Accurate, yet Inconsistent? Consistency Analysis on Language Models

Anonymous ACL submission

Abstract

Consistency, which refers to generating the same predictions for semantically similar contexts, is highly desirable for a sound language model. Although recent pre-trained language models (PLMs) deliver an outstanding performance in various downstream tasks, they should also exhibit a consistent behaviour, given that the models truly understand language. In this paper, we propose a simple framework, called ***c**onsistency **a**nalysis on **l**anguage **m**odels (CALM)*, to evaluate a model's lower-bound consistency ability. Via experiments, we confirm that current PLMs are prone to generate inconsistent predictions even for semantically identical inputs with high confidence. We also observe that multi-task training is of benefit to improve consistency, increasing the value by 17% on average.

1 Introduction

Large-sized pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and GPT2 (Radford et al., 2019), compose the backbone of contemporary natural language processing (NLP) systems, delivering an outstanding performance on many downstream tasks through fine-tuning and in-context learning (Brown et al., 2020). Based on their excellent performance, claims that PLMs can understand language have emerged in the literature (Devlin et al., 2019; Ohsugi et al., 2019; Qiu et al., 2020) and popular press, such as the *Google Blog* post[1] and *Towards Data Science* website[2].

However, recent studies raise questions of PLM's language understanding capacity. Numerous pieces of research demonstrated that PLMs are incapable of identifying the meaning of sentences but rely on the excessive exploitation of statistical cues or syntactic patterns (Habernal et al., 2018; Niven and Kao, 2019; McCoy et al., 2019; Bender and Koller, 2020). Another line of works found that PLMs can memorise frequent word/phrase/knowledge presented in pretraining data but poorly understand unseen expressions and knowledge (Kassner et al., 2020; Ravichander et al., 2020; Hofmann et al., 2021). Moreover, many studies discovered that PLMs are insensitive to sentence order (Pham et al., 2020; Gupta et al., 2021; Sinha et al., 2021) and lack an understanding of negated phrases (Naik et al., 2018; Hossain et al., 2020; Kassner and Schütze, 2020; Ettinger, 2020; Hosseini et al., 2021).

In the spirit of meaning-text theory (MTT), the correspondence between semantic content (*meaning*) and linguistic expressions (*text*) is *many-to-many*, which implies that the *meaning* can be conveyed in various text forms (Mel'čuk and Žolkovskij, 1970; Milićević, 2006). Also, the concept of "*understanding*" is *to focus on the meaning and not the text form* (Krashen, 1982). Therefore, provided a model understands language, it should make consistent decisions in semantically equivalent texts, because *meaning* is a common invariant content that all synonymous texts have. This is the spirit of *consistency*, and the performance of PLMs should be illuminated in terms of consistency, aside from other evaluation metrics like accuracy, to evaluate their language understanding ability.

Many recent studies have investigated PLM's consistency through behavioural testing on augmented data (Ribeiro et al., 2020; Ravichander et al., 2020; Elazar et al., 2021) and text adversarial attacks (Morris et al., 2020; Li et al., 2020a; Garg and Ramakrishnan, 2020; Jin et al., 2020). However, these approaches have several downsides. First, a great human effort or task-specific data production rules are essential for the data augmentation-based investigation. This limitation confined the investigation to a few tasks, such as zero-shot knowledge retrieval (Ravichander et al.,

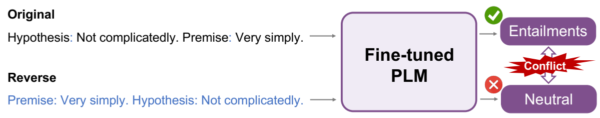
[1] https://www.blog.google/products/search/search-language-understanding-bert/

[2] https://towardsdatascience.com/pre-trained-language-models-simplified-b8ec80c62217

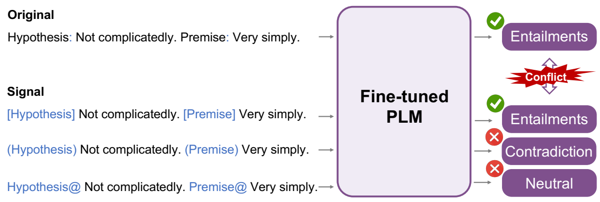
2020; Elazar et al., 2021) and sentiment analysis (Ribeiro et al., 2020), and a certain language, mainly English. Text adversarial attacks aim to lead a model to make inconsistent decisions on adversarial samples, mainly generated by the masked language modelling (MLM) of PLMs to be semantically analogous to the target words (Morris et al., 2020; Li et al., 2020a; Garg and Ramakrishnan, 2020). However, the semantic equivalence of these samples is not guaranteed due to their reliance on PLMs, whose credibility has recently been challenged (Ravichander et al., 2020; Ettinger, 2020; Kassner and Schütze, 2020; Elazar et al., 2021). Also, most text adversarial attack methods use similarity scores of sentence embeddings generated by a pre-trained encoder as a criterion to extract adversarial samples. However, it is questionable whether the encoder trained without *meaning* information can extract semantically similar adversarial samples (Bender and Koller, 2020). Since the semantic equivalence is a prerequisite for evaluating consistency, text adversarial attacks could lead to an overestimation of PLM's inconsistency. Also, additional components, such as synonym dictionaries (Ren et al., 2019) and pre-trained word/sentence embeddings (Hill et al., 2015; Cer et al., 2018), are a core of the adversarial sample generators. This limitation precludes the examination of consistency for other languages where such resources are not available.

