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

Boxi Cao[1,3], **Hongyu Lin**[1], **Xianpei Han**[1,2,4]*, **Fangchao Liu**[1,3], **Le Sun**[1,2]*
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4] Beijing Academy of Artificial Intelligence, Beijing, China
{boxi2020,hongyu,xianpei,fangchao2017,sunle}@iscas.ac.cn

Abstract

Prompt-based probing has been widely used in evaluating the abilities of pretrained language models (PLMs). Unfortunately, recent studies have discovered such an evaluation may be inaccurate, inconsistent and unreliable. Furthermore, the lack of understanding its inner workings, combined with its wide applicability, has the potential to lead to unforeseen risks for evaluating and applying PLMs in real-world applications. To discover, understand and quantify the risks, this paper investigates the prompt-based probing from a causal view, highlights three critical biases which could induce biased results and conclusions, and proposes to conduct debiasing via causal intervention. This paper provides valuable insights for the design of unbiased datasets, better probing frameworks and more reliable evaluations of pretrained language models. Furthermore, our conclusions also echo that we need to rethink the criteria for identifying better pretrained language models[1].

1 Introduction

During the past few years, the great success of pretrained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Raffel et al., 2020) raises extensive attention about evaluating what knowledge do PLMs actually entail. One of the most popular approaches is prompt-based probing (Petroni et al., 2019; Davison et al., 2019; Brown et al., 2020; Schick and Schütze, 2020; Ettinger, 2020; Sun et al., 2021), which assesses whether PLMs are knowledgable for a specific task by querying PLMs with task-specific prompts. For example, to evaluate whether BERT knows the birthplace of Michael Jordan, we could query BERT with "Michael Jordan was born in [MASK]". Recent studies often construct prompt-based probing datasets, and take PLMs' performance on these datasets as their abilities for the corresponding tasks. Such a probing evaluation has been wildly used in many benchmarks such as SuperGLUE (Wang et al., 2019; Brown et al., 2020), LAMA (Petroni et al., 2019), oLMpics (Talmor et al., 2020), LM diagnostics (Ettinger, 2020), CAT (Zhou et al., 2020), X-FACTR (Jiang et al., 2020a), BioLAMA (Sung et al., 2021), etc.

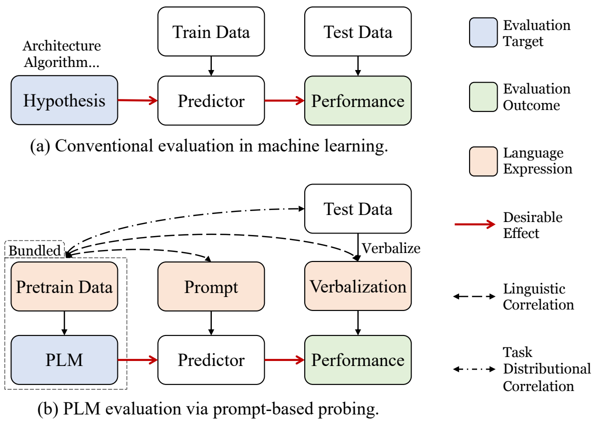

Figure 1: The illustrated procedure for two kinds of evaluation criteria.

Unfortunately, recent studies have found that evaluating PLMs via prompt-based probing could be inaccurate, inconsistent, and unreliable. For example, Poerner et al. (2020) finds that the performance may be overestimated because many instances can be easily predicted by only relying on surface form shortcuts. Elazar et al. (2021) shows that semantically equivalent prompts may result in quite different predictions. Cao et al. (2021) demonstrates that PLMs often generate unreliable predictions which are prompt-related but not knowledge-related.

In these cases, the risks of blindly using prompt-based probing to evaluate PLMs, without understanding its inherent vulnerabilities, are significant. Such biased evaluations will make us overestimate or underestimate the real capabilities of PLMs, mislead our understanding of models, and result in

*Corresponding Authors
[1]We openly released the source code and data at https://github.com/c-box/causalEval.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5796 - 5808
May 22-27, 2022

wrong conclusions. Therefore, to reach a trustworthy evaluation of PLMs, it is necessary to dive into the probing criteria and understand the following two critical questions: *1) What biases exist in current evaluation criteria via prompt-based probing? 2) Where do these biases come from?*

To this end, we compared PLM evaluation via prompt-based probing with conventional evaluation criteria in machine learning. Figure 1 shows their divergences. Conventional evaluations aim to evaluate different hypotheses (e.g., algorithms or model structures) for a specific task. The tested hypotheses are raised independently of the training/test data generation. However, this independence no longer sustains in prompt-based probing. There exist more complicated implicit connections between pretrained models, probing data, and prompts, mainly due to the bundled pretraining data with specific PLMs. These unaware connections serve as invisible hands that can even dominate the evaluation criteria from both linguistic and task aspects. From the linguistic aspect, because pretraining data, probing data and prompts are all expressed in the form of natural language, there exist inevitable *linguistic correlations* which can mislead evaluations. From the task aspect, the pretraining data and the probing data are often sampled from correlated distributions. Such invisible *task distributional correlations* may significantly bias the evaluation. For example, Wikipedia is a widely used pretraining corpus, and many probing data are also sampled from Wikipedia or its extensions such as Yago, DBPedia or Wikidata (Petroni et al., 2019; Jiang et al., 2020a; Sung et al., 2021). As a result, such task distributional correlations will inevitably confound evaluations via domain overlapping, answer leakage, knowledge coverage, etc.

To theoretically identify how these correlations lead to biases, we revisit the prompt-based probing from a causal view. Specifically, we describe the evaluation procedure using a structural causal model (Pearl et al., 2000) (SCM), which is shown in Figure 2a. Based on the SCM, we find that the linguistic correlation and the task distributional correlation correspond to three backdoor paths in Figure 2b-d, which lead to three critical biases:

- **Prompt Preference Bias,** which mainly stems from the underlying linguistic correlations between PLMs and prompts, i.e., the performance may be biased by the fitness of a prompt to PLMs' linguistic preference. For instance, semantically equivalent prompts will lead to different biased evaluation results.

- **Instance Verbalization Bias,** which mainly stems from the underlying linguistic correlations between PLMs and verbalized probing datasets, i.e., the evaluation results are sensitive and inconsistent to the different verbalizations of the same instance (e.g., representing the U.S.A. with the U.S. or America).

- **Sample Disparity Bias,** which mainly stems from the invisible distributional correlation between pretraining and probing data, i.e., the performance difference between different PLMs may due to the sample disparity of their pretraining corpus, rather than their ability divergence. Such invisible correlations may mislead evaluation results, and thus lead to implicit, unaware risks of applying PLMs in real-world applications.

We further propose to conduct causal intervention via backdoor adjustments, which can reduce bias and ensure a more accurate, consistent and reliable probing under given assumptions. Note that this paper not intends to create a "universal correct" probing criteria, but to remind the underlying invisible risks, to understand how spurious correlations lead to biases, and to provide a causal toolkit for debiasing probing under specific assumptions. Besides, we believe that our discoveries not only exist in prompt-based probing, but will also influence all prompt-based applications to pretrained language models. Consequently, our conclusions echo that we need to rethink the criteria for identifying better pretrained language models with the above-mentioned biases.

Generally, the main contributions of this paper are:

- We investigate the critical biases and quantify their risks of evaluating pretrained language models with widely used prompt-based probing, including prompt preference bias, instance verbalization bias, and sample disparity bias.

- We propose a causal analysis framework, which can be used to effectively identify, understand, and eliminate biases in prompt-based probing evaluations.

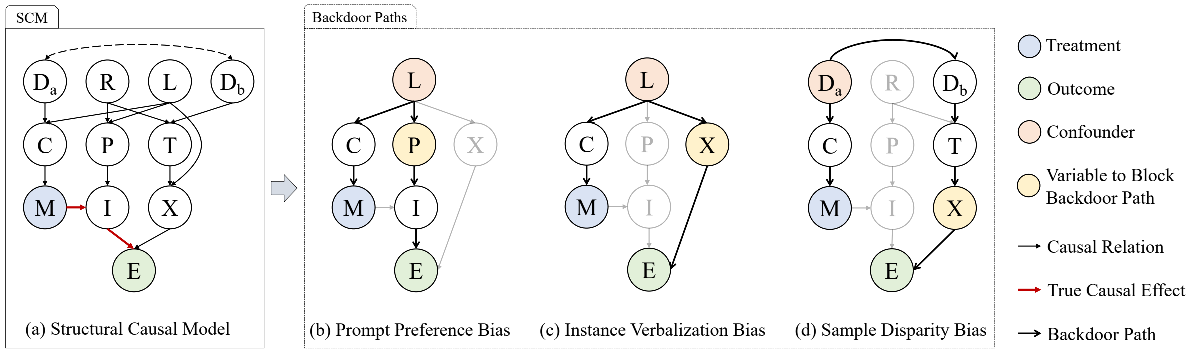

Figure 2: The structural causal model for factual knowledge probing and the three backdoor paths in SCM correspond to three biases.

- We provide valuable insights for the design of unbiased datasets, better probing frameworks, and more reliable evaluations, and echo that we should rethink the evaluation criteria for pretrained language models.

2 Background and Experimental Setup

2.1 Causal Inference

Causal inference is a promising technique for identifying undesirable biases and fairness concerns in benchmarks (Hardt et al., 2016; Kilbertus et al., 2017; Kusner et al., 2017; Vig et al., 2020; Feder et al., 2021). Causal inference usually describes the causal relations between variables via Structural Causal Model (SCM), then recognizes confounders and spurious correlations for bias analysis, finally identifies true causal effects by eliminating biases using causal intervention techniques.

SCM The structural causal model (Pearl et al., 2000) describes the relevant features in a system and how they interact with each other. Every SCM is associated with a graphical causal model $G = \{V, f\}$, which consists of a set of nodes representing variables V, as well as a set of edges between the nodes representing the functions f to describe the causal relations.

Causal Intervention To identify the true causal effects between an ordered pair of variables (X, Y), Causal intervention fixes the value of $X = x$ and removes the correlations between X and its precedent variables, which is denoted as $do(X = x)$. In this way, $\mathcal{P}(Y = y|do(X = x))$ represents the true causal effects of treatment X on outcome Y (Pearl et al., 2016).

Backdoor Path When estimating the causal effect of X on Y, the backdoor paths are the non-causal paths between X and Y with an arrow into X, e.g., $X \leftarrow Z \rightarrow Y$. Such paths will confound the effect that X has on Y but not transmit causal influences from X, and therefore introduce spurious correlations between X and Y.

Backdoor Criterion The Backdoor Criterion is an important tool for causal intervention. Given an ordered pair of variables (X, Y) in SCM, and a set of variables Z where Z contains no descendant of X and blocks every backdoor path between X and Y, then the causal effects of $X = x$ on Y can be calculated by:

$$\mathcal{P}(Y = y|do(X = x)) = \sum_{z} \mathcal{P}(Y = y|X = x, Z = z)\mathcal{P}(Z = z), \quad (1)$$

where $\mathcal{P}(Z = z)$ can be estimated from data or priorly given, and is independent of X.

2.2 Experimental Setup

Task This paper investigates prompt-based probing on one of the most representative and well-studied tasks – factual knowledge probing (Liu et al., 2021b). For example, to evaluate whether BERT knows the birthplace of Michael Jordan, factual knowledge probing queries BERT with "Michael Jordan was born in [MASK]", where `Michael Jordan` is the verbalized subject mention, *"was born in"* is the verbalized prompt of relation `birthplace`, and [MASK] is a placeholder for the target object.

Data We use LAMA (Petroni et al., 2019) as our primary dataset, which is a set of knowledge triples sampled from Wikidata. We remove the N-M relations (Elazar et al., 2021) which are unsuitable for the P@1 metric and retain 32 probing relations in the dataset. Please refer to the appendix for detail.

Pretrained Models We conduct probing experiments on 4 well-known PLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020), which correspond to 3 representative PLM architectures, including autoencoder (BERT, RoBERTa), autoregressive (GPT-2) and denoising autoencoder (BART).

3 Structural Causal Model for Factual Knowledge Probing

In this section, we formulate the SCM for factual knowledge probing procedure and describe the key variables and causal relations.

The SCM is shown in Figure 2a, which contains 11 key variables: 1) **Pretraining corpus distribution** D_a; 2) **Pretraining corpus** C, e.g., Webtext for GPT2, Wikipedia for BERT; 3) **Pretrained language model** M; 4) **Linguistic distribution** L, which guides how a concept is verbalized into natural language expression, e.g., relation to prompt, entity to mention; 5) **Relation** R, e.g., `birthplace`, `capital`, each relation corresponds to a probing task; 6) **Verbalized prompt** P for each relation , e.g, x *was born in* y; 7) **Task-specific predictor** I, which is a PLM combined with a prompt, e.g., <BERT, *was born in*> as a `birthplace` predictor; 8) **Probing data distribution** D_b, e.g., fact distribution in Wikidata; 9) **Sampled probing data** T such as LAMA, which are sampled entity pairs (e.g., <Q41421, Q18419> in Wikidata) of relation R; 10) **Verbalized instances** X, (e.g., <Michael Jordan, Brooklyn> from <Q41421, Q18419>); 11) **Performance** E of the predictor I on X.

The causal paths of the prompt-based probing evaluation contains:

- **PLM Pretraining.** The path $\{D_a, L\} \rightarrow C \rightarrow M$ represents the pretraining procedure for language model M, which first samples pretraining corpus C according to pretraining corpus distribution D_a and linguistic distribution L, then pretrains M on C.
- **Prompt Selection.** The path $\{R, L\} \rightarrow P$ represents the prompt selection procedure, where each prompt P must exactly express the semantics of relation R, and will be influenced by the linguistic distribution L.
- **Verbalized Instances Generation.** The path $\{D_b, R\} \rightarrow T \rightarrow X \leftarrow L$ represents the generation procedure of verbalized probing instances X, which first samples probing data T of relation R according to data distribution D_b, then verbalizes the sampled data T into X according to the linguistic distribution L.
- **Performance Estimation.** The path $\{M, P\} \rightarrow I \rightarrow E \leftarrow X$ represents the performance estimation procedure, where the predictor I is first derived by combining PLM M and prompt P, and then the performance E is estimated by applying predictor I on verbalized instances X.

To evaluate PLMs' ability on fact extraction, we need to estimate $\mathcal{P}(E|do(M = m), R = r)$. Such true causal effects are represented by the path $M \rightarrow I \rightarrow E$ in SCM. Unfortunately, there exist three backdoor paths between pretrained language model M and performance E, as shown in Figure 2b-d. These spurious correlations make the observation correlation between M and E cannot represent the true causal effects of M on E, and will inevitably lead to biased evaluations. In the following, we identify three critical biases in the prompt-based probing evaluation and describe the manifestations, causes, and casual interventions for each bias.

4 Prompt Preference Bias

In prompt-based probing, the predictor of a specific task (e.g., the knowledge extractor of relation `birthplace`) is a PLM M combined with a prompt P (e.g., BERT + *was born in*). However, PLMs are pretrained on specific text corpus, therefore will inevitably prefer prompts sharing the same linguistic regularity with their pretraining corpus. Such implicit prompt preference will confound the true causal effects of PLMs on evaluation performance, i.e., the performance will be affected by both the task ability of PLMs and the preference fitness of a prompt. In the following, we investigate prompt preference bias via causal analysis.

4.1 Prompt Preference Leads to Inconsistent Performance

In factual knowledge probing, we commonly assign one prompt for each relation (e.g., `X was born in Y` for `birthplace`). However, different PLMs may prefer different prompts, and it is unable to disentangle the influence of prompt preference from the final performance. Such invisible

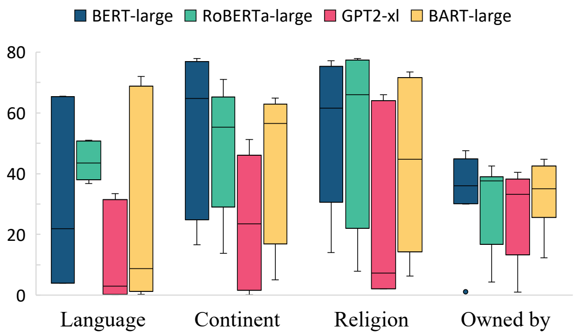

Figure 3: The variances of P@1 performance of 4 PLMs on 4 relations using semantically equivalent prompts. We can see the performance varies significantly.

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

Table 1: The P@1 performance divergence of prompt selection averaged over all relations, we can see prompt preference results in inconsistent performance.

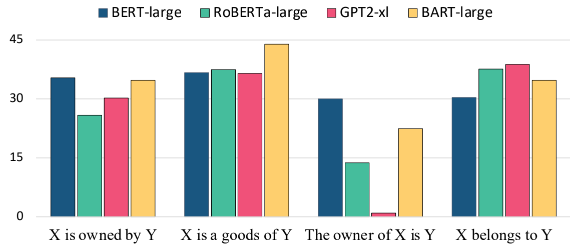

Figure 4: The P@1 performance of 4 PLMs using 4 different prompts of relation `owned by`, where the rank of 4 PLMs is unstable on different prompts: prompt preference leads to 3 distinct "best" models and 3 distinct "worst" models.

prompt preference will therefore lead to inconsistent conclusions.

To demonstrate this problem, we report the performance variance on LAMA using different prompts for each PLM. For each relation, we follow Elazar et al. (2021); Jiang et al. (2020b) and design at least 5 prompts that are semantically equivalent and faithful but vary in linguistic expressions.

Prompt selection significantly affects performance. Figure 3 illustrates the performance on several relations, where the performances of all PLMs vary significantly on semantically equivalent prompts. For instance, by using different prompts, the Precision@1 of relation `languages spoken` dramatically changing from 3.90% to 65.44% on BERT-large, and from 0.22% to 71.94% on BART-large. This result is shocking, because the same PLM can be assessed from "knowing nothing" to "sufficiently good" by only changing its prompt. Table 1 further shows the quantitative results, for BERT-large, the averaged standard deviation of Precision@1 of different prompts is 8.75. And the prompt selection might result in larger performance variation than model selection: on more than 70% of relations, the best and worst prompts will lead to >10 point variation at Precision@1, which is larger than the majority of performance gaps between different models.

Prompt preference also leads to inconsistent comparisons. Figure 4 demonstrates an example, where the ranks of PLMs are significantly changed when applying diverse prompts. We also conduct quantitative experiments, which show that the PLMs' ranks on 96.88% relations are unstable when prompt varies.

All these results demonstrate that the prompt preference bias will result in inconsistent performance. Such inconsistent performance will further lead to unstable comparisons between different PLMs, and therefore significantly undermines the evaluations via prompt-based probing.

4.2 Cause of Prompt Preference Bias

Figure 2b shows the cause of the prompt preference bias. When evaluating the ability of PLMs on specific tasks, we would like to measure the causal effects of path $M \rightarrow I \rightarrow E$. However, because the prompt P and the PLM M are all correlated to the linguistic distribution L, there is a backdoor path $M \leftarrow C \leftarrow L \rightarrow P \rightarrow I \rightarrow E$ between PLM M and performance E. Consequently, the backdoor path will confound the effects of $M \rightarrow I \rightarrow E$ with $P \rightarrow I \rightarrow E$.

Based on the above analysis, the prompt preference bias can be eliminated by blocking this backdoor path via backdoor adjustment, which requires a prior formulation of the distribution $\mathcal{P}(P)$. In Section 7, we will present one possible causal intervention formulation which can lead to more consistent evaluations.

5 Instance Verbalization Bias

Apart from the prompt preference bias, the underlying linguistic correlation can also induce bias in the instance verbalization process. Specifically, an instance in probing data can be verbalized into different natural language expressions (e.g., verbalize

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

Table 2: Different verbalized names of the same entity lead to different predictions on BERT-large.

Q30 in Wikidata into America or the U.S.), and different PLMs may prefer different verbalizations due to mention coverage, expression preference, etc. This will lead to instance verbalization bias.

5.1 Instance Verbalization Brings Unstable Predictions

In factual knowledge probing, each entity is verbalized to its default name. However, different PLMs may prefer different verbalizations, and such underlying correlation is invisible. Because we couldn't measure how this correlation affects probing performance, the evaluation may be unstable using different verbalizations.

Table 2 shows some intuitive examples. When we query BERT "The capital of the U.S. is [MASK]", the answer is Washington. Meanwhile, BERT would predict Chicago if we replace the U.S. to its alias America. Such unstable predictions make us unable to obtain reliable conclusions on whether or to what degree PLMs actually entail the knowledge.

To quantify the effect of instance verbalization bias, we collect at most 5 verbalizations for each subject entity in LAMA from Wikidata, and calculate the *verbalization stability* on each relation, i.e., the percentage of relation instances whose predictions are unchanged when verbalization varies. The results in Figure 5 show the average verbalization stabilities of all four PLMs are < 40%, which demonstrate that the instance verbalization bias will bring unstable and unreliable evaluation.

5.2 Cause of Instance Verbalization Bias

Figure 2c shows the cause of instance verbalization bias: the backdoor path $M \leftarrow C \leftarrow L \rightarrow X \rightarrow E$, which stems from the confounder of linguistic distribution L between pretraining corpus C and verbalized probing data X. Consequently, the observed correlation between M and E couldn't faithfully represent the true causal effect of M on E, but is also mixed up the spurious correlation caused by the backdoor path.

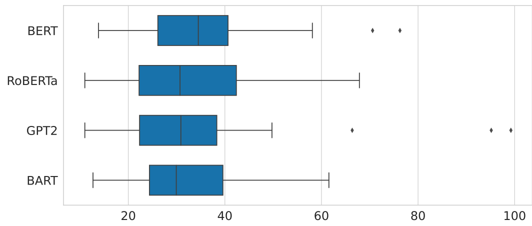

Figure 5: The *verbalization stabilities* of 4 PLMs on all relations, which is measured by the percentage of relation instances whose predictions are unchanged when verbalization varies. We can see that the verbalization stabilities of all 4 PLMs (BERT-large, RoBERTa-large, GPT2-xl, BART-large) are poor.

The instance verbalization bias can be eliminated by blocking this backdoor path via causal intervention, which requires a distribution formulation of the instance verbalization, i.e., $\mathcal{P}(X)$. We will present a possible intervention formulation in Section 7.

6 Sample Disparity Bias

Besides the biases induced by linguistic correlations, the distributional correlations between pretraining corpus and task-specific probing data can also introduce sample disparity bias. That is, the performance difference between different PLMs may due to the sample disparity of their pretraining corpus, rather than their ability divergence.

In conventional evaluation, the evaluated hypotheses are independent of the train/test data generation, and all the hypotheses are evaluated on training data and test data generated from the same distribution. Therefore, the impact of correlations between training data and test data is transparent, controllable, and equal for all the hypotheses. By contrast, in prompt-based probing, each PLM is bundled with a unique pretraining corpus, the correlation between pretraining corpus distribution and probing data distribution cannot be quantified. In the following we investigate this sample disparity bias in detail.

6.1 Sample Disparity Brings Biased Performance

In factual knowledge probing, LAMA (Petroni et al., 2019), a subset sampled from Wikidata,

$\gamma\%$	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

Table 3: The P@1 on LAMA of PLMs whose further pretraining data are with different correlation degrees $\gamma\%$ with LAMA. The BERT-base and GPT2-base both contain 12 layers, while BERT-large and GPT2-medium both contain 24 layers.

is commonly used to compare different PLMs. Previous work claims that GPT-style models are with weaker factual knowledge extraction abilities than BERT because they perform worse on LAMA (Petroni et al., 2019; Liu et al., 2021c). However, because PLMs are pretrained on different pretraining corpus, the performance divergence can stem from the spurious correlation between pretraining corpus and LAMA, rather than their ability difference. For example, BERT's superior performance to GPT-2 may stem from the divergence of their pretraining corpus, where BERT's pretraining corpus contains Wikipedia, while GPT-2's pretraining corpus doesn't.

To verify the effect of sample disparity bias, we further pretrain BERT and GPT-2 by constructing pretraining datasets with different correlation degrees to LAMA, and report their new performances on LAMA. Specifically, we use the Wikipedia snippets in LAMA and collect a 99k-sentence dataset, named WIKI-LAMA. Then we create a series of pretraining datasets by mixing the sentences from WIKI-LAMA with WebText[2] (the pretraining corpus of GPT2). That is, we fix all datasets' size to 99k, and a parameter γ is used to control the mixture degree: for each dataset, there are $\gamma\%$ instances sampled from WIKI-LAMA and $1-\gamma\%$ instances sampled from WebText. Please refer to the appendix for pretraining detail.

Table 3 demonstrates the effect of sample disparity bias. We can see that 1) Sample disparity significantly influences the PLMs' performance: the larger correlation degree γ will result in better performance for both BERT and GPT-2; 2) Sample disparity contributes to the performance difference. We can see that the performance gap between GPT-2 and BERT significantly narrows down when they are further pretrained using the same data. Besides, further pretraining BERT on WebText (γ=0) would significantly undermine its performance. These results strongly confirm that the sample disparity will significantly bias the probing conclusion.

[2] http://Skylion007.github.io/OpenWebTextCorpus

6.2 Cause of Sample Disparity Bias

The cause of sample disparity bias may diverge from PLMs and scenarios due to the different causal relation between pretraining corpus distribution D_a and probing data distribution D_b. Nevertheless, sample disparity bias always exist because the backdoor path will be $M \leftarrow C \leftarrow D_a \rightarrow D_b \rightarrow T \rightarrow X \rightarrow E$ when D_a is the ancestor of D_b, or $M \leftarrow C \leftarrow D_a \leftarrow D_b \rightarrow T \rightarrow X \rightarrow E$ when D_a is the descendant of D_b. Figure 2d shows a common case when the pretraining corpus distribution D_a is an ancestor of probing data distribution D_b. For example, the pretraining data contains Wikipedia and probing data is a sampled subset from Wikipedia (e.g., LAMA, X-FACTR, BioLAMA). As a result, there is a backdoor path between M and E, which will mislead the evaluation.

7 Bias Elimination via Causal Intervention

This section describes how to eliminate the above-mentioned biases by blocking their corresponding backdoor paths. According to the Backdoor Criterion in Section 2.1, we need to choose a set of variables Z that can block every path containing an arrow into M between M and E. Since the linguistic distribution L, pretraining corpus distribution D_a and probing data distribution D_b are unobservable, we choose $Z = \{P, X\}$ as the variable set for blocking all backdoor paths between (M, E) in the SCM by conducting backdoor adjustment:

$$\begin{aligned}&\mathcal{P}(E|do(M=m), R=r) = \\ &\sum_{p\in P}\sum_{x\in X}\mathcal{P}(p,x)\mathcal{P}(E|m,r,p,x).\end{aligned} \quad (2)$$

Equation 2 provides an intuitive solution. To eliminate the biases stemming from the spurious correlations between pretraining corpus, probing data and prompts, we need to consider the natural distribution of prompts and verbalized probing data regardless of other factors. Consequently, the overall causal effects between PLM and evaluation result are the weighted averaged effects on all valid prompts and probing data.

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**

Table 4: The *rank consistencies* over 1000 task samples (each task contains 20 relations from LAMA). For a PLM, the rank consistency is the percentage of its most popular rank in 1000 runtimes. For "Overall Rank", the rank consistency is the percentage of the most popular rank of all PLMs in 1000 runtimes, i.e., the rank of all PLMs remains the same. "Original" means that we use the LAMA's original prompts and verbalized names, "Random" means that we randomly sample prompts and verbalized names every time, "+Intervention" means that we apply causal intervention. We can see that the rank consistency is significantly raised after causal intervention.

Unfortunately, the exact distribution of $\mathcal{P}(x, p)$ is intractable , which needs to iterate over all valid prompts and all verbalized probing data. To address this problem, we propose a sampling-based approximation. Specifically, given a specific assumption about $\mathcal{P}(x, p)$ (we assume uniform distribution in this paper without the loss of generality), we sample K_p prompts for each relation and K_x kinds of verbalization for each instance according to $\mathcal{P}(x, p)$, and then these samples are used to estimate the true causal effects between M and E according to Equation 2.

To verify whether causal intervention can improve the evaluation consistency and robustness, we conduct backdoor adjustment experiments on 8 different PLMs. We randomly sample 1000 subsets with 20 relations from LAMA, and observe whether the evaluation conclusions were consistent and stable across the 1000 evaluation runtimes. Specifically, we use ***rank consistency*** as the evaluation metric, which measures the percentage of the most popular rank of each model in 1000 runtimes. For example, if BERT ranks at 3^{rd} place in 800 of the 1000 runtimes, then the rank consistency of BERT will be 80%.

Table 4 shows the results. We can see that causal intervention can significantly improve the evaluation consistency: 1) The consistency of current prompt-based probing evaluations is very poor on all 8 PLMs: when we randomly select prompts and verbalizations during each sampling, the overall rank consistency is only 5.5%; 2) Causal intervention can significantly improve overall rank consistency: from 5.5% to 68.5%; 3) Casual intervention can consistently improve the rank consistency of different PLMs: the rank of most PLMs is very stable after backdoor adjustment.

The above results verify that causal intervention is an effective technique to boost the stability of evaluation, and reach more consistent conclusions.

8 Related Work

Prompt-based Probing Prompt-based probing is popular in recent years (Rogers et al., 2020; Liu et al., 2021b) for probing factual knowledge (Petroni et al., 2019; Jiang et al., 2020a; Sung et al., 2021), commonsense knowledge (Davison et al., 2019), semantic knowledge (Ettinger, 2020; Sun et al., 2021; Brown et al., 2020; Schick and Schütze, 2020) and syntactic knowledge (Ettinger, 2020) in PLMs. And a series of prompt-tuning studies consider optimizing prompts on training datasets with better performance but may undermine interpretability (Jiang et al., 2020b; Shin et al., 2020; Haviv et al., 2021; Gao et al., 2021; Qin and Eisner, 2021; Li and Liang, 2021; Zhong et al., 2021). Because such prompt-tuning operations will introduce additional parameters and more unknown correlations, this paper does not take prompt-tuning into our SCM, delegate this to future work.

Biases in NLP Evaluations Evaluation is the cornerstone for NLP progress. In recent years, many studies aim to investigate the underlying biases and risks in evaluations. Related studies include investigating inherent bias in current metrics (Coughlin, 2003; Callison-Burch et al., 2006; Li et al., 2017; Sai et al., 2019, 2020), exploring dataset artifacts in data collection and annotation procedure (Lai and Hockenmaier, 2014; Marelli et al., 2014; Chen et al., 2018; Levy and Dagan, 2016; Schwartz et al., 2017; Cirik et al., 2018; McCoy et al., 2019; Liu et al., 2021a; Branco et al., 2021), and identifying the spurious correlations between data and label which might result in catastrophic out-of-distribution robustness of models (Poliak et al., 2018; Rudinger et al., 2018; Rashkin et al., 2018).

