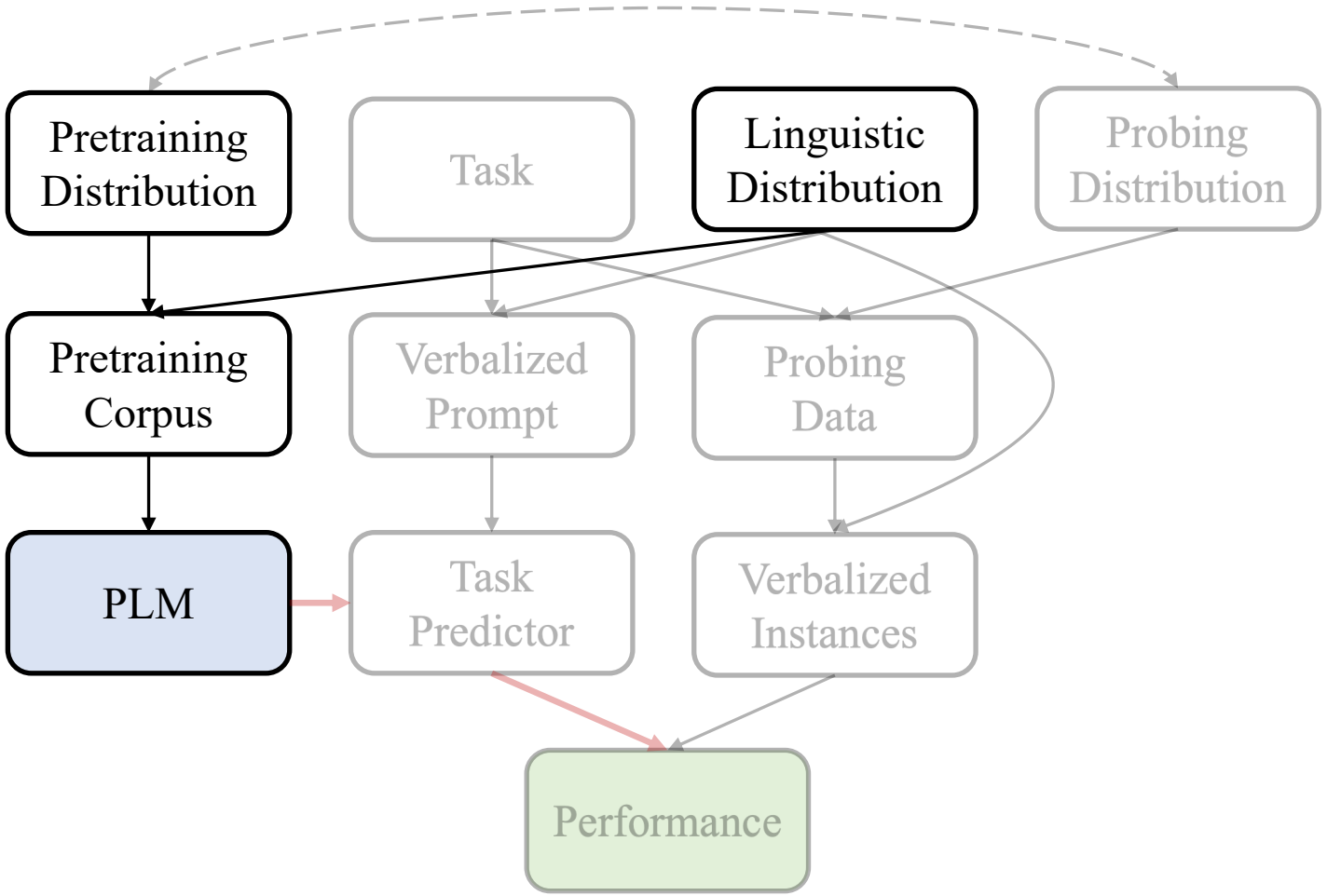
- **PLM evaluation via prompt-based probing:**
  I. PLM are <u>bundled with</u> pretraining corpus.
  II. There exist <u>implicit correlation</u> between pretrain data, prompt and probing data which will <u>mislead evaluation</u>.

Structural Causal Model



- SCM describes the relevant features in a system and how the interact with each other.
- The SCM of prompt-based probing contains 11 key variables and 4 procedures.

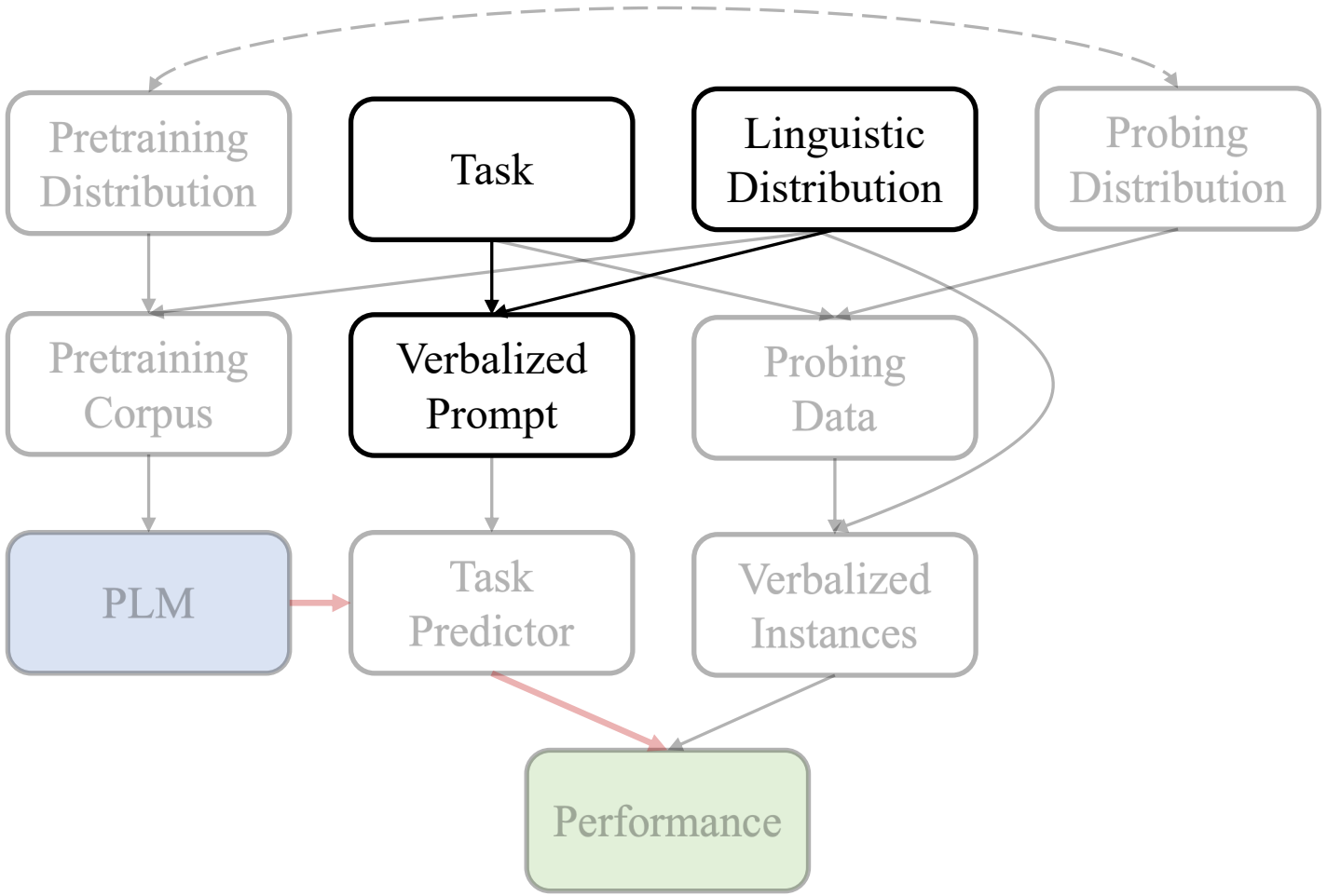
Structural Causal Model



- **PLM Pretraining**
  I. Sample **pretraining corpus** according to **pretraining distribution** and **linguistic distribution**.
  II. Pretrain **language model** on the **pretraining corpus**.

Linguistic Distribution
Guides how concept verbalized into natural language expression
e.g., task-prompt, entity-mention

