BECEL: Benchmark for Consistency Evaluation of Language Models

Myeongjun Jang[1] **Deuk Sin Kwon**[3] **Thomas Lukasiewicz**[2,1]
[1]Department of Computer Science, University of Oxford, UK
[2] Institute of Logic and Computation, TU Wien, Austria
firstname.lastname@cs.ox.ac.uk
[3]Language Super Intelligence Labs, SK Telecom, South Korea
kds0281@gmail.com

Abstract

Behavioural consistency is a critical condition for a language model (LM) to become trustworthy like humans. Despite its importance, however, there is little consensus on the definition of LM consistency, resulting in different definitions across many studies. In this paper, we first propose the idea of LM consistency based on *behavioural consistency* and establish a taxonomy that classifies previously studied consistencies into several sub-categories. Next, we create a new benchmark that allows us to evaluate a model on 19 test cases, distinguished by multiple types of consistency and diverse downstream tasks. Through extensive experiments on the new benchmark, we ascertain that none of the modern pre-trained language models (PLMs) performs well in every test case, while exhibiting high inconsistency in many cases. Our experimental results suggest that a unified benchmark that covers broad aspects (i.e., multiple consistency types and tasks) is essential for a more precise evaluation.

1 Introduction

Human-like behaviour is a critical property that increases a user's trust in an artificial intelligence (AI) agent (De Visser et al., 2016; Jung et al., 2019) by improving the certification process that ascertains whether a system behaves correctly (Huang et al., 2020). [1] Accordingly, despite the outstanding performance of transformer-based PLMs on natural language understanding (NLU) benchmarks, there have been pushbacks in various corners questioning their trustworthiness based on their non-human like behaviours, such as a poor memorisation effect on infrequent information (Kassner et al., 2020; Ravichander et al., 2020; Hofmann et al., 2021), insensitivity to sentence order (Pham et al., 2021; Gupta et al., 2021; Sinha et al., 2021), and a miserable understanding of negation expressions (Hossain et al., 2020; Kassner and Schütze, 2020; Ettinger, 2020; Hosseini et al., 2021; Jang et al., 2022).

In this respect, *behavioural consistency*, a core property of humans, is an important characteristic for a model to be deemed as trustworthy LM. Accordingly, the concept of consistency has been widely discussed in natural language processing (NLP). However, despite its prominent importance, there is little consensus on the precise definition of consistency. Below are examples of different consistency definitions:

- Making consistent decisions in semantically equivalent contexts (Elazar et al., 2021).
- Being consistent on a system's beliefs across various inputs (Li et al., 2019).
- Producing logically or factually accurate statements (Li et al., 2020b).

Hence, different studies on consistency focused on diverging types of consistency but only on certain tasks, primarily natural language inference (NLI) and question answering (QA).

To this end, in this paper, we first define the consistency of an LM based on the concept of *behavioural consistency* and establish a taxonomy by systematically categorising previous works based on our definition. Next, we propose a new benchmark named **be**nchmark for **c**onsistency **e**valuation of **l**anguage models (**BECEL**), a unified dataset for evaluating an LM's consistency, which assesses multiple types of consistency on six different tasks: NLI, semantic textual similarity (STS), words-in-context (WiC), semantic analysis (SA), machine reading comprehension (MRC), and topic classification (TC). Finally, we conduct extensive experiments on our new benchmark to assess the consistent behaviour of widely used PLMs and draw the following meaningful insights.

1. We observe that none of the PLMs coherently shows a consistent behaviour in all test cases.

[1] Trustworthiness = Certification + Explanation (Huang et al., 2020).

Proceedings of the 29th International Conference on Computational Linguistics, pages 3680–3696
October 12–17, 2022.

2. Large-sized models do not necessarily perform better than small-sized models, suggesting that increasing the model size is not the solution to improve consistency.
3. Our experimental results accentuate the necessity of a unified benchmark that enables evaluations from multiple aspects (i.e., various consistency types and downstream tasks).
4. Artefacts in training data have more influence than model design, e.g., training objectives and model structures.

The data of this paper are available at https://github.com/MJ-Jang/BECEL.

2 Language Model's Consistency

Behavioural consistency refers to being consistent in behavioural patterns by adhering to the same principles.[2] Based on this notion, we define the *consistency of an LM* as its ability to make a coherent decision not contradictory to its belief.

This definition of consistency consists of two components. The first one is *belief*, which refers to what a model considers to be true. The second component is *principle*, the property that decides what a coherent decision is. Based on these two components, we classify the various types of consistency in the literature into three large categories: semantic, logical, and factual consistency.

2.1 Semantic Consistency

It is the nature of meaning-text theory (MTT) to consider that the correspondence between linguistic expressions (*text*) and semantic contents (*meaning*) is *many-to-many*, implying that the meaning can be given in different text forms (Mel'čuk and Žolkovskij, 1970; Milićević, 2006). In this regard, a model with a high level of NLU ability should capture the meaning in essence and make the same decisions in semantically identical texts considering the definition of "understanding language"[3], and this is the concept of semantic consistency. The belief and principle become a *model's predictions on semantically identical texts* and *semantic equivalence*, respectively. So, we define the *semantic consistency of an LM* as its ability to make the same decisions on semantically equivalent texts.

Semantic consistency is an indispensable property of LMs regardless of the tasks and data, since it originates from the meaning and the universal nature of language. It is probably the most widely used concept across many studies regarding an LM's consistency. Research on text adversarial attacks showed that several PLMs are susceptible to adversarial samples that are designed to convey a similar meaning to their original counterparts (Jin et al., 2020; Garg and Ramakrishnan, 2020; Li et al., 2020a; Ivgi and Berant, 2021; Li et al., 2021). Ribeiro et al. (2019) investigated semantic consistency of QA models by generating implications that must be *true* considering the model's answer on the original query. Other works observed a discrepancy in the masked language modelling (MLM) predictions of PLMs for queries where the object is replaced with its plural form (Ravichander et al., 2020) and paraphrased queries (Elazar et al., 2021). Also, recent studies introduced the semantic consistency to consistency regularisation for training LMs with improved inductive bias (Wang and Henao, 2021; Zheng et al., 2021; Kim et al., 2021).

2.2 Logical Consistency

Several NLP tasks require the fulfilment of a certain logical property. The predictions that violate this logical property are considered invalid. Therefore, the belief and principle become a *model's predictions regarding instances where the logical property holds* and a *logical property*. Hence, we define the *logical consistency of an LM* as its ability to make decisions without logical contradiction.

Logical consistency can be subdivided according to the required logical properties. Here, we outline four types of logical consistency: *negational*, *symmetric*, *transitive*, and *additive* consistency.

Negational consistency. The core property of negational consistency is the logical negation property (p is true $\Leftrightarrow \neg p$ is false; Aina et al. 2018). That is, an LM's predictions should be different for texts having the opposite meaning if the property holds. Several works observed that PLMs often generate MLM outputs that violate this property, e.g., generating the same predictions for queries like "Birds can [MASK]" and "Birds cannot [MASK]" (Kassner and Schütze, 2020; Ettinger, 2020; Jang et al., 2022). Asai and Hajishirzi (2020) used negational consistency for data augmentation to train QA models.

Symmetric consistency. Provided a function f takes two variables, a symmetric inference is defined as: $f(x, y) = f(y, x)$. Intuitively, this implies that an LM's prediction should be invariant to the input text swap for NLP tasks. Previous

[2] N., Sam M.S., Behavioral Consistency. PsychologyDictionary.org, April 7, 2013, [link].