In this paper, we propose a simple but efficient behavioural testing framework, called **c**onsistency **a**nalysis on **l**anguage **m**odels (CALM), to evaluate the consistency of PLMs. Our approach can be applied to various downstream tasks without additional components and ideally ensures the semantic equivalence and thus measures the lower-bound consistency of PLMs. To be specific, we introduce a free-text sentence type indicator and add perturbations, such as shifting the input sentence ordering (REVERSE) and substituting a special symbol (SIGNAL), which works as a separator, with other symbols (see Figure 1).

Our main contributions are as follows: (i) we propose a behavioural testing framework that perfectly guarantees semantic equivalence (Section 3), (ii) our approach could be easily applied to low-resource languages and various downstream tasks, (iii) we observe that widely used PLMs lack consistency regardless of their training objective and languages (Section 5), (iv) we verify that humans exhibit a very high consistency under the same experimental settings (Section 6), and (v) we show that multi-task training is beneficial to improve consistency (Section 7). We will make our code available after acceptance.



(a) REVERSE case



(b) SIGNAL case

Figure 1: Example of the CALM framework for the MNLI task. The changes in the original free-text inputs are marked in blue.

2 Related Works

Consistency. There have been several attempts to analyse the consistency of language models in various NLP domains. For zero-shot knowledge retrieval tasks, Ravichander et al. (2020) found that PLMs generate different answers if an object of the original query is replaced with its plural form (e.g., 'A robin is a [MASK]' to 'robins are [MASK]'). Elazar et al. (2021) observed a discrepancy in the predictions of PLMs for paraphrase queries and alleviated the issue by fine-tuning the model on the generated paraphrase queries. In question answering (QA), Ribeiro et al. (2019) showed that state-of-the-art QA models generate inconsistent outputs for queries with the same context and used data augmentation to improve consistency. Asai and Hajishirzi (2020) also used data augmentation and, additionally, inconsistency loss, which is designed to penalise inconsistent predictions. Ribeiro et al. (2020) proposed the invariance test to evaluate consistency. For a sentiment analysis task, they changed the named entity presented in a given sentence, because such perturbation does not change the polarity of the sentence. Research on consistency in other domains includes text summarisation (Kryscinski et al., 2020), explanation generation (Camburu et al., 2020), and dialogue generation (Li et al., 2020b). Li et al. (2020b) employed unlikelihood training (Welleck et al., 2019) to improve the consistency of a dialogue model.

Text Adversarial Attack. Text adversarial attacks have a commonality with consistency analysis in that adversarial examples are designed to have a similar meaning with their original counterparts. Jin et al. (2020) proposed a black-box attack approach, called TEXTFOOLER, which replaces important words in an input sentence with synonyms. They used pre-trained word vectors (Hill et al., 2015) to extract synonyms. Li et al. (2020a) used BERT for generating adversarial samples. They first extracted important words for decision-making and replaced them by using the BERT masked language model (MLM). Garg and Ramakrishnan (2020) also used BERT MLM not only for replacing important tokens but also for inserting new tokens. Li et al. (2021) employed MLM for three strategies; "replace", "insert", and "merge" that mask a bi-gram and replace it with a single word. All the approaches presented above leverage the universal sentence encoder (USE, Cer et al. 2018) to extract semantically similar adversarial samples. However, it is doubtful that such an encoder trained using only text form without meaning information can ensure the semantic equivalence between the original and adversarial samples (Bender and Koller, 2020).

3 CALM: Consistency Analysis on Language Models

Behavioural testing refers to examining software systems to assess their capabilities by investigating their behaviour for specially designed inputs (Rim et al., 2021). Our behavioural testing framework evaluates a model's consistency on downstream tasks that infer the relation of two input sentences, such as natural language inference (NLI) and STS tasks. The framework consists of three steps: (1) fine-tune a PLM on the original input format, and inference on development/test dataset, (2) use the fine-tuned PLM to inference on the perturbed input format, and (3) compare the results of the original and perturbed formats. The overall framework of our proposed method is illustrated in Figure 1.

In our experiments, it is crucial to ensure semantic equivalence after perturbation. Inspired by the widely used input formats for human data annotations (Camburu et al., 2018; Kayser et al., 2021) and text-to-text models (Raffel et al., 2020), we introduce a free-text sentence-type indicator to achieve the semantic equivalence. Specifically, we first insert the sentence-type indicator at the beginning of each sentence, followed by a special symbol that acts as a separator (e.g., 'Premise:' and 'Hypothesis:' for the NLI task). These indicator-added input formats are the original data where each model is trained. For the perturbation, we applied the following two methods: REVERSE and SIGNAL.

