Consistency Analysis of ChatGPT

Myeongjun Erik Jang[1] **Thomas Lukasiewicz**[2,1]
[1] Department of Computer Science, University of Oxford, UK
[2] Institute of Logic and Computation, Vienna University of Technology, Austria
myeongjun.jang@cs.ox.ac.uk, thomas.lukasiewicz@tuwien.ac.at

Abstract

ChatGPT has gained a huge popularity since its introduction. Its positive aspects have been reported through many media platforms, and some analyses even showed that ChatGPT achieved a decent grade in professional exams, adding extra support to the claim that AI can now assist and even replace humans in industrial fields. Others, however, doubt its reliability and trustworthiness. This paper investigates the trustworthiness of ChatGPT and GPT-4 regarding logically consistent behaviour, focusing specifically on semantic consistency and the properties of negation, symmetric, and transitive consistency. Our findings suggest that while both models appear to show an enhanced language understanding and reasoning ability, they still frequently fall short of generating logically consistent predictions. We also ascertain via experiments that prompt designing, few-shot learning and employing larger large language models (LLMs) are unlikely to be the ultimate solution to resolve the inconsistency issue of LLMs.

1 Introduction

AI systems can be more reliable and trustworthy, provided that they behave similarly to humans (De Visser et al., 2016; Jung et al., 2019). In this regard, ChatGPT, a LLM that simulates human-like conversations (Fares, 2023), is gaining widespread popularity, reaching 100 million users only two months after its launch (Milmo, 2023). In addition to many convenient functions that it provides, ChatGPT has performed astoundingly well on various professional examination cases, including passing the United States Medical Licensing Examination (Kung et al., 2023), achieving passing grades in four real exams at the University of Minnesota Law School (Choi et al., 2023), and providing decent answers to Operation Management exam questions, which is a core MBA course (Terwiesch, 2023). These surprising results make people believe that LLMs can assist humans even in professional areas and greatly influence diverse academic and industrial fields.

Others, however, question ChatGPT's reliability, pointing out its overconfidence in generating factually incorrect information (Skopeliti and Milmo, 2023), the inability to comprehend the complexity of human language (Bogost, 2022), and imperfect mathematical abilities (Frieder et al., 2023). Even though these mistakes may appear insignificant in normal daily tasks, they can provoke crucial concerns in conservative and risk-sensitive domains, such as law, medicine, and finance.

A correct behaviour is a crucial aspect in deciding models' trustworthiness by improving the certification[1] process (Jang et al., 2022a). In this regard, we mainly investigate the trustworthiness of ChatGPT in terms of logically consistent behaviour. By using the BECEL dataset (Jang et al., 2022a), which is designed to ascertain whether language models satisfy various types of consistency, we analyse ChatGPT's ability to generate logically consistent predictions based on the four properties below:

- Semantic equivalence: $f(X) = f(Y)$ if X and Y mean the same.
- Negation property: $f(X) \neq f(\neg X)$.
- Symmetric property: $f(X, Y) = f(Y, X)$.
- Transitive property: $X \rightarrow Y \wedge Y \rightarrow Z$ then $X \rightarrow Z$.

Our findings suggest that, similarly to previous pre-trained language models (PLMs), ChatGPT is also prone to violate logical consistencies. Furthermore, our results ascertain that employing different prompt designs, few-shot learning, and utilising larger LLMs trained with more data, such as GPT-4, does not necessarily lead to noteworthy enhancements in consistency. Our contributions can be briefly summarised as follows:

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

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15970–15985
December 6-10, 2023

1. We analyse the consistency behaviour of ChatGPT by measuring semantic, negation, symmetric, and transitive consistency.
2. We observe that ChatGPT achieves a certain level of improvements in negation and transitive consistency compared to previous PLMs.
3. We ascertain that ChatGPT and GPT-4 generate different predictions on text inputs conveying the same meaning.
4. We confirm that ChatGPT and GPT-4 are both self-contradictory: they violate semantic consistency for paraphrased inputs generated by themselves.
5. We find that ChatGPT easily violates symmetric consistency, being sensitive to the input sentence order for order-invariant tasks.
6. Our experiments indicate that prompt design, few-shot learning, and the training of larger LLMs like GPT-4 are unlikely to be a fundamental and ultimate solution for improving the model's consistency.

2 Related Works

The consistency of language models has been an important topic in natural language processing (NLP) but conducted under various definitions. The idea of *semantic consistency* is the most widely used concept in consistency analysis, meaning that a model should make consistent decisions in semantically equivalent contexts (Elazar et al., 2021). Semantic consistency is an indispensable property that should be satisfied in every textual data and NLP task. Ravichander et al. (2020) observed that PLMs are likely to generate different masked language modelling predictions when an object in queries is replaced with its plural form. Elazar et al. (2021), on the other hand, found that PLMs generate different masked language modelling predictions when given paraphrased queries. Raj et al. (2022) proposed an enhanced framework that facilitates the assessment of semantic consistency of natural language generation (NLG) outputs by introducing agreement functions. Another line of work employed the idea by introducing a consistency regularisation term for training, which penalises the violation of semantic consistency, to train more robust NLP models (Wang and Henao, 2021; Zheng et al., 2021; Kim et al., 2021).

Symmetric consistency is a consistency type based on symmetric inference, defined as $f(x, y) = f(y, x)$. This implies that a model should be input-order invariant for tasks where the symmetric property holds. Regarding the natural language inference (NLI) task, Wang et al. (2019) believed that symmetric consistency applies to data points with "Not Entailment", i.e., "Contradiction" and "Neutral", as a label. They showed that many deep-learning-based NLI models change their predictions when the premise and hypothesis are switched. On the other hand, Li et al. (2019) only considered "contradiction" labels for their analysis and ascertained that NLI models based on BERT (Devlin et al., 2019) are likely to violate symmetric consistency. Kumar and Joshi (2022) performed a symmetric consistency analysis on NLI and semantic textual similarity (STS) tasks in a more conservative manner, arguing that a model should generate not only the same predictions but also the same confidence scores if it is truly input-order invariant. They also observed that PLM-based models violated symmetric consistency and introduced a consistency regularisation term to compensate for the issue.