Structural Causal Model

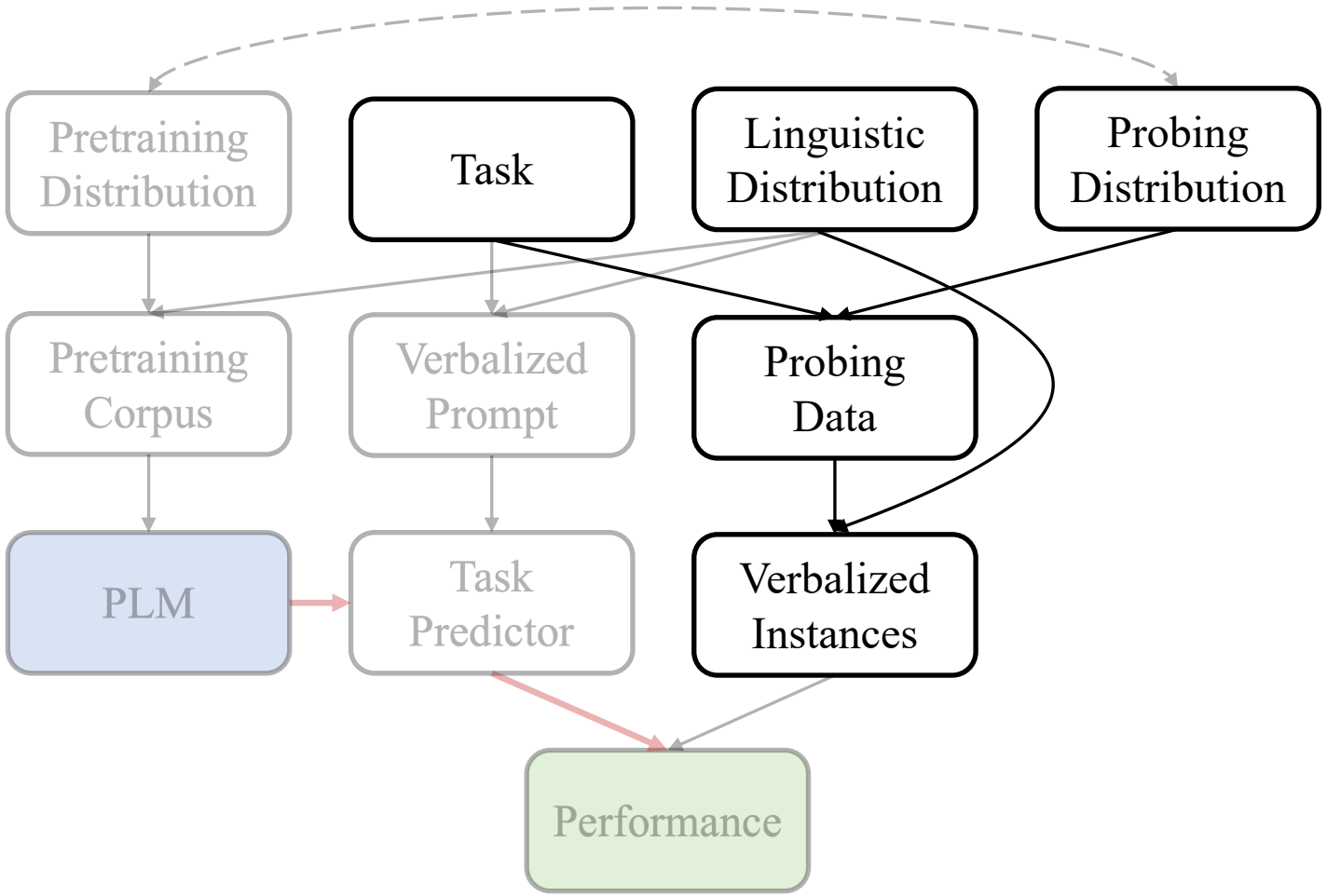


- **Prompt Selection**
  - I. Each **prompt** must exactly express the semantics of evaluated **task**.
  - **II.** **Prompt** will also be influenced by **linguistic distribution**.

Task

In factual knowledge probing, each relation corresponds to a probing task.
e.g., birthplace, capital

Structural Causal Model

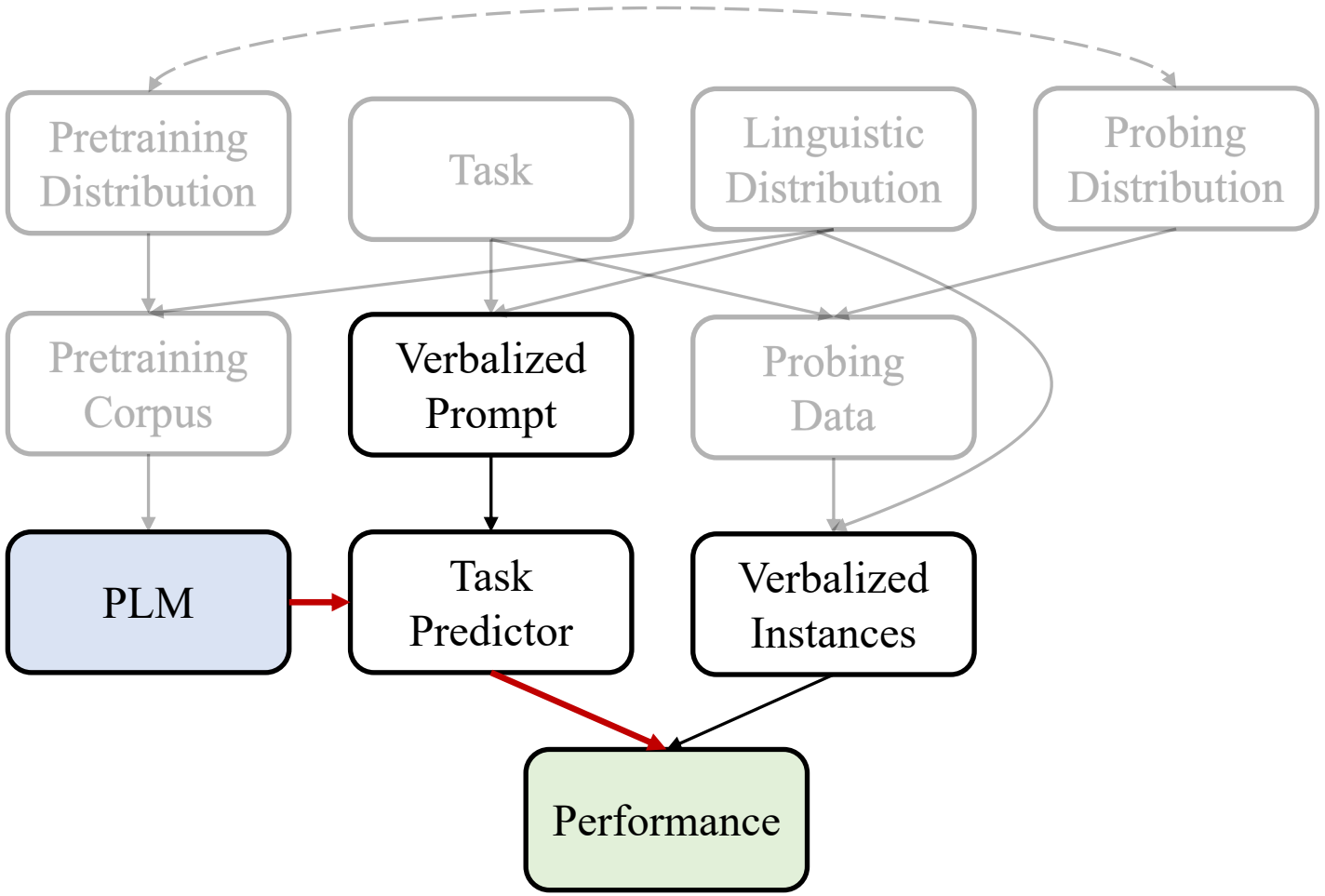


- **Verbalized Instances Generation**
  - I. Sample **probing data** of the **task** according to **probing data distribution**.
  - II. Verbalize **probing data** into **verbalized instances** according to **linguistic distribution**.

Verbalized Instance

e.g., <Michael Jordan, Brooklyn>
from <Q41421, Q18419> of Wikidata

Structural Causal Model

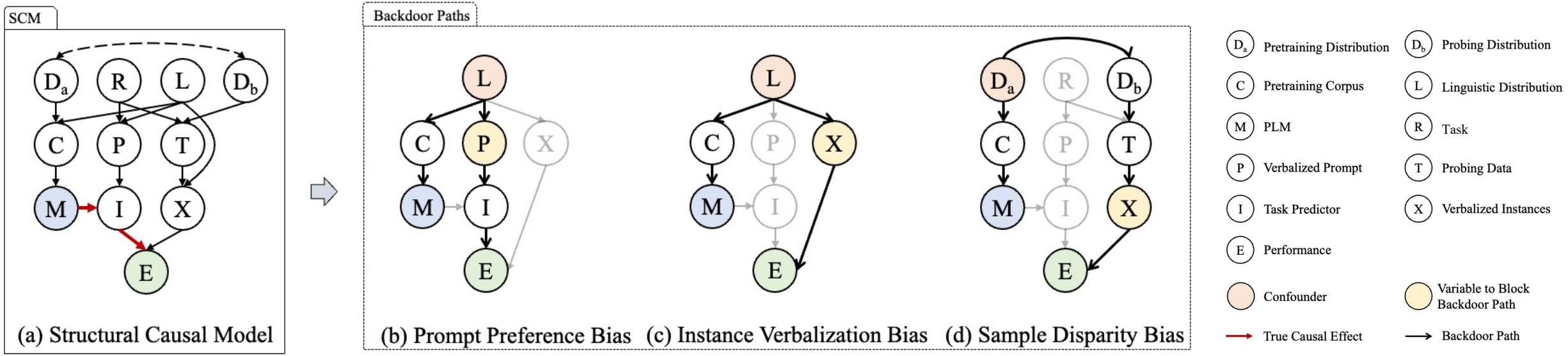


- **Performance Estimation**
  - I. **Task-specific predictor** is derived by combining **PLM** and **prompt**.
  - II. Estimate **performance** by applying **predictor** on **verbalized instance**.

Task-specific Predictor

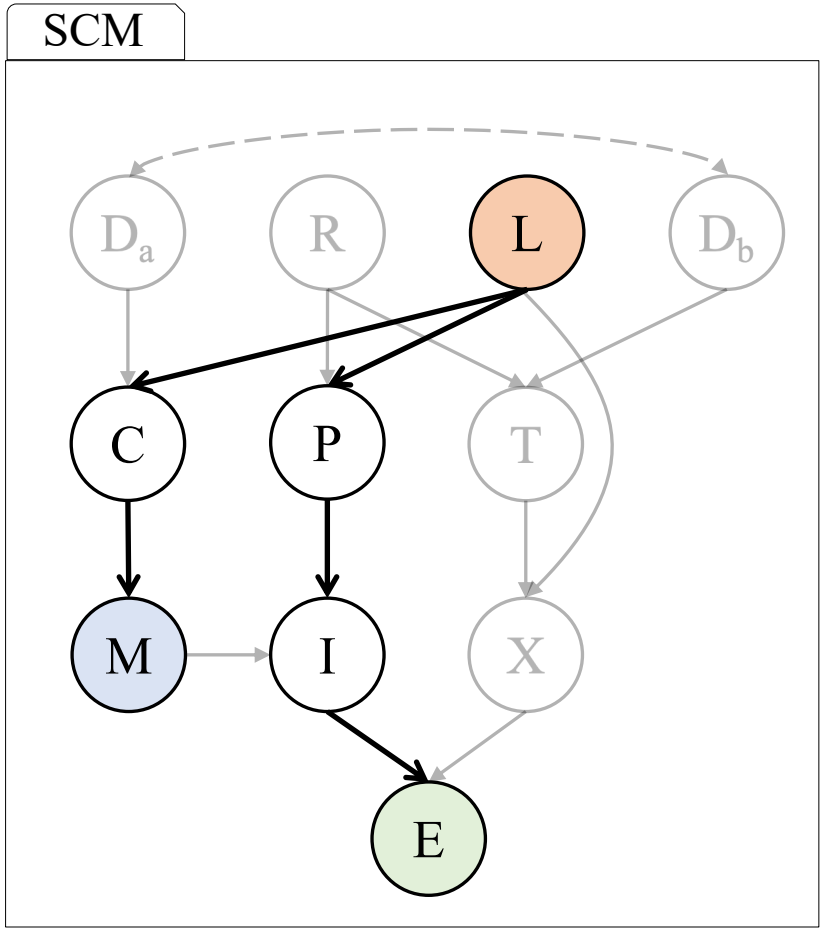
e.g., <BERT, was born in> as a birthplace predictor.

Backdoor Path Corresponds to Bias



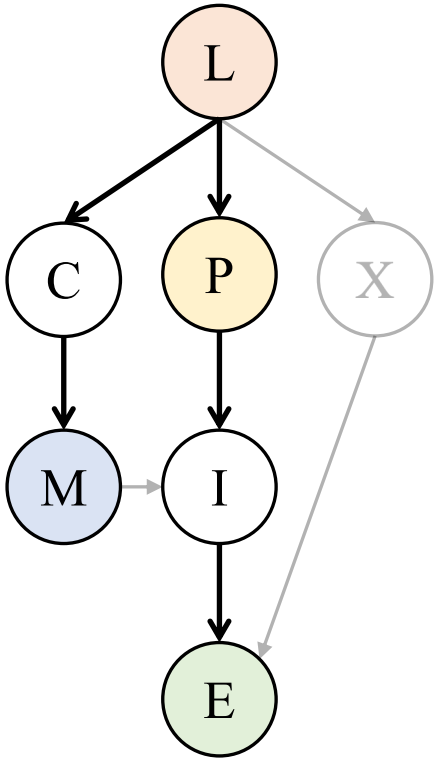
- The true causal effect is represented by path $M \rightarrow I \rightarrow E$.
- However, there exist three backdoor paths between M and E.
- Each backdoor path is corresponding to one critical bias.

Cause of Prompt Preference Bias



- The prompt P and PLM M are all correlated to linguistic distribution L.
- There exist a backdoor path M ← C ← L → P → I → E between PLM M and performance E.
- **The performance will be affected by both the task ability of PLM and the preference fitness of a prompt.**

Prompt Preference Leads to Inconsistent Performance



Prompt Preference Bias

- Design prompts that are <u>semantically equivalent and faithful but vary in linguistic expressions</u>.

`X` is owned by <?>.

`X` is a goods of <?>.

`X` belongs to <?>.

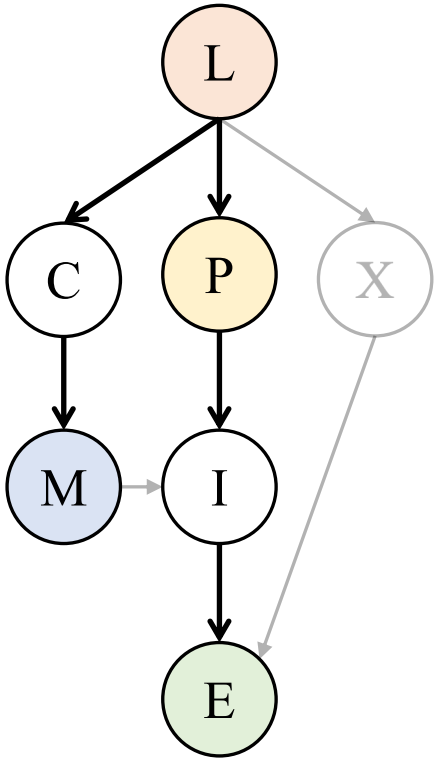
The owner of `X` is <?>.

`X` is an asset of <?>.

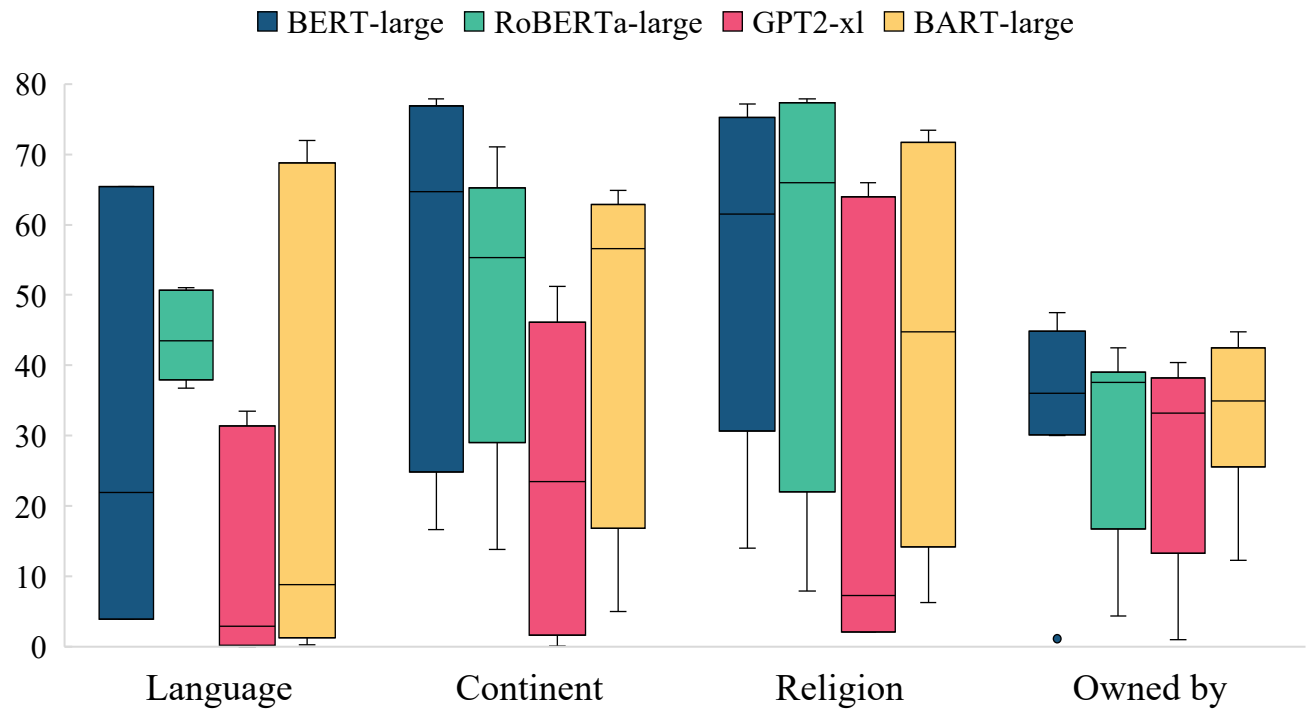
Performances

Semantically equivalent prompts PLMs

Prompt Preference Leads to Inconsistent Performance

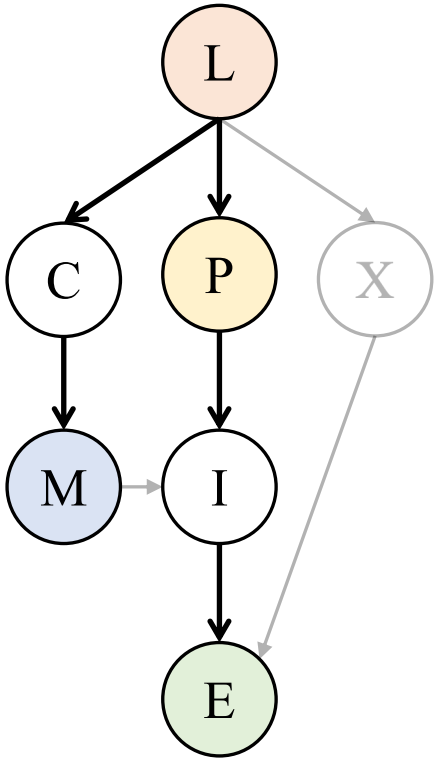


Prompt Preference Bias



- **Prompt selection significantly affects performance.**
  I. The same PLM can be assessed from "knowing nothing" to "sufficiently good" by only changing prompt.

Prompt Preference Leads to Inconsistent Performance

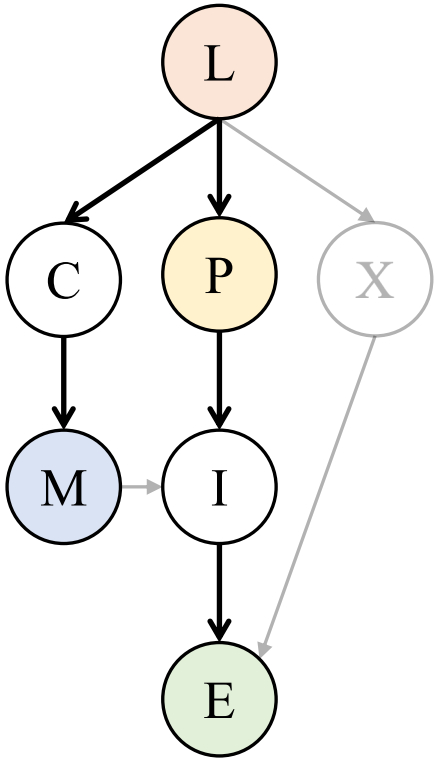


Prompt Preference Bias

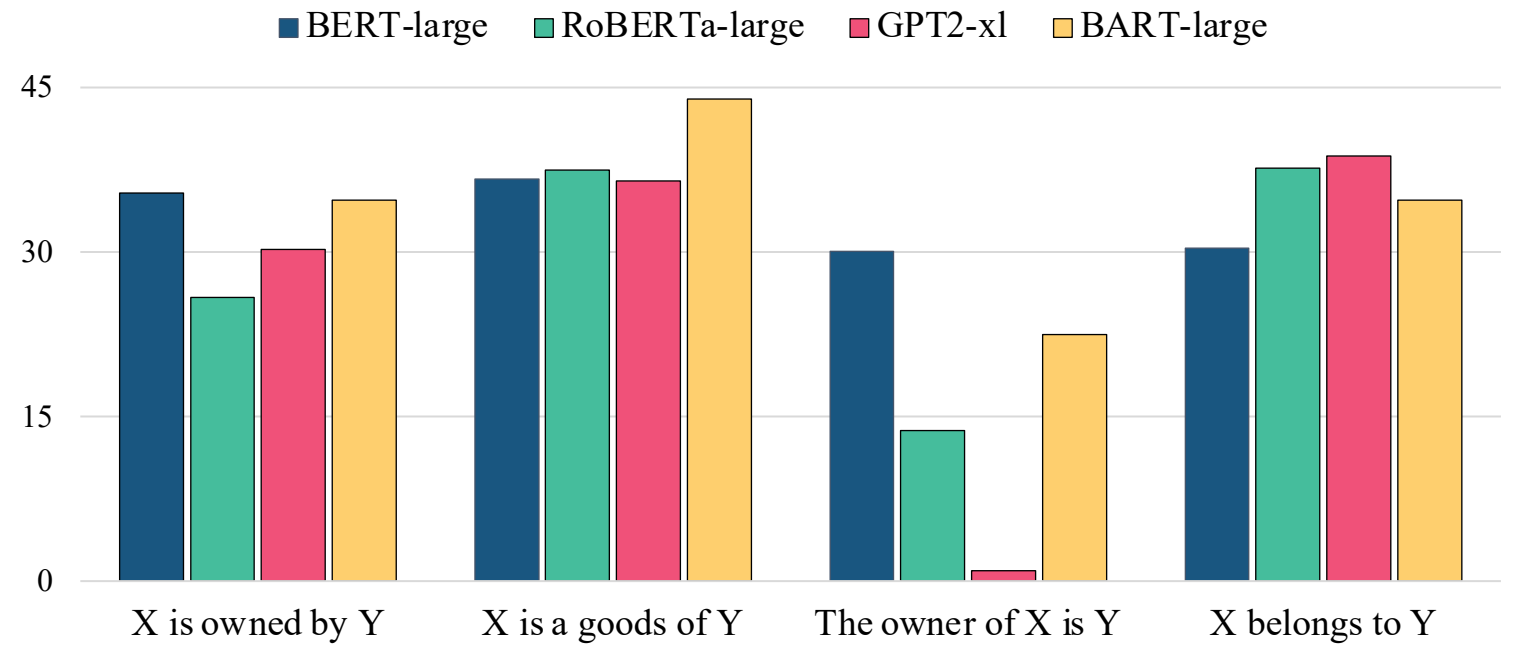
Models	LAMA P@1	Worst P@1	Best P@1	Std
BERT-large	39.08	23.45	46.73	8.75
RoBERTa-large	32.27	15.64	41.35	9.07
GPT2-xl	24.19	11.19	33.52	8.56
BART-large	27.68	16.21	38.93	8.35

