Transitive self-consistency evaluation of NLI models without gold labels

Wei Wu and **Mark Last**
Department of Software and Information Systems Engineering
Ben Gurion University of the Negev
ww@post.bgu.ac.il mlast@bgu.ac.il

Abstract

Natural Language Inference (NLI) is an important task in natural language processing. NLI models are aimed at automatically determining logical relationships between pairs of sentences. However, recent studies based on gold labels assigned to sentence pairs by human experts have provided some evidence that NLI models tend to make inconsistent model decisions during inference. Previous studies have used existing NLI datasets to test the transitive consistency of language models. However, they test only variations of two transitive consistency rules out of four. To further evaluate the transitive consistency of NLI models, we propose a novel evaluation approach that allows us to test all four rules automatically by generating adversarial examples via antonym replacements. Since we are testing self-consistency, human labeling of generated adversarial examples is unnecessary. Our experiments on several benchmark datasets indicate that the examples generated by the proposed antonym replacement methodology can reveal transitive inconsistencies in the state-of-the-art NLI models.

1 Introduction

Transformer-based (Vaswani, 2017) Large Language Models (LLMs) have shown impressive performance on Natural Language Understanding(NLU) tasks such as Natural Language Inference (NLI), or recognizing textual entailment (RTS), which labels a premise sentence as entailing, contradicting or neutral with respect to an associated hypothesis (Aspillaga et al., 2020). NLI models have been used to maintain and improve logical consistency in different NLP tasks, including question-answering, generated dialogue utterances, and text summarization (Mitchell et al., 2022; Song et al., 2020; Laban et al., 2022). However, some recent works demonstrate that NLI models can make inconsistent predictions with entailment relationships (Arakelyan et al., 2024; Nakamura et al., 2023).

Ensuring the logical consistency of Large Language Models (LLMs) has been an important topic in Natural Language Processing (NLP) under various definitions. The logical consistency of an LM is defined as "the ability to make decisions without logical contradictions" (Jang et al., 2022). According to (Jang et al., 2022; Jang and Lukasiewicz, 2023a), logical consistency can be divided into the following categories:

1. **Negational consistency**: $f(X) \neq f(\neg X)$
2. **Symmetric consistency**:$f(X, Y) = f(Y, X)$.
3. **Transitive consistency**: $X \rightarrow Y \wedge Y \rightarrow Z$ then $X \rightarrow Z$
4. **Additive consistency**: $f(X) = f(Y) \rightarrow f(X + Y) = c$, where c is a predicted label.

Existing evaluation methods measure the logical consistency of NLI models either by inference rules from propositional logic adapted from propositional calculus (Nakamura et al., 2023) or by behavioral consistency, including semantic, logical, and factual consistencies (Jang et al., 2022) based on gold labels. However, reliable and trustworthy AI systems should maintain internal *self-consistency*, which requires their predictions across multiple inputs to be logically compatible (Mitchell et al., 2022). This implies that evaluating the logical consistency of NLI models should be based on the labels assigned by the model itself, disregarding the actual correctness of the labels in the real world.

To this end, we propose an automatic transitive self-consistency evaluation procedure of NLI models without using gold labels. In this work, we generate adversarial examples via automatic antonym replacements of specific part-of-speech words in premises and hypotheses extracted from NLI datasets. Then, we use the evaluated NLI

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22638–22654
November 4-9, 2025

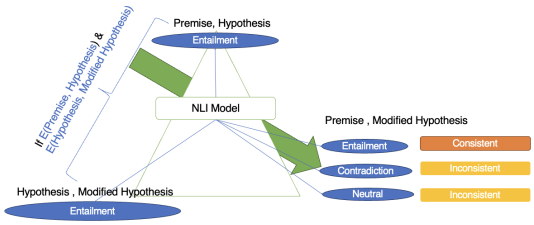

Figure 1: An example type of transitive consistency testing for entailment relationships. If (premise, hypothesis) is an entailment and (hypothesis, modified hypothesis) is an entailment, then (premise, modified hypothesis) should be an entailment by transitive properties. If not, then the model is inconsistent. **E** in the figure denotes Entailment.

models to label the original sentence and the adversarially generated sentence pairs and verify that the model labels satisfy the transitive consistency. Finally, our evaluation experiments show that the examples generated by automatic antonym replacement can test all four rules of transitive self-consistency.

Our contributions are as follows:

1. We propose using antonym replacement to test NLI models with respect to the transitive consistency rules not tested by other methods.

2. We propose a novel antonym replacement methodology to automatically generate a large set of adversarial sentences lexically similar to the original sentences (one-word difference only) while semantically different (contradicting each other).

3. We also propose an evaluation procedure for testing the transitive self-consistency of NLI models using the challenging examples generated by the antonym replacement methodology.

4. Our experiments indicate the proposed evaluation methodology can reveal transitive inconsistencies in state-of-the-art NLI models without needing manually assigned NLI labels.

5. We will release our evaluation dataset of adversarial examples to facilitate future work in NLI research.

2 Related Work

The transitivity of entailment relations, which derives that A entails C from A entails B and B entails C, is incorporated into logic-based Natural Language Inference (NLI) systems using automated theorem proving (Yanaka et al., 2021; Abzianidze, 2015; Mineshima et al., 2015). According to Jang and Lukasiewicz (2023b), pretrained Language Models still suffer from inconsistent behavior, making it extremely important to develop efficient techniques for automated consistency evaluation of Language Models. Li et al. (2019) applied the transitive inference properties to NLI tasks given three related sentences P, H, and Z with respect to the three predicates X, Y and Z. They defined four transitive inference rules:

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

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

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

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

where E, N, and C denote entailment, neutral, and contradiction, respectively. They assessed the transitive consistency of BERT and LSTM models with an evaluation set from MS-COCO (Lin et al., 2014).

Yanaka et al. (2021) introduced an analysis using synthetic and naturalistic NLI datasets with clause-embedding verbs to assess models' transitivity inferences across veridical, entailment, and arbitrary relation types.

Stowe et al. (2022) introduced the IMPLI (Idiomatic and Metaphoric Paired Language Inference) dataset consisting of paired sentences of idioms and metaphors. They construct both sliver pairs (automatic replacement with definition) and gold pairs (human written sentence).