The fundamental idea lying in *negation consistency* is the logical negation property (p is true $\Leftrightarrow$ $\neg p$ is false (Aina et al., 2018)). Intuitively, the main idea is that a model's prediction should differ for text inputs delivering the opposite meaning. Several studies investigated the negation consistency of BERT and found that the model often generates the same outputs when asked negated and non-negated masked queries, e.g., "Birds can lay [MASK]" and "Birds cannot lay [MASK]" (Kassner and Schütze, 2020; Ettinger, 2020). Hosain et al. (2020) created negated versions of NLI datasets and also observed the violation of negation consistency, suggesting that PLMs lack the understanding of negation expressions. To alleviate the issue, several works adopted data augmentation to train a model with abundant data containing negation expressions (Asai and Hajishirzi, 2020; Hosseini et al., 2021). Jang et al. (2022b) introduced the *meaning-matching* task to enhance PLMs' textual understanding ability and observed performance improvements.

Transitive consistency is a consistency type that can measure the deductive reasoning ability. It is derived from transitive inference, represented as $X \rightarrow Y \wedge Y \rightarrow Z$ then $X \rightarrow Z$ for three predicates X, Y, and Z (Gazes et al., 2012; Asai and Hajishirzi, 2020). In the NLI task, Li et al. (2019) employed the concept to generate four transitive inference

rules. For three sentences P, H, and Z, the rules are defined as:

$$E(P,H) \wedge E(H,Z) \rightarrow E(P,Z), \quad (1)$$
$$E(P,H) \wedge C(H,Z) \rightarrow C(P,Z), \quad (2)$$
$$N(P,H) \wedge E(H,Z) \rightarrow \neg C(P,Z), \quad (3)$$
$$N(P,H) \wedge C(H,Z) \rightarrow \neg E(P,Z), \quad (4)$$

where E, N, and C refer to entailment, neutral, and contradiction. Based on the rules, they collected a new evaluation set to assess the transitive consistency of BERT-based NLI models and showed the inconsistency of the models. Other studies investigated the transitive consistency in question answering (QA) (Asai and Hajishirzi, 2020; Mitchell et al., 2022) and WordNet word senses (Lin and Ng, 2022) and ascertained that PLMs lack the ability to perform transitive inference.

Jang et al. (2022a) proposed a universal definition of the language model's consistency and a taxonomy of various consistency types. They also created a new benchmark dataset that enables the evaluation of multiple types of consistencies on various downstream tasks. They assessed diverse PLMs on the new benchmark and confirmed that, like studies stated above, none of PLMs show consistent behaviour on all test cases. All the aforementioned works investigated the consistency of PLMs that emerged before the advent of LLMs like ChatGPT. To our knowledge, this paper is the first evaluation of LLMs from various consistency viewpoints.

3 Experimental Design

3.1 Evaluation Scope

The BECEL dataset provides 19 test sets for assessing five types of consistency on seven downstream tasks. However, we reduced the scope of our experiments mainly because of the competitive usage of OpenAI's LLM API. Specifically, our experiments do not consider the additive consistency, as most PLMs were highly consistent with the additive consistency (Jang et al., 2022a). Factual consistency, which is a popular subject in NLP, is also excluded in our evaluation, as (1) BECEL does not contain test cases for a factual inconsistency analysis, and (2) the factual inconsistency of LLMs is already gaining large attention from many studies (Tam et al., 2023; Zhao et al., 2023), so we decided to focus on exploring relatively less investigated consistency types. As for downstream tasks, we used the SNLI (Bowman et al., 2015), RTE (Candela-Quinonero et al., 2006), MRPC (Dolan and Brockett, 2005), and WiC (Pilehvar and Camacho-Collados, 2019) datasets. The SST2 and AG-News datasets were not included, as they only contain test cases for evaluating the semantic consistency. Table 1 shows the size of the test sets for each downstream task and consistency type.

	SNLI	RTE	MRPC	WiC
semantic	4,406	248	202	140
negation	2,204	153	290	-
symmetric	3,237	1,241	3,668	5,428
transitive	2,375	-	-	3,162

Table 1: Size of the test sets of consistency evaluation data points of the SNLI, RTE, MRPC, and WiC tasks.

3.2 Consistency Evaluation Method

This section briefly demonstrates the process of consistency evaluation by using the BECEL dataset. The evaluation consists of two steps. First, the predictions of the original test set and its corresponding perturbed test set are generated. Next, the predictions of the two test sets are compared to measure the consistency.

For the four downstream tasks in our evaluation scope, Jang et al. (2022a) collected the perturbed test sets for semantic and negation consistency evaluation by modifying "sentence 2" for the RTE, MRPC, and WiC tasks and "hypothesis" for the SNLI task, i.e., generating paraphrase and the opposite meaning sentences for semantic and negation consistency, respectively. They switched the order of the two input texts for symmetric consistency evaluation and created new instances based on existing data examples for assessing transitive consistency. Figure 1 illustrates the overall process for measuring the transitive consistency on the SNLI task and the three remaining consistency types on the MRPC task.

