An Fine-grained Interpretability Evaluation Benchmark for Pre-trained Language Models

Anonymous ACL submission

Abstract

While pre-trained language models (PLMs) have brought great improvements in many NLP tasks, there is increasing attention to explore capabilities of PLMs and interpret their predictions. However, existing works usually focus only on a certain capability of PLMs by testing them with some downstream tasks. There is a lack of datasets for directly evaluating the masked word prediction performance and the interpretability of PLMs. To fill in the gap, we propose a novel evaluation benchmark providing with both English and Chinese annotated data. In order to comprehensively evaluate the capabilities of PLMs, it provides evaluation data in five dimensions, i.e., grammar, semantics, factual knowledge, reasoning and computation. In addition, it provides carefully annotated token-level rationales to evaluate the interpretability of PLM predictions. We conduct experiments on several widely-used PLMs. The results show that they perform very poorly in the dimensions of knowledge and computation. And the rationales provided by them to support predictions are less plausible, especially when they are short. We will release this benchmark at `http://xyz`, hoping it can facilitate the research progress of PLMs.

1 Introduction

Since PLMs such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved significant gains in predictive accuracy on a variety of NLP tasks (Wang et al., 2018), many studies focus on exploring their capabilities (Tenney et al., 2019; Petroni et al., 2019; Brown et al., 2020) and their decision-making mechanisms (Ding and Koehn, 2021; Rethmeier et al., 2020).

Many works have proved that PLMs have learned amounts of knowledge from the massive text corpora (i.e., their training data), such as linguistic knowledge (Tenney et al., 2019; Jawahar et al., 2019) and factual knowledge (Petroni et al., 2019; Pörner et al., 2019). Such learned knowledge has enhanced some capabilities of PLMs, e.g., reasoning (Brown et al., 2020) and computation (Polu and Sutskever, 2020). However, some studies show that PLMs have not captured adequate knowledge and are insufficient in some aspects, e.g., having not learned enough syntactic structures (Wang et al., 2019; Min et al., 2020), having a poor grasp of reasoning over factual knowledge and commonsense (Pörner et al., 2019; Marcus and Davis, 2020), as well as having a poor performance on mathematical problem solving (Hendrycks et al., 2021; Cobbe et al., 2021).

On the other hand, some researchers aim to unveil the decision-mechanism of a PLM, which can help us understand the reasons behind its success and its limitations (Rethmeier et al., 2020; Meng et al., 2022; Mor Geva, 2022). Some works study the inner workings of transformer-based PLMs according to hidden states and their evolutions between layers (Voita et al., 2019; Singh et al., 2019). Other works develop toolkits to capture, analyze and visualize inner mechanisms of PLMs at the level of individual neurons (Rethmeier et al., 2020; Dai et al., 2021; Alammar, 2021; Mor Geva, 2022).

Although lots of studies have been done, it is still unclear what capabilities a PLM has mastered and how much it has mastered. Meanwhile, there lacks quantitative evaluations on PLM's interpretability. To address these problems, we propose a novel evaluation benchmark for PLMs, containing instances with masked words and corresponding human-annotated rationales for masked word predictions. As shown in Table 1, the masked words are used to evaluate model prediction performance, and the rationales are used to evaluate model interpretability. Overall, our contributions include:

1. To our best knowledge, this is the first benchmark that can be used to evaluate both prediction performance and interpretability of PLMs. And it provides both English and Chinese data.

Dimensions	Instances Collected for Evaluation	Instances in Our Evaluation Benchmark	
		Inputs with Masked Words	Answers
Grammar	Aeroflot 's international fleet of 285 **planes** is being repainted and refurbished at Shannon Airport.	Aeroflot 's international fleet of 285 [MASK] is being repainted and refurbished at Shannon Airport.	planes
Semantics	The city of Austin has a total **area** of 703.95 square kilometres.	The city of Austin has a total [MASK] of 703.95 square kilometres.	area
Knowledge	The country **Germany** is located directly to the east of Belgium.	The country [MASK] is located directly to the east of Belgium.	Germany
Reasoning	The man's eyes were stabbed by broken glass, then he went **blind**.	The man's eyes were stabbed by broken glass, then he went [MASK].	blind
Computation	Tony planted a 4 foot tree. The tree grows at a rate of 5 feet every year. It takes **5** years to be 29 feet.	Tony planted a 4 foot tree. The tree grows at a rate of 5 feet every year. It takes [MASK] years to be 29 feet.	5

Table 1: Examples for five evaluation dimensions. The words in the red color are taken as the ground-truth rationale for masked word predictions, and the words in *Answers* column are the golden answers for the masked words.

2. Our evaluation benchmark covers five evaluation dimensions, i.e., grammar, semantics, factual knowledge, reasoning and computation, to comprehensively evaluate the capability of a PLM from multiple perspectives.

3. We conduct experiments on several widely-used PLMs, and the results show that current PLMs have very poor prediction performance in the dimensions of knowledge, reasoning and computation. And the rationales that support predictions are less plausible. We believe this benchmark can help evaluate and improve PLMs.

2 Related Work

In this section, we review studies on exploring PLMs' capabilities and interpretability.

2.1 Capability Analyses of PLMs

While PLMs have been developed rapidly, a large number of studies attempt to explore capabilities of PLMs (Jawahar et al., 2019; Hewitt and Manning, 2019; Pörner et al., 2019).

Grammar and semantics. Some studies prove that PLMs have captured linguistic structures in their representations, with lexical features at low layers, syntactic features at middle layers and semantic features at high layers (Jawahar et al., 2019; Kim et al., 2020; Tenney et al., 2019). Also, some works find that BERT has not learned some syntactic structures and can not perform well on syntax-aware data (Wang et al., 2019; Min et al., 2020).

Knowledge. According to BERT's good performance on answering cloze-style questions about relational facts between entities, Petroni et al. (2019) state that BERT memorizes factual knowledge during pre-training. But Pörner et al. (2019) prove that BERT's impressive performance is partly due to reasoning about the surface form of entity names.

Reasoning. GPT-3 is proved to have a powerful reasoning ability and can generate news articles which are difficult to be distinguished from human-written articles by humans (Brown et al., 2020). However, Marcus and Davis (2020) state that GPT-3 has no idea what it's talking about, and show that it has a poor grasp of reasoning over commonsense.