[3] To focus on the meaning and not the text (Krashen, 1982).

works on symmetric consistency are conducted on the NLI task. Wang et al. (2019) suggested that symmetric consistency holds for instances having *contradiction* and *neutral* as a label, and investigated the change in accuracy after switching the premise and hypothesis. On the contrary, Li et al. (2019) claimed that the property holds if and only if the label is a *contradiction*. They evaluated the symmetric consistency of NLI models on newly constructed data from MS-COCO (Lin et al., 2014). Recently, Kumar and Joshi (2022) expanded the experiments from NLI to STS task and evaluated the consistency in more conservatively by measuring the confidence score difference.

Transitive consistency. Given the three predicates X, Y, and Z, transitive inference is represented as: $X \rightarrow Y \wedge Y \rightarrow Z$ then $X \rightarrow Z$ (Gazes et al., 2012; Asai and Hajishirzi, 2020). Li et al. (2019) applied this property to NLI task. Specifically, for the three related sentences P, H, and Z, they defined four transitive inference rules:

$$E(P,H) \wedge E(H,Z) \rightarrow E(P,Z), \tag{1}$$

$$E(P,H) \wedge C(H,Z) \rightarrow C(P,Z), \tag{2}$$

$$N(P,H) \wedge E(H,Z) \rightarrow \neg C(P,Z), \tag{3}$$

$$N(P,H) \wedge C(H,Z) \rightarrow \neg E(P,Z), \tag{4}$$

where E, N, and C denote entailment, neutral, and contradiction, respectively. They constructed a new evaluation set from MS-COCO (Lin et al., 2014) for assessing the transitive consistency of NLI models. In QA, Asai and Hajishirzi (2020) used the transitive property for augmenting training data by combining two questions (q_1, q_2) where the effect of q_1 is equal to the cause of q_2. (Lin and Ng, 2022) investigated PLMs' transitive inference ability on WordNet word senses and the **IS-A** relation, i.e., if A *is-a* B and B *is-a* C, then A *is-a* C. These works ascertained that PLMs do not fully obey the transitive property.

Additive consistency. We propose a new type of logical consistency that we call additive consistency. For a function f, additive inference is represented as: $f(x) = f(y) = c \rightarrow f(x+y) = c$, where c is a predicted label. For NLP tasks, additive consistency applies to any single-sentence classification task (e.g., SA and TC). Intuitively, this implies that if a model yields the same prediction for different sentences, then the prediction of the combined sentence should also be the same.

Specificity of logical consistency. It is worth mentioning that, unlike semantic consistency, logical consistency is a task- and data-specific condition. It is inapplicable to those where a certain logical property is invalid. For instance, negational consistency cannot be applied to TC. It is obvious that negating the sentence below that belongs to the *Sports* category does not change its category.

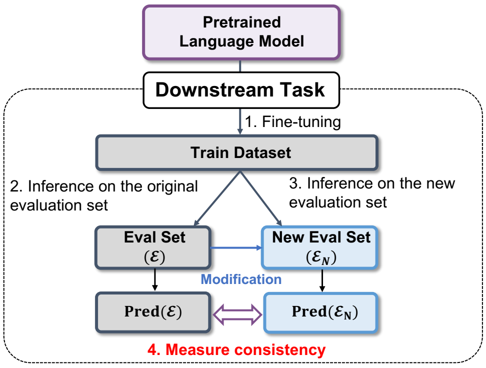


Figure 1: Overall evaluation framework for assessing an LM's consistency.

Tottenham forward Son Heung-min has signed a new four-year contract.

2.3 Factual Consistency

The basic concept of factual consistency is that a model should generate factually accurate outputs. Therefore, the belief is the *model's output*, and the principle becomes *factual correctness*. Hence, we define the *factual consistency of an LM* as its ability to generate outputs not contradictory to the common facts and given context.

By its nature of generating correct facts, factual consistency is closely related to *knowledge grounding* and *reducing hallucinations*. So, most works on factual consistency are on natural language generation (NLG), mainly text summarisation (Kryscinski et al., 2020; Maynez et al., 2020; Wang et al., 2020; Pagnoni et al., 2021), generative open-domain QA (Lewis et al., 2020b; Izacard and Grave, 2021), and dialogue generation (Li et al., 2020b; Shuster et al., 2021; Komeili et al., 2022).

3 BECEL Dataset

3.1 Overview

Figure 1 illustrates the overall framework for evaluating an LM's consistency. A PLM is fine-tuned on a downstream task and generates predictions of its original evaluation set ($\mathcal{E}$) and a new evaluation set ($\mathcal{E}_N$), specially designed from $\mathcal{E}$ to assess a certain type of consistency. Next, we compare the PLM's prediction on $\mathcal{E}$ and $\mathcal{E}_N$ to measure the consistency.

Our **BECEL** dataset contains $\mathcal{E}_N$ of multiple existing downstream tasks for assessing various con-

	BoolQ	SNLI	RTE	MRPC	WiC	SST2	AG-news
semantic	1,076	4,406	248	202	140	187	540
negational	401	2,204	153	290	-	-	-
symmetric	-	3,237	1,241	3,668	5,428	-	-
transitive	-	2,375	-	-	3,162	-	-
additive	-	-	-	-	-	53K	53K

Table 1: Number of new test data points for each downstream task and consistency type.

sistency types. It includes six downstream tasks: SNLI (Bowman et al., 2015) and RTE (Candela-Quinonero et al., 2006) for NLI, MRPC (Dolan and Brockett, 2005) for STS, BoolQ (Clark et al., 2019) for MRC, SST-2 (Socher et al., 2013) for SA, AG-News (Zhang et al., 2015) for TC and WiC (Pilehvar and Camacho-Collados, 2019), for evaluating semantic consistency and four types of logical consistencies.[4] Several data examples are in Figures 5 and 6 in Appendix B. We remove factual consistency from our evaluation scope, as benchmarks and evaluation frameworks for factual consistency are already well-studied across various tasks, such as summarisation (Kryscinski et al., 2020; Wang et al., 2020; Pagnoni et al., 2021), QA (Choi et al., 2018; Rajpurkar et al., 2018; Reddy et al., 2019), and dialogue generation (Dinan et al., 2019; Komeili et al., 2022).

Table 1 illustrates the size of the newly created $\mathcal{E}_N$ for each task and consistency type. In the case where a specific consistency cannot be applied to a particular task, it is excluded from the evaluation. The applicability of each consistency type to various tasks is described in Appendix A.2. In general, we use test sets as $\mathcal{E}$, provided gold labels are available. If not, development sets are used instead. However, training sets are used as $\mathcal{E}$, if two conditions are satisfied: (1) the size of the dev/test sets is small, and (2) new evaluation data can be collected automatically. Specifically, the RTE, MRPC, and WiC tasks for evaluating symmetric and transitive consistency belong to this case.

3.2 Data Collection Schema

Semantic consistency data. $\mathcal{E}_N$ for semantic consistency is a paraphrased version of $\mathcal{E}$. For all tasks, we paraphrase only one text input. Table 9 in Appendix illustrates each task's fixed and modified text inputs for creating $\mathcal{E}_N$. To collect paraphrase sentences, we use the publicly available Quilbot (https://quillbot.com/), as it can generate more natural paraphrases and cover broader linguistic variations compared to model-driven paraphrasing such as text adversarial attacks. In the WiC data, we remove a new data point if the target word does not exist in the paraphrased sentence. We then conduct a human evaluation through Amazon MTurk for the generated paraphrases to improve the data quality. Three annotators are allocated for each instance and asked to score the text similarity of the original and paraphrased sentences from 1 to 5. The instances where the average similarity score is not less than 4 are finally added to $\mathcal{E}_N$.

Negational consistency data. To collect $\mathcal{E}_N$ for negational consistency, we generate the opposite-meaning sentences of the modified variables listed in Table 9 by using two methods: *negation* and *antonym replacement*. For the former, we negate sentences having a single verb by inserting negation expressions like "not". For the latter, we extract adjectives and adverbs and replace only one word at a time with its antonym by using ConceptNet (Speer et al., 2017). Next, we perform the same human evaluation used in semantic consistency but select examples where the average similarity score does not exceed 2. Finally, we conduct a manual review on all instances to remove ambiguous or grammatically incorrect data points.

As mentioned earlier, negational consistency is data-specific. In SNLI, the label changes from "entailment" to "contradiction" if the hypothesis is switched with its opposite-meaning sentence. However, the label alteration is not guaranteed for the other labels, especially for "neutral". So, we only consider the "entailment" label to construct $\mathcal{E}_N$ of SNLI. For the same reason, we only use data points having the label "entailment" for RTE, "equivalent" for MRPC, and "true" for BoolQ to build $\mathcal{E}_N$.

Symmetric consistency data. We swap the text input order of tasks where the symmetric consistency is applicable. For WiC and MRPC, it is valid for every data point. Conversely, for NLI, it only applies to instances having "contradiction" as a label (Li et al., 2019) or "neutral" if the hypothesis is less specific than the premise (Wang et al., 2019). For RTE, we ascertain that the premise is always more specific than the hypothesis, and so data with "not_entailment" label are used to construct $\mathcal{E}_N$. However, it is not guaranteed in SNLI. So, only the data points with "contradiction" label are used.

Transitive consistency data. We construct $\mathcal{E}_N$ for transitive consistency on two tasks: SNLI and WiC. For SNLI, two data points must share the

[4] Brief descriptions of each downstream task are provided in Appendix A.1.

same hypothesis to apply the transitive inference rules described in Section 2.2, but only a premise is shared in the SNLI dataset. Hence, we leverage the symmetric consistency applicable to instances with the "contradiction" label, which enables us to transform the rules 3 and 4 as follows:

$$E(P, H) \wedge C(P, Z) \rightarrow C(H, Z),$$
$$N(P, H) \wedge C(P, Z) \rightarrow \neg E(H, Z).$$

By using the modified rules, we collect $\mathcal{E}_N$ for SNLI automatically. However, since the hypothesis is less specific than the premise in most cases in the SNLI data (Wang et al., 2019), we observe that the modified rules do not apply to several data points. Therefore, we conduct a human evaluation through Amazon MTurk to filter out such instances. Three annotators are allocated to each instance. We add examples to $\mathcal{E}_N$, provided at least two annotations comply with the rules.

For the WiC task, given a target word w and three predicates A, B, and C, the following transitive rules are applicable to every data point:

$$T(A, B|w) \wedge T(B, C|w) \rightarrow T(A, C|w),$$
$$T(A, B|w) \wedge F(B, C|w) \rightarrow F(A, C|w),$$
$$F(A, B|w) \wedge T(B, C|w) \rightarrow F(A, C|w),$$

where T/F implies that the meaning of the word w is used identically/differently in the given two sentences. We use these rules to collect $\mathcal{E}_N$ of WiC.

Additive consistency data. The additive consistency is valid for tasks that take a single-text input. To construct $\mathcal{E}_N$ for each task, we generate all possible combinations of two data points that share the same label and create a new one by merging their text inputs. Next, we remove a new data point if the token length of the merged text exceeds the 75% quantile of that of the training data, because such instances can be considered out-of-distributions that can overestimate the inconsistency issue.

3.3 Evaluation Metrics

Semantic/Symmetric consistency. Assume that $e_i \in \mathcal{E}$, $e_i^N \in \mathcal{E}_N$, and e_i^N is a perturbed version of e_i, and therefore, $|\mathcal{E}| = |\mathcal{E}_N|$. A model M should generate the same predictions for e_i and e_i^N. Therefore, by referencing the robust accuracy (Tsipras et al., 2019; Ivgi and Berant, 2021), we define the inconsistency metric (τ) for semantic and symmetric consistency as follows:

$$\tau = 1 - 1/|\mathcal{E}_N| \sum_{i=1}^{|\mathcal{E}_N|} \mathbb{1}(M(e_i) = M(e_i^N)).$$

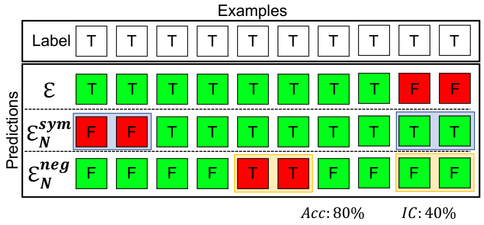


Figure 2: Graphical representation of accuracy and symmetric/negational consistency for binary classification. The blue and yellow boxes denote inconsistent cases for symmetric and negational consistency, respectively.

Negational consistency. Let $e_i \in \mathcal{E}$, $e_i^N \in \mathcal{E}_N$, and e_i^N is a perturbed version of e_i (i.e., $|\mathcal{E}| = |\mathcal{E}_N|$). Contrary to semantic and symmetric consistency, a model M should produce different predictions for e_i and e_i^N, where $e_i^N \in \mathcal{E}_N$ is a new instance designed for measuring negational consistency. Therefore, we define the inconsistency metric for negational consistency as follows:

$$\tau = 1 - 1/|\mathcal{E}_N| \sum_{i=1}^{|\mathcal{E}_N|} \mathbb{1}(M(e_i) \neq M(e_i^N)).$$

Transitive/Additive consistency. For both transitive and additive consistency, a new instance e_i^N is generated from two data points of $\mathcal{E}$. Assume that $e_i^N \in \mathcal{E}_N$, and e_i^N originates from $e_{i,1}, e_{i,2} \in \mathcal{E}$. Including e_i^N where the antecedent is not satisfied, i.e., a model M makes incorrect predictions for $e_{i,1}$ or $e_{i,2}$, can overestimate the inconsistency problem. Therefore, we use a conditional inconsistency as an evaluation metric:

$$\tau = 1 - 1/|\mathcal{C}| \sum_{i=1}^{|\mathcal{C}|} \mathbb{1}(M(c_i) = l_i),$$

where l_i is the label of $c_i \in \mathcal{C}$, and $\mathcal{C} \subset \mathcal{E}_N$ denotes the set of e_i where the model M makes correct predictions for both $e_{i,1}$ and $e_{i,2}$.

3.4 Importance of Measuring Consistency

Previous benchmarks regarding the opposite meanings (Naik et al., 2018; Hossain et al., 2020) or symmetry (Wang et al., 2019) only measure accuracy on the new test suit. It is true that models with low accuracy are likely to be inconsistent, but the high accuracy does not necessarily guarantee high consistency. Figure 2 well illustrates an example case. Although the accuracy is 80% in the original test set and two types of $\mathcal{E}_N$, implying that the model is quite robust on unseen data, the inconsistency is 40%. Therefore, consistency should be treated as an independent evaluation metric.