3.3 Generating Predictions

For test cases of data size above 1K, we sampled 200 data points, keeping a low traffic of the OpenAI API. We conducted zero-shot experiments by using two prompt versions designed by Eleuther AI [2] and Wei et al. (2022) to observe how consistency changes with different prompt design. The

[2] https://github.com/EleutherAI/lm-evaluation-harness

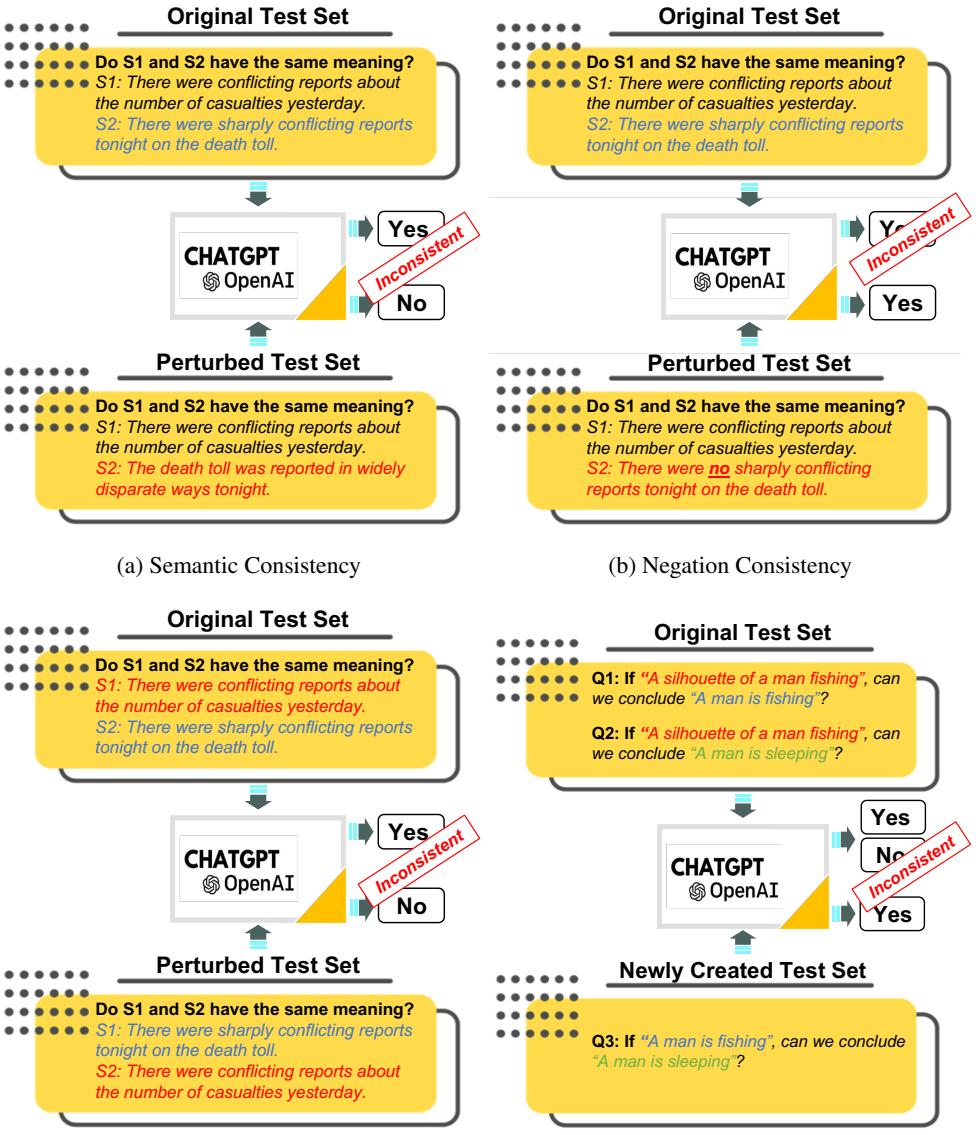

Figure 1: Consistency evaluation process of (a) semantic, (b) negation, and (c) symmetric consistency on MRPC and (d) transitive consistency on SNLI.

prompts of each downstream task and examples are presented in Tables 7 and 8 in Appendix A. Our experiments are conducted with the use of the *24th May* version of the OpenAI API for ChatGPT and the *20th July* version for GPT-4.

Typically, LLMs provided a formatted answer, such as "Yes/No (Equivalent/Not Equivalent)" for MRPC, along with (or without) explanations for the decision. However, we observed a few cases where the output deviated from such structured formats. We manually reviewed and modified such cases to match the desired format.

3.4 Evaluation Metrics

We used the same inconsistency metric as in (Jang et al., 2022a). Specifically, the metric measures the ratio of predictions that violate the target consistency type. Thus, semantic and symmetric inconsistency count the number of predictions where LLMs generate different answers for the original and its corresponding perturbed input. In contrast, negation inconsistency counts the results where the two predictions are the same.

Unlike semantic consistency, which holds unconditionally, negation and symmetric consistencies are conditional properties. For example, negation consistency applies when the label is "Entailment" for the NLI task and "Equivalent" for the STS task.

Regarding symmetric consistency, it applies unconditionally for the STS task, but only to "Not Entailment" for the NLI task. As the BECEL dataset already reflects these conditions, Jang et al. (2022a) calculated the inconsistency metrics using predictions of all test data points, i.e., the condition is determined based on gold labels. However, this can lead to an incorrect estimation of the language model's consistency, as they are not perfectly accurate. For example, consider the below example of the MRPC task:

S1: In the evening, he asked for six pepperoni pizzas and two six-packs of soft drinks, which officers delivered.
S2: In the evening, he asked for six pizzas and soda, which police delivered.
S2-neg: In the evening, he asked for six pizzas and soda, which police did not deliver.

The gold label of the **S1**-**S2** pair is "Equivalent", so predicting the relation between **S1**-**S2-neg** as "Equivalent" is a violation of negation consistency. However, if the model believes that the answer of the **S1**-**S2** pair is "Not Equivalent", then generating "Not Equivalent" as an answer of the **S1**-**S2-neg** pair is hard to be considered as violating negation consistency, because it is the correct answer. Hence, to mitigate the risk of misestimating LLMs' consistency capability, we introduce an additional metric called conditioned inconsistency metric, which only considers data points where LLMs make correct predictions for calculating the metric.

4 Experimental Results

We now present our experimental results on the consistency performance of LLMs. The primary analysis target is the comparison between ChatGPT's performance with Eleuther AI's prompt design and those of fine-tuned PLMs. The BECEL dataset performances of two PLMs (Electra-large (Clark et al., 2020) and T5 (Raffel et al., 2020)) are taken from Jang et al. (2022a). The performance of the prompt design devised by Wei et al. (2022), GPT-4, and few-shot learning (2-shots in our experiments) will be mainly discussed in Section 5.

4.1 Semantic Consistency

It is widely known that ChatGPT can perform various NLP tasks, including summarisation, question answering, and paraphrasing. Therefore, in addition to the original BECEL dataset, we generated paraphrased sentences using ChatGPT and GPT-4

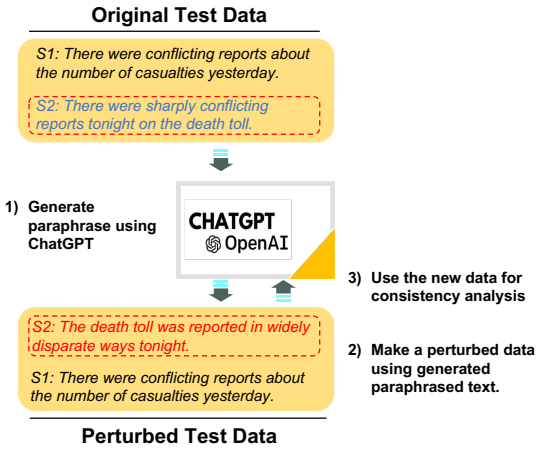

Figure 2: Overall process of measuring semantic consistency by using paraphrases generated by ChatGPT.

and used them for evaluation. For the WiC task, paraphrased instances were removed if they did not contain the target word. The overall procedure of this evaluation is illustrated in Figure 2.

The results are summarised in the second column of Table 2. In the SNLI task using Eleuther AI's prompt, LLMs often fail to distinguish "Neutral" and "Contradiction" classes, potentially leading to an underestimation of consistency. To address this, we integrate the two classes into a unified "Not Entailment" class specifically for this scenario. Our experimental results show that ChatGPT produces much higher levels of inconsistency in the BECEL dataset than fine-tuned PLMs, suggesting ChatGPT's limited capacity for making logically consistent predictions. Moreover, we ascertain that ChatGPT is self-contradictory, i.e., it produces inconsistent outputs for self-generated paraphrase sentences with a probability exceeding 10%. This implies that ChatGPT failed to generate a proper paraphrased sentence or to capture the meaning of texts delivering the same meaning; either case undermines its reliability. Several examples where ChatGPT violates semantic consistency are presented in Table 3. More examples are available in Table 9 in Appendix B.

4.2 Negation Consistency

The third column of Table 2 presents the experimental results of the negation consistency evaluation. Compared to the fine-tuned PLMs, ChatGPT attains a lower negation inconsistency in general. A large improvement has been made in the RTE task and MRPC task compared to the Electra-large model. Also, the conditional inconsistency is exten-

Model	Semantic								Negation					
	MRPC		RTE		SNLI		WiC		MRPC		RTE		SNLI	
	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S	τ	τ_C	τ	τ_C	τ	τ_C
ChatGPT (EAI)	9.4	14.4	11.3	13.7	20.0	20.5	14.3	18.4	26.0	7.0	11.1	2.1	13.0	1.3
ChatGPT (Wei)	10.9	11.4	15.3	17.3	16.5	16.5	14.3	17.2	**23.0**	6.7	15.0	5.2	11.0	3.4
ChatGPT (EAI, 2s)	17.2	13.8	10.5	10.5	17.5	14.0	30.3	34.5	42.0	4.31	15.7	13.7	23.0	20.0
GPT-4 (EAI)	16.3	13.4	12.5	12.5	19.0	22.0	12.9	13.8	33.5	7.6	**7.8**	6.8	8.5	1.7
GPT-4 (Wei)	11.9	10.4	9.7	10.5	23.5	25.5	6.4	7.3	26.5	9.8	12.4	6.3	**4.5**	0.5
Electra-large	5.5	-	8.9	-	**7.9**	-	8.9	-	77.0	-	17.3	-	5.4	-
T5-large	**4.5**	-	**8.6**	-	9.3	-	**8.6**	-	25.2	-	15.9	-	5.8	-

Table 2: Experimental results of the semantic and negation consistency evaluation. "2s" refers to two-shot learning. τ_B and τ_S denote the inconsistency of the BECEL dataset and paraphrases generated by LLMs, respectively. τ and τ_C refer to the original and conditioned negation inconsistency, respectively. The best performance is in bold.

Original Inputs	Perturbed Inputs
Task: SNLI, Paraphrase Type: BECEL	
Premise: Kids play in the water in the middle of the street. Hypothesis: Kids are running from zombies. Prediction: Not Entailment (Contradiction)	Premise: Kids play in the water in the middle of the street. Hypothesis: Children are fleeing from zombies. Prediction: Entailment
Task: MRPC, Paraphrase Type: ChatGPT	
S1: Looking to buy the latest Harry Potter ? S2: Harry Potter 's latest wizard trick ? Prediction: Not Equivalent	S1: Looking to buy the latest Harry Potter ? S2: The newest magical feat of Harry Potter ? Prediction: Equivalent

Table 3: Examples of semantic consistency violation.

sively small compared to the original inconsistency metric, attaining an average of 4.6% and almost perfectly consistent on the SNLI task. The difference with the original inconsistency metric substantiates the introduction of the conservative evaluation metric, i.e., conditional inconsistency, aimed at more precise evaluations. The experimental results suggest that ChatGPT can better understand negation expressions and antonyms, which has been a critical issue for PLMs trained in a self-supervised fashion (Kassner and Schütze, 2020; Ettinger, 2020; Hossain et al., 2020; Hosseini et al., 2021; Jang et al., 2022b). We believe that incorporating human feedback into ChatGPT training (Ouyang et al., 2022) plays a crucial role in learning the meaning of negation expressions and antonyms, compared to previous PLMs that infer their meaning based on the distributional hypothesis by simply relying on the context information. Investigating the impact of providing human feedback on learning textual meaning is an interesting future research direction. Table 5 presents an example of negation consistency violation. More examples can be found in Table 10 in Appendix B.

4.3 Symmetric Consistency

The results of the symmetric consistency evaluation are shown in the second column of Table 4. Compared to the best-performing PLM, ChatGPT produces three times higher symmetric inconsistency in the MRPC task and five times higher in the RTE task. The conditioned inconsistency was lower than the original inconsistency in the RTE and SNLI tasks, and higher in the WiC task. Although the inconsistency rate for the SNLI task might be considered trivial, it should not be overlooked, considering the simple nature of the symmetric property. Consider a medical-domain model that takes a list of symptoms and generates prescriptions. For such a model, which should operate conservatively, its trustworthiness would be significantly undermined if it were to generate entirely different prescriptions whenever the order of symptoms changes, even if the probability of such occurrence is exceedingly low. Hence, an effort should be made to make LLMs satisfy logical consistencies to enhance their reliability and safe usage in real-world applications. Table 5 presents an example of symmetric consistency violations. More examples are presented in Table 10 in Appendix B.

4.4 Transitive Consistency

The third column in Table 4 presents the transitive consistency evaluation results. In contrast to other consistency types where ChatGPT performed similar or worse than fine-tuned PLMs, it shows better

Model	Symmetric								Transitive	
	MRPC		RTE		SNLI		WiC		SNLI	WiC
	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ
ChatGPT (EAI)	13.0	-	45.0	19.7	5.0	1.4	20.0	25.8	**2.1**	**12.3**
ChatGPT (Wei)	14.5	-	45.5	33.3	6.5	3.6	23.5	24.2	2.6	14.9
ChatGPT (EAI, 2s)	31.0	-	55.0	37.2	14.0	2.3	39.5	43.0	5.8	28.9
GPT-4 (EAI)	11.0	-	54.5	20.7	2.0	1.0	6.5	4.1	2.3	3.6
GPT-4 (Wei)	9.0	-	17.0	6.9	**2.5**	2.0	9.5	8.1	2.9	**4.3**
Electra-large	5.3	-	**6.7**	-	6.4	-	7.9	-	2.5	46.5
T5-large	**4.2**	-	8.0	-	8.3	-	**6.3**	-	2.9	45.3

Table 4: Experimental results of the symmetric and transitive consistency evaluation. "2s" refers to two-shot learning. τ and τ_C denote the original and conditioned symmetric inconsistency, respectively. The best performance is in bold.

Original Inputs	Perturbed Inputs
Task: MRPC, Consistency Type: Negation	
S1: He arrives later this week on the first state visit by a US President . S2: Mr Bush arrives on Tuesday on the first state visit by an American President . Prediction: Equivalent	S1: He arrives later this week on the first state visit by a US President . S2: Mr Bush doesn't arrive on Tuesday on the first state visit by an American President. Prediction: Equivalent
Task: SNLI, Consistency Type: Symmetric	
Premise: There is a man climbing as the boy holds the rope Hypothesis: A man holds a rope for a boy who's about to climb a wall. Prediction: Not Entailment (Contradiction)	Premise: A man holds a rope for a boy who's about to climb a wall. Hypothesis: There is a man climbing as the boy holds the rope Prediction: Entailment
Task: WiC, Consistency Type: Transitive	
Original Inputs	
Sentence1: You must carry your camping gear. Sentence2: Sound carries well over water. Word: carry Prediction: Not Equivalent	Sentence1: The airwaves carry the sound. Sentence2: Sound carries well over water. Word: carry Prediction: Equivalent
Newly Created Inputs	
Sentence1: The airwaves carry the sound. Sentence2: You must carry your camping gear. Word: carry Prediction: Equivalent	

Table 5: Examples of negation, symmetric, and transitive consistency violations.

results than fine-tuned PLMs, especially with notable improvements in the WiC dataset. It is very interesting that ChatGPT is better at higher-level logical reasoning like transitive inference, but fails in simpler logical properties, such as the symmetric property. The results indicate that combining the advantages of fine-tuned PLMs and LLMs can pave the way for developing more logically consistent language models. An example that violates transitive consistency is presented in Table 5. More examples are listed in Table 11 in Appendix B.

4.5 ChatGPT's Explainablity

Providing explanations is a core property of trustworthy systems (Huang et al., 2020). It is widely known that generative language models can provide answers with explanations. However, recent studies confirmed that explanations generated by early-stage LLMs like GPT-3 are often logically inconsistent and unreliable (Jung et al., 2022; Wang et al., 2023). We confirmed that the identical problem persists in ChatGPT. We observed several cases that the generated explanations seem plausible but contradictory to other generated explanations with identical context (Camburu et al., 2020), or not faithful. Table 6 presents some examples. For the first example, the explanations of the original and perturbed inputs contradict each other. In the second example, the explanation of the perturbed input is not correct, i.e., the input did mention the age and gender of the person pushing the shopping cart ("boy" and "A young man"). It would be interesting for future work to quantitatively evaluate the quality of explanations provided by LLMs by employing inconsistent explanation evaluation frameworks (Camburu et al., 2020; Jang et al., 2023).

TASK: MRPC, CONSISTENCY TYPE: Symmetric
ORIGINAL QUESTION: "Sentence 1: In the evening, he asked for six pepperoni pizzas and two six-packs of soft drinks, which officers delivered. Sentence 2: In the evening, he asked for six pizzas and soda , which police delivered. Question: Do both sentences mean the same thing? Answer:" ORIGINAL OUTPUT: "Yes, both sentences mean the same thing. They convey that in the evening, the person made a request for six pepperoni pizzas and two six-packs of soft drinks, and that the officers delivered the items. The second sentence is a simplified version of the first one, where 'pizzas' refers to pepperoni pizzas, and 'soda' refers to soft drinks."
PERTURBED QUESTION: "Sentence 1: In the evening, he asked for six pizzas and soda , which police delivered. Sentence 2: In the evening, he asked for six pepperoni pizzas and two six-packs of soft drinks, which officers delivered. Question: Do both sentences mean the same thing? Answer:" PERTURBED OUTPUT: "No, the two sentences do not mean the same thing. The first sentence says that he asked for six pizzas and soda, while the second sentence says that he asked for six pepperoni pizzas and two six-packs of soft drinks. The type of pizza and the specific drinks requested are different in the two sentences."
TASK: SNLI, Semantic-BECEL
ORIGINAL QUESTION: "An enthusiastic boy pushes a shopping card containing a smiling boy. Question: A boy is pushing a shopping cart. True, False or Neither? Answer:" ORIGINAL OUTPUT: "True"
PERTURBED QUESTION: "An enthusiastic boy pushes a shopping card containing a smiling boy. Question: A young man is pushing a shopping cart. True, False or Neither? Answer:" PERTURBED OUTPUT: "Neither (the description does not mention the age or gender of the person pushing the shopping cart, only that the person inside the cart is a smiling boy.)"

Table 6: Examples of ChatGPT's output with explanations.

5 Discussion

Can Prompt Design be a Solution? Prompts are input text consisting of a task demonstration and, for a few-shot task, some examples (Lester et al., 2021). Prompt design has been shown to be an effective method of regulating the behaviour of GPT-3 (Brown et al., 2020). Hence, one might argue that searching for an optimal prompt for each task can improve consistency. However, our experimental results provide sceptical evidence for this claim. When comparing the performance of ChatGPT using Eleuther AI's prompts and those designed by Wei et al. (2022), there was no statistically significant difference in performance across all consistency types and downstream tasks at a confidence level of 0.05. In the case of GPT-4, similar results were observed, where a statistically significant difference at a confidence level of 0.05 was found only in symmetric consistency in the RTE task. We believe that a primary cause contributing to this phenomenon is an inherent characteristic of machine learning: inductive reasoning. The underlying idea behind prompt design is that prompts created by experimenters are not optimal, because language models might have acquired target information from completely different contexts (Jiang et al., 2020). That is, it can resolve the matter of inconsistency if and only if numerous consistency properties are reflected in ChatGPT's inductive bias, which is technically not feasible. Moreover, consistency improvements with prompt design can be considered another violation of semantic consistency, because the prompts will deliver identical semantic meaning, i.e., task description.

Can Few-shot Learning be a Solution? It is widely known that providing few-shot examples generally leads to higher accuracy compared to the zero-shot setting. To ascertain whether this principle extends to consistency, we conducted additional few-shot experiments on ChatGPT by providing two-shot examples and using Eleuther AI's prompt. However, the findings suggest that providing few-shot examples is not beneficial to improving consistency. When compared with the result of ChatGPT employing the same prompt design but under a zero-shot setting, we observe statistically significant increases in inconsistencies across the majority of test cases. The most substantial rise in inconsistency values is evident in the context of symmetric consistency and the WiC task. We even identified several instances where the model altered its decision when the order of two-shot examples was switched. These experimental outcomes indicate that employing a few-shot approach does not represent a definitive solution for enhancing the consistency of LLMs.

Can Increasing Data and Model Size be a Solution? Another possible way to improve consistency is training a larger LLMs with more abundant training data, as larger models generally outperform

smaller ones in numerous NLPs downstream tasks. However, our experimental results reveal the limitations of this approach. Through a comprehensive comparison of ChatGPT and GPT-4 performance employing identical prompt designs, it was ascertained that GPT-4 does not necessarily exhibit superior performance in comparison to ChatGPT from a consistency perspective. While GPT-4 did demonstrate enhanced consistency in several test scenarios, including symmetric and transitive consistency in the WiC task across both prompt designs, symmetric consistency in the RTE task, and negation consistency in the SNLI task using prompts devised by Wei et al. (2022), no discernible improvements were observed in the remaining test cases, which constitutes a portion of 77% of among the total test scenarios. In addition, GPT-4 also exhibited a high level of self-contradiction, just like ChatGPT.

Moreover, increasing the size of the data and model is a technically unsustainable strategy. First, the data collection procedure requires tremendous effort and is challenging to cover all possible variations, especially for consistency types that demand abundant linguistic resources, such as semantic consistency. Second, even if we successfully expand the data, it is doubtful whether we can afford to update an LLM with each new dataset iteration. Considering the ever-changing information, the data expansion and update of an LLM should be performed continuously, as neglecting to do so may raise the concern of outdated information (Zhuo et al., 2023; Wen and Wang, 2023). However, training an LLM entails tremendous financial and environmental costs (Bender et al., 2021). For instance, training a BERT-base model without hyperparameter tuning requires a CO2 emission of 650kg, which is comparable to flying from New York to San Francisco for one passenger (Strubell et al., 2019). A simple expectation of CO2 emission for re-training ChatGPT and GPT-4 would amount to 1,033t and 1,0240t, respectively,[3] while a human is responsible for 5t CO2 emission per year. The continuous emission of such a substantial volume of greenhouse gases would have a detrimental impact on the environment of modern society facing the global climate crisis.

[3]ChatGPT and GPT-4 are approximately 1,590 and 16,000 times larger than BERT-base.

6 Summary and Outlook

The advent of ChatGPT is accelerating the developments in the NLP field driven by LLMs. Its outstanding performance captured considerable attention, resulting in many articles, posts, and analyses highlighting ChatGPT's positive aspects across numerous media. There are others, however, who question its reliability based on the model's faulty behaviours. To this end, this study aims to examine the trustworthiness of ChatGPT in terms of the language model's consistency.

In this paper, we have investigated the consistency behaviour of ChatGPT across four consistency types and downstream tasks. Our experimental results demonstrate that ChatGPT achieves a certain level of enhanced language understanding ability, especially in negation expressions and antonyms. It also exhibits improved deductive reasoning ability with lower transitive inconsistencies compared to the earlier version of PLMs. However, while ChatGPT exhibits enhanced negation and transitive consistency, this does not mean that the model is perfectly consistent, i.e., it still makes mistakes that violate the logical properties, and the frequency of such occurrences is non-negligible. Also, contrary to the widespread belief regarding the outstanding performance of ChatGPT, its performance across various consistency types falls short of expectations. It frequently changes its decision when an input text is replaced with a paraphrased sentence, even though it is generated from ChatGPT itself. Moreover, in input-order invariant tasks, ChatGPT is prone to make a different decision when the order of the input sentences is switched. These issues are also observed in GPT-4 or with the use of different prompt designs and few-shot learning, indicating that these approaches are unlikely to be a fundamental remedy. Given how simple and natural the symmetric and semantic consistencies are in human reasoning, violating these consistencies can be a huge blow to LLMs' trustworthiness. These fallacious behaviours are especially lethal to domains operating conservatively and at high risk. Although LLMs are a revolutionary technique that brought an unprecedented era to NLP, such issues should be resolved before these models are applied in real applications, particularly considering the huge economic and environmental costs consumed for developing LLMs.

Limitations

Limitations of our work include that the data instances of several test cases (e.g., SNLI dataset and symmetric evaluation cases) are sampled due to the heavy usage of ChatGPT. Conducting experiments on whole data points will provide a more precise comparison with baseline models.

We were unable to study the model's consistency on longer documents, e.g., document-level NLI task, because there are no publicly available datasets for evaluating consistency on long documents. We leave this as future work, as a consistency analysis of long documents would be very appropriate for ChatGPT studies.

We performed a qualitative evaluation of explanations generated by ChatGPT, but a quantitative analysis was omitted due to the resources and time required for the evaluations. We leave adopting several explanation quality evaluation frameworks (Camburu et al., 2020; Jang et al., 2023) to ChatGPT's explanations as future work.

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.

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, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Ian Bogost. 2022. ChatGPT is dumber than you think. *The Atlantic*.

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.

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.

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.

Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. 2023. ChatGPT goes to law school. *Minnesota Legal Studies Research Paper*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 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 (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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.

Omar H. Fares. 2023. ChatGPT could be a game-changer for marketers, but it won't replace humans any time soon. *The Conversation*.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of ChatGPT.

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.

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.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2022a. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Myeongjun Jang, Bodhisattwa Prasad Majumder, Julian McAuley, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. KNOW how to make up your mind! adversarially detecting and alleviating inconsistencies in natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 540–553, Toronto, Canada. Association for Computational Linguistics.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022b. 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.

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

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.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. 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.

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.

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.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño,

Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. 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.

Dan Milmo. 2023. ChatGPT reaches 100 million users two months after launch. *The Guardian*.

Eric Mitchell, Joseph J. Noh, Siyan Li, William S. Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D. Manning. 2022. Enhancing self-consistency and performance of pretrained language models with NLI. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

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.

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.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. In *NeurIPS 2022 Workshop on Machine Learning Safety*.

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.

Clea Skopeliti and Dan Milmo. 2023. 'ChatGPT needs a huge amount of editing': users' views mixed on AI chatbot. *The Guardian*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Christian Terwiesch. 2023. Would ChatGPT get a Wharton MBA? A prediction based on its performance in the operations management course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania. Retrieved from: https://mackinstitute. wharton. upenn. edu/wpcontent/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24. pdf [Date accessed: February 6th, 2023].*

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.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jun Wen and Wei Wang. 2023. The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. *Clinical and Translational Medicine*, 13(3):e1207–e1207.

Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. 2023. Can ChatGPT-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*.

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.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity.

A Prompt Design

SNLI	
Format	{*s1*} Question: {s2} True, False or Neither? Answer:
Example	A land rover is being driven across a river. Question: A Land Rover is splashing water as it crosses a river. True, False, or Neither? Answer:
RTE	
Format	{*s1*} Question: {s2} True or False? Answer:
Example	The harvest of sea-weeds is not allowed in the Puget Sound because of marine vegetation's vital role in providing habitat to important species. Question: Marine vegetation is harvested. True or False? Answer:
MRPC	
Format	Sentence 1: {*s1*} Sentence 2: {s2} Question: Do both sentences mean the same thing? Answer:
Example	Sentence 1: The increase reflects lower credit losses and favorable interest rates. Sentence 2: The gain came as a result of fewer credit losses and lower interest rates. Question: Do both sentences mean the same thing? Answer:
WiC	
Format	Sentence 1: {*s1*} Sentence 2: {s2} Question: Is the word {*word*} used in the same way in the two sentences above? Answer:
Example	Sentence 1: It was the deliberation of his act that was insulting. Sentence 2: It was the deliberation of his act that was insulting. The deliberations of the jury. Question: Is the word deliberation used in the same way in the two sentences above? Answer:

Table 7: Eleuther AI prompt formats and their example used in our experiments for each downstream task.

SNLI	
Format	If {*s1*}, can we conclude {s2}? \n Yes, It's impossible to say, No \n Answer:
Example	If A land rover is being driven across a river, can we conclude A Land Rover is splashing water as it crosses a river? \n Yes, It's impossible to say, No \n Answer:
RTE	
Format	{*s1*} \n Based on the paragraph above can we conclude that {s2}? Yes, No \n Answer:
Example	The harvest of sea-weeds is not allowed in the Puget Sound because of marine vegetation's vital role in providing habitat to important species. \n Based on the paragraph above can we conclude that Marine vegetation is harvested? \n Yes, No \n Answer:
MRPC	
Format	Here are two sentences: \n {*s1*} \n {s2} \n Do they have the same meaning? \n Yes, No \n Answer:
Example	Here are two sentences: \n The increase reflects lower credit losses and favorable interest rates. \n The gain came as a result of fewer credit losses and lower interest rates. \n Do they have the same meaning? \n Yes, No \n Answer:
WiC	
Format	In these two sentences (1) {*s1*} (2) {s2} , does the word {*word*} mean the same thing? \n Yes, No \n Answer:
Example	In these two sentences (1) It was the deliberation of his act that was insulting. (2) It was the deliberation of his act that was insulting, does the word deliberation mean the same thing? \n Yes, No \n Answer:

Table 8: Prompt formats designed by Wei et al. (2022) and their examples used in our experiments for each downstream task.

B Examples

Original Inputs	Perturbed Inputs
Task: RTE, Paraphrase Type: BECEL	
S1: Note that SBB, CFF and FFS stand out for the main railway company, in German, French and Italian. S2: The French railway company is called SNCF. Prediction: Not Entailment	S1: Note that SBB, CFF and FFS stand out for the main railway company, in German, French and Italian. S2: SNCF is the French railway company. Prediction: Entailment
Task: SNLI, Paraphrase Type: ChatGPT	
Premise: A person swimming in a swimming pool. Hypothesis: A person enjoying the waters. Prediction: Not Entailment (Neutral)	Premise: A person swimming in a swimming pool. Hypothesis: An individual is relishing the water. Prediction: Entailment
Task: WiC, Paraphrase Type: BECEL	
Word: glaze S1: Glaze the bread with eggwhite. S2: The potter glazed the dishes. Prediction: Equivalent	Word: glaze S1: Glaze the bread with eggwhite. S2: The dishes were glazed by the potter. Prediction: Not Equivalent

Table 9: More examples of semantic consistency violation.

Original Inputs	Perturbed Inputs
Task: MRPC, Consistency Type: Negation	
S1: The dead cavalry have been honored for more than a century with a hilltop granite obelisk and white headstones . S2: The dead cavalrymen are honored with a hilltop granite obelisk and white headstones . Prediction: Equivalent	S2: The dead cavalry have been honored for more than a century with a hilltop granite obelisk and white headstones . S2: The dead cavalrymen are honored with a hilltop granite obelisk and black headstones. Prediction: Equivalent
Task: SNLI, Consistency Type: Negation	
Premise: A young man wearing goggles is putting some liquid in a beaker while a young girl wearing blue gloves looks down while holding a pen. Hypothesis: Two people are in the same area. Prediction: Entailment	Premise: A young man wearing goggles is putting some liquid in a beaker while a young girl wearing blue gloves looks down while holding a pen. Hypothesis: Two people are not in the same area. Prediction: Entailment
Task: MRPC, Consistency Type: Symmetric	
S1: In 2001 , the diocese reached a $ 15 million settlement involving five priests and 26 plaintiffs . S2: The diocese reached a settlement in 2001 involving five priests and 26 plaintiffs for an undisclosed sum . Prediction: Not Equivalent	S1: The diocese reached a settlement in 2001 involving five priests and 26 plaintiffs for an undisclosed sum . S2: In 2001 , the diocese reached a $ 15 million settlement involving five priests and 26 plaintiffs . Prediction: Equivalent
Task: WiC, Consistency Type: Symmetric	
Word: master S1: One of the old masters. S2: A master of the violin. Prediction: Equivalent	Word: master S1: A master of the violin. S2: One of the old masters. Prediction: Not Equivalent

Table 10: More examples of negation and symmetric consistency violations.

TASK: WiC, CONSISTENCY TYPE: Transitive	
ORIGINAL INPUTS	
SENTENCE1: Strike a medal.	SENTENCE1: Strike coins.
SENTENCE2: Strike coins.	SENTENCE2: A bullet struck him.
WORD: strike	WORD: strike
PREDICTION: Equivalent	PREDICTION: Not Equivalent
NEWLY CREATED INPUTS	
SENTENCE1: Strike a medal SENTENCE2: A bullet struck him. WORD: strike	
PREDICTION: Equivalent	
TASK: SNLI, CONSISTENCY TYPE: Transitive	
ORIGINAL INPUTS	
PREMISE: A performance group is staged in one collective motion.	PREMISE: A performance group is staged in one collective motion.
HYPOTHESIS: The group is separate from one another.	HYPOTHESIS: There's a performance group doing something together.
PREDICTION: Contradiction	PREDICTION: Entailment
NEWLY CREATED INPUTS	
PREMISE: The group is separate from one another. HYPOTHESIS: There's a performance group doing something together.	
PREDICTION: Entailment	

Table 11: More examples of transitive consistency violations.