- **Prompt selection significantly affects performance.**
  - I. The same PLM can be assessed from "knowing nothing" to "sufficiently good" by only changing prompt.
  - II. The prompt selection might result in larger performance variation than model selection.

Prompt Preference Leads to Inconsistent Performance



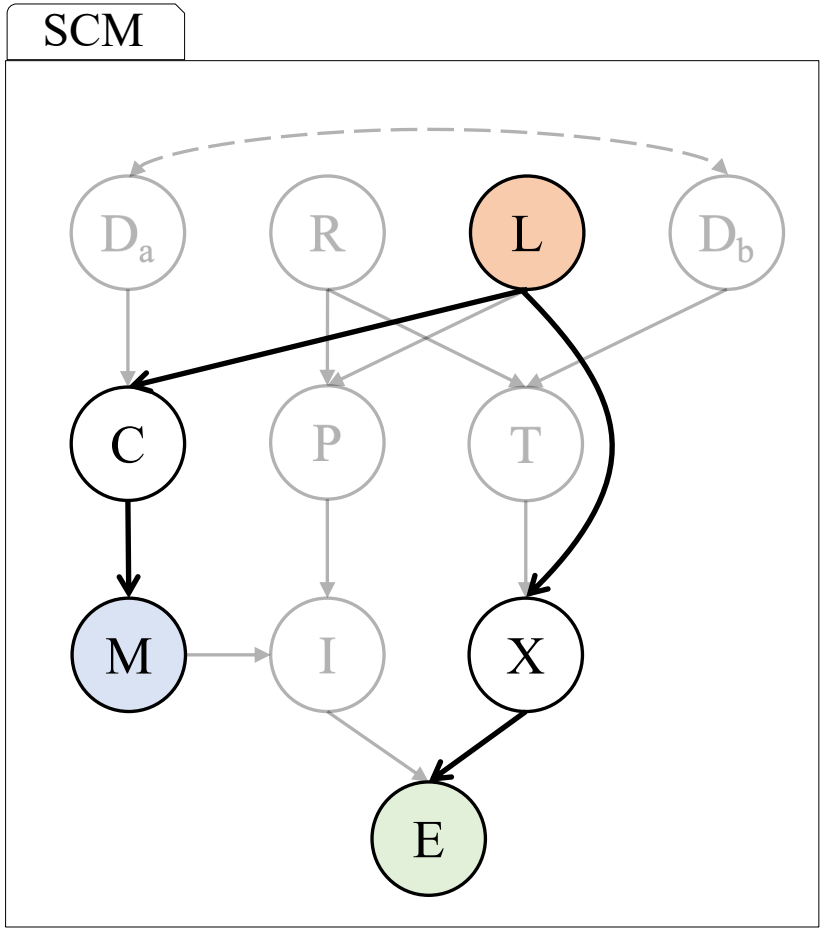
Prompt Preference Bias



Best	BERT	BART	BERT	GPT2
Worst	RoBERTa	GPT2	GPT2	BERT

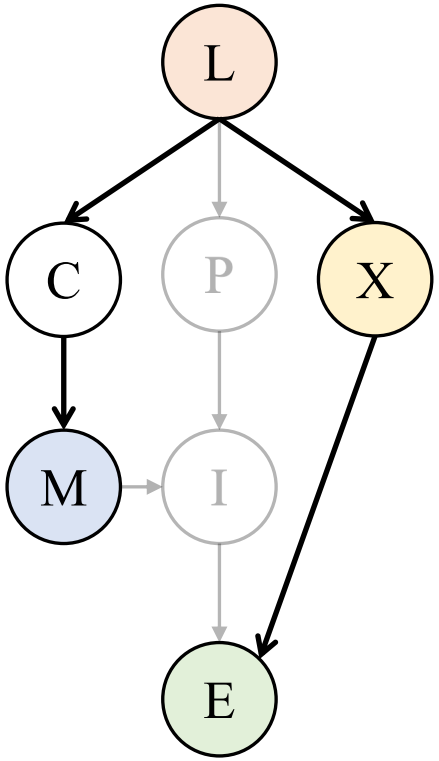
- **Prompt preference leads to inconsistent comparisons.**
  - I. On relation "owned by", prompt preference leads to 3 distinct "best" model and 3 distinct "worst" model.
  - II. PLMs' ranks on 96.88% relations are unstable when prompt varies.

Cause of Instance Verbalization Bias

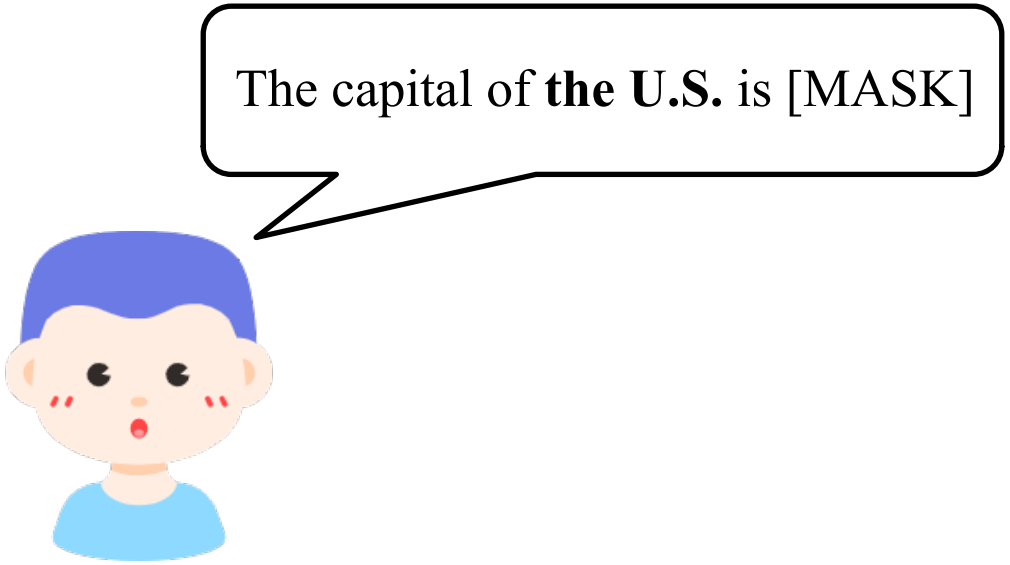


- The verbalized probing data X and PLM M are all correlated to linguistic distribution L.
- There exist a backdoor path M ← C ← L → X → E between PLM M and performance E.
- **Different PLMs may prefer different verbalizations due to mention coverage, expression overlap, etc.**

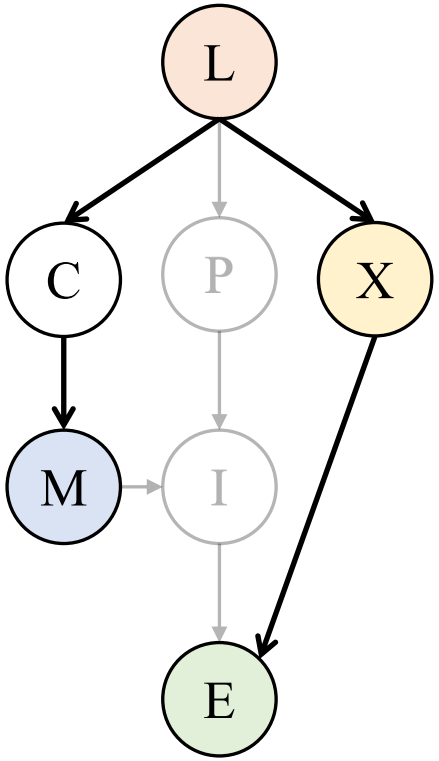
Instance Verbalizations Bring Unstable Predictions