Model		BoolQ		MRPC			RTE			SNLI				SST2		WiC			AG-News	
		τ_{sem}	τ_{neg}	τ_{sem}	τ_{neg}	τ_{sym}	τ_{sem}	τ_{neg}	τ_{sym}	τ_{sem}	τ_{neg}	τ_{sym}	τ_{trn}	τ_{sem}	τ_{add}	τ_{sem}	τ_{sym}	τ_{trn}	τ_{sem}	τ_{add}
BERT	base	20.5	87.2	16.6	90.3	7.6	15.8	76.9	17.8	11.0	15.9	12.1	4.0	5.2	0.2	7.1	8.9	46.8	2.8	1.6
	large	16.5	77.3	12.5	90.8	6.8	12.3	75.8	15.8	9.9	11.7	10.2	3.6	3.3	0.1	8.4	7.0	49.3	3.0	1.7
RoBERTa	base	13.5	43.5	13.2	83.5	4.7	12.8	56.9	18.6	9.6	9.5	9.3	3.3	4.5	0.1	10.1	6.9	50.8	3.1	3.1
	large	10.2	40.8	8.4	84.2	4.3	9.8	24.6	11.6	7.9	5.9	9.7	3.5	2.3	0.1	9.3	7.3	46.6	2.7	1.1
Elelctra	base	7.1	63.7	8.8	86.6	7.1	9.4	32.8	9.8	9.2	7.7	9.5	3.3	3.0	0.0	10.1	5.1	48.0	2.8	2.4
	large	6.8	42.3	5.5	77.0	5.3	8.9	17.3	6.7	7.9	5.4	6.4	2.5	4.0	0.1	8.9	7.9	46.5	2.6	1.0
ERNIE2.0	base	13.3	62.4	6.3	79.6	6.6	13.2	35.0	13.2	10.1	13.2	9.5	3.3	5.2	0.1	5.1	5.1	51.1	3.2	2.7
	large	7.6	66.8	7.3	62.7	6.4	9.8	37.1	22.8	9.0	7.5	7.3	3.0	3.5	0.0	9.0	6.9	46.7	3.5	1.7
GPT2	base	12.8	85.8	18.4	87.2	14.5	18.1	75.3	33.3	16.3	30.0	23.0	10.4	19.6	0.8	14.1	13.1	47.4	2.7	2.2
	large	23.3	75.3	14.6	89.5	10.6	13.9	52.3	15.8	11.5	13.9	12.0	4.9	6.2	0.1	13.4	12.5	49.8	3.0	4.3
BART	base	13.4	71.2	12.2	84.4	5.6	11.4	70.5	18.3	10.8	10.9	14.4	4.7	4.7	0.1	8.7	7.7	53.0	3.0	3.5
	large	7.9	58.2	11.4	82.2	4.6	10.2	29.7	27.2	8.7	6.6	7.5	2.8	3.0	0.1	6.9	5.4	53.1	2.5	3.4
T5	base	12.9	29.4	8.1	39.8	3.7	11.7	18.5	16.8	10.9	7.2	10.6	3.6	4.1	0.2	16.0	7.9	46.3	2.1	0.3
	large	10.9	19.7	4.5	25.2	4.2	8.6	15.9	8.0	9.3	5.8	8.3	2.9	3.0	0.1	8.6	6.3	45.3	1.7	0.2

Table 2: The average of semantic (τ_{sem}), negational (τ_{neg}), symmetric (τ_{sym}), transitive, (τ_{trn}), and additive inconsistency (τ_{add}). All the metrics are lower the better. We repeat each experiments for five times.

4 Experiments and Analysis

4.1 Experimental Design

Model candidates. We evaluated the consistency of the below widely used PLMs (both *base* and *large* size models) on our new benchmark suit.

- **Encoder models**: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), Electra (Clark et al., 2020), and ERNIE 2.0 (Sun et al., 2020).
- **Decoder models**: GPT2 (Radford et al., 2019).
- **Encoder-Decoder models**: BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020).

Training details. At fine-tuning, we use AdamW optimiser (Loshchilov and Hutter, 2017) and a linear learning rate scheduler decaying from 1e-3. All models are trained for 10 epochs, and the early stopping method is used during the training. The batch size and learning rate are different across model size and tasks. See Appendix A.3 for more details. For T5, we apply text-to-text multitask training by using the free-text input format used by Raffel et al. (2020). We repeat the experiments for each model and task for five times and report their average values. Our best validation performance is almost close to the reported results in previous works (see Table 6 in the appendix).

4.2 Semantic Consistency Results

The results are in Table 2. We ascertain that PLMs show a different consistency across diverse tasks. Specifically, τ_{sem} is extremely low in the SST2 and AG-News tasks. We conjecture that a leading cause is that these tasks have a high correlation between labels and certain words, such as sentiment words and proper nouns, and therefore, are hardly affected by paraphrases. Among the other tasks, the PLMs are relatively more consistent in MRPC than the others but still make many mistakes considering that the STS task is designed to focus on semantic equivalence. We also observe that GPT2 and BERT are highly inconsistent. T5 and Electra show the lowest τ_{sem}, but the difference to the others, apart from GPT2 and BERT, are marginal. The results suggest that a model's training objective somewhat affects its semantic consistency.

	BoolQ	MRPC	SST2	RTE
BAE	12.4 (-1.1)	7.9 (-5.3)*	5.9 (+3.6)*	11.7 (-1.1)
TextFooler	11.2 (-2.3)*	8.9* (-6.7)*	6.4 (+4.2)*	11.3 (-1.5)

Table 3: The inconsistency results of the adversarial training experiments. The value written in parenthesis is the difference compared to the original RoBERTa-base model. The difference is statistically significant with p value < 0.05 (*).

Can adversarial training be a solution? Adversarial training is widely used to improve robustness by providing models with original and adversarial samples (Jin et al., 2020). We investigate whether it is beneficial to improve semantic consistency. We apply two text attack methods, BAE (Garg and Ramakrishnan, 2020) and TextFooler (Jin et al., 2020), to the RoBERTa-base model by using TextAttack (Morris et al., 2020). Five adversarial samples are generated for each data point.

The results are in Table 3. We confirm that adversarial training is not always beneficial. The improvement is marginal, except for MRPC and even backfired in SST2. We speculate that a leading cause is that the attack methods are likely to generate incorrect paraphrase sentences (see Appendix 10 for examples). Moreover, adversarial training is vulnerable to instances that the attack method cannot generate. It has been observed that about 45% of inconsistent predictions contain examples that have different sentence structures (e.g., changing active to passive), which synonym-replacement-based methods like BAE and TextFooler are unable to produce. The results suggest that adversarial training cannot be an ultimate solution to improve semantic consistency.

4.3 Negational Consistency Results

Table 2 presents the results of the negational consistency experiments. It is astonishing that τ_{neg} is very high across all tasks apart from SNLI, suggesting that the fine-tuned PLMs entirely fail to understand the opposite meaning. For SNLI, we strongly believe that the leading cause of low τ_{neg} are superficial cues in the training data. It is well known that there is a strong correlation between negation expressions and "contradiction" labels in the SNLI data (Gururangan et al., 2018), and we confirm that almost 68% of training instances with negation expressions in the hypothesis have "contradiction" labels. So, achieving high consistency in SNLI is easy, as our new evaluation set originates from instances with "entailment" labels, as illustrated in Section 3.2. The relatively low τ_{neg} of the T5 models, which can benefit from the SNLI data through multi-task training, also support our claim.

Model design vs. superficial cues. Similarly to the semantic consistency experiments, GPT2 and BERT perform worst in general. To compare the impact of model designs (e.g., training objectives, model structure) and superficial cues, we use the following metric for the model M on the task T:

$$\rho_T^M = (\tau_T^M - \tau_{SNLI}^M)/(\tau_T^M - \tau_T^*),$$

where τ_T^M implies the negational inconsistency of the model M on the task T. τ_T^* denotes the best inconsistency of task T among similar-size models (e.g., *base*). Intuitively, the metric implies that the performance gap with SNLI (i.e., effect of superficial cues) is ρ times higher than that with the best PLM (i.e., effect of model designs).

We measure ρ of BERT, GPT2, and BART, because their performance does not rank at the top across all tasks (ρ becomes larger if the model's inconsistency is close to the best performance). Single-task trained models are considered for deciding the best performance. The results are in Table 4. We observe that ρ is greater than 1 in every case. RTE has relatively low values, as it shares the same superficial cues with SNLI, but their total amount is much less. This suggests that superficial cues have a greater effect than model designs.

Model	BoolQ		MRPC		RTE	
	base	large	base	large	base	large
BERT	1.63	1.80	6.95	2.81	1.38	1.10
GPT2	1.32	1.78	7.53	2.82	1.07	1.10
BART	2.18	1.99	15.31	3.09	1.58	1.31

Table 4: ρ values of BERT, GPT2, and BART.