REVERSE: This method changes the order of the two input sentences. An example case of this method is illustrated in Figure 1a. Without the sentence-type indicator, a model might be unable to distinguish between the first and second input sentence after the ordering alteration. The existence of the indicator will prevent confusion by specifying the input types and, there, can maintain the meaning of inputs after the alteration. We verify that humans are insensitive to this perturbation through human evaluation (see Section 6).

Let $O = \{o_1, ..., o_N\}$ and $R = \{r_1, ..., r_N\}$ denote a set of the original and REVERSE inputs, respectively, and M is a model that we will evaluate. Then, the consistency of the REVERSE case is calculated as follows:

$$C_R = \frac{1}{N}\sum_{t=1}^{N} \mathbb{1}(M(o_t) = M(r_t)),$$

where $M(x)$ denotes the prediction of model M on the data point x. Intuitively, the metric implies the accuracy between the prediction of the original and REVERSE inputs.

SIGNAL: This method changes a special symbol in the sentence-type indicator. An example case of this method is illustrated in Figure 1b. The substitution of the special symbol does not change the meaning of the inputs, because it conveys no specific semantic content. Therefore, a model should make a consistent prediction after the perturbation. In our experiments, we replace a colon in the original input format with multiple other symbols.

Let us assume $S_t = \{s_1^t, ..., s_k^t\}$ be the set of perturbed inputs of the SIGNAL case for t-th data point (o_t). First, we define the pass rate (p_t) of the o_t as follows:

$$p_t = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}(M(s_i^t) = M(o_t)).$$

Next, we consider that the model is consistent on the o_t if $p_t \geq \theta$, where θ is a pre-defined threshold.

As a result. the consistency of the SIGNAL case is defined as follows:

$$C_S = \frac{1}{N} \sum_{t=1}^{N} \mathbb{1}(p_t \geq \theta).$$

For the experiments, we use ten different special symbols, such as a square bracket and semi-colon, for replacement (i.e., $k = 10$). The details of the used symbols can be found in Table 7 in the appendix. We set the θ to 1.0, because, ideally, a model should make the same predictions for all the perturbations.

4 Experimental Design

4.1 Datasets

For the experiments, we select the NLI and STS tasks from the GLUE benchmark (Wang et al., 2019). For the NLI tasks, we use **MNLI** (MultiNLI, Multi-Genre Natural Language Inference, Williams et al., 2018), **QNLI** (Question Natural Language Inference, Rajpurkar et al., 2016) and **RTE** (Recognising Textual Entailment, Candela-Quinonero et al., 2006); they are composed of two sentence pairs and a label indicating whether the sentence pairs are entailed or not. For the STS tasks, we use **QQP** (Quora Question Pairs[3]) and **MRPC** (Microsoft Research Paraphrase Corpus, Dolan and Brockett, 2005), which consist of two sentence pairs and a label indicating whether the two sentences share an identical meaning.

We also evaluate the Korean datasets to show the general applicability of our framework to other languages. For the NLI task, **KorNLI** (Ham et al., 2020) and **KLUE-NLI** (Park et al., 2021) are selected, and for the STS task, **KLUE-STS** (Park et al., 2021) is used. The three Korean datasets do not provide test sets. Therefore, we randomly sampled test sets from the validation set for the **KLUE** datasets and from the training set for the **KorNLI** dataset. The basic statistics of the datasets are given in Table 6 in Appendix A.1.

4.2 Model candidates

We conduct experiments on various types of PLMs having different sizes. For the English tasks, we select the encoder-based models ($RoBERTa$ (Liu et al., 2019b) and $ELECTRA$ (Clark et al., 2020)), the decoder-based models ($GPT2$ (Radford et al., 2019)), and the Seq2Seq models ($BART$ (Lewis et al., 2020) and $T5$ (Raffel et al., 2020)). For the Korean tasks, $KoBERT$ and $KoElectra$ are used as the encoder-based models. For the decoder-based and Seq2Seq models, we use $KoGPT2$ and $KoBART$, respectively. We leverage the pre-trained PLMs from the HuggingFace transformers (Wolf et al., 2020) library.

[3] https://www.kaggle.com/c/quora-question-pairs/data

4.3 Training Details

Apart from the $T5$ models, a classification head is added on top of each model, and all weights are updated while optimising the classification objective function. We fine-tune each of our candidate backbone models on individual tasks. Meanwhile, fine-tuning for $T5$ models is not performed, because the HuggingFace $T5$ models are already trained on the datasets used in our experiments through multi-task training.

At fine-tuning, we use the AdamW optimiser (Loshchilov and Hutter, 2017) and a linear learning rate scheduler decaying from $1e$-3. We fine-tune the models for 10 epochs with a learning rate of $1e$-5 and batch size of 64 and apply an early stopping method during the training. More detailed information regarding the hyperparameter search is presented in Appendix A.2.

5 Experimental Results

5.1 Experiments on English Datasets

The experimental results for the English datasets are summarised in Table 1. In general, all models except for the $T5$ models exhibit the same trend. In the REVERSE case, they show a relatively high consistency on STS tasks. However, all PLMs fall short of expectations on the NLI tasks, making consistent predictions on only 40∼50% of the evaluation data. In the SIGNAL case, the PLMs record a high consistency in most of the cases, but it should be not overlooked that they make inconsistent predictions on roughly 4∼7% of data points despite the minor alteration of a single special symbol. The result implies that PLMs could provide wrong predictions even with meaningless typos, which could result in a negative consequence in practical applications, especially in risk-sensitive domains.