Computation. Some studies aim to make PLMs solve theorem proving and quantitative reasoning problems, such as GPT-f (Polu and Sutskever, 2020) and Minerva (Lewkowycz et al., 2022). Other studies release corresponding datasets to measure model capabilities on mathematical problem solving, and find that these enormous PLMs fail to achieve high test performance, while a bright middle school student could solve every problem (Hendrycks et al., 2021; Cobbe et al., 2021).

2.2 Interpretability of PLMs

Interpretation Methods Recent studies to interpret PLM's predictions are mainly classified into three categories. First, some works use input saliency methods (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017) to assign an importance score for each input token, representing the token's impact on model predictions (Ding and Koehn, 2021). Second, hidden states and their evolutions between layers are always used to glean information about the inner workings of a PLM (Singh et al., 2019; Voita et al., 2019). Third, examination of neuron activations is used to trace and analyze model processes in quantifying knowledge changes by extracting underlying patterns of neuron firings (Rethmeier et al., 2020; Dai et al., 2021). Meanwhile, several tools are released to capture, analyze, visualize, and interactively explore inner mechanisms of PLMs (Dalvi et al., 2019; Alammar, 2021; Mor Geva, 2022).

Dimensions	English	Chinese
Grammar	Penn Treebank 3.0	Chinese Treebank 8.0, Chinese Dependency Treebank 1.0
Semantics	Wikipedia, WebNLG (Moryossef 2019), WSC (Levesque 2012)	Baidu Baike, DuIE (Li 2019), CLUEWSC2020 (Xu 2020)
Knowledge	FreebaseQA (Jiang 2019)	CKBQA
Reasoning	COPA (Roemmele 2011)	XCOPA (Ponti 2020)
Computation	Alg514 (Kushman 2014), Dolphin18K (Huang 2016)	Math23K (Wang 2017)

Table 2: Datasets used to create our datasets. Please see Appendix A for details.

Evaluation Datasets While attention to construct interpretability evaluation datasets for specific NLP tasks keeps increasing (DeYoung et al., 2020; Wang et al., 2021; Camburu et al., 2018), there is a lack of evaluation datasets for PLMs. Ding and Koehn (2021) create four datasets with token-level rationales for two grammar tasks, i.e., subject-verb number agreement and pronoun-antecedent gender agreement. Ekin Akyürek (2022) propose a fact tracing dataset with instance-level rationales to evaluate model capability on fact learning. He et al. (2022) provide a simile property probing dataset to evaluate a PLM's performance on interpreting similes.

Evaluation Metrics Plausibility and faithfulness are often used to evaluate interpretability (Doshi-Velez and Kim, 2017; Jacovi and Goldberg, 2020; Adebayo et al., 2020), where the former measures how much the rationales provided by models align with human-annotated ones, and the latter measures the degree to which the rationales in fact influence the corresponding predictions. With token-level rationales, token F1-score is often used to evaluate plausibility (Mathew et al., 2021; Wang et al., 2022); sufficiency and comprehensiveness (DeYoung et al., 2020), consistency under perturbations (Ding and Koehn, 2021; Wang et al., 2022), as well as sensitivity and stability (Yin et al., 2022) are used to evaluate faithfulness.

In this work, we provide a novel benchmark to evaluate both predictive accuracy and interpretability of PLMs. This benchmark provides evaluation data in multiple dimensions and provides human-annotated token-level rationales.

3 Evaluation Dataset Construction

Our evaluation dataset is constructed in three steps: 1) data collection; 2) perturbed data construction; 3) iterative rationale annotation and checking. We first introduce the five evaluation dimensions in Section 3.1. Then we describe the annotation process in Section 3.2-3.4. Finally, we give our data statistics.

3.1 Evaluation Dimensions

According to the abilities that a PLM should have for predicting the right answer, we define five evaluation dimensions, as described below. The corresponding examples are shown in Table 1.

- **Grammar**. The instances in this dimension are designed to evaluate what lexical linguistic knowledge a PLM has learned, such as the tense of a verb, the gender of a pronoun, and the number of a noun. As shown by the first example in Table 1, the noun right after the number "285" must be plural if it is countable.
- **Semantics**. This dimension aims to test whether a PLM has learned syntactic and semantic features or not, such as entity types and semantic co-reference rules. The second example in Table 1 requires the model to master the concept of "city" and the corresponding property "area".
- **Knowledge**. The instances in this type are used to evaluate the extent to which a PLM has learned real-world factual knowledge. As shown by the third example in Table 1, the prediction of "Germany" requires the model to learn and memorize related knowledge.
- **Reasoning**. This dimension measures the inferential capability of a PLM over open-domain commonsense. The forth example in Table 1 states that the model should deduce "blind" according to the premise that the eye was hurt.
- **Computation**. These instances test the quantitative reasoning ability of a PLM on handling mathematical problems, as illustrated by the last example in Table 1.

3.2 Data Collection

In order to create high-quality evaluation datasets, we construct our datasets on the basis of some existing human-annotated datasets, as shown in Table 2. Please see Appendix A for details. Our collection process consists of the following two steps: instance construction and masked word selection.

Instance Construction Each input in our evaluation datasets is a sentence or a paragraph consisting of multiple sentences, as shown in Table 1. Since the input forms of some existing datasets in Table 2 are not as we want them to be in our datasets, we have them manually modified. Specially, we create inputs for the following three dimensions.

For *knowledge*, the original input consists of a question and an answer. We replace the wh-phrase in the question with the answer to form a new input. For example, the third example in Table 1 was originally "*Which country is located directly to the east of Belgium?*" with the answer "*Germany*".

For *reasoning*, each original input contains a given premise and two plausible alternatives for either the cause or the effect of the premise. We concatenate the premise and its cause with some appropriate conjunctions such as "since" and "as" to build a new instance. Similarly, we use conjunctions such as "then" and "so" to connect the premise and the proper effect to create a new input.

For *computation*, the original instance consists of a question, an equation and answer. We use the answer to replace the wh-phrase in the question to construct a new input. The original question for the fifth example in Table 1 is "*Tony planted a 4 foot ... How many years will it take to be 29 feet?*".