Jang et al. (2022) defined the consistency of an LM and proposed a new dataset BECEL (Benchmark for Consistency Evaluation of Language Models), a unified dataset, which assessed four types of consistency (negation, symmetric, transitive and additive) several downstream tasks including NLI and WiC (Words-in-Context) and test transitive consistency on them. They leverage symmetric consistency applicable to instances with "contradiction" label as only a premise is shared in SNLI, so that they transform the transitive properties to: $X \rightarrow Y \wedge X \rightarrow Z$ then $Y \rightarrow Z$, with $X \rightarrow Z$ as a contradiction. Therefore, they have only tested variations of two transitive consistencies out of four.

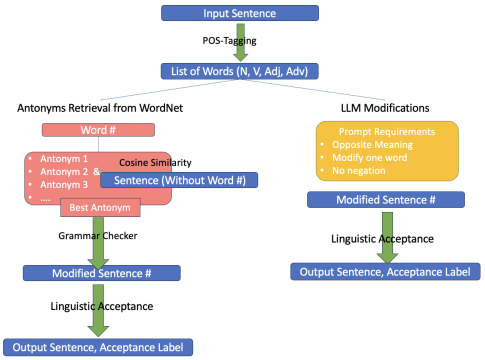

Figure 2: The antonym replacement pipeline using Wordnet and LLMs. The acceptance label is automatically assigned by a sentence classification model.

Other works investigated the trustworthiness of AI systems regarding logically consistent behaviors. Jang and Lukasiewicz (2023b) measured the semantic, negation, symmetric, and transitive consistency of ChatGPT and GPT-4. They evaluate LLMs based on whether their answers are "Entailment" or "Not Entailment". Yamamoto et al. (2024) evaluated encompassing metrics (accuracy, consistency, and logical coherence) of LLMs. They find that ChatGPT excelled in accuracy and logical coherence, Gemini showed superior consistency and Claude's performance hightlighted for improvement in complex symmetrical reasoning tasks.

Another approach to consistency evaluation is performing adversarial attacks on NLI models (Williams et al., 2022; Nie et al., 2020; Chien and Kalita, 2020). Regarding logic-based techniques, Nakamura et al. (2023) proposed a LogicAttack, to attack NLI models on logical consistency using various logical forms of premise and hypothesis, with a range of inference rules from propositional logic (such as Modus Tollens and Bidirectional Dilemma). They leverage the inference rules to perturb pairs having entailment relationships but fool many models into inferring otherwise.

Arakelyan et al. (2024) proposed a framework to measure the extent of semantic sensitivity. They evaluate NLI models on adversarially semantic-preserving generated examples. They selected correctly labeled sentence pairs, generated adversarial examples via conditional text generation requiring symmetric equivalence entailment, then replaced original hypotheses with these candidates to evaluate the models and found that semantic sensitivity performance degradations over both in- and out-of-domain settings.

The above-mentioned methods evaluate the logical consistency of pre-trained NLI models based on expensive gold labels assigned to sentence pairs by human experts. In contrast, our work measures the logical consistency of the model outputs without needing gold labels so that we can create large benchmark datasets more efficiently. Instead of generating semantic-preserving adversarial examples or focusing on entailment relationships, we generate semantic-altering examples to further evaluate more types of transitive consistency of NLI models.

3 Methods

3.1 Evaluation Dataset Construction

To evaluate the transitive consistency of language models, we propose to generate adversarial examples from NLI datasets containing premise-hypothesis sentence pairs by modifying the original premises and hypotheses with antonyms. Our method can ignore the ground-truth NLI labels of sentence pairs in the dataset since it evaluates the self-consistency of model-assigned labels rather than their correctness in the real world.

Given a sentence, we select all words with specific part-of-speech tags (namely, verbs, adjectives, adverbs, and nouns), and for each word chosen, generate a lexically similar but semantically different sentence by replacing the selected word with its antonym.

We can select a word's replacement antonym directly from the Wordnet (Miller, 1995) or by running LLM. Both antonym replacement pipelines are illustrated in figure 2.

After retrieving the list of a word's antonyms from the Wordnet (Miller, 1995), our Wordnet-based pipeline retains only the antonyms having the same part of speech as the original word and replaces it with an antonym having the highest cosine similarity to the rest of the sentence's content. After replacement, we verify that the modified sentence remains grammatically correct and coherent using automatic grammar checking tools. Our Wordnet-based antonym replacement pipeline skips a selected word in the following cases:

1. No word antonyms are found on the Wordnet.
2. No antonyms match the word's part-of-speech tag

3. The modified sentence is not grammatically correct.

We present the complete Wordnet-based antonym replacement procedure in Algorithm 1.

The LLM-based antonym replacement pipeline uses the following prompt: *Change the meaning of the following sentence to the opposite by replacing one word in quotation marks with its antonym, do not use negation, replacing one word at most. Please explain your change.*, where the quotation marks indicate the word to be replaced. Contrary to the Wordnet-based approach, we assume that LLMs take care of the modified sentence's grammaticality, implying that the LLM-based pipeline does not need a grammar checker.

To further ensure the modified sentences are semantically meaningful, we utilize a BERT model fine-tuned on CoLA (Corpus of Linguistic Acceptability) (Warstadt et al., 2019), which labels the modified sentence as "acceptable" if it is morphologically, syntactically, and semantically meaningful and "unacceptable" if not. All the unacceptable modified sentences are filtered out from the evaluation dataset.

After antonym replacement by one of the pipelines, we can automatically generate adversary examples from NLI benchmarks, where each entry contains an original premise, an original hypothesis, and a modified hypothesis. The gold labels can be included in the dataset for completeness to contribute to other evaluations. The evaluation dataset built in our experiments will be available online.

Dataset	bart-l	roberta-l	deberta-l	mBERT
MNLI	90.6%	90.6 %	91.1%	87.9%
SNLI	86.8%	85.9 %	87.9%	86.8%
WANLI	63.1%	61.9%	66.1%	66.8%
ANLI-r1	41.3%	45.8%	50.3%	54.2%
ANLI-r2	38.5%	35.3%	32.1%	48.7%
ANLI-r3	32.4%	21.2%	28.6%	41.0%

Table 1: The accuracy of the modified NLI datasets for the models from Wordnet-based replacemnt. -r1, -r2,-r3 represent different rounds in ANLI dataset. mBERT is an abbreviation of the ModernBERT model.