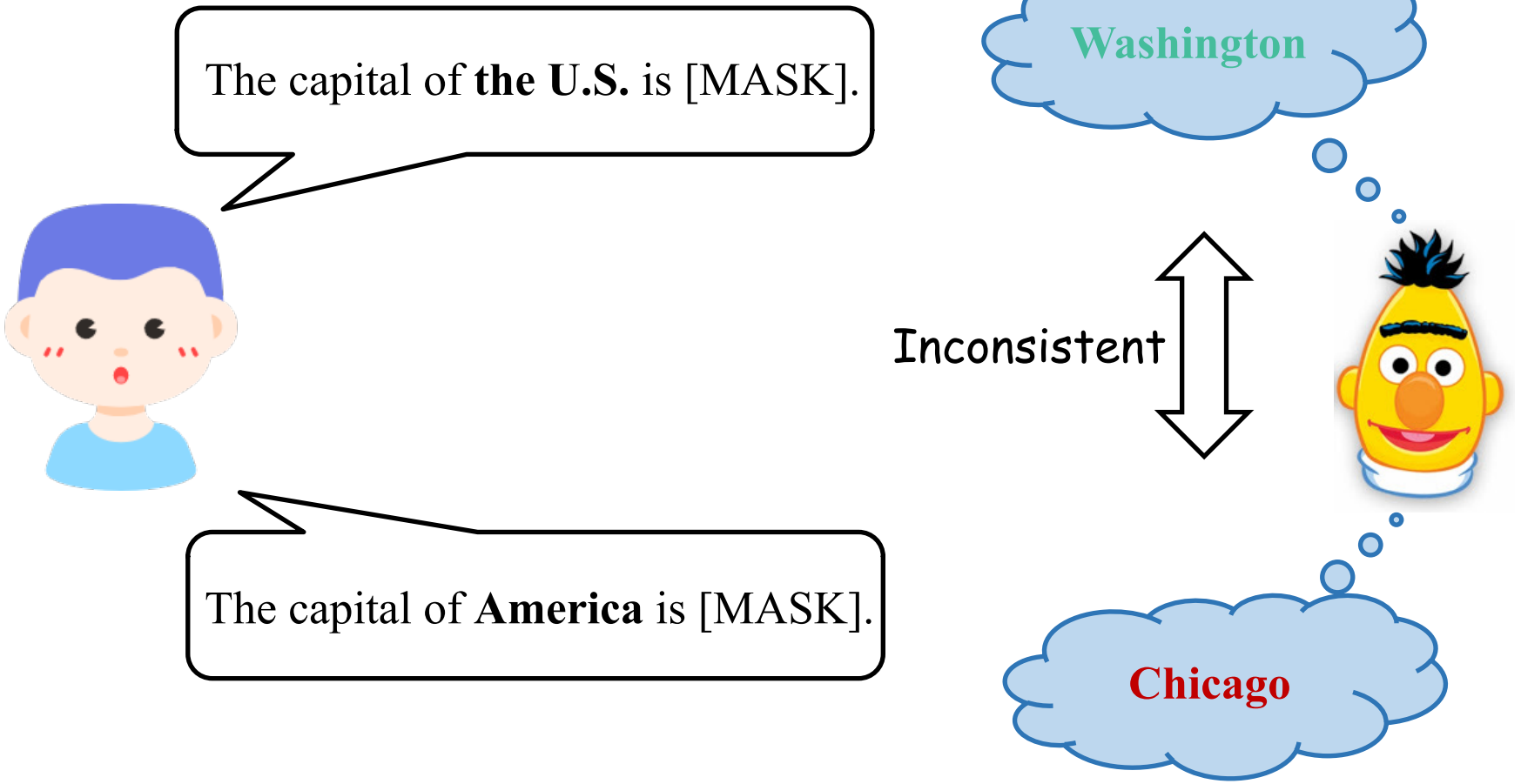
Instance Verbalization Bias



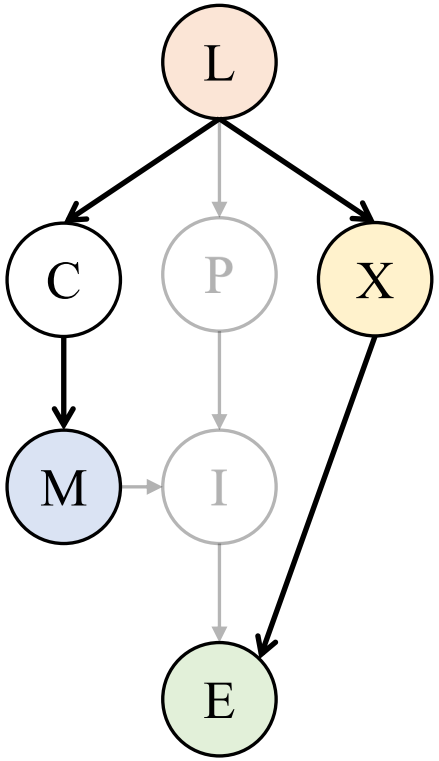
Instance Verbalizations Bring Unstable Predictions



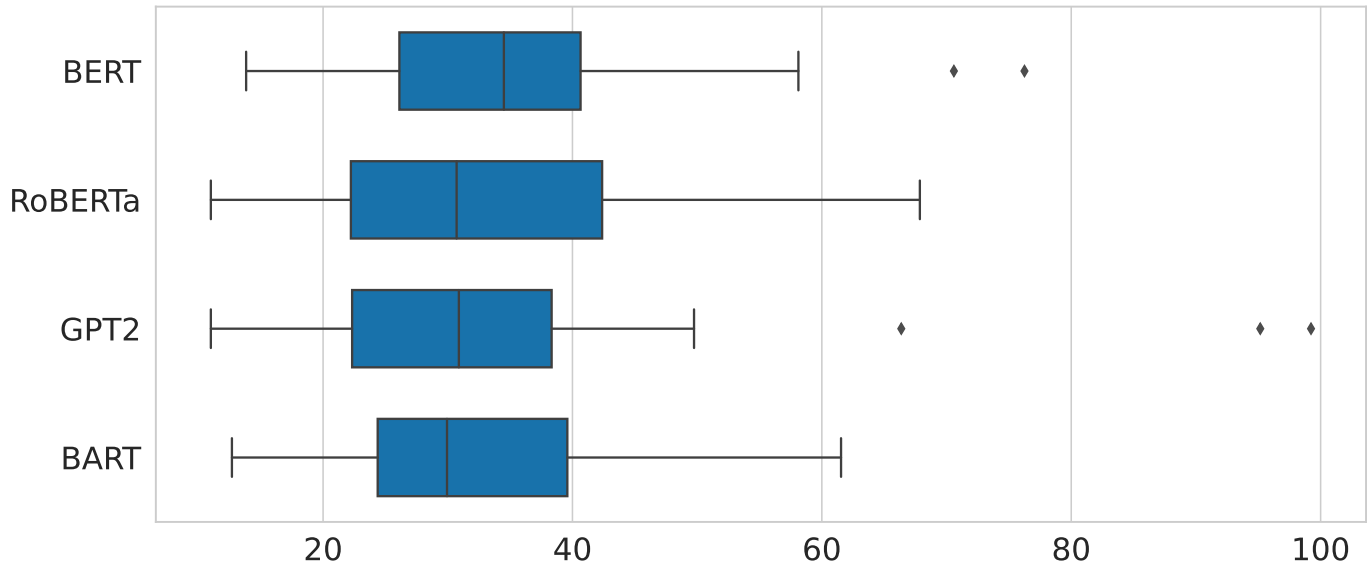
Instance Verbalization Bias



Instance Verbalizations Bring Unstable Predictions



Instance Verbalization Bias

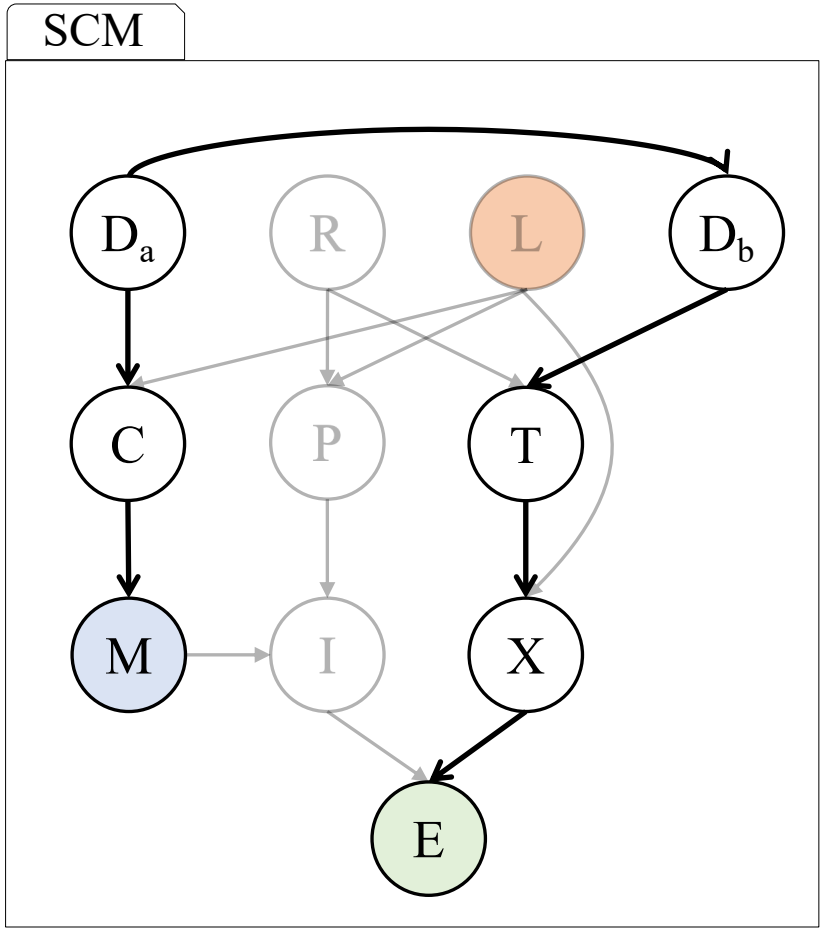


➢ **Instance verbalization bias brings unstable evaluation.**

I. Verbalization stability: the percentage of unchanged predictions when verbalization varies.

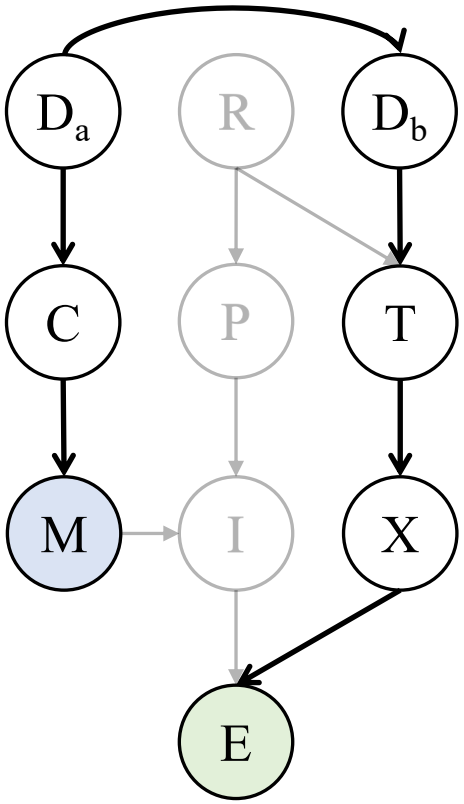
II. The average verbalization stability of 4 PLMs are all < 40%.