Masked Word Selection For each created input, we select an appropriate word or phrase to mask, denoted as w_m. Then the masked sentence is input to a PLM, which will output a prediction for the masked position. And w_m is the golden answer for the masked position. To precisely evaluate model prediction performance, the golden answer for the masked position should be as unique as possible. We take the uniqueness of answer as one criteria for selecting masked words. In the dimensions of "knowledge" and "computation", the question's answer is selected as w_m. For the other three dimensions, annotators need to select appropriate masked words according to the uniqueness principle.

Besides, we make some rules to ensure the diversity of masked words. For example, the masked words in grammar dimension cover all parts-of-speech, and those in semantics dimension cover different entity types and relations.

Then for each masked position and the corresponding golden answer, three annotators rate their confidences on a 4-point scale by judging *whether the golden answer is unique* (4), *among the top 3 predictions* (3), *among the top 5 predictions* (2), or *none of the above* (1). The masked position is considered to be appropriate if the confidence of each annotator is no less than 3, i.e., its golden answer is unique or among the top 3 predictions.

3.3 Perturbed Data Creation

Recent studies (Ding and Koehn, 2021; Wang et al., 2022) propose to evaluate the model faithfulness using the consistency of rationales under perturbations that are not supposed to change the model decision mechanism. Following them, we construct perturbed examples for each original input.

Perturbation Criteria In our work, perturbations should not change the model prediction and the internal decision mechanism. Please note that the influence of perturbations on model's prediction and decision mechanism comes from human's basic intuition. Based on the literature (Jia and Liang, 2017; McCoy et al., 2019; Ribeiro et al., 2020), we define three perturbation types.

- **Alteration of dispensable words (*Dispens.*)**. Insert, delete or replace words that should have no effect on model predictions, e.g., inserting the word "*unfortunately*" at the beginning of the sentence "*the man's eyes were stabbed by broken glass, then he went blind*".

- **Alteration of important words (*Import.*)**. Replace important words which could affect model predictions with their synonyms or related words, e.g., replacing "*stabbed*" with "*pierced*". In this situation, the rationale will change, but the prediction may not change.

- **Syntactic transformation (*Trans.*)**. Transform the syntactic structure of an instance without changing its meaning, e.g., "*the man's eyes were stabbed by broken glass*" is transformed into "*the broken glass stabbed the man's eyes*".

We create at least one perturbed example for each original input. We ask two annotators to create perturbed examples, and ask other annotators to review and modify the created examples.

3.4 Iterative Rationale Annotation

Given an input with a mask and the golden answer for the mask, the annotators highlight important input tokens that support the mask prediction as the rationale. Then we introduce the criteria and annotation process of rationales.

Rationale Criteria As discussed in recent studies of natural language understanding tasks, a rationale should satisfy sufficiency, compactness and comprehensiveness (Lei et al., 2016; Yu et al., 2019). As the comprehensiveness is not suitable for the rationales of PLMs' predictions, we use sufficiency and compactness as the criteria.

- **Sufficiency**. A rationale is sufficient if it contains enough information for humans to make the correct prediction. In other words, humans can make the correct prediction only based on tokens in the rationale.
- **Compactness**. A rationale is compact if all of its tokens are indeed required in making a correct prediction. That is to say, when any token is removed from the rationale, the prediction will change or become difficult to make.

Annotation Process To ensure data quality, following Wang et al. (2022), we also adopt an iterative annotation workflow, including three steps.

Step 1: rationale annotation. Given the input and the corresponding golden answer, the annotators label all critical tokens that are needed for the prediction of the golden answer based on their intuition on the model decision mechanism.

Step 2: rationale scoring. The checkers double-check the annotations according to the annotation criteria. For each rationale, the checkers rate their confidences for sufficiency by judging *whether they are unable* (1), *probably able* (2), or *definitely able* (3) to make the correct prediction only based on it, and rate their confidences for compactness by judging *whether it contains redundant tokens* (1), *contains disturbances* (2), *is probably concise* (3), or *is very concise* (4).

A rationale is of high-quality if its average score on sufficiency and compactness is equal to or greater than 3 and 3.6 respectively. All unqualified data whose average score on a property is lower than the corresponding threshold goes to step 3.

Step 3: rationale modification. Low-quality rationales are given to the annotators for correction.

Then the corrected rationales are scored by checkers again. This iterative annotation-scoring process runs for three iterations and the unqualified data is discarded after that.

3.5 Data Statistics

Table 3 shows the detailed statistics of our benchmark. We can see that the number of pairs and the length ratio of rationale vary with evaluation dimensions. As discussed in "masked word selection", the instances in grammar dimension cover as much lexical knowledge as possible. The English grammar dataset is larger than the Chinese one as there are less agreement rules in Chinese grammar. The Chinese dataset for semantics is larger than the English one as there are more available data in Chinese. The rationale length ratio affects interpretability results, as discussed in Section 5.3.

Dimensions	English		Chinese	
	Size	RRL(%)	Size	RRL(%)
Grammar	1,365	29.8	701	20.7
Semantics	793	31.6	1,210	27.1
Knowledge	295	45.8	300	51.5
Reasoning	300	48.5	300	43.4
Computation	307	59.7	400	54.5

Table 3: Statistics of our benchmark. "Size" means the number of original/perturbed pairs. "RRL" represents the ratio of rationale length to its input length.

4 Metrics

Following previous works (Jacovi and Goldberg, 2020; DeYoung et al., 2020; Wang et al., 2022), we evaluate interpretability from the aspects of plausibility and faithfulness.

Plausibility Plausibility measures how well the rationale provided by the model aligns with the human-annotated rationale (Jacovi and Goldberg, 2020; DeYoung et al., 2020; Ding and Koehn, 2021). We use token F1-score for plausibility evaluation, as shown in Equation 1. For each prediction, we select the top K important tokens as its rationale, where the token importance score is assigned by a specific saliency method.

$$\begin{aligned} \texttt{Token-F1} &= \frac{1}{N}\sum_{i=1}^{N}(2 \times \frac{P_i \times R_i}{P_i + R_i}) \\ \texttt{where} \quad P_i &= \frac{|S_i^p \cap S_i^g|}{|S_i^p|} \ \texttt{and} \ R_i = \frac{|S_i^p \cap S_i^g|}{|S_i^g|} \end{aligned} \tag{1}$$

where S_i^p and S_i^g represent the predicted rationale and human-annotated rationale of the i-th instance respectively; N represents the number of instances.

Faithfulness Faithfulness evaluates to what extent the rationale provided by the model truly affects the model prediction (Jacovi and Goldberg, 2020; Ding and Koehn, 2021). A variety of metrics have been proposed to evaluate faithfulness,

e.g. sufficiency and comprehensiveness (DeYoung et al., 2020), consistency under perturbations (Ding and Koehn, 2021; Wang et al., 2022), sensitivity and stability (Yin et al., 2022). Most of these metrics are only applicable to classification models.

Considering the characteristics of PLMs, we use the consistency of rationales under perturbations to evaluate faithfulness. Specifically, we adopt Mean Average Precision (MAP) (Wang et al., 2022) and Pearson Correlation Coefficient (PCC) (Ding and Koehn, 2021) as evaluation metrics.

MAP, as defined in Equation 2, evaluates the consistency of two rationales by calculating the order consistency of their corresponding sorted token lists. The higher the MAP is, the more faithful the rationale is.

$$\texttt{MAP} = \frac{\sum_{i=1}^{|X^p|}(\sum_{j=1}^{i} G(x_j^p, X_{1:i}^o))/i}{|X^p|} \quad (2)$$

where X^o and X^p are the sorted token lists of the original and perturbed inputs respectively. $|X^p|$ represents the token number of X^p. $X_{1:i}^o$ contains the top-i important tokens of X^o. The function $G(x, Y)$ determines whether the token x belongs to the list Y, i.e., $G(x, Y) = 1 \texttt{ iff } x \in Y$.

PCC, as shown in Equation 3, measures the linear correlation between token importance scores of the original instance and the perturbed one. Based on perturbation types defined in Section 3.3, we first align the two importance score lists. Specifically, the unaligned tokens, such as the deleted ones and inserted ones, will be aligned to a virtual token with importance score 0. As it is difficult to align the perturbed instance with its original instance under the perturbation type of *Trans.*, we do not perform PCC calculation on pairs of this type. A high PCC score[1] represents a faithful rationale.

$$\texttt{PCC} = \frac{\sum_{i=1}^{n}(v_i^o - \bar{v}^o)(v_i^p - \bar{v}^p)}{\sqrt{\sum_{i=1}^{n}(v_i^o - \bar{v}^o)^2}\sqrt{\sum_{i=1}^{n}(v_i^p - \bar{v}^p)^2}} \quad (3)$$

where v_i^o and v_i^p represent the i-th elements in the importance score vectors of the original and perturbed instance, respectively. $\bar{v}^o$ and $\bar{v}^p$ are the mean of two score vectors respectively.

From the definitions of MAP and PCC, it can be seen that MAP measures the association of two token lists based on importance order, and PCC assesses the association of two token lists according to their importance values.

[1]In our experiments, our reported PCC values are computed on pairs with $\texttt{p-value} < 0.05$, where p-value represents the significance level of the linear correlation.

5 Experiments

5.1 Experimental Setting

In order to evaluate model performance on our benchmark, we adopt several widely-used PLMs and interpretation methods as our baseline models and methods respectively. We only provide high-level descriptions for them and refer to the respective papers and source codes for details.

Evaluated PLMs We take BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as baseline models for English. And we adopt BERT-base-chinese (BERT-base) (Devlin et al., 2019), RoBERTa-wwm-ext (RoBERTa-base) (Cui et al., 2021) and ERNIE (Sun et al., 2019) as baseline models for Chinese[2].

Interpretation Methods We use attention (ATT) based (Jain and Wallace, 2019) and integrated gradient (IG) based saliency methods (Sundararajan et al., 2017) to assign an importance score for each input token. For the prediction of each input, we select the top-K important tokens to compose the rationale. In our experiments, K is the product of the average length ratio (i.e., RRL in Table 3) and the current input length.

In ATT based method, the attention weights in the last layer are taken as token importance scores for predictions, where the attention weight of token i on token j is denoted as $s_{i,j}$. Then we use $s_{i,m} = \sum_{j \in m} s_{i,j}$ to represent the impact of token i on the prediction of the masked segment m which may contain multiple tokens.

In IG based method, token importance is determined by integrating the gradient along the path from a defined baseline x_0 to the original input, where x_0 is set as a sequence of "[CLS] [PAD] ... [SEP]" and has the same token number as the original input. The step size is set to 100.

Evaluation Metrics We evaluate model prediction performance on the first (Top1) and the first three (Top3) predicted answers, using the predictive accuracy, i.e., the percentage of inputs whose predictions exactly match the golden answers. And we use the metrics described in Section 4 to evaluate PLM's interpretability.

[2]Since the parameters in the masked-LM layers of BERT-large and RoBERTa-large are not released with models, which are required in our experiments, we do not conduct experiments on large versions of Chinese BERT and RoBERTa.

Model + TopN	Grammar			Semantics			Knowledge			Reasoning			Computation		
	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb
BERT-base + Top1	61.3	61.3	61.2	43.9	45.3	42.4	**7.2**	**8.3**	**6.0**	20.5	21.7	19.3	**1.1**	**1.5**	0.8
RoBERTa-base + Top1	67.6	**68.8**	66.3	51.0	53.1	48.9	5.0	5.3	4.7	23.8	25.0	22.7	1.0	0.8	**1.3**
ERNIE-base + Top1	64.3	65.5	63.1	50.6	51.6	49.7	4.5	4.0	5.0	25.0	26.7	23.3	0.3	0.3	0.3
ERNIE-large + Top1	**67.7**	68.5	**66.9**	**53.0**	**54.4**	**51.6**	4.3	4.0	4.7	**30.5**	**32.0**	**29.0**	0.5	0.3	0.8
BERT-base + Top3	79.2	80.3	78.2	61.7	63.2	60.2	12.2	**13.0**	11.3	32.8	35.0	30.7	**3.3**	**4.0**	2.5
RoBERTa-base + Top3	81.7	82.2	81.2	71.9	72.4	71.4	10.0	8.7	11.3	43.5	43.0	44.0	2.9	3.3	2.5
ERNIE-base + Top3	81.7	82.6	80.9	71.7	72.3	71.1	**12.3**	12.0	**12.7**	45.5	47.0	44.0	1.6	1.5	1.8
ERNIE-large + Top3	**84.0**	**84.9**	**83.0**	**73.6**	**74.6**	**72.6**	11.2	10.3	12.0	**50.2**	**52.7**	**47.7**	2.8	2.5	**3.0**

Table 4: Masked word prediction performance of base PLMs on original inputs, perturbed inputs and all inputs.