3.2 Evaluating Transitive Consistency

After generating k variations of each hypothesis in the NLI dataset, we test the transitive consistency of NLI models following the definitions by (Li et al., 2019): For three related sentences, P(premise), H (hypothesis), H'(modified hypothesis), the following four transitive inference rules are defined:

$$E(P,H) \wedge E(H,H') \rightarrow E(P,H'), (1)$$

$$E(P,H) \wedge C(H,H') \rightarrow C(P,H'), (2)$$

$$N(P,H) \wedge E(H,H') \rightarrow \neg C(P,H'), (3)$$

$$N(P,H) \wedge C(H,H') \rightarrow \neg E(P,H'), (4)$$

E, N, and C denote entailment, neutral, and contradiction, respectively.

Dataset	bart-l	roberta-l	deberta-l	mBERT
MNLI	61.45%	59.10%	61.27%	58.66%
SNLI	65.36%	64.15%	65.56%	61.51%
WANLI	65.46%	63.41%	64.13%	57.10%
ANLI-r1	63.87%	63.87%	67.09%	65.80%
ANLI-r2	66.77%	66.77%	67.30%	64.10%
ANLI-r3	63.26%	59.59%	60.00%	57.95%

Table 2: The rate of mutual **contradictions** predicted by the NLI models between (hypothesis, modified hypothesis) and (modified hypothesis, hypothesis) of Wordnet-based replacement.

Dataset	bart-l	roberta-l	deberta-l	mBERT
MNLI	20.57%	28.87%	19.42%	13.93%
SNLI	15.34%	26.05%	19.61%	7.91%
WANLI	18.14%	25.41%	19.18%	12.81%
ANLI-r1	12.90%	25.80%	16.77%	12.25%
ANLI-r2	16.02%	21.79%	17.30%	10.89%
ANLI-r3	20.81%	29.79%	21.42%	15.51%

Table 3: The rate of mutual **entailments** predicted by the NLI models between (hypothesis, modified hypothesis) and (modified hypothesis, hypothesis) of Wordnet-based replacement.

To evaluate the consistency automatically, all NLI labels are provided by the model. To ensure the equivalence relationship between the original and the modified hypothesis, we use a bidirectional labeling process, where:

$$M(H,H') = \hat{y}_{(E,C)} = M(H',H)$$

, so that they are logically equivalent, where M denotes the NLI model and $\hat{y}$ is either entailment or contradiction. For consistency evaluation, we ignore all cases that are not mutually equivalent or neutral. We can test all four types of transitive inference rules listed above with the adversarially generated dataset. In contrast, other works (Jang et al., 2022; Jang and Lukasiewicz, 2023b; Arakelyan et al., 2024) tested either inference rules (2) and (3) or rule (1) only. An NLI model is inconsistent if the predicted labels of some generated (premise, modified hypothesis) pairs contradict the corresponding logic rules.

Model	Dataset	E&E → !E	E&C → !C	N & E → C	N&C → E
bart-large	MNLI	8.02%	29.15%	3.39%	0.88%
	SNLI	15.45%	28.64%	1.10%	3.22%
	WANLI	11.24%	14.39%	11.85%	2.16%
	ANLI-r1	0.00%	35.29%	6.25%	6.06%
	ANLI-r2	50.00%	25.74%	9.09%	5.55%
	ANLI-r3	19.04%	52.30%	16.27%	2.56%
roberta-large	MNLI	8.79%	26.95%	3.74%	1.22%
	SNLI	5.01%	22.01%	2.00%	2.79%
	WANLI	7.05%	11.56%	6.91%	2.01%
	ANLI-r1	0.00%	37.14%	4.16%	8.33%
	ANLI-r2	17.64%	34.04%	0.00%	0.00%
	ANLI-r3	8.82%	34.78%	18.18%	2.79%
deberta-large	MNLI	6.00%	26.34%	2.05%	0.80%
	SNLI	5.55%	14.66%	1.79%	1.61%
	WANLI	7.97%	11.61%	8.01%	1.62%
	ANLI-r1	33.33%	33.33%	0.00%	0.00%
	ANLI-r2	0.00%	23.07%	13.33%	1.78%
	ANLI-r3	17.24%	47.88%	1.11%	4.51%
modernBERT	MNLI	9.87%	35.48%	4.80%	1.80%
	SNLI	19.44%	16.55%	5.71%	0.84%
	WANLI	11.53%	17.45%	8.87%	3.48%
	ANLI-r1	25.00%	39.13%	0.00%	6.45%
	ANLI-r2	28.57%	38.09%	0.00%	3.84%
	ANLI-r3	25.92%	45.78%	13.33%	3.78%

Table 4: The **Inconsistency Rate** of NLI models on transitive consistency. The results are from Wordnet-based examples. E, C, N denote Entailment, Contradiction and Neutral, respectively and ! denotes a negation. Each column represents one of the violations of transitive inference rules.

4 Experimental Results

4.1 Datasets

NLI Datasets. In our experiments, we utilize four benchmark NLI datasets for the generation of adversarial examples: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI ((Nie et al., 2020), and WANLI (Liu et al., 2022). The amounts of acceptable modifications from each NLI benchmark are **5566**, **5209**, **801**, and **6822** for SNLI, MNLI, ANLI, and WANLI, respectively.

The word vectors are generated using the Word2Vec model and part-of-speech tags are assigned using spaCy [1]. We check the grammar of the modified sentence using the Python package *language-tool-python*[2] based on US English. We generate variations for the hypothesis sentences using Wordnet-based and LLM-based antonym replacement pipelines.

4.2 Antonym Replacement Procedure

We compare the adversarial examples generated by the WordNet-based antonym replacement algorithm to the following three LLMs: **o4-mini** [3], **DeepSeek-R1** [4] (Guo et al., 2025) and **flan-t5**: *google/flan-t5-large*[5]. We randomly select 100 sentences from the datasets for modification and choose the same word for replacement by WordNet and LLMs.

Table 6 demonstrates the manual evaluation results of comparing WordNet to LLM-generated examples. Overall, the performance of the LLMs and the Wordnet-based antonym replacement is similar in terms of precision and recall. However, the

[1] https://pypi.org/project/spacy/
[2] https://pypi.org/project/language-tool-python/
[3] https://platform.openai.com/docs/models/o4-mini
[4] https://deepseek-r1.com
[5] https://huggingface.co/google/flan-t5-large

Assigned Label	Adversarial Examples
Contradiction	Go *downwards* / *upwards* to one of the gates, all of which will lead you into the cathedral.
Contradiction	Leading organizations want to be sure their employees are *safe* / *dangerous*.
Contradiction	It's *impossible* / *possible* to have a plate hand-painted to your own design in Hong Kong.
Contradiction	Jobs *never* / *ever* held onto an idea for long.
Entailment	It's not *likely* / *improbable* you'll forget that, right?
Entailment	You can *ride* /*walk* a roller coaster ther that goes high up.

Table 5: Generated Adversarial Examples of both "mutual Entailment" and "mutual Contradiction" of Wordnet-based replacement. Labels are assigned by the ModernBERT model.

Model	Acc.	Precision	Recall	CPU-T/sec.	GPU-T/sec.
o4-mini	0.91	0.95	0.96	–	96
DeepSeek-R1	0.90	0.93	0.97	–	211
Flan-t5	0.95	0.97	0.98	65	387
Wordnet-AR	0.91	0.91	1.00	71	71

Table 6: The evaluation results of LLM-generated examples automatically labeled as accepted sentences. Accuracy, Precision, and Recall are calculated based on manual labels. CPU-T and GPU-T denote the Running Time on a CPU or a GPU processor in seconds.

Model	Verification	Antontym Diff.
o4-mini	97	58
DeepSeek-R1	98	60
Flan-t5	69	–

Table 7: Verification: the number of sentences matches the requirements in the prompt. Antonym Diff. - The number of different antonyms selected by the Wordnet-based antonym replacement algorithm.

Wordnet-based algorithm is much cheaper computationally than the LLMs.

4.3 Linguistic Acceptability

The linguistic acceptability checks whether a given sentence contains morphological, syntactic, and semantic violations. However, it excludes sentences with pragmatical anomalies, unavailable meanings, prescriptive rules, or nonce words (Warstadt et al., 2019). We utilize the **textattack/bert-base-uncased-CoLA** [6] model. To evaluate the performance of the automatic linguistic acceptability tool, we manually labeled 200 sentences modified by Wordnet-based antonym replacement. The overall accuracy is 0.82, with a precision of 0.82 and a recall of 0.96 for the acceptable sentences, implying that the automatic acceptability model helps to remove most of the semantically meaningless sentences generated by the algorithm. Table 8 demonstrate examples of acceptable and unacceptable sentences modified by Wordnet-based antonym replacement algorithm.

4.4 Transitive Inconsistency

NLI models. We tested transformer-based models that are fine-tuned on the MNLI dataset. We evaluate one encoder-decoder model, two decoder models, and an encoder model. Specifically, we choose: **bart-large** (Lewis et al., 2020): *facebook/bart-large-mnli* [7],**roberta-large** (Liu, 2019): *FacebookAI/roberta-large-mnli* [8],**deberta-large** (He et al., 2020): *microsoft/deberta-large-mnli* [9], **ModernBERT-base-nli** (Warner et al., 2024): *tasksource/ModernBERT-base-nli* [10]

Table 1 explores the accuracy of (premise, hypothesis) pairs of the NLI models on the modified NLI datasets. It shows that the models perform well on SNLI and MNLI datasets but not on the ANLI dataset, while the performance on WANLI is in the middle. The gold labels are used for testing only, and they are ignored during the further model evaluation process.

The results of modified sentence labeling by NLI models after antonym replacement are shown in Tables 2 and 3. These tables demonstrate the rate of mutual contradictions and entailments assigned by the NLI models to the (hypothesis, modified hypothesis) and (modified hypothesis, hypothesis) sentence pairs. Table 5 shows the generated adver-

[6]https://huggingface.co/textattack/bert-base-uncased-CoLA

[7]https://huggingface.co/facebook/bart-large-mnli

[8]https://huggingface.co/FacebookAI/roberta-large-mnli

[9]https://huggingface.co/microsoft/deberta-large-mnli

[10]https://huggingface.co/tasksource/ModernBERT-base-nli

Acceptability	Modified Sentence
Acceptable	The deep end of the pool is good for swimming.
Acceptable	The computer industry is invulnerable to terrorist attacks.
Unacceptable	The shallow end of the pool is good for sink.
Unacceptable	She'll be sure to arrive the lights off for you.
Unacceptable	There differ a widespread feeling that it is important to have a job.

Table 8: Examples of acceptable and unacceptable sentences in WANLI dataset modified by the Wordnet-based antonym replacement algorithm. Acceptability labels are assigned by the model.

sarial examples and the mutual NLI labels.

Intuitively, as expected, since the modified sentences are replaced by antonyms, the majority of around 60% are labeled as Mutual Contradiction, whereas 20% are labeled as Entailment. The rest are labeled as Neutral, and since we could not apply the transitive inference rules to Neutral cases, we would not discuss them here.

The rate of mutual relationship indicates that even though antonym replacement methods aim to generate contradictions, several examples result in entailment because the replacement does not affect the relationship of the sentence pair.

Table 4 demonstrates the Inconsistency Rate of NLI models on four types of transitive inference rules. We found the most common inconsistency where (premise, hypothesis) is labeled as an **entailment** and (hypothesis, modified hypothesis) is labeled as a **contradiction**, but (premise, modified hypothesis) is labeled as **not a contradiction**. It is less likely that (premise, hypothesis) is labeled as **neutral**, and (hypothesis, modified hypothesis) is labeled as **contradiction**, whereas (premise, modified hypothesis) is labeled as an **entailment**. The violation of the other two transitive inference rules varies across the models and the datasets.

5 Discussion

5.1 Antonym Replacement

5.1.1 Wordnet vs. LLM

We compare the sentences modified by the Wordnet-based antonym replacement algorithm and by the LLMs. Table 7 indicates the number of generated sentences that met the requirements in the prompt and the number of antonym selections different from the Wordnet-based algorithm. We utilize the web interfaces of o4-mini and DeepSeek to use their searching and reasoning modules.

We noticed that o4-mini and DeepSeek could follow the instructions in the prompt, whereas flan-t5 failed a lot. The antonyms LLMs select for the same word mostly differ from the Wordnet selections. However, many words selected by different methods are synonyms, such as big/large, small/little, and stupid/foolish. Some of them are different but correct choices of noun antonyms, for example, parent/adult, and girl/woman. Several mistakes remain the same (have-lack and function/malfunction).

Moreover, there are some limitations to Wordnet itself. Not every selected word has antonyms on the Wordnet. We skip the selected word in the sentence in the Wordnet-based replacement if no antonyms are found on the Wordnet. For LLM-based replacement, a small-scale experiment is discussed in Appendix F. Some antonyms, such as (am/is/are - differ), apply only to some uses of the original word. For example, the word "differ" can only be used as a replacement for "am/is/are" when these words express equivalence or identity. Consequently, such nonsensical sentences are included as modifications, but are still labeled acceptable by the acceptability model. The number of antonyms per word on the Wordnet is also limited.