On the contrary, the $T5$ models show the opposite pattern. They exhibit a relatively high consistency in the REVERSE case but entirely fail in the SIGNAL case. Unlike PLMs, humans achieved a very high consistency level on both the REVERSE

Model	MNLI			QNLI			RTE			QQP			MRPC		
	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S
$RoBERTa_{base}$	87.1	61.2	96.1	92.5	65.3	95.6	66.9	52.1	91.0	90.5	97.3	97.3	88.5	94.9	96.3
$RoBERTa_{large}$	**90.0**	65.5	96.9	94.1	64.2	97.8	74.7	54.4	90.6	91.1	96.8	98.2	87.6	94.6	93.3
$Electra_{base}$	88.4	62.3	93.4	92.1	56.7	93.4	74.2	52.4	84.4	90.9	96.9	96.5	88.4	93.1	90.9
$Electra_{large}$	89.7	63.2	95.2	**94.6**	52.4	96.4	83.3	57.9	94.2	**91.4**	97.3	93.4	**90.3**	93.9	96.6
$GPT2_{base}$	79.6	46.6	83.8	86.5	49.0	89.8	57.4	64.6	43.7	87.9	92.8	92.0	70.9	91.0	54.7
$GPT2_{large}$	85.8	56.8	91.4	91.4	51.8	93.7	69.4	39.9	66.5	90.6	93.8	95.1	81.7	89.4	79.9
$BART_{base}$	85.7	54.3	96.1	91.5	51.6	97.1	62.6	54.2	79.4	90.2	96.8	97.6	75.7	97.1	94.3
$T5_{base}$	85.9	60.3	3.2	93.2	87.4	0.0	66.4	82.1	0.0	90.6	97.4	54.9	85.8	96.4	0.0
$T5_{large}$	89.8	85.7	73.5	93.9	94.5	48.5	79.4	87.1	0.0	91.3	97.7	21.4	87.7	97.5	0.0
Human	80.9	**97.1**	**97.1**	89.2	**98.7**	**98.7**	**85.2**	**95.7**	**97.1**	85.4	**98.7**	**98.7**	67.0	**98.0**	**100.0**

Table 1: Results for the consistency evaluation on the English datasets. Acc_{val} denotes an accuracy on the validation dataset. C_R and C_S stands for the consistency for the REVERSE and SIGNAL cases, respectively. We trained each model five times and recorded the average of each metric. The best values are in bold.

Dataset	Type	Input 1	Input 2	Prediction
RTE	Original	Sentence1: Microsoft was established in Italy in 1985.	Sentence2: Microsoft was established in 1985.	entailment
	Signal	[Sentence1] Microsoft was established in Italy in 1985.	[Sentence2] Microsoft was established in 1985.	not_entailment
MRPC	Original	Sentence1: Spinnaker employs roughly 83 people ; NetApp employs 2,400.	Sentence2: Spinnaker employs 83 people, most of whom are engineers.	equivalent
	Reverse	Sentence2: Spinnaker employs 83 people, most of whom are engineers.	Sentence1: Spinnaker employs roughly 83 people ; NetApp employs 2,400.	not_equivalent
QNLI	Original	Question: With what word was Tesla's sociability described?	Sentence: Tesla was asocial and prone to seclude himself with his work.	entailment
	Reverse	Sentence: Tesla was asocial and prone to seclude himself with his work.	Question: With what word was Tesla's sociability described?	not_entailment

Table 2: Examples of inconsistent predictions of the $RoBERTa_{large}$ model.

and SIGNAL cases, reaching almost 100%. More detailed analyses of the experimental results are demonstrated in the following sections. We also describe several examples of inconsistent predictions in Tables 2 and 3. More examples are available in Tables 9, 10, and 11 in the appendix.

5.2 Analysis and Discussion

Models are more consistent on STS tasks. In the REVERSE case, we observe that the consistency of STS tasks outperforms that of NLI tasks by a considerable margin. We conjecture a leading cause is a difference between the training objective of each task. The objective of the STS tasks is to identify whether two sentences with different wordings are semantically identical. Therefore, models trained on such tasks can capture the intrinsic meaning of sentences better and are thus more robust to the meaning-preserving perturbations. In the SIGNAL case, when comparing the tasks with similar training data sizes (**MRPC** with **RTE** and **QQP** with **MNLI**), the consistency of the SIGNAL case is also higher than that of the REVERSE case, but the difference is marginal.

More data higher consistency. We find that the number of training data plays an important role in improving consistency. For the NLI tasks, both the REVERSE and SIGNAL consistencies of the **RTE** dataset are generally lower than those of the **MNLI** and **QNLI** datasets. Similarly, for the STS tasks, the consistency of the **MRPC** dataset are lower than those of the **QQP** dataset. Through a paired t-test, we confirm a statistical significance under the significance level of 0.1.