Model + Method	Grammar			Semantics			Knowledge			Reasoning			Computation		
	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*
BERT-base + ATT	0.35	0.85	0.93 / 0.90	0.39	0.82	0.94 / 0.91	0.68	0.73	0.92 / 0.90	0.53	0.73	0.88 / 0.84	0.64	0.83	0.87 / 0.89
BERT-base + IG	0.35	0.70	0.88 / 0.74	0.30	0.67	0.89 / 0.75	0.61	0.59	0.83 / 0.71	0.49	0.55	0.80 / 0.62	0.59	0.76	0.88 / 0.78
RoBERTa-base + ATT	0.38	**0.86**	0.93 / **0.91**	0.37	**0.84**	0.94 / **0.92**	0.65	0.71	0.92 / 0.90	0.50	0.76	0.90 / 0.86	0.64	0.85	0.89 / 0.90
RoBERTa-base + IG	0.30	0.72	0.88 / 0.80	0.31	0.68	0.87 / 0.81	0.63	0.63	0.82 / 0.86	0.50	0.65	0.82 / 0.81	0.60	0.80	0.86 / 0.87
ERNIE-base + ATT	**0.50**	**0.86**	**0.95 / 0.91**	**0.49**	0.82	**0.96** / 0.91	**0.71**	0.74	**0.96 / 0.92**	**0.66**	**0.78**	**0.95 / 0.88**	**0.69**	**0.86**	**0.94 / 0.91**
ERNIE-base + IG	0.38	0.66	0.89 / 0.71	0.36	0.61	0.88 / 0.69	0.64	0.54	0.84 / 0.68	0.54	0.53	0.82 / 0.59	0.61	0.76	0.91 / 0.77
ERNIE-large + ATT	0.40	0.83	0.92 / 0.88	0.41	0.80	0.90 / 0.89	**0.71**	**0.76**	0.87 / 0.88	0.54	0.69	0.87 / 0.81	0.66	0.83	0.86 / 0.89
ERNIE-large + IG	0.33	0.56	0.70 / 0.64	0.33	0.51	0.69 / 0.64	0.62	0.51	0.76 / 0.70	0.50	0.47	0.67 / 0.61	0.60	0.68	0.72 / 0.73

Table 5: Interpretability evaluation of base PLMs with two interpretation methods. As illustrated in Section 4, the metric PCC is not performed on all inputs. For inputs suitable for PCC calculation, we compute MAP* on them.

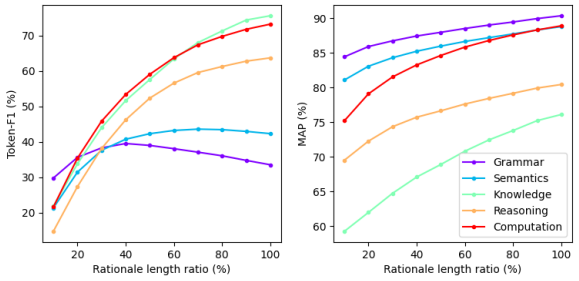

Figure 1: Plausibility (F1) and faithfulness (MAP) of RoBERTa-base with ATT based interpretation method over different rationale length ratios.

5.2 Main Results

Model Prediction Performance Table 4 shows model performance on masked word predictions. It can be seen that all models perform well on instances of grammar and semantics dimensions, which proves that these PLMs have learned enough linguistic knowledge from the large-scale corpus (Hewitt and Manning, 2019; Jawahar et al., 2019; Tenney et al., 2019). However, in the other three dimensions, all models show a poor prediction performance, especially on knowledge and computation. Existing studies also show that PLMs have no such abilities (Pörner et al., 2019; Hendrycks et al., 2021; Cobbe et al., 2021).

From the comparisons between evaluated PLMs, we get two interesting findings. First, RoBERTa and ERNIE perform better than BERT in dimensions of grammar, semantics and reasoning. Furthermore, ERNIE large outperforms ERNIE base in these three dimensions. We think there are two reasons, i.e., the larger size of training corpus and the larger size of parameters. Second, BERT and ERNIE base have better performance in knowledge and computation. As discussed above, the abilities in these two dimensions have not been learned by PLMs from the current training corpus and learning objectives. We think the relevant learning objectives need to be designed and the corresponding training data needs to be created.

Model Interpretability Table 5 gives results on interpretability of different models and methods. There are three main findings. Firstly, with ATT interpretation methods, all the evaluated PLMs have a relatively strong faithfulness, indicating that they are robust under perturbations. As shown in Table 4, compared with predictive accuracy on the original data, the predictive accuracy on the perturbed data has not decreased too much. For example, in the dimension of grammar, the accuracies of most PLMs are reduced by about 2%. Secondly, across all evaluated PLMs, ATT method outperforms IG both in plausibility and faithfulness. We think this is because the interactions between words are more important for word generation based on the context. Thirdly, token F1-score (plausibility) and MAP (faithfulness) are positively correlated with the length ratio of extracted rationale. Compared

Data	Grammar			Semantics			Knowledge			Reasoning			Computation		
	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.
Original	67.7	72.0	63.9	57.5	51.5	47.0	8.4	1.7	4.8	15.7	30.3	26.1	0.0	0.0	1.7
Perturbed	66.4	70.3	60.1	55.8	51.8	31.7	6.3	1.7	4.8	15.7	30.3	15.9	1.6	1.7	0.0
	(-1.3)	(-1.7)	(**-3.8**)	(-1.7)	(+0.3)	(**-15.3**)	(**-2.1**)	(0.0)	(0.0)	(0.0)	(0.0)	(**-10.2**)	(+1.6)	(**+1.7**)	(**-1.7**)

Table 6: Predictive accuracy of RoBERTa-base over different perturbation types.

Dimension	Dispens.		Import.		Trans.	
	$F1^{\Delta}$	MAP	$F1^{\Delta}$	MAP	$F1^{\Delta}$	MAP
Grammar	0.003	0.906	**0.011**	0.834	0.001	0.810
Semantics	0.008	0.881	0.004	0.868	**0.014**	0.700
Knowledge	0.003	0.769	0.007	0.775	0.002	0.648
Reasoning	**0.012**	0.809	**0.026**	0.786	**0.020**	0.663
Computation	0.003	0.912	0.009	0.897	**0.012**	0.833

Table 7: Interpretability results under different perturbation types. $F1^{\Delta}$ represents the F1-score difference between original input and perturbed input, where bold values indicate large differences.

with performance on rationales, the performance on predictions is much poor. How to improve model prediction on plausible rationale is the future work.

Comparing PLMs with different training data size, ERNIE which is trained on a larger corpus performs better in plausibility with two interpretation methods, and RoBERTa has a higher MAP in the dimensions of grammar and semantics. Comparing PLMs with different parameter size, we find that ERNIE base is superior to ERNIE large on both faithfulness and plausibility in all dimensions except for knowledge. This shows that larger parameter size may not lead to higher interpretability.

Finally, we compare two metrics for faithfulness, i.e., MAP and PCC, where MAP relies on token importance order and PCC relies on token importance values, as discussed in Section 4. From Table 5, we can see that the two metrics of the same model have the similar trend over different interpretation methods. But the gap Between PCC values is smaller than that between MAP values.

5.3 Analysis

We give an in-depth analysis about the impacts of extracted rationale length and perturbation type on model interpretability. Due to space limitation, we take the results of RoBERTa-base with ATT based method for example, and results of other PLMs and methods have the similar trend.

Impacts of Rationale Length As shown in Figure 1, in the dimensions of knowledge, reasoning and computation, where the length ratio of rationale is about 0.5, both plausibility and faithfulness increase with the increase of rationale length ratio. This states that the most important words provided by the model and the interpretation method perform poorly on interpretability in these three evaluation dimensions. In the dimensions of grammar and semantics, where the rationale length ratio is about 0.3, plausibility achieves the highest F1 score when the extracted rationale length ratio is about 0.5; and MAP increases much slowly with the increase of rationale length. This shows that PLMs perform well in these two dimensions.

Impacts of Perturbation Types In Table 6, we give model prediction accuracy over three perturbation types. It can be seen that the prediction accuracy alters significantly on syntactically transformed perturbations (*Trans.*). And the model is relatively robust on the other two perturbation types.

Meanwhile, we further analyze interpretability results over different perturbation types, as shown in Table 7. It can be seen that faithfulness under *Trans.* type is significantly lower than those under the other two perturbation types. Meanwhile, the perturbation types of *Trans.* and *Import.* have a larger influence on plausibility in the dimensions of semantics, reasoning and computation. Correspondingly, *Dispens.* has little influence on model interpretability, just as it has little effect on model predictions.

6 Conclusion

To comprehensively evaluate PLMs, we construct a novel evaluation benchmark to evaluate both model prediction performance and interpretability from five dimensions, i.e., grammar, semantics, knowledge, reasoning and computation. We conduct experiments on several popular PLMs, and the results show that they perform very poorly in some dimensions, such as knowledge and computation. Meanwhile, the results show that the rationales they provided for predictions are less plausible, especially with a short rationale. Finally, the evaluated PLMs have a strong robustness under perturbations, but they are less robust on syntax-aware data. We will release this evaluation benchmark, and hope it will facilitate the research progress of PLMs.

References

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. CoRR, abs/2011.05429.

J Alammar. 2021. Ecco: An open source library for the explainability of transformer language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 249–257, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. CoRR, abs/2110.14168.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3504–3514.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. CoRR, abs/2104.08696.

Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. Neurox: A toolkit for analyzing individual neurons in neural networks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):9851–9852.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5034–5052, Online. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Frederick Liu Binbin Xiong Ian Tenney Jacob Andreas Kelvin Guu Ekin Akyürek, Tolga Bolukbasi. 2022. Tracing knowledge in language models back to the training data.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7875–7887, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 887–896, Berlin, Germany. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. 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 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. CoRR, abs/2002.00737.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 271–281, Baltimore, Maryland. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In Thirteenth international conference on the principles of knowledge representation and reasoning.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. arXiv preprint arXiv:2206.14858.

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye Jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu. 2019. Duie: A large-scale chinese dataset for information extraction. In CCF International Conference on Natural Language Processing and Chinese Computing, pages 791–800. Springer.

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. CoRR, abs/1907.11692.

Gary Marcus and Ernest Davis. 2020. Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. Technology Review.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In AAAI.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in gpt. arXiv preprint arXiv:2202.05262.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2339–2352, Online. Association for Computational Linguistics.

Guy Dar Paul Roit Shoval Sadde Micah Shlain Bar Tamir Yoav Goldberg Mor Geva, Avi Caciularu. 2022. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Stanislas Polu and Ilya Sutskever. 2020. Generative language modeling for automated theorem proving. CoRR, abs/2009.03393.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2362–2376, Online. Association for Computational Linguistics.

Nina Pörner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised QA. CoRR, abs/1911.03681.

Nils Rethmeier, Vageesh Kumar Saxena, and Isabelle Augenstein. 2020. Tx-ray: Quantifying and explaining model-knowledge transfer in (un-)supervised nlp. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), volume 124 of Proceedings of Machine Learning Research, pages 440–449. PMLR.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In AAAI spring symposium: logical formalizations of commonsense reasoning, pages 90–95.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. CoRR, abs/1706.03825.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. CoRR, abs/1904.09223.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 3319–3328. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. CoRR, abs/1909.01380.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. 2021. Dutrust: A sentiment analysis dataset for trustworthiness evaluation. arXiv preprint arXiv:2108.13140.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. A fine-grained interpretability evaluation benchmark for neural nlp. arXiv preprint arXiv:2205.11097.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 845–854, Copenhagen, Denmark. Association for Computational Linguistics.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. 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 1061–1070, Hong Kong, China. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In Proceedings of

the 28th International Conference on Computational Linguistics, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in NLP. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2631–2647, Dublin, Ireland. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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 4094–4103, Hong Kong, China. Association for Computational Linguistics.

A Datasets for Building Our Benchmark

In order to construct a high-quality evaluation benchmark, we build our datasets based on some existing human-annotated datasets. As shown in Table 2, for each evaluation dimension, we select several datasets to build our evaluation dataset.

To test the grammatical competence of PLMs, we collect instances from some linguistic datasets which show both lexical and syntactic knowledge. Specifically, we adopt Penn Treebank 3.0[3] to build English dataset, as well as Chinese Treebank 8.0[4] and Chinese Dependency Treebank 1.0[5] to build Chinese dataset. All of these three datasets are from LDC, and we have been authorized.

To evaluate the performance of PLMs on semantic understanding, our dataset covers multiple types of conceptual knowledge, such as conceptual senses of words, concept properties, relationships between concepts and semantic co-reference rules. For concept properties and relationships, we adopt Wikipedia[6] and WebNLG (Moryossef et al., 2019) for English dataset, as well as Baidu Baike[7] and DuIE (Li et al., 2019) for Chinese dataset. For semantic co-reference, we task WSC (Levesque et al., 2012) and CLUEWSC2020 (Xu et al., 2020) for English and Chinese respectively.

To test the capability of PLMs on grasping factual knowledge, we take the datasets for the knowledge based question answering task as base datasets, i.e., Freebase QA dataset (Jiang et al., 2019) for English, and CKBQA[8] for Chinese. Our questions cover single-hop and multi-hop questions. Meanwhile, we filter out questions with multiple answers to ensure the uniqueness of the prediction.

To evaluate the reasoning ability of PLMs on real-world commonsense, we utilize the COPA dataset (Roemmele et al., 2011) and the XCOPA dataset (Ponti et al., 2020) to build our English and Chinese datasets respectively.

For testing the ability of PLMs on solving mathematical word problems, we use Alg514 (Kushman et al., 2014) and Dolphin18K (Huang et al., 2016) for English, and Math23K (Wang et al., 2017) for Chinese. And we only select simple questions whose equations have no more than two operators.

[3] https://catalog.ldc.upenn.edu/LDC99T42
[4] https://catalog.ldc.upenn.edu/LDC2013T21
[5] https://catalog.ldc.upenn.edu/LDC2012T05
[6] https://huggingface.co/datasets/wikipedia
[7] https://baike.baidu.com
[8] https://github.com/pkumod/CKBQA, the dataset for knowledge-based question answering task in CCKS 2019.

Data cleaning. In the process of collection, we ask annotators to discard instances that contain: 1) offensive content, 2) information that names or uniquely identifies individual people, 3) discussions about politics, guns, drug abuse, violence or pornography.

B Other Annotation Details

We give more details about annotator information, annotation training and payment, and instructions for data usage.

Annotator information. We have two annotators for each dimension, and three checkers for all dimensions. The annotators annotate the rationales and modify the rationales according to the scores from the checkers. They are college students majoring in languages. Our checkers are full-time employees, and perform quality control. Before this work, they have lots of experience in annotating data for NLP tasks.

Annotation training and payment. Before real annotation, we train all annotators for several times so that they understand the annotation task, rationale criteria, etc. During real annotation, we have also held several meetings to discuss common mistakes and settle disputes. All annotators were paid for their work based on the quality and quantity of their annotations. According to their annotation time, the average salary per hour is 31.25 RMB.

Instructions of data annotation and usage. Before annotation, we provide a full instruction to all annotators, including the responsibility for leaking data, disclaimers of any risks, and screenshots of annotation discussions. Meanwhile, our datasets are only used for interpretability evaluation. And we will release a license with the release of our benchmark.

C Limitation Discussion

We provide an evaluation benchmark to evaluate capabilities and interpretability of PLMs. There are three limitations in our work.

- How to automatically and effectively evaluate the quality of human-annotated rationales is still open. We have three annotators to perform quality control. However, this manner heavily relies on human intuitions and experiences.

- Due to resource limitation, we do not conduct experiments on capability-specific PLMs, such

Model + TopN	Grammar			Semantics			Knowledge			Reasoning			Computation		
	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb	All	Original	Perturb
BERT-base + Top1	52.2	52.5	51.9	64.7	65.6	63.8	0.7	0.7	0.7	24.0	23.7	24.3	0.5	0.3	0.7
BERT-large + Top1	58.8	58.9	58.8	67.9	69.1	66.7	0.7	1.0	0.3	28.7	28.0	29.3	1.3	1.0	1.6
RoBERTa-base + Top1	59.5	59.3	59.6	62.7	62.9	62.4	2.9	3.1	2.7	29.7	29.3	30.0	1.1	1.0	1.3
RoBERTa-large + Top1	**72.0**	**71.5**	**72.5**	**73.6**	**74.0**	**73.1**	**5.1**	**6.4**	**3.7**	**40.5**	**40.0**	**41.0**	**2.0**	**2.0**	**2.0**
BERT-base + Top3	69.0	69.1	68.9	80.5	81.3	79.7	1.0	1.4	0.7	36.5	35.7	37.3	2.8	2.3	3.3
BERT-large + Top3	73.2	73.4	73.0	83.9	84.9	83.0	1.2	2.0	0.3	42.0	41.3	42.7	4.4	4.2	4.6
RoBERTa-base + Top3	73.2	73.6	72.9	79.7	80.3	79.1	5.4	6.8	4.1	46.5	46.3	46.7	5.1	4.6	5.5
RoBERTa-large + Top3	**83.3**	**83.3**	**83.4**	**89.0**	**89.5**	**88.4**	**8.8**	**11.2**	**6.4**	**57.8**	**59.7**	**56.0**	**6.7**	**7.2**	**6.2**

Table 8: Masked word prediction performance of baseline PLMs on English dataset, where performance is evaluated on all inputs, original inputs and perturbed inputs respectively.