The advantage of using LLMs is that the generated sentences are acceptable (coherent and semantically meaningful) without grammar-checking tools. However, by analyzing the antonym selection choices, LLMs do not have a significant advantage in replacing a word with more proper antonyms than Wordnet, as they repeat the same replacement mistakes, while being more expensive computationally.

5.1.2 Sentence Acceptability Analysis

The unacceptable sentences can be classified into four major error types.

1. **Invalid Substitution**: Antonyms selected lead to incoherent/nonsensical sentences.

2. **Grammar Contradictions**: Syntactic errors caused with one word change.

3. **Polysemous**: The replaced word has more than one meaning.

4. **Fixed Phrases**: The replacement of one word in a fixed phrase.

With the acceptability model (bert-base-uncased-CoLA), we can automatically remove most of the unacceptable sentences caused by invalid antonym substitutions and grammar contradictions, which helps to improve significantly for filtering the modified sentences. However, several sentences are mistakenly accepted by the model due to polysemous and fixed phrases.

When we replace adjectives, adverbs, verbs, or nouns with their antonyms, the following outcomes are possible: The modified sentence is incoherent. The modified sentence is coherent and contradicts the original sentence. The modified sentence is coherent and it is entailed by the original sentence. The modified sentence is coherent and unrelated to the original sentence ("neutral"). We sample 100 mutual entailments labeled by modernBERT and BART models (Appendix G), and find that many sentences entailed because the replaced antonyms does not affect the relationship (here/there, man/civilian, ride/walk). However, 20% of entailments are mislabeled due to double negation, fixed phrases or polysemous. It indicates that the NLI models are not sensitive to semantic changes and are vulnerable to adversarial examples. The models successfully detect a semantic contradiction with antonym replacement for over 60% of labeled mutual contradictions.

With regard to the datasets, most of the replaced words in SNLI are nouns, where almost all pronoun contradictions happened. It became more diverse and more meaningful in the MNLI, ANLI and WANLI datasets, but more polysemous and fixed phrase issues were involved in the generated examples.

5.2 Transitive Inconsistency

In our experiments, we could reveal more examples that are inconsistent with respect to transitive inference rules than the previous evaluation studies (Jang et al., 2022; Jang and Lukasiewicz, 2023b). According to our results, the violations of rules (1) and (2) are more frequent than those of rules (3) and (4).

Disregarding the correctness of the NLI models, the violations of the inference rules reveal self-inconsistency of transitivity. It suggests that the models find it harder to maintain the rules with the chains of entailment (rules 1 and 2) than neutral cases (rules 3 and 4). There are more inconsistencies in rule (2), where an entailment and contradiction lead to a non-contradiction. It implies the models failed to maintain a self-consistent decision on the generated adversarial examples.

As the models are fine-tuned on MNLI dataset, we assume they would perform better. However, the results indicate that a model does not necessarily become more consistent after fine-tuning. The models are least inconsistent for WANLI and most inconsistent for ANLI-r3. Because challenging examples are adversarially generated by human experts in ANLI, but the models are more consistent for r2 than for r1. WANLI contains machine-generated examples filtered and revised by human coworkers. MNLI is derived from SNLI but is more complicated. The difficulty of the examples in each dataset relates to the model performance on accuracy (Table 1), but does not correlate to the inconsistency rate.

The results reveal critical limitations of the model's logical reasoning abilities. They show that all models struggle with not only transitive self-inconsistencies but also semantic inconsistencies. This indicates that the model labeling of sentences relies on superficial cues of the sentence instead of fully understanding its meaning.

5.3 Self-Consistency and Correctness

Our study builds upon the assumption that in a multi-step reasoning task, the logical consistency requirement is independent of the actual correctness of NLI labels. Correctness refers to whether the NLI model's labeling of a given sentence pair matches the ground truth. On the other hand, self-consistency evaluates whether the model's labeling of multiple sentence pairs is logically coherent across a multi-step reasoning chain (transitive inference rules).

Consider a multi-step reasoning chain for the three sentence pairs: (premise, hypothesis), (hypothesis, modified hypothesis), and (premise, modified hypothesis), where the modified hypothesis is generated by an antonym replacement algorithm. Based on the gold label for (premise, hypothesis), we can evaluate the correctness of the pair (hypothesis, modified hypothesis) and then check the consistency of the reasoning chain according to the transitive rules.

In Table 10, we show examples of (premise, hypothesis) pairs, which are correctly labeled as entailment by NLI models. After modifying the

hypothesis by an antonym replacement algorithm, we ask NLI model to label two additional sentence pairs: (hypothesis, modified hypothesis) and (premise, modified hypothesis). We consider the following cases:

1. The NLI models' labeling is correct and self-consistent. In this case, the respective labels of the above sentence pairs should be either (Entailment, Entailment) or (Contradiction, Contradiction), to match the transitive rules (1) and (2). (Examples 1 and 5)

2. The labeling of the above sentence pairs is correct but inconsistent. In this case, the labels may be either (Entailment, Contradiction/Neutral) or (Contradiction, Entailment/Neutral) (Examples 2 and 6).

3. One of the labels assigned to the above sentence pairs is incorrect; however, the labeling is self-consistent, as both pairs are labeled as either (Entailment, Entailment) or (Contradiction, Contradiction). (Examples 3 and 7)

4. One of the assigned labels is incorrect, and the labeling is also inconsistent, since the labels may be either (Entailment, Contradiction/Neutral) or (Contradiction, Entailment/Neutral). (Examples 4 and 8)

Ideally, we are expecting the reasoning chains of NLI models to provide correct labels and to be logically consistent. The alignment of correctness and consistency represents a valid semantic understanding and logical reasoning capability. If the NLI model labels each sentence pair correctly, its reasoning chain can still be logically inconsistent, making the models less reliable from the NLI user's viewpoint. On the other hand, the NLI model may label a sentence pair incorrectly, even though its reasoning chain remains logically consistent, suggesting a limited semantic understanding of labeled sentences. In the worst case, the NLI model's labeling can be both incorrect and inconsistent, suggesting that the model struggles with semantic understanding and logical reasoning.