Cause of Sample Disparity Bias

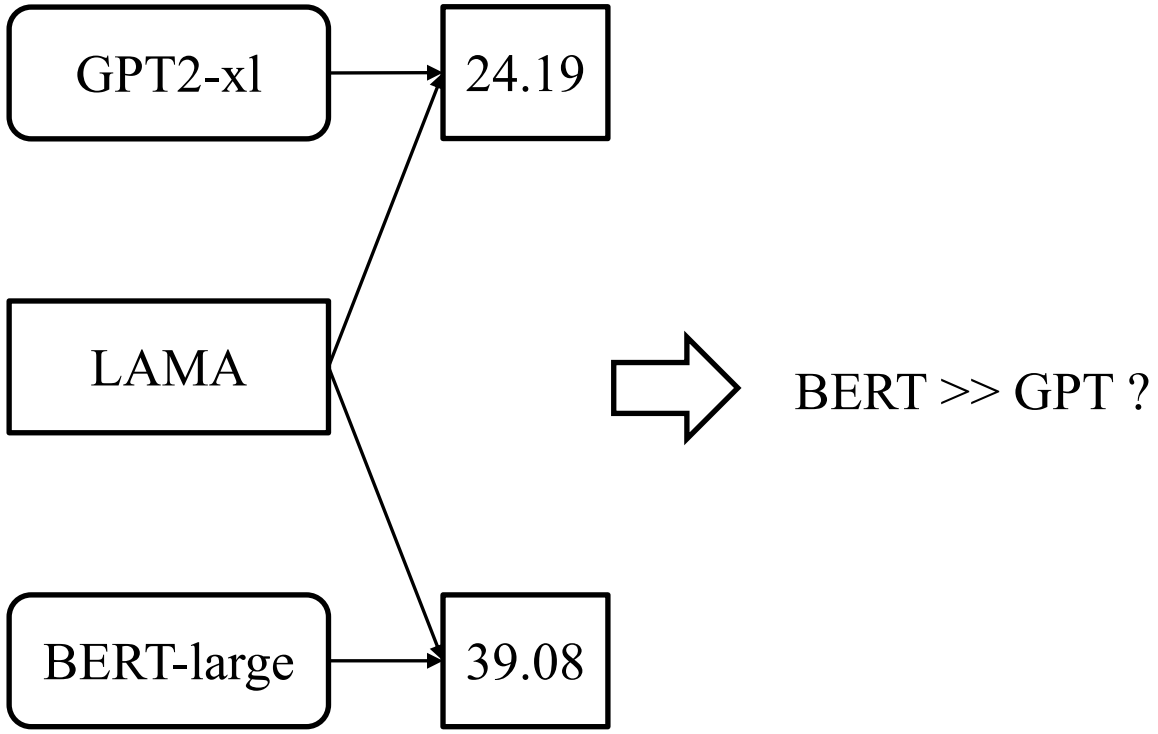


- There exist task-specific correlation between pretraining corpus distribution D_a and probing data distribution D_b.
- The backdoor path always exist no matter the causal relations between D_a and D_b.
- **The performance difference between different PLMs may due to the sample disparity of their pretraining corpus, rather than the ability divergence.**

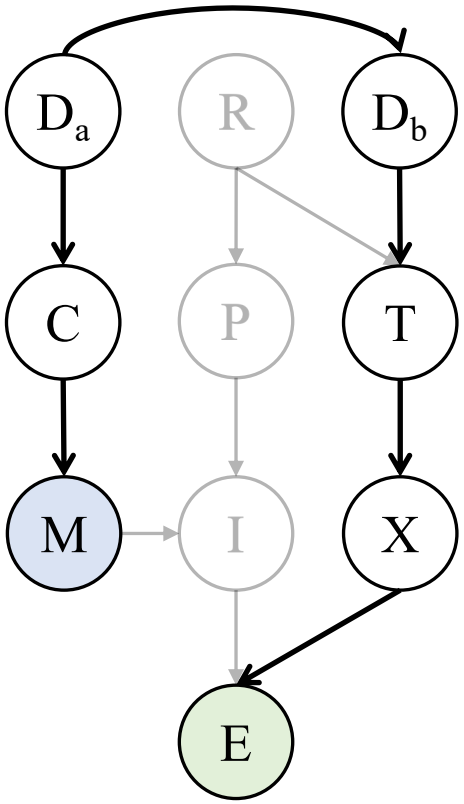
Sample Disparity Brings Biased Performance



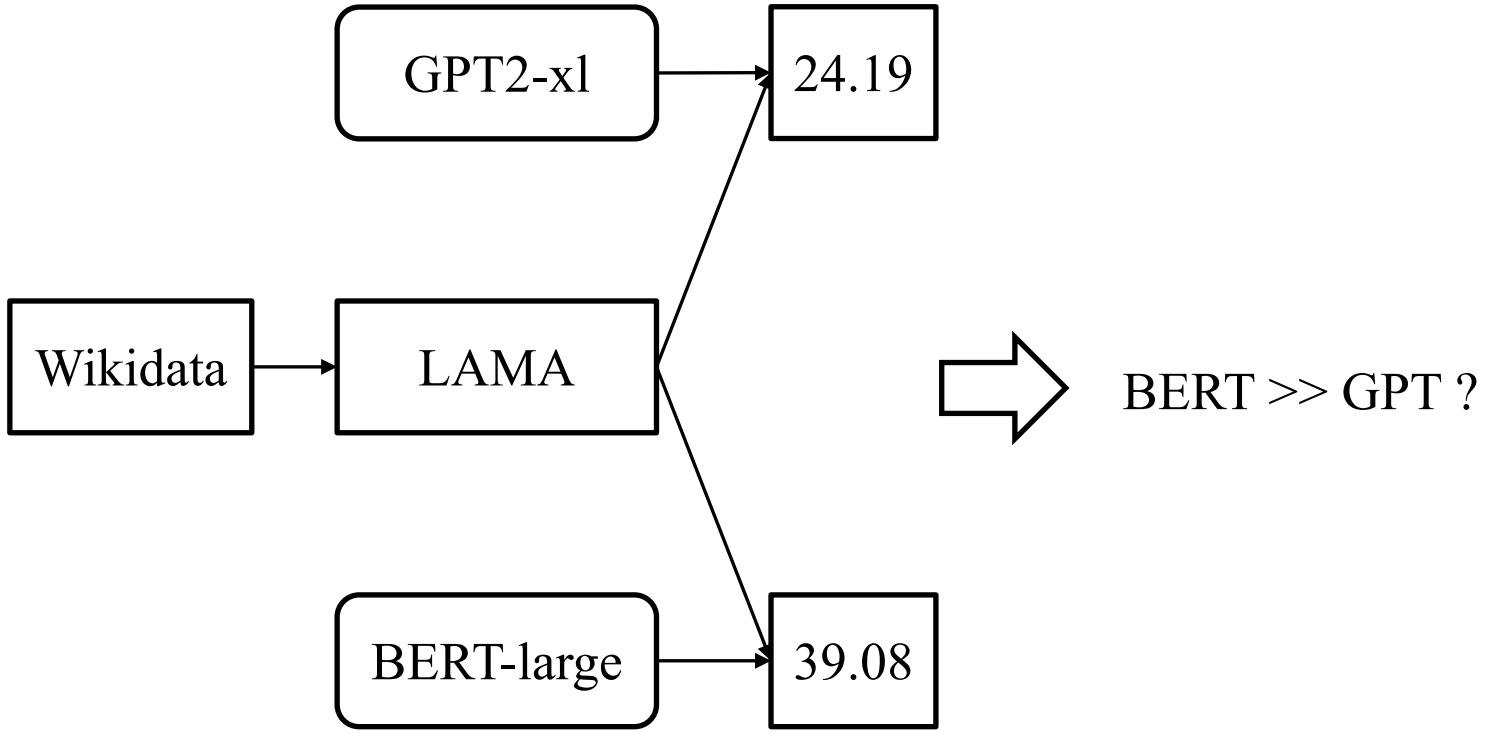
Sample Disparity Bias



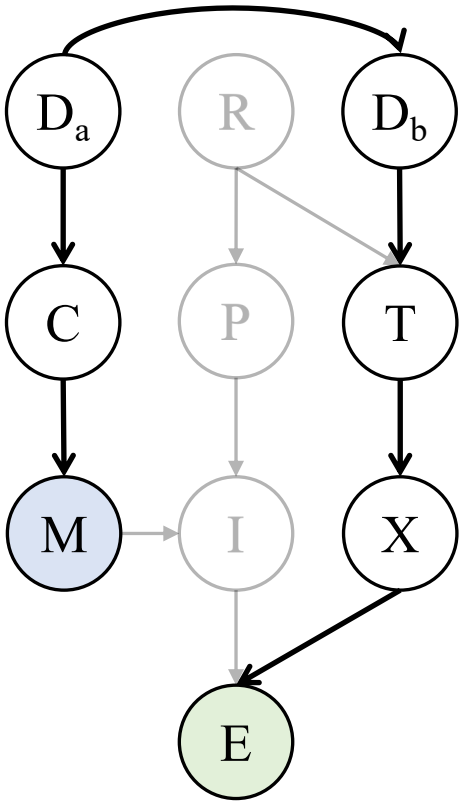
Sample Disparity Brings Biased Performance



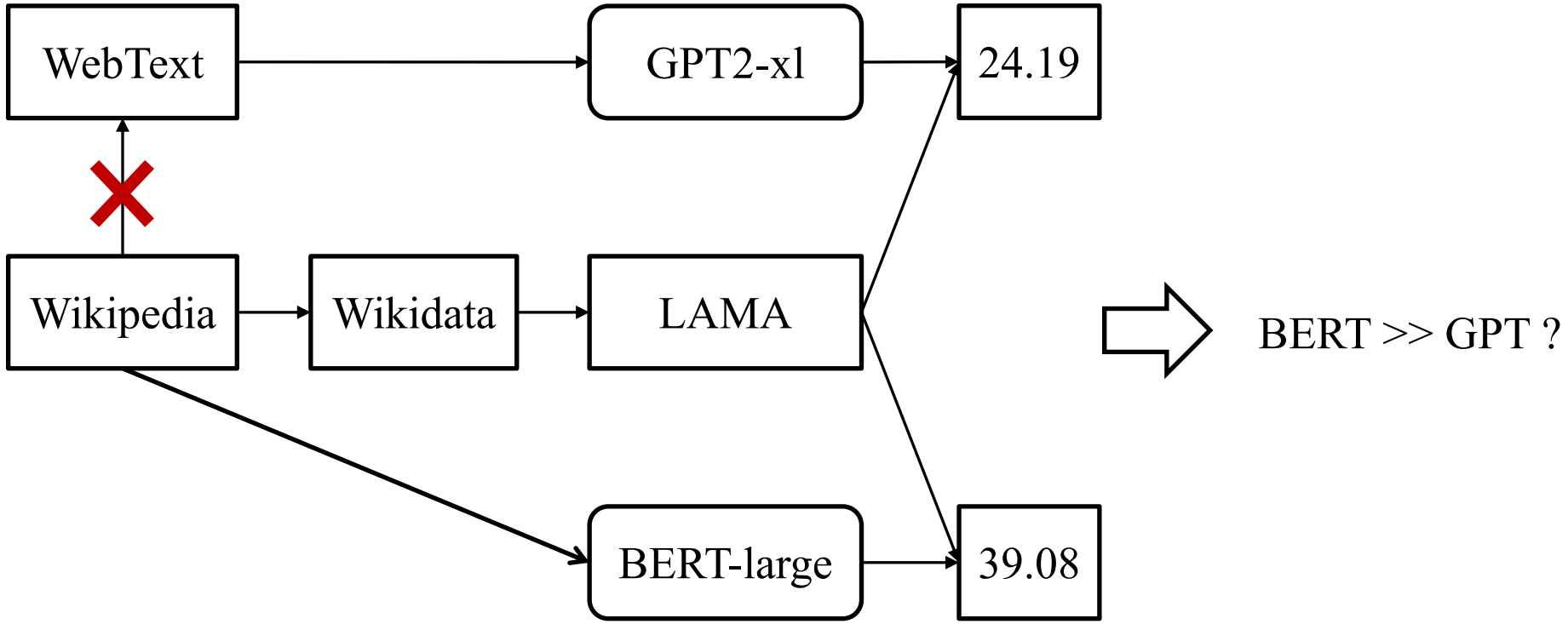
Sample Disparity Bias



Sample Disparity Brings Biased Performance

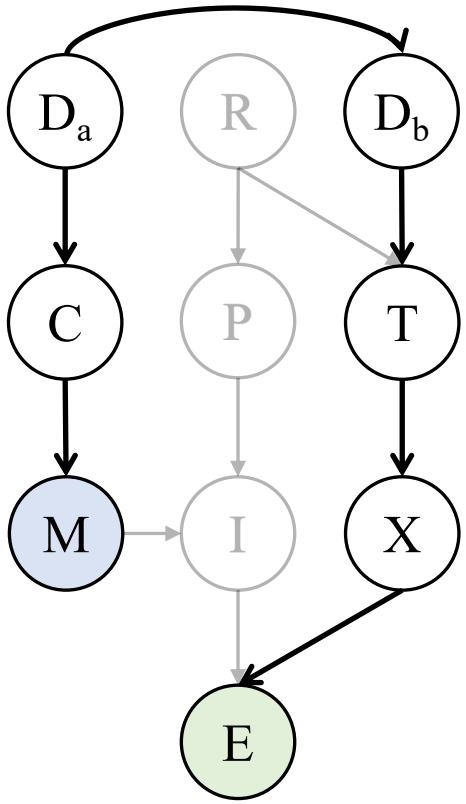


Sample Disparity Bias



- BERT's superior performance to GPT-2 may stem from the divergence of their pretraining corpus, where BERT's pretraining corpus contains Wikipedia, while GPT2's pretraining corpus doesn't.

Sample Disparity Brings Biased Performance



Sample Disparity Bias

LAMA

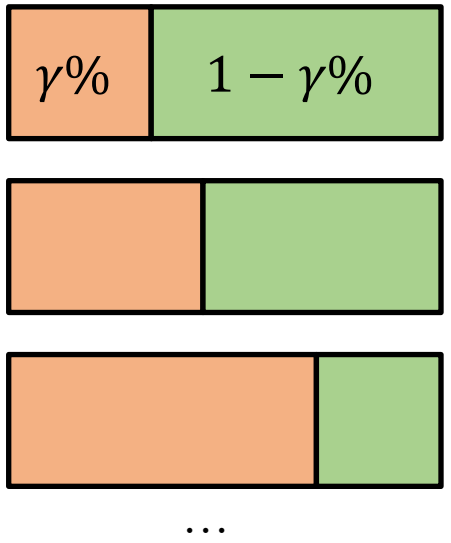
e.g., (Jordan, birthplace, Brooklyn)

WIKI-LAMA

e.g., Jordan was born at Cumberland Hospital in Fort Greene, Brooklyn…

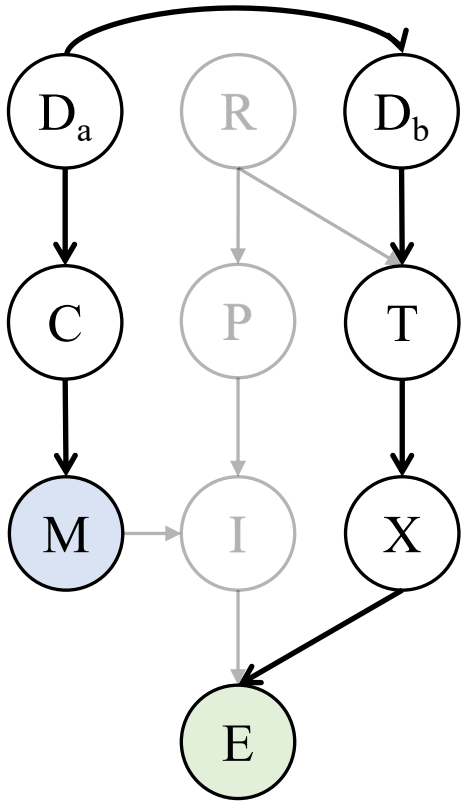
+ WebText

Further Pretrain



γ: The correlation degree between further pretrain data and probing data.

Sample Disparity Brings Biased Performance

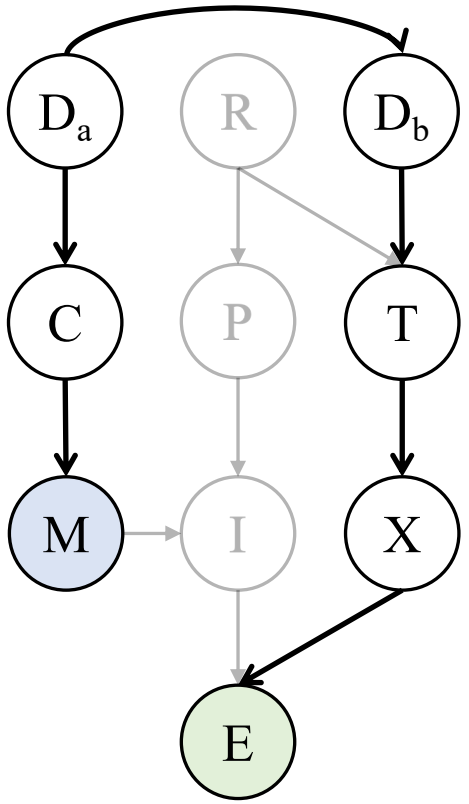


Sample Disparity Bias

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

- **Sample disparity significantly influences performance.**
  I. The larger correlation degree γ will result in better performance for both BERT and GPT2.

Sample Disparity Brings Biased Performance

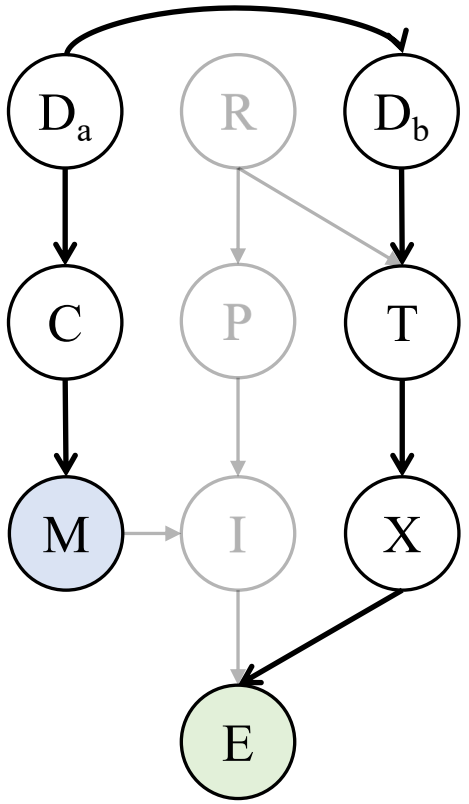


Sample Disparity Bias

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

- **Sample disparity contributes to performance difference.**
  I. The performance gap between GPT-2 and BERT significantly narrows down when further pretrained on the same data.