Model + Method	Grammar			Semantics			Knowledge			Reasoning			Computation		
	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*	F1	MAP	PCC/MAP*
BERT-base + ATT	0.47	**0.91**	**0.99 / 0.93**	**0.44**	**0.84**	**0.97** / 0.90	0.59	**0.73**	**0.99** / 0.85	0.53	**0.87**	**0.99 / 0.89**	0.57	**0.90**	**0.95 / 0.92**
BERT-base + IG	0.34	0.73	0.89 / 0.75	0.42	0.67	0.87 / 0.71	**0.66**	0.51	0.83 / 0.63	0.56	0.64	0.83 / 0.68	0.62	0.79	0.85 / 0.80
BERT-large + ATT	0.47	0.87	0.95 / 0.91	0.40	0.83	0.94 / 0.89	0.58	0.64	0.94 / 0.85	0.59	0.84	0.96 / 0.87	0.60	0.90	0.94 / 0.90
BERT-large + IG	0.37	0.42	0.45 / 0.48	0.41	0.38	0.44 / 0.46	0.64	0.40	0.54 / 0.61	0.57	0.41	0.39 / 0.51	**0.66**	0.56	0.48 / 0.60
RoBERTa-base + ATT	**0.55**	0.88	0.95 / 0.91	**0.44**	0.83	0.94 / 0.89	0.58	0.63	0.90 / 0.85	**0.63**	0.83	0.92 / 0.88	0.62	0.87	0.91 / 0.90
RoBERTa-base + IG	0.39	0.66	0.78 / 0.73	0.37	0.56	0.72 / 0.68	0.56	0.41	0.73 / 0.67	0.56	0.59	0.79 / 0.69	0.63	0.73	0.81 / 0.78
RoBERTa-large + ATT	0.53	0.90	0.96 / **0.93**	0.43	0.82	0.94 / **0.91**	0.55	0.63	0.90 / **0.86**	0.56	0.83	0.90 / 0.88	0.58	0.87	0.92 / 0.90
RoBERTa-large + IG	0.37	0.57	0.74 / 0.67	0.37	0.50	0.73 / 0.65	0.56	0.45	0.67 / 0.60	0.54	0.54	0.74 / 0.67	0.63	0.67	0.74 / 0.75

Table 9: Interpretability results of base PLMs with two interpretation methods on English dataset. As illustrated in Section 4, the metric PCC is not performed on all inputs. For inputs suitable for PCC calculation, we also compute MAP on them, denoted as MAP*.

Dimension	Dispens.		Import.		Trans.	
	$F1^{\Delta}$	MAP	$F1^{\Delta}$	MAP	$F1^{\Delta}$	MAP
Grammar	0.000	0.913	0.002	0.821	0.001	0.837
Semantics	0.001	0.854	0.001	0.856	**0.016**	0.644
Knowledge	**0.016**	0.783	**0.017**	0.750	**0.057**	0.512
Reasoning	0.002	0.874	0.002	0.821	**0.013**	0.753
Computation	0.001	0.921	0.004	0.877	**0.019**	0.888

Table 10: Interpretability results of RoBERTa-base with ATT based method under different perturbation types. $F1^{\Delta}$ represents the F1-score difference between original input and perturbed input, where bold values indicate large differences.

as GPT-f for computation, and those PLMs with enormous parameter size, such as GPT-3.

- What is the relationship between linguistic knowledge learned by PLMs and their interpretations for masked word predictions? Such analysis is as the future work.

D English results

In this section, we show results on English dataset, as shown in Table 8 - Table 10. Similarly, we give analyses from the perspectives of model prediction performance and interpretability.

Model Prediction Performance Table 8 shows the predictive accuracy of evaluated PLMs on English dataset. Generally, the performance in different dimensions has the similar trend with that on Chinese dataset. Firstly, all evaluated PLMs perform very poorly in the dimensions of knowledge and computation. Secondly, both for BERT and RoBERTa, the large-size model outperforms the base-size one. Thirdly, comparing models with the same size of parameters, RoBERTa which is trained on a larger corpus outperforms BERT on most of dimensions.

Impacts of perturbation types. As shown in Table 11, in most of dimensions, RoBERTa-base is less robust under perturbation types of *Trans.* and *Impor.*. Meanwhile, RoBERTa-base is less robust under the perturbation type of *Dispens.* in some dimensions, such as semantics, knowledge and reasoning, while Chinese RoBERTa-base is robust under *Dispens.* type in all dimensions.

Model Interpretability Table 9 shows the interpretation results of the evaluated PLMs on the English dataset. Most of the conclusions on the Chinese dataset (illustrated in Section 5.2) are applicable to the English dataset. One difference is that ATT based method not always performs better than IG based method on plausibility evaluation in the dimensions of knowledge and computation. But model prediction performance is very poor in these two dimensions, which may affect the interpretability performance of PLMs.

Impacts of perturbation types. Table 10 shows

Data	Grammar			Semantics			Knowledge			Reasoning			Computation		
	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.	Dispens.	Import.	Trans.
Original	59.8	58.6	54.3	60.2	73.9	63.2	4.3	1.6	3.1	25.6	30.0	35.5	2.6	1.1	0.0
Perturbed	60.4 (+0.6)	59.0 (+0.4)	51.2 (**-3.1**)	59.2 (-1.0)	71.8 (**-2.1**)	64.6 (+1.4)	2.9 (-1.4)	4.9 (**+3.3**)	1.8 (-1.3)	28.1 (+2.5)	30.5 (+0.5)	32.3 (**-3.2**)	2.6 (0.0)	0.0 (**-1.1**)	0.0 (0.0)

Table 11: Predictive accuracy of RoBERTa-base over different perturbation types on English dataset.

interpretability results of RoBERTa-base under different perturbation types. It can be seen that in the dimensions of semantic, knowledge and reasoning, the perturbation type of syntactical transformation (*Trans.*) brings a significant drop on faithfulness (MAP). Meanwhile, in most dimensions, *Trans.* causes a large F1-score difference between the original intput and the perturbed input. This proves that the evaluated PLMs are less robust to perturbations in *Trans.* type.