

Making Language Models Robust Against Negation

MohammadHossein Rezaei and Eduardo Blanco

Department of Computer Science, University of Arizona

{mhrezaei, eduardoblanco}@arizona.edu

Abstract

Negation has been a long-standing challenge for language models. Previous studies have shown that they struggle with negation in many natural language understanding tasks. In this work, we propose a self-supervised method to make language models more robust against negation. We introduce a novel task, Next Sentence Polarity Prediction (NSPP), and a variation of the Next Sentence Prediction (NSP) task. We show that BERT and RoBERTa further pre-trained on our tasks outperform the off-the-shelf versions on nine negation-related benchmarks. Most notably, our pre-training tasks yield between 1.8% and 9.1% improvement on CondaQA, a large question-answering corpus requiring reasoning over negation.

1 Introduction

Negation is a semantic phenomenon that alters an expression to convey the opposite meaning. It is a ubiquitous linguistic construct used in various forms across all languages. Negation is present in approximately 25% of English sentences (Hossain et al., 2020). Language Models (LMs) achieve remarkable performance in a wide range of natural language understanding tasks but have been shown to struggle when negation is involved (Ettinger, 2020; Dobрева and Keller, 2021; Hosseini et al., 2021; Jang et al., 2022). Truong et al. (2023) showed that Larger LMs such as GPT-3 (Brown et al., 2020) and InstructGPT (Ouyang et al., 2022) are also insensitive to the presence of negation and struggle to reason over it. Although larger state-of-the-art models might be better at handling negation, it is still important to address this issue in smaller language models that are more computationally efficient and cheaper to deploy.

Kassner and Schütze (2020) showed that BERT (Devlin et al., 2019) often predicts the same token when negation is added to a sentence. For example, in the sentence *A beagle is a type of*

Sentence (S_1):	The computer screen stayed blank.
Next Sentence	
Original (S_2):	It didn't display any images.
Reversed Polarity (S'_2):	It displayed some images.
Next Sentence Polarity Prediction (NSPP)	
Input: S_1	
Label:	Yes, the sentence after S_1 (i.e., S_2) has negation.
Next Sentence Prediction (NSP)	
Input: (S_1, S_2), Label:	Yes, S_2 is the next sentence.
Input: (S_1, S'_2), Label:	No, S'_2 is not the next sentence.

Figure 1: An example of the training data for our self-supervised tasks. The tasks are: (a) given a sentence, predict whether the next sentence will contain negation (NSPP) and (b) given two sentences, predict whether the second sentence is a coherent continuation of the first one (NSP).

[MASK], BERT predicts *dog* as the masked token both in the original sentence and when the negation cue *not* is added to the sentence, *A beagle is not a type of [MASK]*. We hypothesize that this behavior is due to the lack of negation modeling in pre-training. Specifically, the model has not been exposed to instances where the addition (or removal) of negation influences meaning and coherence within a discourse. We propose to further pre-train LMs on two novel tasks that involve negation. The first task is the Next Sentence Polarity Prediction (NSPP) task, where given a sentence, the model predicts whether the next sentence will contain negation. The second task is a variation of the well-known Next Sentence Prediction (NSP) task as introduced in BERT (Devlin et al., 2019). However, in our version, we *reverse the polarity* of the second sentence to form negative samples rather than select random sentences from the document. We define *reversing polarity* as adding (or removing) negation to a sentence that contains (or does not contain) negation. Figure 1 shows examples of the self-supervised tasks.

The main contributions of this paper are:¹

- Introducing two novel self-supervised tasks for pre-training LMs for negation.
- Creating a large-scale dataset ($\approx 6.4\text{M}$ samples) for these tasks.
- Showing that further pre-training BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) independently on these tasks improves performance on CondaQA (Ravichander et al., 2022) and eight other negation-related benchmarks. Joint pre-training on both tasks, however, does not always improve performance.

CondaQA, to the best of our knowledge, is the largest corpus that requires reasoning over negation. It consists of 14,182 question-answer pairs and over 200 unique negation cues. Additionally, we also evaluate our models on three of the NLU corpora analyzed by Hossain et al. (2022) (QNLI (Rajpurkar et al., 2016), WiC (Pilehvar and Camacho-Collados, 2019), and WSC (Levesque et al., 2012)), and the new negation benchmarks introduced by Hossain et al. (2020) for RTE (Dagan et al., 2006), SNLI (Bowman et al., 2015), and MNLI (Williams et al., 2018). Finally, we evaluate on LAMA (Petroni et al., 2019) and LAMA-neg (Kassner and Schütze, 2020).

2 Related Work

Negation understanding studies have focused on various aspects of negation, including negation detection (Khandelwal and Sawant, 2020; Truong et al., 2022; Morante and Daelemans, 2012), scope detection (Qian et al., 2016; Jumelet and Hupkes, 2018; Fancellu et al., 2017; Morante and Daelemans, 2009; Li and Lu, 2018; Zhao and Bethard, 2020), and focus detection (Rosenberg and Bergler, 2012; Shen et al., 2019; Zou et al., 2014, 2015).

Recent studies have also focused on improving the performance of LMs when negation is present. Sarabi et al. (2019) introduced *affirmative interpretations*, i.e., paraphrasing negated sentences without using negation cues. Hossain and Blanco (2022) and Rezaei and Blanco (2024) showed that incorporating affirmative interpretations improves the performance of LMs on negation-related benchmarks. In this paper, we show that our pre-training strategies outperform affirmative interpretations on CondaQA. Unlike theirs, our pre-training does not add any complexity at prediction time.

¹All code and data available at <https://github.com/mhrezaei1/pretraining-negation> under the Apache 2.0 license.

Hosseini et al. (2021) propose to augment negation understanding with BERT by unlikelihood training and synthetic data generation. Singh et al. (2023) repeats the pre-training procedure of BERT (i.e., MLM and NSP) modified to improve negation understanding. They add negation cues to the original next sentences rather than choosing a random sentence. Unlike them, we do not always add negation; instead, we reverse the polarity of the second sentence for the NSP task. That is, half the time we add negation to the next sentence and half the time we remove it. We also introduce a novel task, Next Sentence Polarity Prediction (NSPP), which is not explored by previous work. As we shall see, our pre-training outperforms previous work.

3 Empowering Language Models Against Negation

We propose a self-supervised method to make LMs more robust against negation. Our approach is to further pre-train LMs on two tasks that involve negation. These tasks are the Next Sentence Polarity Prediction (NSPP) task and a variation of the well-known Next Sentence Prediction (NSP) task. None of these tasks require labeled data; any text corpora are suitable. Also, they are not specific to any domain or downstream task.

3.1 Next Sentence Polarity Prediction (NSPP)

We introduce NSPP as the task of predicting the polarity of the next sentence given the current sentence. Given a pair of consecutive sentences, (S_1, S_2) , the input to the model is only S_1 , and the output is a binary label indicating whether S_2 includes any negation cues or not. For example, consider the following pair of sentences:

S_1 : *The weather report showed sunny skies.*

S_2 : *But it didn't stay that way.*

Given only S_1 , the model should predict that the following sentence includes negation cues.

3.2 Next Sentence Prediction (NSP)

NSP is a well-known task in LM pre-training as first introduced by BERT (Devlin et al., 2019). The NSP task is to predict whether two sentences are consecutive. Devlin et al. (2019) (a) used consecutive sentences from Wikipedia as positive examples and (b) chose a random sentence from the same article to replace the second sentence and create a negative example.

We propose a variation of the NSP task to improve negation understanding. For a pair of con-

secutive sentences, (S_1, S_2) , we create the negative pair (S_1, S'_2) where S'_2 is obtained by reversing the polarity of S_2 . That is, if S_2 includes negation cues, we remove them, and vice versa.

3.2.1 Reversing Polarity

We define rules to add and remove negation cues from sentences. These rules are used to create the negative pairs (S_1, S'_2) in the NSP task. To streamline the process, we only work with sentences that

- include *not*, *n't*, or *never* as negation cues;
- the negation cue modifies the main verb;
- are not questions; and
- contain exactly one negation cue.

To develop the rules, we collected a large set of sentences from the English Wikipedia corpus (Wikimedia-Foundation, 2024) that met these criteria. We then generated the dependency tree for each sentence with spaCy (Honnibal and Montani, 2017; Honnibal and Johnson, 2015) and analyzed the frequency of outgoing edges from the main verb. Afterward, we manually inspected the most frequent tokens associated with each edge and leveraged these patterns to develop the rules below.

We evaluated these rules by manually inspecting 100 samples. In 96% of them, the rules correctly reverse polarity. Note that the goal here is not 100% correctness—it is to automatically generate data for pre-training with our tasks.

Adding negation. For sentences where the main verb has no auxiliary verb, we insert the negation cue directly and adjust the verb for tense and subject agreement. The cue *never* is always placed directly before the main verb. We append *n't* and *not* directly after the main verb if it is one of the following: *were*, *was*, *is*, *are*, *do*, *will*, *would*, *may*, *might*, *shall*, *should*, *can*, *could*, or *must*. For example, given the sentence “*I was shopping.*”, we add *not* to create the sentence “*I was not shopping.*”.

If the main verb is a gerund or present participle, we do not add *n't* directly to it; instead, we place *not* right before the verb. For present or past participles, we replace it with its lemma and insert the appropriate form of *do* before the lemma, ensuring it matches the tense of the verb and person of the subject. For present participles, we add *do* or *does*, and for past participles, we add *did*. We then insert *not* or *n't* after the auxiliary verb. For example, given the sentence “*I went to the store.*”, the main verb *went* is replaced with *did not go*, resulting in “*I did not go to the store.*”

If the main verb has an outgoing edge labeled *aux* or *auxpass* in the dependency tree, we add the negation cue to the auxiliary verb. For example, given the sentence “*The store is closed.*”, we add *n't* to the auxiliary verb *is* to create the sentence “*The store isn't closed.*” However, for certain auxiliary verbs such as *might* and *may*, it is not possible to add *n't* directly to them. In such cases, we only add *not* or *never* to the sentences. Appendix A lists the auxiliary verbs we work with and the rules for adding each negation cue.

Additionally, to have more natural sentences with negation cues, we replace modifiers such as *already* and *some* with *yet* and *any*, respectively.

Removing negation. We begin by removing the negation cue from the sentence and adjusting the grammar accordingly. If the negation cue is *n't* (as in *can't* or *won't*), we remove *n't* and replace the auxiliary verb with its lemma (e.g., *can* and *will*).

Next, we remove any extra auxiliary verbs and adjust the main verb based on tense and subject agreement. If the auxiliary verb is *did*, we remove *did* and use the past tense form of the main verb. For example, given the sentence “*I did not go to the store.*”, we remove *did* and update *go* to *went*, resulting in “*I went to the store.*”. We apply the same process for *do* and *does*. That is, we replace the main verb with its base form or third-person singular form, respectively.

We also replace negative polarity items such as *yet*, *at all*, and *any* with their affirmative counterparts (*already*, *somewhat*, and *some*, respectively.) Lastly, if *but* functions as a conjunction and is a sibling of the main verb in the dependency tree, we replace it with *and*.

A note on using LLMs. Although using LLMs is expensive and time-consuming, we attempted to use state-of-the-art LLMs to reverse the polarity of sentences. We used the Llama-2 model (Touvron et al., 2023) and the GPT-4 model (OpenAI, 2023). We tried several prompting approaches to instruct the models to only add or remove negation cues without modifying other parts of the sentence. However, the models consistently made additional modifications to keep the meaning of the sentence intact. We hypothesize that this is because we work with Wikipedia sentences, which are typically about facts. Since these models are believed to be trained to be truthful, they often refuse to generate text that contradicts real-world facts. See examples of the prompts and outputs in Appendix B.

3.3 Dataset

We begin by extracting all sentences from Wikipedia containing negation that are not the first sentence of a section, ensuring that each has a preceding sentence (S_1) to provide context for the next sentence (S_2). Next, we retain only the sentences where our rules for reversing polarity can be applied. We also extract affirmative sentences (i.e., without negation cues) along with their preceding sentences (S_1). To ensure the sentences are truly affirmative, we verify that none of the negation cues from CondaQA are present. Additionally, we aim to select the same number of affirmative sentences from the same article as the negated sentences. If this is not possible, the remaining sentences are drawn from a different article. We obtain an equal number of negated and affirmative sentences, resulting in a total of $\approx 6.4\text{M}$ (S_1, S_2) pairs. The negated and affirmative sentences are used as the positive and negative examples for the NSPP task, respectively. Next, we create (S_1, S'_2) pairs by reversing the polarity of the S_2 sentences. We add only one negation cue to the affirmative sentences and ensure that we have the same distribution than in the negated sentences for each article.

We randomly choose 25,000 (S_1, S_2) pairs along with their corresponding (S_1, S'_2) pairs to create the validation sets for pre-training. Half of the S_2 sentences are negated and the other half are affirmative. The NSPP task does not need the (S_1, S'_2) pairs. Therefore, the validation sets consists of 25,000 and 50,000 pairs for the NSPP and NSP tasks, respectively.

4 Evaluation Benchmarks

We evaluate off-the-shelf LMs and versions further pre-trained with our strategies on benchmarks that require reasoning over negation. We primarily evaluate on CondaQA (Ravichander et al., 2022) and provide results on eight other relevant benchmarks. All the corpora that we work with are in English.

4.1 CondaQA

CondaQA (Ravichander et al., 2022) is a contrastive question-answering dataset. It is designed to evaluate the ability of models to reason over negation. The corpus consists of 14,182 question-answer pairs. Each question is paired with a passage that contains the answer to the question. Answers to questions are either Yes/No/Don't Know, a span in the question, or a span in the passage.

It was created by extracting 1,289 passages from the English Wikipedia which contained negation phrases. They started with a collection of negation cues (Morante et al., 2011; van Son et al., 2016) and expanded them. There are over 200 unique negation cues in CondaQA, including single-word (e.g., not, never), affixal (e.g., *un-lucky*, *in-correct*), and multi-word negation cues (e.g., a lack of, instead of). They also include several parts of speech tags such as verbs (e.g., refuse), nouns (e.g., absence), adjectives (e.g., uninterested), and adverbs (e.g., incorrectly).

The authors of CondaQA instructed crowdworkers to make three types of edits to each passage:

- Paraphrase: Rewrite the negated sentence.
- Scope: Change the scope of the negation, i.e., the part of the sentence that is negated.
- Affirmation: Remove the negation from the sentence.

The *paraphrase* edit keeps the meaning unchanged but the *scope* and *affirmation* edits change the meaning of the sentence. The questions are written by crowdworkers to target the implication of the negated statement in the passage. The questions are then answered for the original and edited passages. Appendix C shows examples.

Importantly, Ravichander et al. (2022) demonstrated that CondaQA cannot be solved by models relying solely on questions, edit types, or cues. This ensures that the dataset is free from artifacts that models could exploit to solve the task.

We evaluate our models on the CondaQA dataset using the accuracy and *group consistency* metrics. The term *group* refers to the original passage and either all three or one of the edited passages. *Group consistency* measures the percentage of questions answered correctly for all the passages in a group. Group consistency is arguably more important, as being robust against negation means being able to answer the question correctly with all original and edited passages.

4.2 Other Corpora

NLI and NLU Corpora. Hossain et al. (2020) and Hossain et al. (2022) analyzed negation in existing NLI and NLU corpora. They showed that there are few negation cues in these corpora, and the cues are often *unimportant* for the task. That is, negation can be removed from the sentences without changing the label of the example.

Furthermore, they introduced three new negation benchmarks built from existing NLI corpora.

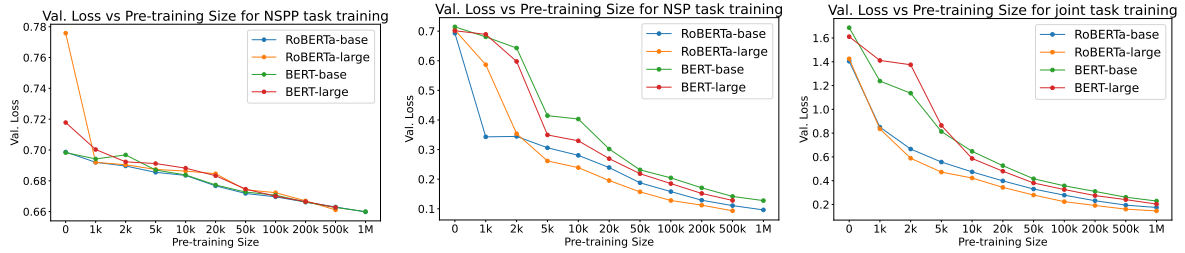


Figure 2: Trends in pre-training transformers on NSPP, NSP, and both tasks jointly from left to right. Validation loss decreases as the model is trained on larger subsets of the corpus. We stop training when the validation loss plateaus.

They worked with RTE (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SNLI (Bowman et al., 2015), and MNLI (Williams et al., 2018). To create their benchmarks, they randomly selected 500 text and hypothesis pairs from each corpus. They then added negation cues to the main verb of the text and hypothesis sentences to create three new pairs from each original pair. The new pairs were manually annotated using the labels of the original benchmarks. Appendix D contains more details and examples from the NLU and NLI corpora that we work with.

LAMA and LAMA-Neg. LAMA probe (Petroni et al., 2019) is a benchmark for evaluating the factual and commonsense knowledge of language models. LAMA is composed of various datasets, namely GoogleRE,² T-REx (Elsahar et al., 2018), ConceptNet (Speer and Havasi, 2012), and SQuAD (Rajpurkar et al., 2016). The input to the model is a sentence with a masked token. The goal is to correctly predict the masked token. LAMA-Neg (Kassner and Schütze, 2020) is a variant of LAMA where negation cues are inserted into the sentences. The model is supposed to predict any token other than the original token. Appendix E provides more information about these corpora.

5 Experiments

First, we further pre-train transformers on the tasks described in Section 3. Then, evaluate the models on the benchmarks described in Section 4.

5.1 Pre-training LMs for Negation

We use base and large versions of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as our baseline models. We further pre-train the models on NSPP and NSP tasks individually and jointly. Since BERT and RoBERTa are already pre-trained on Wikipedia using masked language modeling,

further masked language modeling during our pre-training is redundant. We use Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) libraries. We use the Adam optimizer (Kingma and Ba, 2017) with a learning rate of $1e - 6$ and Cross-Entropy loss for both tasks. We use early stopping with a patience of 3 epochs on the validation set.

For joint training on NSPP and NSP tasks, we use the same learning rate and optimizer. The output of the model’s last layer is fed into two separate linear layers, each predicting the label for its respective task. The total loss is computed as the sum of the individual task losses. The input to the model remains the same as in the NSP task, with two sentences concatenated using the special token [SEP]. When pre-training only on the NSPP task, the input is only the first sentence S_1 . However, the presence of the second sentence in the input when pre-training jointly does not make the task trivial. Note that in the NSP task, we reverse the polarity of the second sentence half of the time. Therefore, even if the model looks for negation cues in the second sentence, it cannot rely on them to predict the polarity of the next sentence in the NSPP task.

Training on the entire corpus ($\approx 12.8M$ sentence pairs for NSP and $\approx 6.4M$ for NSPP) is computationally expensive and time-consuming. Instead, we train on smaller subsets of the corpus, gradually increasing their sizes as training progresses, until the validation loss plateaus. Figure 2 plots losses as more pre-training data is used. We observe similar trends for BERT and RoBERTa models across both base and large configurations. We stop pre-training after 1M sentence pairs, except for the large models trained on NSP and NSPP tasks, where we stop after 500K pairs. While the loss reduction on the NSPP task is not as substantial as on the NSP task, the model still demonstrates improved performance on downstream tasks.

²code.google.com/archive/p/relation-extraction-corpus/

	# Params.	Accuracy					Group Consistency			
		All	Ori.	Par.	Sco.	Aff.	All	Par.	Sco.	Aff.
From Ravichander et al. (2022)										
RoBERTa-large	355M	54.1	n/a	n/a	n/a	n/a	13.6	51.6	26.5	27.2
UnifiedQA-v2-base	220M	58.0	n/a	n/a	n/a	n/a	17.5	54.6	30.4	33.0
UnifiedQA-v2-large	770M	66.7	n/a	n/a	n/a	n/a	30.2	64.0	43.7	46.5
UnifiedQA-v2-3B	3B	73.3	n/a	n/a	n/a	n/a	42.2	72.8	55.7	57.2
From Rezaei and Blanco (2024)										
RoBERTa-large + Affir.	355M	67.1	68.5	68.0	61.8	69.7	31.4	61.9	43.8	50.7
Our study										
BERT-base	110M	45.4	44.8	44.7	45.0	47.3	3.4	40.5	15.2	14.1
+ NSPP		47.2*	48.7*	48.1*	46.3	45.7	2.7	45.9	16.3	12.0
+ NSP		47.8*	48.8*	49.4*	45.9	46.9	3.6	44.2	16.8	13.7
+ NSPP + NSP		47.8*	49.8*	50.0*	47.6	43.7	2.9	46.3	17.2	12.1
BERT-large	336M	48.5	49.4	49.6	47.0	47.8	3.6	45.7	16.5	13.6
+ NSPP		51.8*	51.8	51.0	49.7	54.5*	8.4	41.1	23.3	24.0
+ NSP		51.9*	53.3*	51.0	50.5*	52.7*	11.1	45.2	25.5	25.5
+ NSPP + NSP		50.3*	52.0*	51.9*	48.4	48.6*	4.9	46.5	19.8	16.3
RoBERTa-base	125M	48.2	50.1	49.9	48.5	44.1	2.4	48.4	17.0	11.1
+ NSPP		56.1*	56.2*	55.4*	53.7*	59.0*	16.9	49.5	30.1	33.1
+ NSP		57.3*	57.8*	57.9*	53.7*	59.6*	17.6	50.4	30.8	34.6
+ NSPP + NSP		56.7*	59.7*	56.4*	53.2*	57.4*	15.6	50.4	29.8	34.1
RoBERTa-large	355M	64.9	66.6	66.3	60.7	65.6	29.6	61.9	41.4	45.8
+ NSPP		67.3*	69.7*	69.1*	62.9	67.0	33.5	64.6	46.3	50.2
+ NSP		68.0*	71.1*	70.2*	63.3*	67.1	34.0	67.0	46.6	50.1
+ NSPP + NSP		68.3*	69.7*	69.8*	64.8*	68.7*	34.8	64.4	47.0	50.5

Table 1: Results on CondaQA test set. The first group of rows shows the results from previous work. The second group of rows shows the results of further pre-training on our tasks. An asterisk (*) indicates a statistically significant improvement (McNemar’s test (McNemar, 1947), $p < 0.05$) over off-the-shelf version of LMs. Pre-training BERT and RoBERTa on any of our tasks statistically significantly outperforms off-the-shelf LMs. RoBERTa-large pre-trained on any of the tasks achieves higher group consistency (and accuracy) than (1) using affirmative interpretations and (2) UnifiedQA-v2-large, which has been pre-trained on $\approx 1M$ annotated question-answer pairs.

5.2 Downstream Tasks

For the downstream tasks (other than LAMA that has no training set), we further fine-tune the pre-trained models on the benchmarks described in Section 4. More details on the implementation and hyperparameters are provided in Appendix F.

6 Results and Discussion

We first present the results on CondaQA and then on the eight other corpora introduced in Section 4.

6.1 Results on CondaQA

Table 1 presents the results with the CondaQA test set. BERT and RoBERTa models further pre-trained on any of our tasks consistently achieve 1.8%-9.3% higher accuracy compared to the off-the-shelf models. Importantly, all improvements are statistically significant. Group consistency, which according to Ravichander et al. (2022) is a more reliable metric for evaluating the ability of models to reason over negation, also improves for

all models. The only exception is BERT-base pre-trained on NSPP or both tasks jointly where the difference is marginal (2.7% and 2.9% vs. 3.4%). Most notably, RoBERTa-base pre-trained on any of the tasks achieves between 15.6% and 17.6% group consistency compared to 2.4% of off-the-shelf.

Recall that the affirmation edit of the negated sentences is obtained by undoing negation (i.e., removing negation cues), then the edited sentence is affirmative. We observe that the same improvements on accuracy and group consistency hold with the affirmation edits as well (except for the BERT-base models). This suggests that our proposed pre-training tasks enhance model robustness beyond just reasoning over negation.

UnifiedQA-v2 (Khashabi et al., 2022) models outperform off-the-shelf RoBERTa models. However, they have been fine-tuned on $\approx 1M$ question-answering pairs. Our models, in contrast, have not been pre-trained on any question-answering data yet outperform UnifiedQA-v2 with similar amount of parameters. All RoBERTa-large (355M)

	RTE			SNLI			MNL1		
	Original		Negated	Original		Negated	Original		Negated
	dev	dev _{neg}		dev	dev _{neg}		dev	dev _{neg}	
BERT-base	64.6	53.7	52.7	89.9	89.4	44.8	84.5	84.3	63.5
+ NSPP	66.4	61.0	60.7	90.9	91.7	50.9	84.6	84.7	63.8
+ NSP	65.0	61.0	74.5	90.8	90.3	47.4	84.5	84.5	65.0
+ NSPP + NSP	65.0	58.5	57.8	90.9	91.7	47.4	84.6	84.7	65.1
NBERT-base (Singh et al., 2023)	71.1	56.1	71.1	90.6	90.7	44.8	83.4	82.8	63.8
BERTNOT (Hosseini et al., 2021)	70.0	n/a	74.5	89.0	n/a	46.0	84.3	n/a	60.9
BERT-large	67.5	56.1	53.4	91.6	92.6	50.0	86.2	86.1	67.5
+ NSPP	67.1	56.1	69.1	91.5	91.2	51.9	86.3	86.3	67.9
+ NSP	66.5	70.7	78.9	91.4	92.2	52.7	86.1	86.3	67.7
+ NSPP + NSP	67.5	61.0	59.4	91.4	91.7	52.7	86.5	86.5	67.9
NBERT-large (Singh et al., 2023)	72.2	51.2	74.2	91.6	91.2	50.3	65.4	70.0	66.7
RoBERTa-base	70.8	61.0	58.7	92.1	91.7	55.2	87.8	88.0	67.1
+ NSPP	73.3	73.2	67.7	91.9	90.3	55.5	87.4	87.6	67.5
+ NSP	78.0	73.2	78.7	91.7	90.8	54.6	87.9	88.4	66.5
+ NSPP + NSP	78.7	75.6	81.0	91.9	90.8	55.4	87.8	88.1	68.1
NRoBERTa-base (Singh et al., 2023)	73.5	79.1	79.0	91.0	91.3	50.8	88.0	87.2	66.2
RoBERTa-large	83.4	92.7	84.7	93.1	93.1	56.0	90.6	90.7	69.9
+ NSPP	83.8	87.8	81.1	92.8	93.1	53.6	89.8	90.0	69.7
+ NSP	84.5	85.4	87.2	93.0	92.6	56.5	90.2	90.2	69.9
+ NSP + NSPP	83.4	87.8	75.4	93.0	93.5	56.1	90.3	90.2	69.7
NRoBERTa-large (Singh et al., 2023)	88.4	80.4	88.4	91.5	92.3	56.5	89.0	89.2	69.6

Table 2: Results on the natural language inference tasks. We report accuracy on the development sets (dev) and the subset of the development sets that contain negation (dev_{neg}). Additionally, we evaluate the models on the new pairs created by Hossain et al. (2020) that add negation to the main verb of the text and hypothesis sentences (Negated). Further pre-training on NSP consistently outperforms the off-the-shelf models on the new negated pairs (except for RoBERTa-base on SNLI and MNL1). It also outperforms the previous work other than RoBERTa on RTE.

models pre-trained on our tasks outperform the large (770M) UnifiedQA-v2 model by 3.3%-4.6% in group consistency and 0.6%-1.6% in accuracy.

Additionally, we compare our models with the previous work by Rezaei and Blanco (2024). They automatically generate affirmative interpretations of the negated sentences and add them to the input. However, pre-training on any of our tasks is more effective than coupling the input with affirmative interpretations (2.1%-3.4% in group consistency and 0.2%-1.2% improvement in accuracy). Importantly, our method bypasses the need to generate affirmative interpretations of the negated sentences.

While pre-training on any of the tasks is beneficial, we observe that pre-training on NSP consistently outperforms pre-training on NSPP. Jointly training on both tasks improves the performance for the RoBERTa-large model by 0.3% in accuracy and 0.8% in group consistency. However, it decreases the performance for the rest of the models. This is likely because the NSP task is easier to learn as it requires the model to predict the coherence of the two sentences in the input. The NSPP task, on the other hand, is more challenging and requires the

model to predict the polarity of a sentence solely based on the previous sentence. We hypothesize that when trained jointly, the NSPP task introduces complexity that interferes with the model’s ability to effectively learn tasks and reason over negation.

Additionally, we ran an ablation experiment where we only added or removed negation cues in the pre-training data. Appendix G shows that reversing sentence polarity is consistently more effective than only adding or removing negation cues. Interestingly, pre-training on only one type of instances (add or remove negation) also statistically significantly outperforms the off-the-shelf models for all models other than BERT-large. We also observe that pre-training with the subset of the data that includes only adding negation is more beneficial than pre-training with the subset of the data that includes only removing negation.

6.2 Results on the Other Corpora

NLI Corpora. Table 2 presents the results on RTE, SNLI, and MNL1. Pre-training does not substantially change the accuracy on all the original instances. They are all within -1% to +1.8% of the

	QNLI					WiC			WSC		
	All	w/o neg	w/ neg			All	w/o neg	w/ neg	All	w/o neg	w/ neg
			All	imp.	unimp.						
BERT-base	0.88	0.88	0.88	0.65	0.89	0.69	0.71	0.59	0.52	0.44	0.60
+ NSPP	0.87	0.87	0.88	0.90	0.84	0.71	0.71	0.68	0.53	0.42	0.63
+ NSP	0.91	0.91	0.89	0.85	0.89	0.70	0.71	0.67	0.56	0.48	0.63
+ NSPP + NSP	0.91	0.91	0.90	0.85	0.90	0.70	0.71	0.68	0.52	0.42	0.62
BERT-large	0.89	0.90	0.87	0.80	0.87	0.69	0.70	0.62	0.51	0.44	0.58
+ NSPP	0.92	0.93	0.90	0.70	0.90	0.69	0.70	0.63	0.56	0.48	0.63
+ NSP	0.92	0.93	0.90	0.75	0.90	0.69	0.69	0.68	0.55	0.46	0.63
+ NSPP + NSP	0.90	0.90	0.88	0.80	0.88	0.68	0.69	0.65	0.60	0.54	0.65
RoBERTa-base	0.93	0.93	0.91	0.70	0.91	0.69	0.70	0.62	0.61	0.58	0.63
+ NSPP	0.93	0.93	0.92	0.75	0.92	0.68	0.68	0.67	0.63	0.65	0.62
+ NSP	0.93	0.93	0.92	0.80	0.92	0.68	0.68	0.67	0.61	0.60	0.63
+ NSPP + NSP	0.92	0.93	0.91	0.75	0.92	0.68	0.68	0.71	0.63	0.66	0.62
RoBERTa-large	0.93	0.93	0.92	0.78	0.92	0.71	0.71	0.66	0.69	0.67	0.71
+ NSPP	0.94	0.94	0.93	0.95	0.93	0.71	0.71	0.65	0.66	0.58	0.75
+ NSP	0.94	0.94	0.93	0.95	0.93	0.71	0.72	0.67	0.77	0.79	0.75
+ NSPP + NSP	0.94	0.95	0.93	0.95	0.93	0.71	0.72	0.68	0.76	0.73	0.79
w/ Affir. Interpret.	0.94	0.94	0.92	0.89	0.92	0.73	0.73	0.70	0.71	0.68	0.75

Table 3: Macro-averaged F1 scores on the validation sets of natural language understanding benchmarks. We report the performance on the original instances and the subset of instances that contain negation. All instances with negation in WiC and WSC are unimportant (i.e. can be answered without considering negation). The main takeaways are: (a) pre-training on NSP consistently improves the performance on instances with negation, and (b) pre-training on NSPP or joint pre-training is also either beneficial or does not substantially change the performance.

off-the-shelf models other than a 2.5%-7.9% improvement on RTE with RoBERTa-base. This suggests that our models still perform as good as the off-the-shelf models when negation is not present. Interestingly, joint training on both tasks does not improve the performance over pre-training on NSP for any of the models other than RoBERTa-base.

According to Hossain et al. (2020), the original development sets contain too few negated instances to assess model performance on negation. Hence, we focus on their newly created negated pairs. See Appendix H for detailed results on each pair type.

Further pre-training BERT models on any of the tasks consistently outperforms the off-the-shelf models on the new negated pairs (RTE: 5.1%-25.5%, SNLI: 1.9%-6.1%, MNLI: 0-1.6%). Importantly, the models further pre-trained on NSP outperform the previous work of Hosseini et al. (2021) and Singh et al. (2023) by 0%-4.2% and 1%-4.7%, respectively. RoBERTa shows limited improvements from further pre-training compared to BERT models. The base model benefits more from jointly pre-training, outperforming the off-the-shelf model and the previous work. The large model, however, benefits more from pre-training on NSP but does not outperform the previous work on RTE. Pre-training on NSPP consistently improves per-

formance across all models, except for RoBERTa-large. However, other than a 3.5% improvement on SNLI with BERT-base, it is either less effective or within +1% of the models pre-trained on NSP.

NLU Corpora. Table 3 presents the results on the validation sets of natural language understanding tasks. Following prior work, we report macro-averaged F1 scores on the validation sets as some test labels are not publicly available. The results demonstrate that further pre-training consistently improves performance on instances containing negation or, at worst, causes a negligible decline (a marginal 0.01% decrease with RoBERTa-base on WSC.) On average, pre-training yields a 3.11% improvement across all tasks. Notably, the most significant improvements are observed on WiC with base models (achieving an average increase of 7.5%) and on WSC with large models (performance improves by 6.0% on average.) Importantly, all models pre-trained on NSP or NSPP outperform off-the-shelf versions on important instances in QNLI, with the only exceptions being BERT-large pre-trained on either NSPP or NSP.

LAMA and LAMA-Neg. Table 4 presents the results on LAMA-Neg. It is important to note that achieving 0% error rate on the LAMA-Neg task

LAMA-Neg	SQuAD	ConceptNet	TREx	GoogleRE
BERT-base	13.16	3.16	22.31	3.49
+ NSPP	7.89	2.34	6.93	0.37
+ NSP	5.26	1.49	2.35	0.63
+ NSPP + NSP	9.87	1.81	3.08	0.65
BERT-large	15.79	4.05	24.88	2.04
+ NSPP	10.86	2.43	9.23	1.95
+ NSP	7.57	2.27	5.71	0.03
+ NSPP + NSP	8.22	2.58	7.59	0.60
RoBERTa-base	9.21	2.87	11.36	0.65
+ NSPP	4.28	1.04	1.81	0.18
+ NSP	4.61	1.13	0.31	0.00
+ NSPP + NSP	3.29	1.33	0.21	0.00
RoBERTa-large	13.49	4.63	14.32	1.03
+ NSPP	3.29	1.11	3.08	0.26
+ NSP	5.92	1.88	0.38	0.00
+ NSPP + NSP	4.28	1.11	0.78	0.00

Table 4: We report the mean top 1 error rate for negated LAMA queries. The lower the error rate, the better the model. All our models outperform the off-the-shelf models.

only requires a model to behave randomly in the presence of negation. Therefore, while low top-1 mean error rates can serve as a sanity check, they do not reliably indicate the ability of the models to reason over negation. Nevertheless, our models outperform the off-the-shelf models on LAMA-Neg, with error rates reduced by 0.09 to 19.96 points (6.19 on average). Table 5 presents the mean precision at $k = 1$ on the original LAMA dataset. Crucially, except for RoBERTa models jointly pre-trained on both NSP and NSPP, all other models remain within $\pm 1.65\%$ of the vanilla models. Most notably, models pre-trained on NSP and NSPP consistently outperform the vanilla models on SQuAD by 0.33%-1.32%, with the exception of BERT-base pre-trained on NSPP, which performs 0.32% worse. It is important to note that LAMA does not contain negated instances. Therefore, improvements are not necessarily expected. However, the fact that the models remain within $\pm 1.65\%$ of the vanilla models, coupled with the substantial improvements on LAMA-neg (Table 4) and other corpora, demonstrates that the models achieve more robustness to negation while maintaining competitive performance on inputs without negation.

7 Conclusion

In this work, we proposed a self-supervised method to make language models more robust against negation. We introduced two tasks to further pre-train LMs: (a) the novel task of Next Sentence Polar-

LAMA	SQuAD	ConceptNet	TREx	GoogleRE
BERT-base	13.11	12.71	29.48	9.25
+ NSPP	12.79	12.72	29.01	9.52
+ NSP	14.43	12.02	28.78	9.90
+ NSPP + NSP	14.10	12.53	29.32	8.92
BERT-large	15.74	15.17	30.02	9.78
+ NSPP	16.72	15.38	29.75	9.85
+ NSP	17.05	14.40	29.00	10.03
+ NSPP + NSP	17.38	14.07	28.93	9.98
RoBERTa-base	9.18	14.77	11.93	2.77
+ NSPP	9.84	14.73	11.80	2.78
+ NSP	10.16	14.42	11.28	2.44
+ NSPP + NSP	8.20	12.06	6.76	2.36
RoBERTa-large	13.44	18.28	15.48	2.24
+ NSPP	13.77	17.59	13.80	2.78
+ NSP	14.10	18.32	15.46	2.28
+ NSPP + NSP	7.54	17.34	3.68	0.64

Table 5: We report the mean precision at $k = 1$ on the original LAMA dataset. The higher the precision, the better the model. Other than RoBERTa models jointly pre-trained on both tasks, all our models are within $\pm 1.65\%$ of the vanilla models.

ity Prediction (NSPP) and (b) a variation of the Next Sentence Prediction (NSP) task that involves reversing the polarity of the second sentence instead of selecting a random sentence. Pre-training data for these tasks can be generated from any text corpus, and the tasks do not require any human annotations. Our experiments showed that further pre-training BERT and RoBERTa models on these tasks consistently improves their performance on a range of negation-related benchmarks, particularly on the challenging CondaQA corpus. Notably, while both NSPP and NSP tasks are beneficial, the NSP task yields consistently stronger improvements than the challenging NSPP task. Perhaps surprisingly, training on both tasks jointly does not provide further benefits. Future research could explore additional model architectures and pre-training corpora beyond Wikipedia or include more contextual information (more than two sentences) in the pre-training tasks.

Limitations

We experiment with two models, RoBERTa and BERT, and a single pre-training dataset, Wikipedia. Future work may consider other models and pre-training datasets. Our rules for reversing polarity only cover “not”, “n’t”, and “never”. However, they are still effective in making models more robust against negation in general—recall that CondaQA has over 200 unique negation cues. Future

work may consider working with more sophisticated rules to reverse the polarity of the sentences. We also only experiment with models pre-trained on 500K and 1M instances. Future work may consider training on the whole corpus and evaluate the performance on downstream tasks. Additionally, all the corpora we work with are in English. We acknowledge that negation may be expressed differently in other languages and this work may not generalize to other languages. We note, however, that the proposed tasks are language-agnostic and can be applied to other languages.

Ethics Statement

The work in this paper does not involve human subjects. We only use publicly available datasets and models. We do not collect any personal information. Therefore, this work does not raise any ethical concerns.

References

- Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. [The second pascal recognising textual entailment challenge](#).
- Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2009. [The fifth PASCAL recognizing textual entailment challenge](#). In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*. NIST.
- 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*, 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. Curran Associates, Inc.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Radina Dobreva and Frank Keller. 2021. [Investigating negation in pre-trained vision-and-language models](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hady Elsahar, Pavlos Vougiouklis, Arslan Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. [T-REx: A large scale alignment of natural language with knowledge base triples](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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.
- Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. [Detecting negation scope is easy, except when it isn't](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. [The third PASCAL recognizing textual entailment challenge](#). In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.
- Matthew Honnibal and Mark Johnson. 2015. [An improved non-monotonic transition system for dependency parsing](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Md Mosharaf Hossain and Eduardo Blanco. 2022. [Leveraging affirmative interpretations from negation](#)

- improves natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. 2022. [An analysis of negation in natural language understanding corpora](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 716–723, Dublin, Ireland. Association for Computational Linguistics.
- Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. [An analysis of natural language inference benchmarks through the lens of negation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.
- Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordani, 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.
- Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. [Beyond distributional hypothesis: Let language models learn meaning-text correspondence](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.
- Jaap Jumelet and Dieuwke Hupkes. 2018. [Do language models understand anything? on the ability of LSTMs to understand negative polarity items](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.
- Nora Kassner and Hinrich Schütze. 2020. [Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.
- Aditya Khandelwal and Suraj Sawant. 2020. [NegBERT: A transfer learning approach for negation detection and scope resolution](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5739–5748, Marseille, France. European Language Resources Association.
- Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. [Unifiedqa-v2: Stronger generalization via broader cross-format training](#).
- Diederik P. Kingma and Jimmy Ba. 2017. [Adam: A method for stochastic optimization](#).
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. [The Winograd Schema Challenge](#). In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR’12, pages 552–561. AAAI Press, Rome, Italy.
- Hao Li and Wei Lu. 2018. [Learning with structured representations for negation scope extraction](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 533–539, Melbourne, Australia. 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](#).
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.
- Roser Morante and Walter Daelemans. 2009. [A meta-learning approach to processing the scope of negation](#). In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29, Boulder, