Consistency Analysis of ChatGPT

Anonymous ACL submission

Abstract

ChatGPT, a question-and-answer dialogue system based on a large language model, has gained 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, including the law, medical, and finance domains, adding extra support to the claim that AI now can assist and, even, replace humans in industrial fields. Others, however, doubt its reliability and trustworthiness. In this paper, we investigate ChatGPT's trustworthiness regarding logically consistent behaviours. Our findings suggest that, although ChatGPT seems to achieve an improved language understanding ability, it still fails to generate logically correct predictions frequently. Hence, while it is true that ChatGPT is an impressive and promising new technique, we conclude that its usage in real-world applications without thorough human inspection requires further consideration, especially for risk-sensitive areas.

1 Introduction

AI systems can be more reliable and trustworthy provided they behave in a similar manner to humans (De Visser et al., 2016; Jung et al., 2019). In this regard, ChatGPT, a large language model (LLM) that simulates human-like conversations (Fares, 2023), is gaining widespread popularity, reaching 100 million users only two months after its launch (Milmo, 2023). It offers many convenient features to users, such as summarising documents, writing essays, answering questions, and programming computer languages. Also, ChatGPT has performed astoundingly well on various 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 various academic and industrial fields.

Others, however, question ChatGPT's reliability, pointing out its overconfidence in generating factually incorrect information (Skopeliti and Milmo, 2023), 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, e.g., drafting an email, they provoke crucial concerns in conservative and risk-sensitive domains, such as law, medicine, and finance.

In this article, we investigate the reliability and trustworthiness of ChatGPT in terms of the language model's consistency. 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 three properties: semantic equivalence, logical negation, and symmetricity. Our experimental results show that although ChatGPT understands negation expressions and antonyms much better than previous pre-trained language models (PLMs) like BERT (Devlin et al., 2019), it still violates semantic equivalence and symmetricity quite frequently. Our contributions can be briefly summarised as follows:

1. We analyse the consistency behaviour of ChatGPT by measuring semantic, negation, and symmetric consistency.
2. We observe that ChatGPT achieves a much lower negation inconsistency compared to other PLMs, proving its improved understanding of negation expressions and antonyms.
3. We ascertain that ChatGPT is likely to generate different predictions on text inputs that

deliver the same meaning, i.e., paraphrased inputs.

4. Even worse, we confirm that ChatGPT is self-contradictory, meaning that it violates semantic consistency for paraphrased inputs generated by ChatGPT itself.
5. We find that ChatGPT is extremely sensitive to the input sentence order for order-invariant tasks, e.g., semantic textual similarity (STS).

Hence, we conclude that despite its favourable reputation and positive media coverage, ChatGPT is not completely reliable, suggesting that using ChatGPT without human confirmation would be hazardous, particularly in a highly risky industry.

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. 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 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 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 behind it 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). Hossain 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) expanded the evaluation scope from negation expressions to antonyms and ascertained the same tendency in recent PLMs. They proposed a new training task named *meaning-matching* 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, respectively. Based on the rules,

	SNLI	RTE	MRPC
semantic	4,406	248	202
negation	2,204	153	290
symmetric	3,237	1,241	3,668

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

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 a consistency viewpoint.

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 extremely competitive usage of ChatGPT. Specifically, our experiments do not consider additive and transitive consistency, because most PLMs were highly consistent on the former (Jang et al., 2022a), and the latter requires much more difficult reasoning ability compared to other consistency types. Regarding downstream tasks, we used the SNLI (Bowman et al., 2015), RTE (Candela-Quinonero et al., 2006), and MRPC (Dolan and Brockett, 2005) datasets, which contain test cases for measuring semantic, negation, and symmetric consistency. Table 1 shows the size of the test sets for each downstream task and consistency type.

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 three 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 and MRPC 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. Figure 1 illustrates the overall process for measuring the three consistency types on the MRPC task.