In this work, we focus on self-consistency rather than labeling correctness. While NLI models are designed to identify logical relationships between two statements, the transitivity logic requires the models to reason consistently across multiple statement pairs, representing a deeper level of semantic understanding. We believe that both correctness and self-inconsistency are essential for making NLI models trustworthy and acceptable.

6 Conclusion and Future Work

In this paper, we propose an automatic procedure for the evaluation of transitive consistency in NLI models without the need for manually assigned labels. We propose a novel algorithm for generating an adversarial test dataset by replacing specific part-of-speech words with their antonyms. Finally, we test the self-consistency of several NLI models on the adversarially generated sentence pairs. The experimental results demonstrate that all evaluated NLI models make inconsistent decisions on transitivity. In the future, one can evaluate other types of logical consistency (negational, symmetrical) using the antonym replacement approach. Using a similar approach, one can also evaluate the logical consistency of AI systems such as ChatGPT and Gemini on NLI tasks.

Limitations

As appears in tables 2 and 3, some antonym replacements introduce entailments rather than contradictions. It indicates that the procedure of antonym replacement requires further exploration and possible enhancement. Wordnet as the source of antonyms is also limited. The quality and diversity of the generated adversarial examples can be enhanced with other sources. The generated benchmark can be further enhanced with a coherence checker and a filtering method in the pipeline. Another limitation of this work is that it focuses only on transitive consistency. The proposed framework can be extended to other types of logical consistency. Finally, our current experiments are limited to English-based models and datasets.

Acknowledgments

The authors acknowledge financial support from China Scholarships Council.

References

Lasha Abzianidze. 2015. A tableau prover for natural logic and language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2502.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of nli models. In *Proceedings of the 18th Conference of the European*

Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 432–444.

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1882–1894.

Samuel 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*, pages 632–642.

Tiffany Chien and Jugal Kalita. 2020. Adversarial analysis of natural language inference systems. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 1–8. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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

Myeongjun Jang and Thomas Lukasiewicz. 2023a. Consistency analysis of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15970–15985.

Myeongjun Jang and Thomas Lukasiewicz. 2023b. Consistency analysis of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15970–15985, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. 2015. Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2061.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768.

Mutsumi Nakamura, Santosh Mashetty, Mihir Parmar, Neeraj Varshney, and Chitta Baral. 2023. Logicattack: Adversarial attacks for evaluating logical consistency of natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13322–13334.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8878–8885.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Alex Warstadt, Amanpreet Singh, and Samuel Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. Anlizing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54.

Shouko Yamamoto, Kohaku Kobayashi, and Ran Tanaka. 2024. An empirical automated evaluation and analysis of symmetrical reasoning in large language models. *Authorea Preprints*.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural nli models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934.

A Model Inconsistency Examples

Below are several transitive inconsistency examples from the NLI models, where P is the premise, H is a hypothesis, and H' is the modified hypothesis. All labels are generated by the NLI models.

1. **Premise**: We are looking for a girl who is not at all fat.
   Hypothesis: We are looking for a girl who is not very **fat**.
   Modified: We are looking for a girl who is not very **thin**.
   (P,H): Entailment
   (H,H'): Entailment
   (P,H'): Contradiction

2. **Premise**: The entire economy received a massive jump-start with the outbreak of the Korean War, with Japan ironically becoming the chief local supplier for an army it had battled so furiously just a few years earlier.
   Hypothesis: Japan became the **local** supplier for Korea.
   Modified: Japan became the **national** supplier for Korea.
   (P,H): Entailment
   (H,H'):Contradiction
   (P,H'): Entailment

3. **Premise**: Two economists at Virginia Commonwealth University–yes, here are the economists again, but this time making a more plausible argument–studied millions of auto-accident claims filed between 1989 and 1993.
   Hypothesis: Two **qualified** researchers looked into vehicular accident insurance claims.
   Modified: Two **unqualified** researchers looked into vehicular accident insurance claims.
   (P,H):Entailment
   (H,H'):Contradiction
   (P,H'): Neutral

4. **Premise**: But we don't rule out regulation in the future if industry fails to do a good job of policing itself.
   Hypothesis: Regulation is a possibility for the **future**.
   Modified: Regulation is a possibility for the **past**.
   (P,H): Entailment
   (H,H'): Contradiction
   (P,H'): Entailment

5. **Premise**: As the brain becomes more powerful, the need for it to be physically smaller becomes more urgent.
   Hypothesis: The brain needs to be smaller to be **more** powerful.
   Modified Hypothesis: The brain needs to be smaller to be **less** powerful.
   (P,H): Entailment
   (H,H'): Contradiction
   (P,H'): Entailment

6. **Premise**: The sun's heat warms the earth, and the earth's heat warms the water, and the water's heat warms the air, and the air's heat warms the human body.
   Hypothesis: The **human** body is warmed by the sun.
   Modified: The **nonhuman** body is warmed by the sun.
   (P,H): Entailment
   (H,H'): Entailment
   (P,H'): Neutral

7. **Premise**: He was a gregarious man, always ready to make friends.
   Hypothesis: He was a **shy** man, always ready to make friends.
   Modified: He was a **confident** man, always ready to make friends.
   (P,H): Neutral
   (H,H'): Contradiction
   (P,H'): Entailment

8. **Premise**: Their supplies scarce, their harvest meager, and their spirit broken, they abandoned the fort in 1858.
   Hypothesis: Their **supplies** remained very low and hard to maintain.
   Modified: Their **demands** remained very low and hard to maintain.
   (P,H): Neutral
   (H,H'): Entailment
   (P,H'): Contradiction

B Model Consistency Examples

Below are several transitive consistency examples from the NLI models, where P is the premise, H is a hypothesis, and H' is the modified hypothesis. All labels are generated by the NLI models.

1. **Premise**: If you're a good swimmer, it's a good idea to try the shallow end of the pool.
   Hypothesis: The shallow **end** of the pool is good for swimming.
   Modified: The shallow **beginning** of the pool is good for swimming.
   (P,H): Entailment
   (H,H'): Entailment
   (P,H'):Entailment

2. **Premise**: The entire city was surrounded by open countryside with a scattering of small villages.
   Hypothesis: The whole countryside is scattered with **small** villages.
   Modified: The whole countryside is scattered with **large** villages.
   (P,H): Entailment
   (H,H'): Contradiction
   (P,H'): Contradiction

3. **Premise**: Taking an ecumenical tack, nation officials in Chicago recently issued edicts commanding preachers to back off their anti-Semitic rhetoric.
   Hypothesis: Nation officials in Chicago are **involved** in religious issues.
   Modified: Nation officials in Chicago are **uninvolved** in religious issues.
   (P,H): Entailment
   (H,H'): Contradiction
   (P,H'): Contradiction

4. **Premise**: The following list contains the aircraft used by the Royal Danish Air Force or its predecessors, the Danish Army Air Corps and Royal Danish Naval Aviation. During the Nazi occupation from 1940 to 1945, Danish military aviation was prohibited. The aircraft currently in use are highlighted in blue.
   Hypothesis: Denmark was **occupied** for half a decade in the middle of the twentieth century.
   Modified: Denmark was **unoccupied** for half a decade in the middle of the twentieth century.
   (P,H): Entailment
   (H,H'): Contradiction
   (P,H'): Contradiction

5. **Premise**: Zimbabwe is a landlocked country in southern Africa lying wholly within the tropics. It straddles an extensive high inland plateau that drops northwards to the Zambezi valley where the border with Zambia is and similarly drops southwards to the Limpopo valley and the border with South Africa.
   Hypothesis: Zimbabwe has a **large**, diverse population.
   Modified: Zimbabwe has a **small**, diverse population. (P,H): Neutral
   (H,H'): Contradiction
   (P,H'):Neutral

6. **Premise**: Spoons was a comedy sketch show first broadcast on the United Kingdom's Channel 4 from 30 September 2005. In the United States, "Spoons" is broadcast on BBC America. The relationship themed show combined recent trends in sketch shows—dark content, strong language, and recurring catchphrases.
   Hypothesis: Spoons was **popular** in the United Kingdom, but not in the United States.
   Modified: Spoons was **unpopular** in the United Kingdom, but not in the United States.
   (P,H):Neutral
   (H,H'):Contradiction
   (P,H'): Neutral

C Wordnet-based Antonym Replacement Algorithm

Algorithm 1 Wordnet-based antonym replacement algorithm

procedure ANTONYM REPLACEMENT($s \in [p, h]$)
  $words \leftarrow tokenize(s)$
  $kwList \leftarrow [V, Adj, Adv, N] \in words$
  for $kw \in kwList$ **do**
    $aList \leftarrow Wordnet(kw)$
    for $a \in aList$ **do**
      if $POS(a) == POS(kw)$ **then**
        $Words \leftarrow kw \notin words$
        $sim \leftarrow CosSim(a, Words)$
        $A \leftarrow max(sim)$
        $newSent \leftarrow replace(s, kw, A)$
        if $grammar(newSent)$ & $coherence(newSent)$ **then**
          $newSentList \leftarrow newSent$
  return $newSentList$
Notations:
s:Input sentence either premise or hypothesis
$kwList, kw$: List of selected words, Selected word
$aList, a$: List of antonyms, An antonym
POS: Part-of-Speech tag
sim: Similarity Scores
$CosSim()$:Cosine Similarity Function
A: Selected Antonym
$replace()$: Function to replace kw in s with A
$grammar()$: Grammar Checking Tool
$coherence()$ Coherence Checking Tool

D LLM and Wordnet-based Generated Examples

Several examples of LLM and Wordnet-based modifications are demonstrated below. The selected examples are after verifying the requirements (no negation & one word change) for each model.

Below are the LLM and Wordnet-based modified examples *original* / *modified*: **ChatGPT generated examples**

1. There are jobs that are not *right* / *wrong* for the right people.
2. Kings *frequently* / *rarely* founded orders than can still be found today.
3. I went to bed *early* / *late*.

DeepSeek generated examples

1. Two little boys are pretending they are climbing a *large* / *small* mountain.
2. The young people are taking a picture *inside* / *outside*.
3. The man with the sword is not a *good* / *poor* protector, no matter what he says.

google/flan-t5

1. Experienced pilots pilot these *high-drafted* / *low-drafted* craft.
2. Everybody visible is dressed extremely formally / Everybody in a hurry is extremely formally dressed.
3. Participants thought auditing should be *less* / *more* confrontational and more/less collaborative.

Wordnet-based

1. Kings *frequently* / *infrequently* founded orders that can still be found today.
2. Terrorists will *find* / *lose* new methods of performing transactions.
3. People who have the most to *lose* / *keep* are the ones who are willing to take the biggest risks.

E Acceptability

Dataset	Acceptable	Total	Rate
MNLI	5209	6530	79.77%
SNLI	5566	7672	72.54%
WANLI	6822	7784	87.63%
ANLI-r1	155	196	79.08%
ANLI-r2	156	204	76.64%
ANLI-r3	490	660	74.24%

Table 9: The sizes of the number and percentage of acceptable sentences by Wordnet-based algorithm in each dataset.

Table 9 indicates the number and percentage of acceptable sentences for each dataset. About 87% of modifications in WANLI are acceptable by

the model. However, about 75% of the modifications were acceptable in other datasets. We conduct manual evaluation of 200 sentences modified by Wordnet-based algorithm.

We demonstrate the following types of errors (*original* / *modified*) in both Entailments and Contradictions:

1. **Invalid substitution**: Many invalid antonyms were selected which would lead to incoherent/nonsensical sentences.

   a. There *are* / *differ* many fine beaches along he the shallow bays.

   b. The church *has* / *lacks* cracks in the ceiling.

   c. There are two *little* / *large* boys smiling.

   d. People are *outside* / *inside* in a park.

2. **Polysemous**: Another common error is due to polysemous, where the word has more than one meaning.

   a. You can pay using the US dollar when buy goods from the *duty-free* / *duty-bound* shops.

   b. Baronets were called sir in their *day* / *night*.

   c. One (not the only one) of the very lucky numbers of Monday's drawing was 5H, another *just* / *inequitable* as lucky as 5S.

3. **Grammar Contradictions**: Since we only replace one word in each sentence, the pronouns may contradict after replacement.

   a. A *man* / *woman* is talking to **his** daughter.

   b. A *father* / *mother* and **his** mother are taking a walk.

4. **Fixed Phrases**: The replacement of one word in a fixed phrase is also problematic.

   a. Steve Jobs came *back* / *advance* to Apple.

   b. We should take a *break* / *make* from this.

   c. The effects of climate *change* / *stay* can be seen.