Models are highly confident. The inconsistency issue might be less concerned provided the predictions are made by chance, i.e., high entropy. Therefore, we investigate the entropy of each model's predictive distribution on the inconsistent predictions. The results are illustrated in Figure 2. Note that, in the binary classification, the entropy of confidence scores 0.7 and 0.9 are 0.88 and 0.47, respectively. The results demonstrate that all PLMs

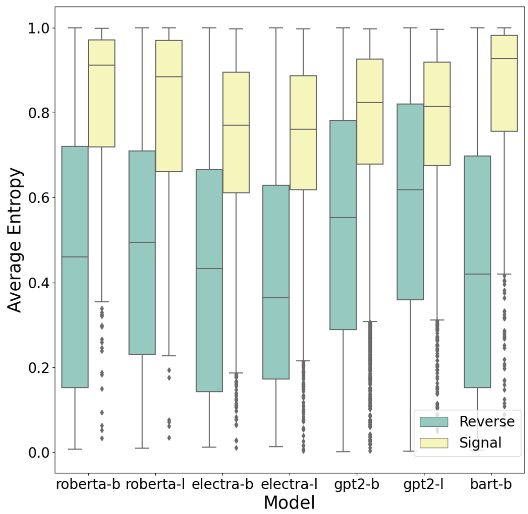


Figure 2: The average entropy of each English model's predictive distribution on inconsistent predictions. "b" and "l" denotes "base" and "large", respectively.

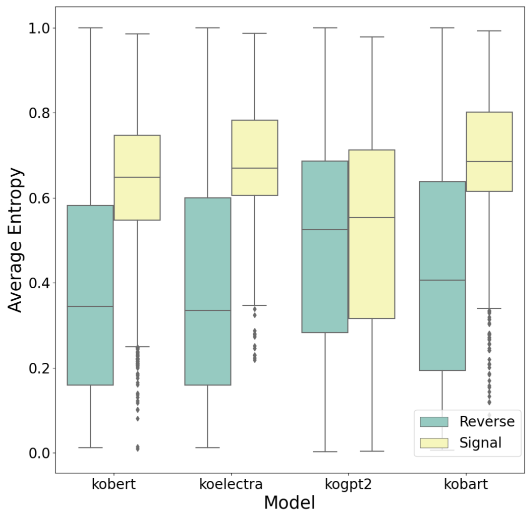


Figure 3: The average entropy of each Korean model's predictive distribution on inconsistent predictions.

are quite confident in the inconsistent predictions, particularly for the REVERSE case. Although they are less confident in the SIGNAL case, the predictive distributions are still distant from the uniform distribution. Furthermore, there exist numerous instances having extremely low entropy values that are less than the 25th percentile.

Impact of pre-training objectives. The PLMs used in our experiments are trained based on different training objectives. $RoBERTa$ (Liu et al., 2019b) uses a dynamic word-based MLM, $ELECTRA$ (Clark et al., 2020) uses a replaced token detection (RTD), $GPT2$ (Radford et al., 2019) uses an autoregressive language modelling, $BART$ (Lewis et al., 2020) uses a span-based MLM, and $T5$ uses both span-based MLM and multi-task training. We observe that $GPT2$ models exhibit a significantly low consistency compared to the other models, even in the STS tasks. The result suggests that the autoregressive LM is a less effective training objective in terms of consistency. Also, our results presented in Table 1 show that no models are perfectly consistent despite their differences. These findings reveal a potential downside of modern language understanding systems.

Is a large model more consistent? In Table 1, large-sized models outperform their corresponding base-sized models in terms of accuracy, just as in the previous studies. However, no such pattern is evident when it comes to consistency. We perform a paired t-test between the base and large models of $RoBERTa$, $ELECTRA$, and $GPT2$, and find no significant difference in both the REVERSE and SIGNAL cases. These results suggest that accuracy is not a sufficient criterion and raise the need to evaluate the model's performance from other lenses, such as consistency.

Analysis of the T5 Models. Compared to the other models that showed relatively high performance in the SIGNAL case, the consistency of the $T5$ models in the SIGNAL case falls short of expectation. One of the strong reasons is that the input formats of the $T5$ models for diverse training tasks use a colon as a separator (Raffel et al., 2020), which is the same format as that of our original case. Because $T5$ models are trained in multi-task fashion on many downstream tasks based on the colon-separator input formats, our SIGNAL case inputs became a completely new distribution to the model. As a result, the desired texts (i.e., labels) were not properly generated. It would be a severe issue for a text-to-text framework, provided the model generates entirely wrong predictions on inputs with such minor changes. Several generated examples of the $T5$ models are provided in Table 3. More examples are available in Table 11 in the appendix.

Meanwhile, the $T5$ models, especially the large model, outperform the others in the REVERSE

Original Inputs: mrpc sentence1: The best-performing stock was Altria Group Inc., which rose more than 27 percent to close at $42.31 a share. sentence2: Altria Group Inc. MO.N fell 50 cents, or 1.2 percent, to $41.81.
Signal Inputs: mrpc **sentence1;** The best-performing stock was Altria Group Inc., which rose more than 27 percent to close at $42.31 a share. **sentence2;** Altria Group Inc. MO.N fell 50 cents, or 1.2 percent, to $41.81.

Original Prediction	Signal Prediction
not_equivalent	sentence2; Altria Group Inc. MO.N fell 50

Original Inputs: mrpc sentence1: The Toronto Stock Exchange opened on time and slightly lower. sentence2: The Toronto Stock Exchange said it will be business as usual on Friday morning.
Signal Inputs: mrpc **sentence1#** The Toronto Stock Exchange opened on time and slightly lower. **sentence2#** The Toronto Stock Exchange said it will be business as usual on Friday morning.

Original Prediction	Signal Prediction
not_equivalent	acceptable

Original Inputs: mrpc sentence1: "It was a little bit embarrassing the way we played in the first two games, "Thomas said. "We're in the Stanley Cup finals, and it was a little bit embarrassing the way we played in the first two games.
Signal Inputs: mrpc **sentence1@** "It was a little bit embarrassing the way we played in the first two games, "Thomas said. **sentence2@** "We're in the Stanley Cup finals, and it was a little bit embarrassing the way we played in the first two games.

Original Prediction	Signal Prediction
equivalent	sentence2@ "We're in the Stanley Cup finals

Table 3: Examples of inconsistent predictions of the $T5_{large}$ model on the Signal case of the MRPC dataset. The changes made in the Signal case inputs are in bold.

Model	KorNLI			KLUE-NLI			KLUE-STS		
	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S
$KoBERT$	85.5	53.2	91.8	71.7	48.2	76.3	73.1	90.2	80.7
$KoElectra$	86.1	52.8	94.5	78.4	55.6	89.6	75.0	94.6	92.4
$KoGPT2$	83.9	49.0	76.1	64.3	48.4	63.6	76.7	83.9	73.3
$KoBART$	85.2	54.6	93.7	71.9	51.9	83.0	76.7	87.4	91.1
Human	**87.3**	**94.0**	**96.0**	**86.0**	**98.0**	**98.0**	**88.0**	**100.0**	**100.0**

Table 4: Results for the consistency evaluation on the Korean datasets. Acc_{val} denotes an accuracy on the validation dataset. C_R and C_S stand for the consistency for the REVERSE and SIGNAL cases, respectively. We trained each model 5 times and recorded the average of each metric. The best values are in bold.

case. We speculate a leading cause is that the $T5$ models are simultaneously trained with multiple tasks, including STS tasks, which are regarded to have a positive influence in obtaining a high consistency according to our experimental results.

5.3 Experiments on Korean Datasets

The experimental results for the Korean datasets are demonstrated in Table 4. Interestingly, the results for the Korean datasets exhibit a similar trend with those for the English datasets. The consistency of the SIGNAL case is considerably higher than that of the REVERSE case. Also, models trained on the STS tasks mark a high consistency in both the REVERSE and SIGNAL cases, while those trained on the NLI tasks completely failed in the REVERSE case. Moreover, KoGPT2 generally delivered a lower consistency in both the Reverse and Signal cases than the other models, such as KoBERT and KoElectra. Finally, just as in the English datasets, all models are highly confident in inconsistent predictions (see Figure 3). The results indicate that, for the inconsistency issue of PLMs, we do not have to blame languages but the models themselves.

6 Human Evaluation

We also evaluate the human consistency ability. Five human annotators native to each language are asked to solve the individual tasks for the English and Korean tasks. We provide 30 samples of the original input format extracted from the validation data and their corresponding perturbed examples for the Reverse and Signal cases for each annotator.

In Tables 1 and 4, the results demonstrate that humans can make consistent decisions regardless of tasks, perturbation types, and languages. However,

Model	MNLI			QNLI			RTE		
	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S	Acc_{val}	C_R	C_S
$RoBERTa_{large}$	90.0	65.5	**96.9**	94.1	64.2	**97.8**	74.7	54.4	90.6
$RoBERTa_{large}$-multi	**90.3**	**73.4**	95.6	**94.3**	**81.4**	96.8	**85.2**	**92.8**	**91.1**
$GPT2_{base}$	**79.6**	46.6	**83.8**	86.5	49.0	**89.8**	57.4	64.6	43.7
$GPT2_{base}$-multi	78.4	**52.4**	81.3	**86.6**	**56.0**	88.8	**60.9**	**71.0**	**72.4**

Table 5: Results for the consistency evaluation on multi-task training. Acc_{val} denotes an accuracy on the validation dataset. C_R and C_S stands for the consistency for the REVERSE and SIGNAL cases, respectively. We trained each model five times and recorded the average of each metric. The best value is in bold.

in the English datasets, the accuracy of humans is generally lower than that of fine-tuned models. A leading cause of this result is the small sample size for the human evaluation that degrades the performance considerably even for a single mistake. Another reason is that the average input length of the English datasets is quite long (28 words), which make annotators hardly concentrate on the evaluation. On the contrary, it is easier to focus on the Korean tasks whose average input length is much shorter (14 words). As a result, human performance on the Korean datasets outperforms that of the LMs. Also, we find that the labels of several samples of the **MRPC** dataset seem incorrect, which causes a decrease in human accuracy. We list such examples in Table 8 in Appendix A.3.