3.3 Generating ChatGPT Predictions

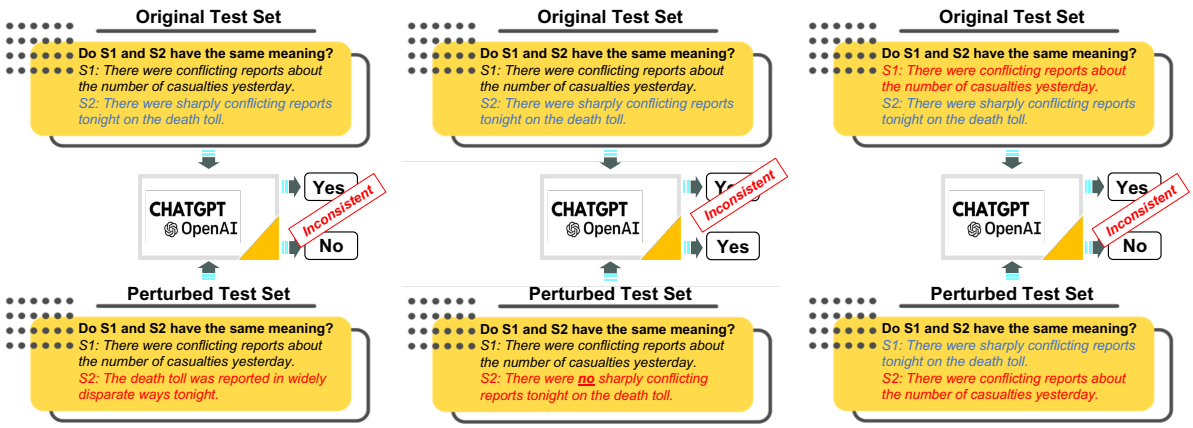
For test cases where the size of data exceeds 1K, e.g., SNLI task and symmetric consistency of RTE and MRPC, we sampled 200 data points due to the heavy usage of ChatGPT. We conducted zero-shot experiments by using the same prompts designed by Eleuther AI [1]. The prompts of each downstream task and examples are presented in Table 2. Our experiments are conducted on the *30 Jan* version of ChatGPT by using the ***pyChatGPT*** package, an unofficial Python wrapper for OpenAI's ChatGPT API [2].

Normally, ChatGPT gave us an answer "True/False/Neither (Entailment/Contradiction/Neutral)" for SNLI, "Yes/No (Equivalent/Not Equivalent)" for MRPC, and "True/False (Entailment/Not Entailment)" along with (or without) explanations for the decision. However, we observed a few cases where the output does not follow the aforementioned format but just gives an explanation. We reviewed such cases and gave the correct answers.

3.4 Evaluation Metrics

Basically, 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

[1] https://github.com/EleutherAI/lm-evaluation-harness
[2] https://github.com/terry3041/pyChatGPT



(a) Semantic Consistency (b) Negation Consistency (c) Symmetric Consistency

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

	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:

Table 2: Format and example of prompts used in our experiments for each downstream task.

where ChatGPT generates 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 gold label is "Entailment" for the NLI task and "Equivalent" for the STS task. Regarding symmetric consistency, it applies unconditionally for the STS task and only to "Not Entailment" for the NLI task. As the BECEL dataset already reflects these conditions, Jang et al. (2022a) calculated the inconsistency metrics based on all test data points. However, it can exaggerate the inconsistency of language models if their performance is insufficient. 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". However, if the model believes that the answer 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. We observed that the zero-shot accuracy of ChatGPT is much lower than that of fine-tuned PLMs reported by Jang et al. (2022a). Therefore, we introduce a conditioned inconsistency metric, which only uses data points where ChatGPT makes correct predictions.

4 Experimental Results

We now present our experimental results on the performance of ChatGPT and do a comparison with fine-tuned PLMs. The BECEL dataset perfor-

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S	τ_B	τ_S
BERT-large	12.5	-	12.3	-	9.9	-	-	-
RoBERTa-large	8.4	-	9.8	-	**7.9**	-	-	-
Electra-large	5.5	-	8.9	-	**7.9**	-	-	-
T5-large	**4.5**	-	**8.6**	-	9.3	-	-	-
ChatGPT	29.7	9.9	11.3	10.5	28.0	21.0	15.0	11.0

Table 3: Experimental results of semantic consistency evaluation. τ_B and τ_S denote the inconsistency of the BECEL dataset and paraphrases generated by ChatGPT, respectively. The best performance is formatted in bold.

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ_C
BERT-large	90.8	-	75.8	-	11.7	-	-	-
RoBERTa-large	84.2	-	24.6	-	5.9	-	-	-
Electra-large	77.0	-	17.3	-	5.4	-	-	-
T5-large	25.2	-	15.9	-	5.8	-	-	-
ChatGPT	**21.3**	4.6	**10.5**	6.9	**5.0**	0.0	9.0	0.0

Table 4: Experimental results of the negation consistency evaluation. τ and τ_C denote the original and conditioned negation inconsistency, respectively. The best performance is in bold.

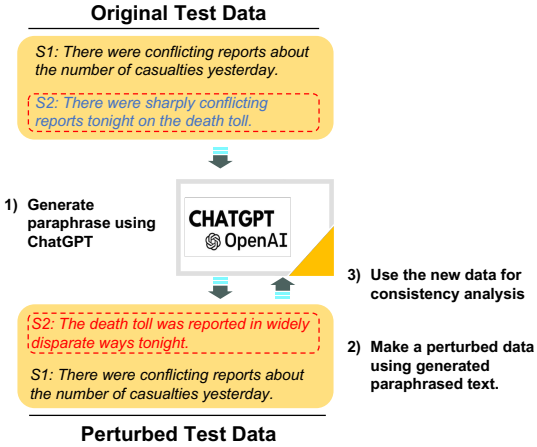


Figure 2: Overall process of measuring semantic consistency by using paraphrases generated by ChatGPT.

mances of four PLMs (BERT-large (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), Electra-large (Clark et al., 2020), and T5 (Raffel et al., 2020)) are taken from Jang et al. (2022a).

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 used them for evaluation. The overall procedure of this evaluation is illustrated in Figure 2.

The results are summarised in Table 3. In the SNLI task, ChatGPT often fails to distinguish "Neutral" and "Contradiction", which can underestimate its consistency. Therefore, we integrate the two classes into the "Not Entailment" class, which is noted as **SNLI-2C** in the table. Our experimental results show that ChatGPT produces much higher levels of inconsistency in the BECEL dataset than fine-tuned PLMs, yielding 3.8 times higher inconsistency on average than the best-performing fine-tuned PLM. This suggests that ChatGPT is not completely trustworthy regarding semantic consistency. Moreover, we ascertain that ChatGPT is self-contradictory, i.e., it even produces inconsistent outputs for paraphrased inputs generated by itself with a probability of more than 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 5.

4.2 Negation Consistency

Table 4 presents the experimental results of the negation consistency evaluation. Compared to the fine-tuned PLMs, ChatGPT attains lower negation inconsistency in all three downstream tasks, improved by 19% on average than the best-performing fine-tuned PLM, and considerably outperformed BERT-large model. In addition, the conditional inconsistency is 3.8% on average and perfectly consistent on the SNLI task. The 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 by simply relying on the context information based on the distributional hypothesis. Investigating the impact of providing human feedback on learning textual meaning is an interesting future research direction. Several examples of negation consistency violation are presented in Table 7.

TASK: RTE, PARAPHRASE TYPE: BECEL	
ORIGINAL INPUTS	PERTURBED INPUTS
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: BECEL	
ORIGINAL INPUTS	PERTURBED INPUTS
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	
ORIGINAL INPUTS	PERTURBED INPUTS
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
TASK: SNLI, PARAPHRASE TYPE: ChatGPT	
ORIGINAL INPUTS	PERTURBED INPUTS
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

Table 5: Examples of semantic consistency violation.

4.3 Symmetric Consistency

The results of symmetric consistency evaluation are described in Table 6. There is a surprising degree of inconsistency in ChatGPT compared to fine-tuned PLMs. 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, even in conditioned inconsistency, which reduces the overestimation of inconsistent behaviour. We observe that the extremely high inconsistency for the SNLI task is mainly because ChatGPT fails to distinguish between the labels of "Neutral" and "Contradiction". However, the model is still not completely consistent even in SNLI-2C, which integrates "Neutral" and "Contradiction" into the same class. Although the inconsistency rate might be considered trivial, especially in the SNLI-2C case, the issue should not be overlooked, considering the simple nature of the symmetric property. Consider a model that takes a list of symptoms and generates prescriptions. For such a model that should operate conservatively, it would greatly undermine the model's trustworthiness if it generates entirely different prescriptions whenever the order of symptoms changes, even if such an error occurs with a probability of 2%. Hence, an effort should be made to make LLMs satisfy logical consistencies to enhance their reliability and safe usage in real-world applications. Table 7 presents examples of symmetric consistency violations.

Model	MRPC		RTE		SNLI		SNLI-2C	
	τ	τ_C	τ	τ_C	τ	τ_C	τ	τ_C
BERT-large	6.8	-	15.8	-	10.2	-	-	-
RoBERTa-large	4.3	-	11.6	-	9.7	-	-	-
Electra-large	5.3	-	**6.7**	-	**6.4**	-	-	-
T5-large	**4.2**	-	8.0	-	8.3	-	-	-
ChatGPT	12.5	-	35.5	32.6	40.5	49.23	3.0	2.52

Table 6: Experimental results of the symmetric consistency evaluation. τ and τ_C denote the original and conditioned symmetric inconsistency, respectively. The best performance is in bold.

4.4 ChatGPT's Explainablity

Providing explanations is a core property of trustworthy systems (Huang et al., 2020). It is widely known that generative language models like ChatGPT can provide answers with explanations. However, we observed that while ChatGPT generates plausible explanations, those explanations are not perfectly reliable. Table 8 presents some examples. For the first example, the explanations of the original and perturbed inputs contradict each other. Regarding 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"). These wrong explanations also contribute to undermining ChatGPT's trustworthiness.

5 Discussion

Can Prompt Design be a Solution? Prompts are input text consisting of a task demonstration and,

TASK: MRPC, CONSISTENCY TYPE: Negation	
ORIGINAL INPUTS	PERTURBED INPUTS
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: MRPC, CONSISTENCY TYPE: Negation	
ORIGINAL INPUTS	PERTURBED INPUTS
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	
ORIGINAL INPUTS	PERTURBED INPUTS
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: MRPC, CONSISTENCY TYPE: Symmetric	
ORIGINAL INPUTS	PERTURBED INPUTS
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

Table 7: Examples of negation and symmetric consistency violations.

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, we are sceptical of this claim. The consistency metrics could be improved with different prompts, but we believe that it cannot fundamentally resolve the inconsistency problem, because prompt design cannot go beyond inductive reasoning. The underlying idea behind prompt design is that prompts created by experimenters might not be optimal, because language models might have acquired target information from completely different contexts (Jiang et al., 2020). That is, prompt design can be regarded as maximising the generalisation effect by searching for the most closely related prompts to perform the target task during training. As a result, no matter how prompt design allows us to find the best prompt that maximises the generalisation effect, it cannot resolve the issue, as our experimental results suggest that various consistency properties are not reflected in ChatGPT's inductive bias. 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.

Data Augmentation is Not Sustainable. Creating new data points based on certain consistency types and using them for training (Asai and Hajishirzi, 2020; Hosseini et al., 2021) or consistency regularisation (Wang and Henao, 2021; Zheng et al., 2021; Kim et al., 2021) is the most widely used approach to reflect logical consistency in the model's inductive bias. This remedy, however, is unsustainable. First, the data augmentation process requires a tremendous effort. For simple consistency types, e.g., symmetric and negation consistency, generating or collecting data points is relatively simple, but for complex consistency types, such as transitive and semantic consistency, it can be extremely challenging to cover all possible variations. Second, even if we successfully expand the data, it is doubtful whether we can afford to update an LLMs on the new dataset. Considering the ever-changing character of language, the data expansion and update of an LLM should be performed continuously. However, training an LLM entails tremendous financial and environmental costs (Bender et al., 2021). For instance, training a BERT-base model without hyperparameter tuning, which is 1590 times smaller than ChatGPT, requires a CO2 emission of 650kg, which is comparable to flying

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 8: Examples of ChatGPT's output with explanations.

from New York to San Francisco for one passenger (Strubell et al., 2019). A simple expectation of CO2 emission for re-training ChatGPT is 1033t, while a human is responsible for 5t CO2 emission per year. Therefore, it is desirable to enlarge our viewpoint beyond LLMs to implement sustainable remedies that can fundamentally solve the inconsistency problem, particularly in a modern society facing the global climate crisis.

6 Summary and Outlook

The recent 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.

We have investigated the consistency behaviour of ChatGPT across three consistency types and downstream tasks. Our experimental results demonstrated that ChatGPT achieves a certain level of enhanced language understanding ability, especially in negation expressions and antonyms, showing considerable improvements in negation consistency compared to the earlier version of PLMs. However, contrary to the widespread belief regarding the outstanding performance of ChatGPT, its overall consistency 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, i.e., the model is self-contradictory. Moreover, in input-order invariant tasks, ChatGPT is likely to make a different decision when the order of the input sentences is switched. Given how simple and natural the symmetric consistency is in human reasoning, violating symmetric consistency is a huge blow to ChatGPT's reliability. These fallacious behaviours are 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 ChatGPT is used in real applications, particularly considering the huge economic and environmental costs for training and inference of LLMs.

7 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. Also, evaluations on logical consistencies requiring more delicate reasoning ability (e.g., transitive consistency) are omitted. We leave this as future work.

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.

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 *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.

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, 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.

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.

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*.

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, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

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.

Christian Terwiesch. 2023. Would Chat GPT 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.

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.

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.