The linguistic acceptability model helps to filter out most of the nonsensical modified sentences (mostly invalid substitution and grammar contradictions). However, the model excludes the unacceptable sentences of pragmatical anomalies, unavailable meanings, prescriptive rules, or nonce words (Warstadt et al., 2019). Thus, such cases are acceptable to the model. But with the conducted manual evaluation of the 200 sentences modified by Wordnet-based antonym replacement, these acceptable nonsensical sentences were minor, as the accuracy was 82%. Within these minor cases, most of them are mutual entailments, others are neutral cases. Only specific modifications (have -lack, am/is/are - differ) are mutual contradictions. It reveals that the NLI models are semantically inconsistent because they label nonsensical sentences as entailments.

Moreover, the main issue is with fixed phrases. Since we are generating lexical similar modifications by changing one word only, although the overall meaning shifts, the phrase itself become nonsensical. However, these nonsensical pairs will be further filtering out because they would not be labeled as mutual relationships (entailment/contradiction).

F LLMs for Missing Antonyms

Not every word selected has antonyms on the Wordnet and we skip such words in the Wordnet-based antonym replacement algorithm. We are experimenting with 10 words (Noun, Adjective, Adverb, Verbs) in 10 sentences to find if LLMs could fill in this gap. We use the same LLMs for modifications. The results show that 4 out of 10 sentences are coherent while others are not. Below are the examples of LLM modifications (*original/ modified*):

1. **Coherent**: Culture is **very / slightly** important to human behaviour. (o4-mini)

2. **Coherent**: The Fed is the ultimate **regulator / disruptor** of the banking system. (o4-mini)

3. **Coherent**: The **ban / permit** on headcarves will make women safer. (DeepSeek-R1)

4. **Coherent**: Culture is **very / barely** important to human behaviour. (DeepSeek-R1)

5. **Incoherent**: The NFPB is worried that the current tax **code / loophole** is unfair. (o4-mini)

6. **Incoherent**: High levels of stress can cause **memory / creativity** loss. (o4-mini)

7. **Incoherent** : High levels of stress can cause **memory / ignorance** loss. (DeepSeek-R1)

8. **Incoherent**: Do you **mean / deny** that? (DeepSeek-R1).

9. **Incoherent**: Culture is **very / non** important to human behaviour. (flan-t5)

For the examples presented above, the LLMs could find antonyms while no antonyms are in the Wordnet. However, many of the antonyms replaced result in incoherent sentences. It might because of the specific word selected for replacement, but the LLMs did not provide any useful explanations on the choices. It suggests that the LLMs could fill in the gap of the Wordnet but not necessarily better, because skipping these words would not affect much on the modifications.

G Modified Mutual Entailments and Contradictions

Examples of modifications in the generated dataset. The "Entailment" is assigned if the pair of (original, modified) and (modified, original) are both "Entailment" and is "Contradiction" if the pairs are both "Contradiction" assigned by the NLI models.

Datasets:

- **SNLI**
- **Entailment**:

1. There is a statue that not *many* / *few* people seem to be interested in.
2. A mother is helping her *child* / *parent* complete the experiment.

Contradiction:

1. Tons of people are *gathered* / *uncollected* around the statue.
2. A young boy is playing in the field because his *mother*/*father*.

- **MNLI**
- **Entailment**:

1. When states provide better coverage than private plans,any employers will drop *dependent* / *independent*.
2. evern said the people were always welcome *there* /*here*.

Contradiction:

1. The Government Executive articles housed on the website are not *able* / *unable* to be searched.
2. Go *downwards* / *upwards* to one of the gates, all of which will lead you into the cathedral.

- **WANLI**

Entailment:

- 1. The national *security* / *insecurity* apparatus is not to blame for the fact that the CIA is in the business of trying to protect its agents from harm.
  2. The United States did not *experience* / *inexperience* the same economic growth as other countries.

Contradiction:

- 1. This town is *known* / *unknown* for its dairy products, and people love the cheese and ice-cream.
  2. The Japanese have been *concerned* / *unconcerned* with the environment for many years.

- **ANLI**

Entailment:

- 1. Robert Anthony Stoops *attended* /*unaccompanied* the University of Oklahoma.
  2. A *former* / *latter* chief of pathology of the Minneapolis VA Medical Center predicted Gleason would die in 2008.

Contradiction:

1. The BSF is funded by the country directly **north** /*south* of Mexico.
2. The National Register of Historic Places was *established* / *unestablished* in 1984.

Premise	Hypothesis/Modified	GoldL(P,H)	Label(P,H')	Label(H,H')	Correctness	Consistency
1.To give the impression that he was an up-and-coming artist, he wore a pair of sunglasses.	He was trying to *give* / *take* the impression that he was a rising artist.	Entailment	Entailment	Entailment	Correct	Consistent
2."That's all right", said the Doctor, "I'll be back in a minute."	The Doctor was not going to be gone *long* / *short*.	Entailment	Neutral	Entailment	Correct	Inconsistent
3.Some of the students found the course very difficult and some found it very easy.	Some of the students found the course very easy, and some found it very *difficult* / *easy*.	Entailment	Entailment	Entailment	Incorrect	Consistent
4.A king should never be so unwise as to attack a king.	A king should *never* / *ever* attack another king.	Entailment	Contradiction	Entailment	Incorrect	Inconsistent
5.The entire city was surrounded by open countryside with a scattering of small villages.	The whole countryside is scattered with *small* / *large* villages.	Entailment	Contradiction	Contradiction	Correct	Consistent
6.Taking an ecumenical tack, nation officials in Chicago recently issued edicts commanding preachers to back off their anti-Semitic rhetoric.	Nation officials in Chicago are involved in *religious* / *secular* issues.	Entailment	Entailment	Contradiction	Correct	Inconsistent
7.I think it behooves Slate, in its effort to take over the public-opinion industry to make a thorough effort to uncover the truth behind this unnatural connection.	Slate should make an effort uncover the *truth* / *inaccuracy*.	Entailment	Contradiction	Contradiction	Incorrect	Consistent
8.Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury N.C. *think* / *forget* it sounds convincing.	Entailment	Neutral	Contradiction	Incorrect	Inconsistent

Table 10: Examples of correct/incorrect labeling of (Hypothesis, Modified Hypothesis) and consistent/inconsistent labeling of (Premise, Modified Hypothesis). P, H, H' denotes premise, hypothesis, modified hypothesis, respectively. The correctness of labeling is manually assigned for the labeling of (hypothesis, modified hypothesis). The words replaced in the hypothesis are in *original* / *modified*.