7 The Effect of Multi-Task Training on Consistency

From the earlier experiments, we observed that the $T5$ text-to-text models trained on multiple downstream tasks are very consistent in the REVERSE case but fail in the SIGNAL case. On the contrary, all classification-based models showed an opposite pattern. Therefore, we hypothesise that training classification-based models on multiple downstream tasks can attain high consistencies in both the REVERSE and SIGNAL cases.

To train the PLMs on multiple tasks simultaneously, we leverage the MT-DNN structure (Liu et al., 2019a), which shares the encoder but has individual classifiers for each task. We select $RoBERTa_{large}$ and $GPT2_{base}$ as backbone model candidates and train them on the English datasets.

Through experiments, we ascertain that our hypothesis is valid. Table 5 demonstrates the experimental results. We record the results of the NLI tasks, because all PLMs achieved a quite high consistency in the STS tasks. As in the previous study (Liu et al., 2019a), the accuracy of the multi-task models improved in general, and the enhancements are substantial for the tasks with less training data (i.e., **RTE**). Also, we observe that multi-task training improves not only accuracy but also consistency. Specifically, all PLMs achieve great improvements in the REVERSE case, obtaining a 21% increase on average. Similar to the trend observed on the accuracy, the improvement is considerable in the RTE task, recording a 40% increase on average. In the SIGNAL case, different patterns are observed depending on the size of the training data. For the **MNLI** and **QNLI**, the consistency slightly decreased by 1.6% on average, but the drop is marginal considering the complete failure of the $T5$ models. On the contrary, the consistency is improved in the **RTE** dataset, especially for $GPT2_{base}$, which is increased by 65%. Our experimental results suggest that attaining good representations through multiple language understanding tasks could be a remedy to improve consistency, especially for the small-sized datasets.

8 Summary and Outlook

Consistency is a highly desirable property that a good language understanding model should possess to obtain a human-level language understanding capability. In this paper, we proposed a simple yet efficient framework, called **CALM**, that measures a lower-bound consistency of PLMs. Through experiments, we ascertained that PLMs exhibit cases of inconsistent behaviour regardless of the pre-training objective and language despite their excellent accuracy on downstream tasks. We also confirmed that multi-task training has a positive impact on improving consistency. Our findings suggest that high accuracy is not a sufficient criterion to evaluate PLMs' language understanding abilities, and it is time to assess language models from various points of view.

References

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 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.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

Joaquin Candela-Quinonero, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc. 2006. Evaluating predictive uncertainty, visual objects classification and recognising textual entailment: selected proceedings of the first pascal machine learning challenges workshop.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sunga, Brian Stropea, and Ray Kurzweila. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *arXiv preprint arXiv:2102.01017*.

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.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12946–12954.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Improving bert's interpretations of complex words with derivational morphology. *arXiv preprint arXiv:2101.00403*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. *arXiv preprint arXiv:2105.03761*.

Stephen D Krashen. 1982. *Principles and practice in second language acquisition*. Language teaching methodology series. Pergamon, Oxford.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. 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 pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020a. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020b. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

I. Loshchilov and F. Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

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.

I. A. Mel'čuk and A. K. Žolkovskij. 1970. Towards a functioning 'meaning-text' model of language. *Linguistics*, 8(57):10–47.

Jasmina Milićević. 2006. A short guide to the meaning-text linguistic theory. *Journal of Koralex*, 8:187–233.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with bert for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

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.

Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert, and Naoaki Okazaki. 2021. Behavioral testing of knowledge graph embedding models for link prediction. In *3rd Conference on Automated Knowledge Base Construction*.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021. UnNatural Language Inference. 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 7329–7346, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

	# of classes	Train set size	Validation set size	Test set size
MNLI	3	393K	9.8K	9.8K
QNLI	2	105K	5.5K	5.5K
RTE	2	2.5K	277	3K
QQP	2	364K	40K	391K
MRPC	2	3.7K	408	1.7K
KorNLI	3	53K	10K	10K
KLUE-NLI	3	25K	1.5K	1.5K
KLUE-STS	2	1.2K	260	259

Table 6: Descriptions of datasets for the experiments.

A Appendix

A.1 Dataset Statistics

Table 6 shows the statistics of the datasets that we used for the experiments. The number of data points in the RTE, MRPC, and KLUE-STS tasks is considerably smaller than in the others.

A.2 Hyperparameter Search

We investigated the following range of hyperparameter values to decide the optimal values for the fine-tuning:

- Batch size: 32, 64, 128,
- Learning rate: $5e$-5, $1e$-5, $5e$-6.

Datasets with a large amount of training data, e.g., **MNLI** and **QQP**, are insensitive to the hyperparameter values. Therefore, we select hyperparameter values that generally perform well on small-sized datasets, such as **RTE** and **MRPC**.

A.3 Samples of MRPC data

Table 8 shows several examples of the **MRPC** data that are considered to have incorrect answers. It seems that most of the human annotators made correct predictions. We believe such samples decreased the human accuracy on the **MRPC** dataset, because we used only a few instances for the human evaluation.

Symbols	[]	{}	()	<>	;	#	!	@	∼	-
Examples	[Premise]	{Premise}	(Premise)	<Premise>	Premise;	premise#	Premise!	Premise@	Premise∼	Premise-

Table 7: The special symbols that we use for the SIGNAL case. The examples illustrate the alteration of the "Premise:" indicator.

Inputs	User1	User2	User3	User4	User5	Label
Sentence1: "Sanitation is poor,,, there could be typhoid and cholera," he said. Sentence2: "Sanitation is poor, drinking water is generally left behind... there could be typhoid and cholera."	1	1	1	1	1	0
Sentence1: The only announced Republican to replace Davis is Rep. Darrell Issa of Vista, who has spent $1.71 million of his own money to force a recall. Sentence2: So far the only declared major party candidate is Rep. Darrell Issa, a Republican who has spent $1.5 million of his own money to fund the recall.	0	0	1	0	0	1

Table 8: An example of human answers on MRPC data points that seem to have wrong labels. 0 and 1 implies 'not_equivalent' and 'equivalent', respectively.

Dataset	Type	Input 1	Input 2	Prediction
RTE	Original	Sentence1: These folk art traditions have been preserved for hundreds of years.	Sentence2: Indigenous folk art is preserved.	entailment
	Signal	Sentence1! These folk art traditions have been preserved for hundreds of years.	Sentence2! Indigenous folk art is preserved.	not_entailment
MRPC	Original	Sentence1: The initial report was made to Modesto Police December 28.	Sentence2: It stems from a Modesto police report.	equivalent
	Reverse	Sentence2: It stems from a Modesto police report.	Sentence1: The initial report was made to Modesto Police December 28.	not_equivalent
QNLI	Original	Question: What is essential for the successful execution of a project?	Sentence: For the successful execution of a project, effective planning is essential.	entailment
	Reverse	Sentence: For the successful execution of a project, effective planning is essential.	Question: What is essential for the successful execution of a project?	not_entailment

Table 9: Examples of inconsistent predictions of $Electra_{large}$.

Dataset	Type	Input 1	Input 2	Prediction
RTE	Original	Sentence1: In 1900 Berlin's arterial roads ran across Potsdam Square - Potsdamer Platz.	Sentence2: Postdam Square is located in Berlin.	not_entailment
	Reverse	Sentence2: Postdam Square is located in Berlin.	Sentence1: In 1900 Berlin's arterial roads ran across Potsdam Square - Potsdamer Platz.	entailment
MRPC	Original	Sentence1: Both are being held in the Armstrong County Jail.	Sentence2: Tatar was being held without bail in Armstrong County Prison today.	equivalent
	Signal	<Sentence1> Both are being held in the Armstrong County Jail.	<Sentence2> Tatar was being held without bail in Armstrong County Prison today.	<extra_id_0>...
QNLI	Original	Question: What fueled Luther's concept of Christ and His Salvation?	Sentence: His railing against the sale of indulgences was based on it.	not_entailment
	Signal	[Question] What fueled Luther's concept of Christ and His Salvation?	[Sentence] His railing against the sale of indulgences was based on it.	entailment

Table 10: Examples of inconsistent predictions of $T5_{large}$. The generated output of the SIGNAL case in the MRPC dataset is "<extra_id_0>1 County Jail.<extra_id_1>.<extra_id_2> sentence1>".

<table>
<tr><td colspan="2">Original Inputs: rte sentence1: At least 50 animals died in a late December avalanche. sentence2: Humans died in an avalanche.
Original Inputs: rte [sentence1] At least 50 animals died in a late December avalanche. [sentence2] Humans died in an avalanche.</td></tr>
<tr><td>Original Prediction</td><td>Signal Prediction</td></tr>
<tr><td>not_entailment</td><td><extra_id_0> <extra_id_1> [sentence1] At least 50 animals died in an</td></tr>
<tr><td colspan="2">Original Inputs: rte sentence1: Microsoft denies that it holds a monopoly. sentence2: Microsoft holds a monopoly power.
Signal Inputs: rte sentence1! Microsoft denies that it holds a monopoly. sentence2! Microsoft holds a monopoly power.</td></tr>
<tr><td>Original Prediction</td><td>Signal Prediction</td></tr>
<tr><td>not_entailment</td><td>rte sentence1! Microsoft denies that it holds a</td></tr>
<tr><td colspan="2">Original Inputs: rte sentence1: An earthquake has hit the east coast of Hokkaido, Japan, with a magnitude of 7.0 Mw. sentence2: An earthquake occurred on the east coast of Hokkaido, Japan.
Signal Inputs: rte {sentence1} An earthquake has hit the east coast of Hokkaido, Japan, with a magnitude of 7.0 Mw. {sentence2} An earthquake occurred on the east coast of Hokkaido, Japan.</td></tr>
<tr><td>Original Prediction</td><td>Signal Prediction</td></tr>
<tr><td>entailment</td><td><extra_id_0>e<extra_id_1>e {sentence2} An earthquake has hit the east</td></tr>
</table>

Table 11: More examples of inconsistent predictions of $T5_{base}$ on the Signal case of the RTE dataset. The changes made on the Signal case inputs are in bold.