Sample Disparity Brings Biased Performance

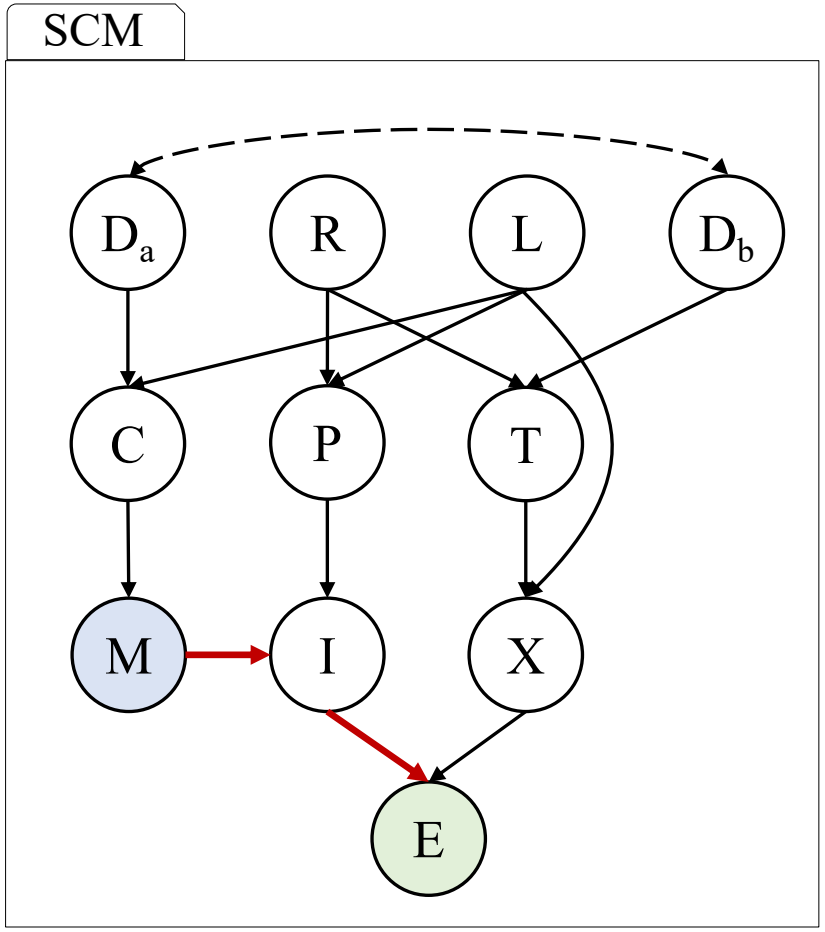


Sample Disparity Bias

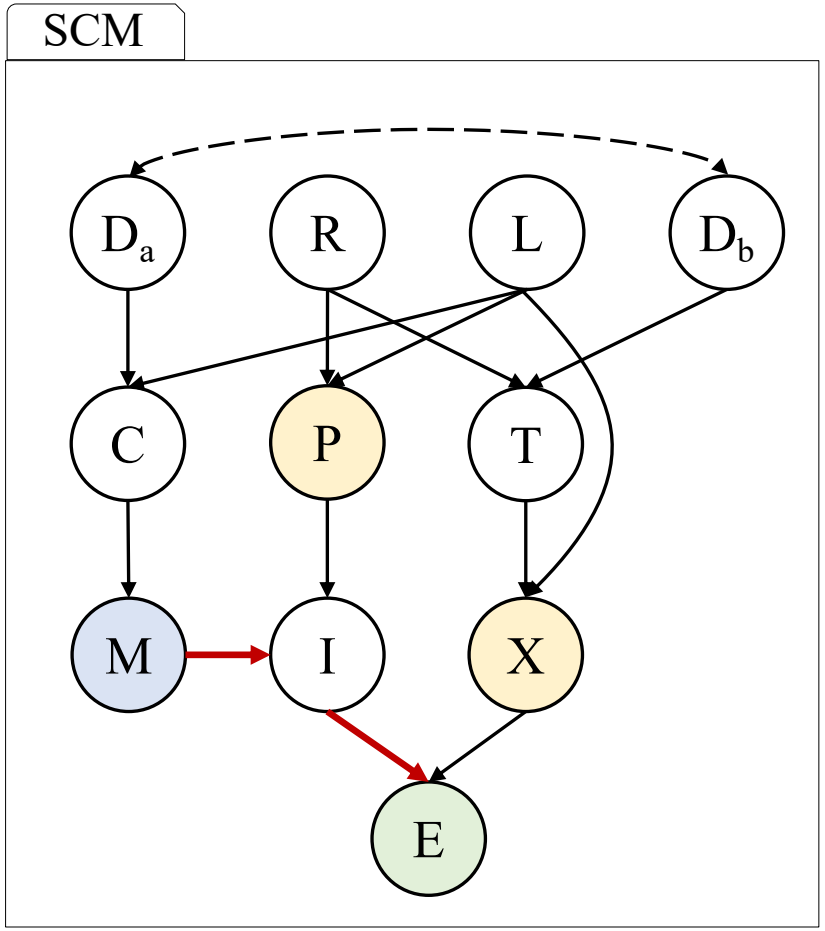
γ%	BERT-base	BERT-large	GPT2-base	GPT2-medium
0%	30.54	33.08	15.22	22.11
20%	35.77	39.56	22.02	28.21
40%	38.68	39.75	24.32	30.29
60%	38.72	40.68	25.42	31.16
80%	39.79	41.48	25.65	31.88
100%	40.15	42.51	26.82	33.12
None	37.13	39.08	16.88	22.60

- **Sample disparity contributes to performance difference.**
  - I. The performance gap between GPT-2 and BERT significantly narrows down when further pretrained on the same data.
  - II. Further pretrain BERT on WebText will significantly undermines the performance.

Bias Elimination via Causal Intervention

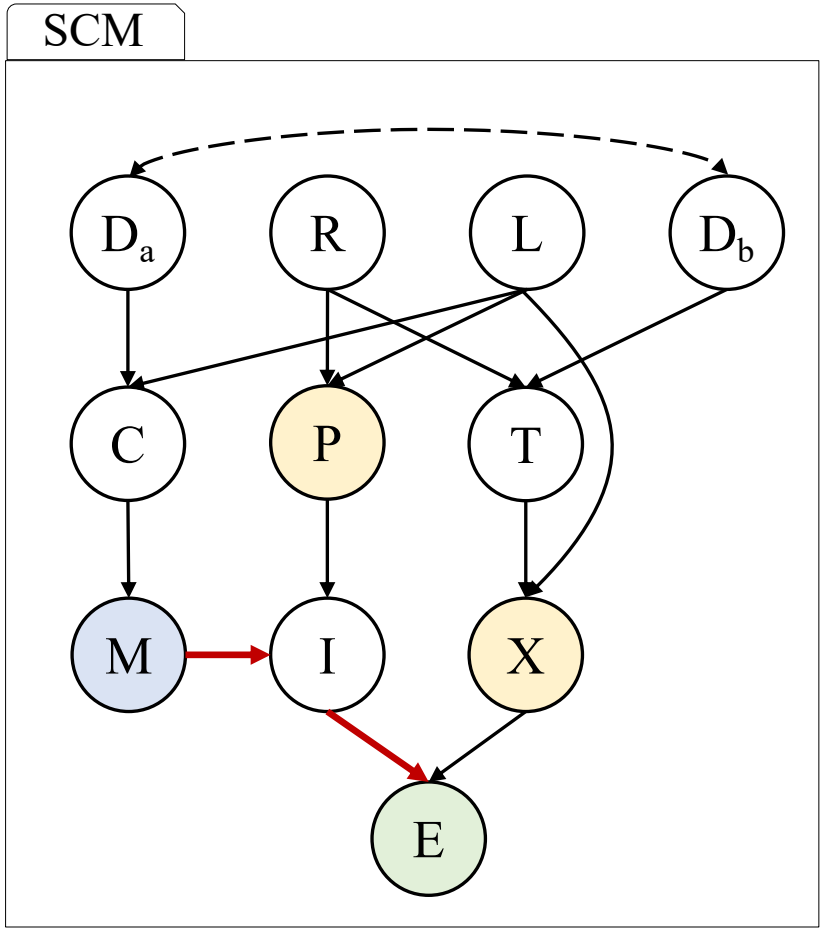


Bias Elimination via Causal Intervention



- **Backdoor Criterion**
  I. Find a variable set that block each backdoor paths.

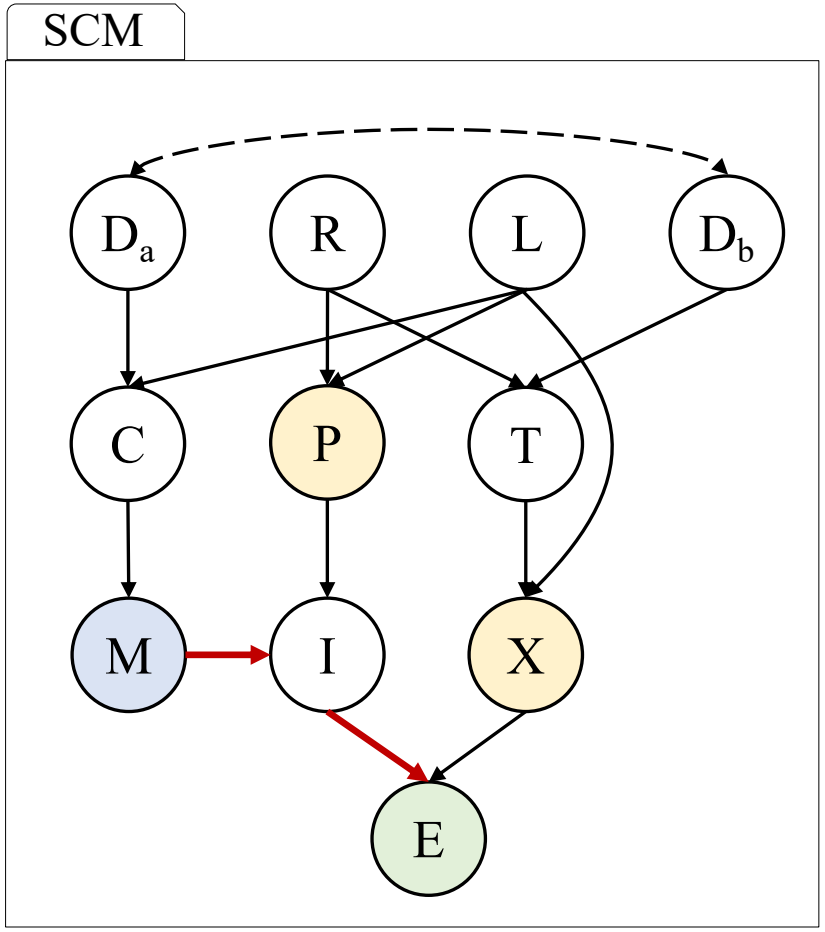
Bias Elimination via Causal Intervention



- **Backdoor Criterion**
  I. Find a variable set that block each backdoor paths.
  II. Conduct backdoor adjustment:

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

Bias Elimination via Causal Intervention



- **Backdoor Criterion**
  I. Find a variable set that block each backdoor paths.
  II. Conduct backdoor adjustment:

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

  III. Given a special assumption about $\mathcal{P}(p,x)$, we sample prompts and verbalizations.

Bias Elimination via Causal Intervention

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**

- Randomly sample 1000 subsets with 20 relations from LAMA.
- **Rank Consistency** for a PLM: percentage of most popular rank in 1000 runtimes.
- **Overall Rank**: the percentage of most popular rank that all PLMS remain the same.
- Original: prompts and verbalizations from LAMA.
- Random: randomly sample prompts and verbalizations every time.

Bias Elimination via Causal Intervention

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**

- **Causal intervention can significantly improve the evaluation consistency.**
  I. The consistency of current prompt-based probing is poor.

Bias Elimination via Causal Intervention

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**

- **Causal intervention can significantly improve the evaluation consistency.**
  - I. The consistency of current prompt-based probing is poor.
  - II. Causal intervention can significantly improve the overall rank consistency.
  - III. The rank of most PLMs is stable after causal intervention.

Discussions and Conclusions

- A causal analysis framework is proposed to effectively identify, interpret and eliminate evaluation biases with a theoretical guarantee.
- Can be extended and adapted to other evaluation settings in a principled manner.
- Our conclusions echo that we need to rethink the criteria for identifying better pretrained language models.

Thanks!