Most previous studies demonstrate the evaluation biases empirically, and interpret the underlying reasons intuitively. However, intuitive explanations are also difficult to critical and extend. In contrast, this paper investigates the biases in prompt-based probing evaluations from a causal view. Based on the causal analysis framework, we can identify, understand, and eliminate biases theoretically, which can be extended and adapted to other evaluation settings in a principled manner[3]. We believe both the causal analysis tools and the valuable insights can benefit future researches.

9 Conclusions and Discussions

This paper investigates the critical biases and quantifies their risks in the widely used prompt-based probing evaluation, including prompt preference bias, instance verbalization bias, and sample disparity bias. A causal analysis framework is proposed to provide a unified framework for bias identification, interpretation and elimination with a theoretical guarantee. Our studies can promote the understanding of prompt-based probing, remind the risks of current unreliable evaluations, guide the design of unbiased datasets, better probing frameworks, and more reliable evaluations, and push the bias analysis from empirical to theoretical.

Another benefit of this paper is to remind the evaluation criteria shifts from conventional machine learning algorithms to pretrained language models. As we demonstrate in Figure 1, in conventional evaluation, the evaluated hypotheses (e.g., algorithms, architectures) are raised independently of the train/test dataset generation, where the impact of correlations between training data and test data is transparent, controllable, and equal for all the hypotheses. However, in evaluations of pretrained language models, the pretraining corpus is bundled with the model architecture. In this case, it is significant to distinguish what you need to assess (architecture, corpus, or both), as well as the potential risks raised by the correlations between pretraining corpus and test data, which most current benchmarks have ignored. Consequently, this paper echoes that it is necessary to rethink the criteria for identifying better pretrained language models, especially under the prompt-based paradigm.

In the future, we would like to extend our causal analysis framework to fit prompt-tuning based probing criteria and all PLM-based evaluations.

Acknowledgments

We sincerely thank all anonymous reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. 62122077, the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27020200, and the National Natural Science Foundation of China under Grants no. 62106251 and 62076233.

[3]Greatly inspired by the reviewer's valuable comments.

Ethics Consideration

This paper has no particular ethic consideration.

References

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *ArXiv preprint*, abs/2109.00725.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 656–666.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Alice Lai and Julia Hockenmaier. 2014. Illinois-LH: A denotational and distributional approach to semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 329–334, Dublin, Ireland. Association for Computational Linguistics.

Omer Levy and Ido Dagan. 2016. Annotating relation inference in context via question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–255, Berlin, Germany. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv preprint*, abs/2107.13586.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021c. GPT understands, too. *ArXiv preprint*, abs/2103.10385.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating ADEM: A deeper look at scoring dialogue responses. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6220–6227. AAAI Press.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020. A survey of evaluation metrics used for NLG systems. *ArXiv preprint*, abs/2008.12009.

Timo Schick and Hinrich Schütze. 2020. Few-Shot Text Generation with Pattern-Exploiting Training. *ArXiv preprint*, abs/2012.11926.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *ArXiv preprint*, abs/2107.02137.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9733–9740. AAAI Press.

A Datasets Construction Details

Instance Filtering We follow the data construction criteria as LAMA, we remove the instances whose object is multi-token or not in the intersection vocabulary of these 4 PLMs.

Relation Selection We remove all the N-M relations in LAMA such as "share border with" or "twin city". Because in these relations, there are multiple object entities corresponding to the same subject entity. In that case, the metric Precision@1 is not suitable for evaluating PLMs in such relations. In addition, due to the completeness limitation of knowledge bases, it's impossible to find all the correct answers for each subject. Therefore, we do not include these relations in our experiments.

Prompt Generation Because of the difference between the pretraining tasks of these 4 PLMs (autoencoder, autoregressive and denoising autoencoder), we design prompts where the placeholder for the target object is at the end, e.g., *The birthplace of x is y* instead of *y is the birthplace of x*. We follow the instruction from Wikidata, combine the prompts from Elazar et al. (2021) and Jiang et al. (2020b), and manually filter out the prompts with inappropriate semantics. All the prompts are created before any experiments and fixed afterward.

B Further Pretraining Details

We further pretrain BERT with masked language modeling (mask probability=15%) and GPT2 with autoregressive language modeling task respectively. Training was performed on 8 40G-A100 GPUs for 3 epochs, with maximum sequence length 512. The batch sizes for BERT-base, BERT-large, GPT2-base, GPT2-medium are $256, 96, 128, 64$ respectively. All the models is optimized with Adam using the following parameters: $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$ and the learning rate is $5e-5$ with warmup ratio=0.06.