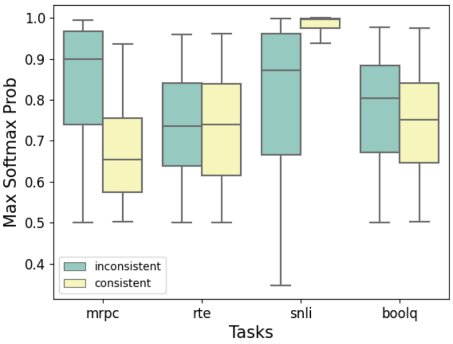


Figure 3: Box plot of maximum softmax probability of RoBERTa-base for negational consistency experiments.

Overconfident inconsistent predictions. Negational inconsistency would be less concerning, if the predictions are made by change (i.e., high entropy). However, we observe that models are very confident regarding their inconsistent decisions, generating similar or higher softmax probabilities than the consistent predictions in most cases (see, e.g., Figure 3). The confidence score seems reasonable only in the SNLI task, which contains superficial cues. The results suggest that fine-tuned PLMs are hard to trust, considering their overconfidence in incorrect and inconsistent predictions.

4.4 Symmetric Consistency Results

Table 2 shows the experimental results of symmetric consistency. In terms of the model, GPT2 again, performs worst in most cases, implying that decoder-only auto-regressive models are not suitable for achieving high consistency. The inconsistency is not significantly different for the other models, but Electra outperforms the others in general.

Compared to the NLI tasks, the inconsistency is much lower in WiC and MRPC, which are designed to focus on semantic equivalence, suggesting that achieving high symmetric consistency might be possible by making PLMs capturing the latent meaning of the texts. Although the inconsistency is fairly low, it should not be overlooked, because symmetry is an uncomplicated property that requires a simple reasoning ability. For this reason, humans are likely to show an extremely low inconsistency. We conduct a brief human evaluation on the MRPC task by asking five human annotators 30 questions each and observe that humans are highly consistent on symmetry, achieving $\tau_{sym} = 0.7$.

4.5 Transitive Consistency Results

The transitive consistency results are in Table 2. Interestingly, they are entirely different in the two tasks. In the SNLI task, which is designed to infer the logical relationship between two given sentences, all PLMs show a strong performance. However, in the WiC task, which focuses on the word's meaning, the inconsistency is very low even though the evaluation data originate from the training set. The results suggest that the transitive reasoning ability is highly contingent on the purpose of downstream tasks.

Does the training data size matter? SNLI and WiC have two major differences: (1) task objective and (2) data size (i.e., approximately 500K and 6K for SNLI and WiC, respectively). To ascertain whether more training data help achieving a high consistency, we conduct an additional experiment by down-sampling the training data size of SNLI to 6K. The Electra models that record the best τ_{trn} are used for this experiment. The results are in Table 5. The inconsistency increases after the down-sampling, but is still lower than that of WiC, and the validation accuracy is impaired, especially in the base-size model. The results suggest that small training data can cause high inconsistency, as a model becomes less accurate, but the task objective affects much more than the training data size.

Model		SNLI		SNLI-6K		WiC	
		$\mathcal{A}_{val}$	τ_{trn}	$\mathcal{A}_{val}$	τ_{trn}	$\mathcal{A}_{tr}$	τ_{trn}
Electra	base	91.8	3.3	64.8	10.0	81.6	48.0
	large	93.5	2.5	85.6	3.3	80.7	46.5

Table 5: Results of the down-sampled SNLI experiments. $\mathcal{A}_{val}$ and $\mathcal{A}_{tr}$ denote the validation and training accuracy, respectively. We report $\mathcal{A}_{tr}$ of WiC, because its $\mathcal{E}_N$ originates from the training data.

4.6 Additive Consistency Results

It is noteworthy that this experiment is a very easy task, because the input is a combination of two sentences that belong to the same category, so the model has more evidence to make the correct decision. The results of the additive consistency experiments are in Table 2. All the PLMs are highly consistent in SST2 but make some mistakes in AG-News except for the T5 models. The average τ_{add} of 2.3 in AG-News is not a low score considering the task difficulty. To become trustworthy, PLMs need to be more consistent on the additive property.

5 Discussion

Are large models more consistent? It is well-known that large-size models consistently outperform small-size models in terms of accuracy. Does the same trend occur from a consistency perspective? Figure 4 illustrates the portion of the three cases: the performance of the large models are better, worse, or show no statistical difference. Interestingly, the case where there is no statistical difference in consistency between the large- and base-size models accounts for a large portion, and sometimes base-size models even perform better. This pattern is hardly seen in accuracy-based evaluation metrics, suggesting that additional evaluation metrics such as consistency other than the accuracy should be considered for a precise evaluation.

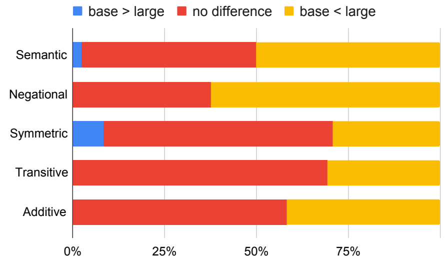


Figure 4: Portion of experimental cases where the large-size models are more or less consistent than the base-size models. A t-test under the significance level of 0.05 confirmed the statistical difference in performance.

Necessity of a unified benchmark. Our experimental results highlight the importance of evaluating models in a wide spectrum. We verify that none of the PLMs performs coherently well in every experiment, suggesting that focusing on a certain task or consistency type contains the risk of reaching a wrong conclusion. For instance, we might conclude that PLMs are fairly consistent if we only consider semantic consistency. If we conduct experiments only in the NLI tasks like extant studies, the conclusion might be distorted, since the results of all inconsistency types seem reasonable in the NLI tasks, especially in SNLI. Our new dataset, however, prevents us from drawing such a fallacious conclusion by allowing us to assess models across multiple consistency types and tasks, demonstrating its importance to have a unified benchmark covering a wide array of topics including different evaluation criteria and task types.

Uncontrollable AI. Due to the nature of inductive reasoning, the inductive bias of machine learning and deep learning models is greatly affected by the patterns in the training data. Although this is well-known and widely accepted (Alzubi et al., 2018; Katsaros et al., 2019; Anagnostis et al., 2020; Xu et al., 2020; Thielen et al., 2020; Ma et al., 2021), our experimental results show that the artefacts in data are a more influential factor than the model design in deciding its inductive bias (Section 4.3). The problem is that we have a control over the model design but not the artefacts, as it is difficult to review and manipulate all data points with an enormous size. This evokes a critical concern: *uncontrollable AI*. However elaborate the model that we design with highly advanced training objectives and model structures, we might not have a full control over the model, as the ungovernable effect of the artefacts in data remains. It is thus imperative to take appropriate actions to address the data-driven faulty behaviour of the model, such as the generation of ethically problematic outputs (Nangia et al., 2020). To overcome such issues and move forward to developing more trustworthy and safer AI, perhaps it is time to think beyond inductive reasoning.

6 Summary and Outlook

In this work, we first defined LM consistency based on the concept of behavioural consistency: a core property that a sound LM should obey. Next, we categorised various previous studies regarding consistency into three types: semantic, logical, and factual consistency. Finally, we designed a benchmark suite to assess various types of consistency on multiple downstream tasks.

Through extensive experiments, we observed that none of the PLMs shows perfectly consistent outputs in all test cases. Our experimental results highlight the essence of evaluation schema in multiple spectrums to avoid reaching a distorted conclusion. We also revealed that the impact of spurious artefacts presented in training data is greater than that of model design, such as model size and learning objective. This finding raises concerns about uncontrollable AI, as we have no control over the artefacts in tremendous amounts of data. Our work suggests that we should probably go beyond neural models, which only allow inductive reasoning, to develop trustworthy and safe AI.

Acknowledgements

This work was partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1 and by the AXA Research Fund. We also acknowledge the use of Oxford's ARC facility, of the EPSRC-funded Tier 2 facility JADE II (EP/T022205/1), and of GPU computing support by Scan Computers International Ltd.

References

Laura Aina, Raffaella Bernardi, and Raquel Fernández. 2018. A distributional study of negated adjectives and antonyms. In *CEUR Workshop Proceedings*, volume 2253.

Jafar Alzubi, Anand Nayyar, and Akshi Kumar. 2018. Machine learning from theory to algorithms: An overview. In *Journal of Physics: Conference Series*, volume 1142, page 012012. IOP Publishing.

Athanasios Anagnostis, Elpiniki Papageorgiou, and Dionysis Bochtis. 2020. Application of artificial neural networks for natural gas consumption forecasting. *Sustainability*, 12(16):6409.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, Lisbon, Portugal. 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.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

In *International Conference on Learning Representations*.

Ewart J. De Visser, Samuel S. Monfort, Ryan McKendrick, Melissa A. B. Smith, Patrick E. McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3):331.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

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. Erratum: Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1407–1407.

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.

Regina Paxton Gazes, Nicholas W. Chee, and Robert R. Hampton. 2012. Cognitive mechanisms for transitive inference performance in rhesus monkeys: Measuring the influence of associative strength and inferred order. *Journal of Experimental Psychology: Animal Behavior Processes*, 38(4):331.

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. 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 3594–3608, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270.

Maor Ivgi and Jonathan Berant. 2021. Achieving model robustness through discrete adversarial training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1529–1544, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.

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.

Eun-Soo Jung, Suh-Yeon Dong, and Soo-Young Lee. 2019. Neural correlates of variations in human trust in human-like machines during non-reciprocal interactions. *Scientific Reports*, 9(1):1–10.

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.

Dimitrios Katsaros, George Stavropoulos, and Dimitrios Papakostas. 2019. Which machine learning paradigm for fake news detection? In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 383–387. IEEE.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. 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 6130–6141, Online. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

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.

Ashutosh Kumar and Aditya Joshi. 2022. Striking a balance: Alleviating inconsistency in pre-trained models for symmetric classification tasks. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1887–1895, Dublin, Ireland. Association for Computational Linguistics.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Ruixi Lin and Hwee Tou Ng. 2022. Does BERT know that the IS-a relation is transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99, Dublin, Ireland. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. 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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *ArXiv*, abs/1711.05101.

Zeyu Ma, Han Yu, Jinyao Xia, Chunzhi Wang, Lingyu Yan, and Xianjin Zhou. 2021. Network traffic prediction based on seq2seq model. In *2021 16th International Conference on Computer Science & Education (ICCSE)*, pages 710–713. IEEE.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. 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.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Nils Thielen, Dominik Werner, Konstantin Schmidt, Reinhardt Seidel, Andreas Reinhardt, and Jörg Franke. 2020. A machine learning based approach to detect false calls in SMT manufacturing. In *2020 43rd International Spring Seminar on Electronics Technology (ISSE)*, pages 1–6. IEEE.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Haohan Wang, Da Sun, and Eric P Xing. 2019. What if we simply swap the two text fragments? a straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7136–7143.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinghui Xu, Yu Wen, Chun Yang, and Dan Meng. 2020. An approach for poisoning attacks against RNN-based cyber anomaly detection. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1680–1687. IEEE.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. 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 3403–3417, Online. Association for Computational Linguistics.

A Appendix

A.1 Task Descriptions

BoolQ (Clark et al., 2019) is a dataset for machine reading comprehension (MRC) with yes/no questions. Each data point consists of a triplet such as question, passage, and answer, requiring a broad range of inference capacities to solve questions. SNLI (Bowman et al., 2015) and RTE (recognising textual entailment) (Candela-Quinonero et al., 2006) are datasets for natural language inference (NLI). Each data point is composed of a sentence pair and a label indicating the relationship between the pair (i.e., "entailment", "neural", and "contradiction"). MRPC (Microsoft Research Paraphrase Corpus) (Dolan and Brockett, 2005) is a dataset for semantic textual similarity (STS). Each data point consists of a sentence pair and a label indicating whether the two paraphrased sentences are semantically equivalent. SST-2 (Stanford Sentiment Treebank) (Socher et al., 2013) is a dataset for sentiment analysis (SA). Each data point is composed of a phrase and a binary sentiment label (i.e., positive and negative). AG-News (Zhang et al., 2015) is a dataset of new articles for topic classification (TC). Each data point is composed of a title and a description of an article, and a label related to one of the four topics of the article (i.e., "World", "Sports", "Business", and "Sci/Tech"). WiC (Word-in-Context) (Pilehvar and Camacho-Collados, 2019) is a dataset for identifying the intended meaning of words. Each data point consists of two sentences containing the same specific word and a label indicating whether the word is used with the same meaning in different contexts.

A.2 Applicability of Logical Consistencies to Downstream Tasks

Negational Consistency Applicability. Among our downstream tasks, negational consistency is invalid for the TC and WiC tasks. Regarding the TC task, negated sentences normally belong to the same category as their original version, as illustrated in the example in Section 2.2. Similarly, the labels are preserved in the WiC task, because the meaning of the target word does not change in the perturbed sentence.

Although negational consistency is theoretically applicable to the SA task, we remove it from our evaluation scope for a practical reason. We observe that our method for generating the opposite meaning sentence does not suit well on the spoken language (e.g., movie reviews) that constitute the SST2 dataset.

Symmetric Consistency Applicability. For symmetry to hold, the following two conditions are necessary:

Condition 1. The input should consists of two sentences.

Condition 2. The hierarchy between the two sentences should be equivalent.

The TC and SA tasks violate the first condition. Regarding the MRC task (i.e., BoolQ), the question is dependent on the passage, and, therefore, it violates the second condition. As a result, the three tasks are removed from our scope for evaluating symmetric consistency.

Transitive Consistency Applicability. Theoretically, transitive consistency is valid for the downstream tasks where symmetric consistency holds. However, it requires one more condition for practical reasons: the two data points must have a common sentence, e.g., the same hypothesis in the NLI task. Only the SNLI and WiC datasets satisfy this condition among our candidate tasks. Although it is possible to construct new data for the MRPC and RTE datasets, it can cause a distribution shift issue that could exaggerate the inconsistency problem. Therefore, we conducted the transitive consistency evaluation only on the SNLI and WiC datasets.

Additive Consistency Applicability. Additive consistency always holds for tasks that take a single sentence as an input. However, it is not guaranteed if a downstream task requires more than two sentences as an input. Table 7 shows the example of the violation in the SNLI task. Thus, we tested additive consistency only for the SA and TC tasks.

A.3 Training Hyperparameters

Table 8 describes the batch-size per GPU, input sentence length (i.e., number of tokens), and learning rates used for training models for each dataset. Similarly to previous works, we confirm that the datasets with large training data (e.g., SNLI, SST2, and AG-news) were insensitive to hyperparameter values.

A.4 Human Annotation

We used Amazon Mechanical Turk (https://www.mturk.com/) for annotating our data. We employed Anglophone annotators with an acceptance rate of at least 98% and the number of HITs greater than

Model		BoolQ	MRPC	RTE	SNLI	SST2	AG-News	WiC	
		$\mathcal{F}_{val}$	$\mathcal{F}_{val}$	$\mathcal{F}_{val}$	$\mathcal{F}_{val}$	$\mathcal{F}_{val}$	$\mathcal{F}_{val}$	$\mathcal{F}_{tr}$	$\mathcal{F}_{val}$
BERT	base	66.6 (1.0)	81.8 (1.9)	62.1 (2.0)	90.1 (0.2)	90.5 (0.3)	93.2 (0.2)	62.5 (13.4)	53.0 (7.4)
	large	70.3 (1.4)	82.0 (1.4)	64.4 (3.3)	91.0 (0.2)	92.4 (0.4)	93.9 (0.2)	70.0 (9.0)	57.7 (2.9)
RoBERTa	base	75.8 (1.0)	86.4 (0.9)	71.5 (1.8)	91.5 (0.0)	92.9 (0.4)	94.1 (0.1)	78.6 (3.9)	63.8 (1.8)
	large	84.9 (0.4)	88.6 (1.0)	81.5 (2.1)	93.0 (0.1)	95.9 (0.3)	94.3 (0.2)	77.0 (6.2)	66.0 (2.0)
Electra	base	73.8 (2.6)	88.3 (0.2)	75.3 (2.5)	91.8 (0.1)	93.9 (0.2)	93.2 (0.2)	80.4 (4.3)	66.5 (5.9)
	large	87.1 (0.5)	90.1 (0.6)	86.7 (1.2)	93.5 (0.3)	95.4 (2.2)	93.8 (0.4)	80.7 (1.8)	69.0 (1.2)
ERNIE2.0	base	76.2 (1.2)	86.8 (0.9)	73.4 (2.6)	91.1 (0.2)	93.6 (0.2)	93.5 (0.1)	62.8 (13.1)	56.0 (4.6)
	large	82.6 (0.7)	86.6 (1.2)	76.7 (1.1)	92.1 (0.0)	95.1 (0.2)	94.0 (0.2)	67.1 (9.5)	55.2 (10.1)
GPT2	base	62.9 (1.7)	77.3 (1.0)	65.3 (2.5)	84.7 (1.1)	90.9 (0.5)	92.9 (0.1)	73.4 (3.8)	63.5 (2.6)
	large	75.3 (0.8)	80.4 (1.2)	69.0 (2.8)	90.8 (0.2)	94.1 (0.4)	94.1 (0.2)	87.4 (5.7)	64.5 (2.2)
BART	base	64.8 (2.4)	85.6 (1.1)	70.7 (1.0)	90.8 (0.2)	93.0 (0.3)	93.8 (0.3)	78.8 (5.5)	56.0 (1.7)
	large	78.4 (3.9)	81.6 (8.3)	74.9 (3.1)	93.1 (0.1)	95.9 (0.2)	94.0 (0.7)	77.3 (3.5)	58.9 (2.9)
T5	base	79.9 (0.2)	86.8 (0.9)	77.6 (0.2)	90.1 (0.1)	94.0 (0.2)	92.1 (0.2)	82.3 (0.3)	64.5 (1.1)
	large	83.8 (0.6)	89.3 (0.9)	88.0 (0.6)	92.1 (0.2)	95.8 (0.1)	92.5 (0.4)	84.6 (1.5)	70.3 (0.8)

Table 6: Our validation performance of the PLMs on the seven downstream tasks; $\mathcal{F}_{tr}$ and $\mathcal{F}_{val}$ denote F1 score on the training and validation set, respectively. We report the training performance of the WiC task, because the gap between training and validation performance is large compared to the other tasks. We report the average of five repetitions. The values written in parenthesis imply a standard deviation.

EXAMPLE 1
Premise: Two women are embracing while holding to go packages.
Hypothesis: Two woman are holding packages.
Label: entailment
EXAMPLE 2
Premise: Two men on bicycles competing in a race.
Hypothesis: People are riding bikes.
Label: entailment

MERGED EXAMPLE
Premise: Two women are embracing while holding to go packages. Two men on bicycles competing in a race.
Hypothesis: Two woman are holding packages. People are riding bikes.

Table 7: Example of SNLI data where negational consistency does not hold. The label of the merged example cannot be "entailment", because two women are not riding bikes.

	BoolQ	SNLI	RTE	MRPC	WiC	SST2	AG-news
b-size	8	64	8	8	64	32	32
s-len	512	128	256	128	128	128	256
lr	$2e^{-5}$	$1e^{-5}$	$1e^{-5}$	$2e^{-5}$	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$

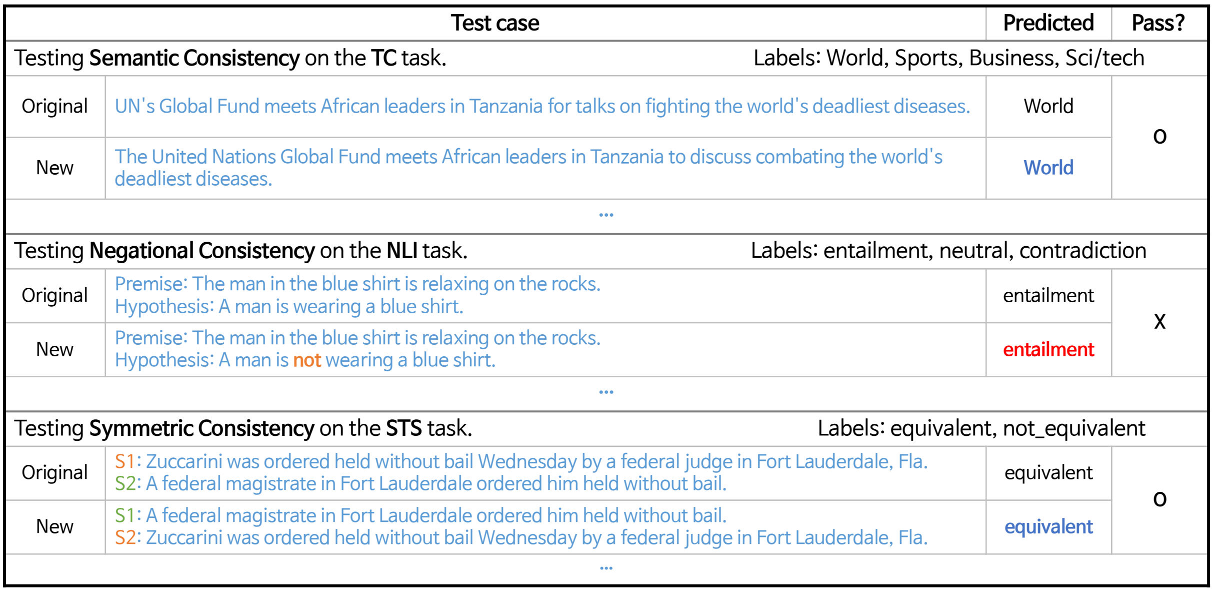
Table 8: Batch-size, sentence length, and learning rates used for the BECEL benchmark experiments.

	Fixed Variable	Modified Variable
BoolQ	passage	question
SNLI	premise	hypothesis
RTE	premise	hypothesis
MRPC	sentence1	sentence2
WiC	word, sentence1	sentence2
SST2	-	text
AG-news	-	text

Table 9: Modified variables of each dataset for collecting $\mathcal{E}_N$ for semantic and negational consistency.

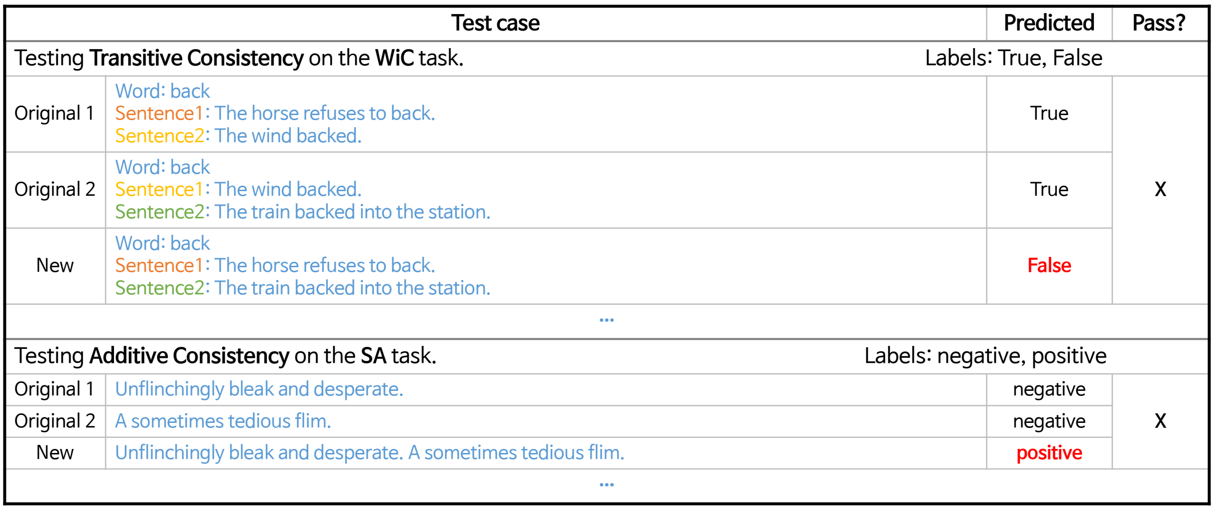
1,000. The representative snapshot of the UI for the human annotation is shown in Figure 7.

B Examples



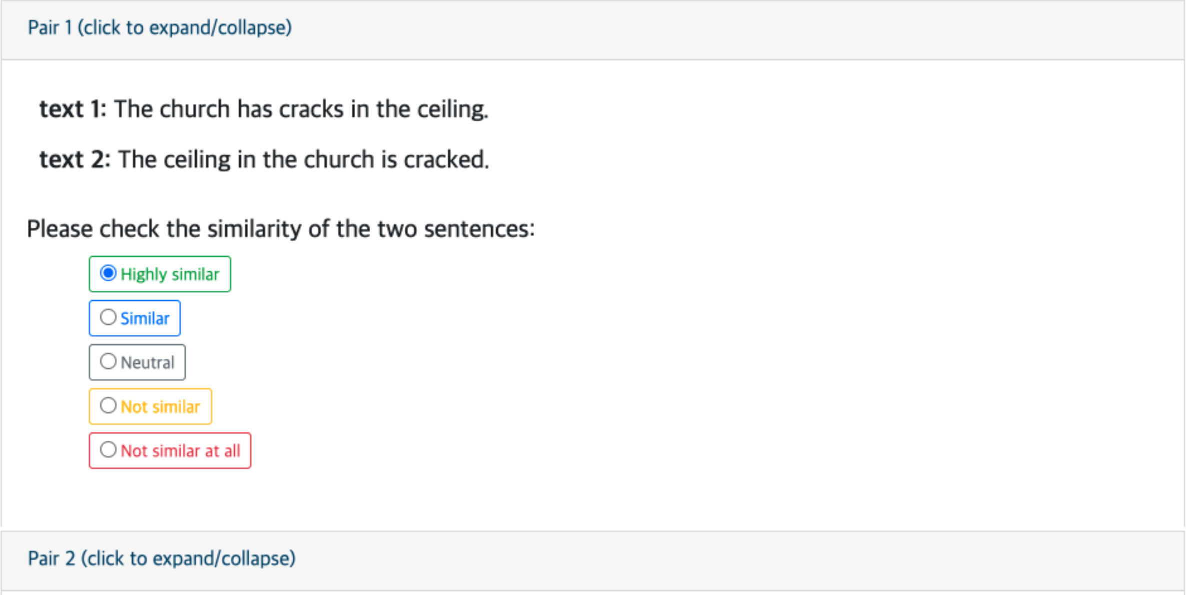
Test case		Predicted	Pass?
Testing **Semantic Consistency** on the **TC** task.	Labels: World, Sports, Business, Sci/tech		
Original	UN's Global Fund meets African leaders in Tanzania for talks on fighting the world's deadliest diseases.	World	O
New	The United Nations Global Fund meets African leaders in Tanzania to discuss combating the world's deadliest diseases.	World	
...			
Testing **Negational Consistency** on the **NLI** task.	Labels: entailment, neutral, contradiction		
Original	Premise: The man in the blue shirt is relaxing on the rocks. Hypothesis: A man is wearing a blue shirt.	entailment	X
New	Premise: The man in the blue shirt is relaxing on the rocks. Hypothesis: A man is **not** wearing a blue shirt.	entailment	
...			
Testing **Symmetric Consistency** on the **STS** task.	Labels: equivalent, not_equivalent		
Original	S1: Zuccarini was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla. S2: A federal magistrate in Fort Lauderdale ordered him held without bail.	equivalent	O
New	S1: A federal magistrate in Fort Lauderdale ordered him held without bail. S2: Zuccarini was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.	equivalent	
...			

Figure 5: Data examples of semantic, negational, and symmetric consistency evaluation.



Test case		Predicted	Pass?
Testing **Transitive Consistency** on the **WiC** task.	Labels: True, False		
Original 1	Word: back Sentence1: The horse refuses to back. Sentence2: The wind backed.	True	X
Original 2	Word: back Sentence1: The wind backed. Sentence2: The train backed into the station.	True	
New	Word: back Sentence1: The horse refuses to back. Sentence2: The train backed into the station.	False	
...			
Testing **Additive Consistency** on the **SA** task.	Labels: negative, positive		
Original 1	Unflinchingly bleak and desperate.	negative	X
Original 2	A sometimes tedious flim.	negative	
New	Unflinchingly bleak and desperate. A sometimes tedious flim.	positive	
...			

Figure 6: Data examples of transitive and additive consistency evaluation.



Pair 1 (click to expand/collapse)

text 1: The church has cracks in the ceiling.

text 2: The ceiling in the church is cracked.

Please check the similarity of the two sentences:

- ● Highly similar
- ○ Similar
- ○ Neutral
- ○ Not similar
- ○ Not similar at all

Pair 2 (click to expand/collapse)

Figure 7: Snapshot of our human annotation UI for annotating semantic consistency evaluation data.

<table>
<tr><td colspan="2"><b>ORIGINAL SAMPLE</b>
Sentence 1: The stupendous power of the Tevatron made possible the 1995 discovery of the top quark - the last of six flavors of quarks predicted by the standard model theory of particle physics.
Sentence 2: The top quark is the last of six flavors of quarks predicted by the standard model theory of particle physics.
<b>ADVERSARIAL SAMPLE</b>
Sentence 1: The stupendous power of the Tevatron made possible the 1995 discovery of the top quark - the top of six flavors of quarks predicted by the standard model theory of particle physics.
Sentence 2: The top quark is the last of six flavors of quarks predicted by the standard model theory of particle physics.</td></tr>
<tr><td>ORIGINAL SAMPLE LABEL
entailment</td><td>ADVERSARIAL SAMPLE LABEL
entailment</td></tr>
<tr><td colspan="2"><b>ORIGINAL SAMPLE</b>
Sentence 1: Rockweed has been harvested commercially in Nova Scotia since the last 1950's and is currently the most important commercial seaweed in Atlantic Canada.
Sentence 2: Marine vegetation is harvested.
<b>ADVERSARIAL SAMPLE</b>
Sentence 1: Rockweed has been introduced commercially in Nova Scotia since the last 1950's and is currently the most important commercial seaweed in Atlantic britain.
Sentence 2: Marine vegetation is harvested.</td></tr>
<tr><td>ORIGINAL SAMPLE LABEL
entailment</td><td>ADVERSARIAL SAMPLE LABEL
entailment</td></tr>
</table>

Table 10: Examples of degenerated adversarial samples of BAE (Garg and Ramakrishnan, 2020) for the RTE dataset. The words that changed in the adversarial samples are underlined in both original and adversarial samples. It is hard to consider that the label of the adversarial samples is the same as the original label.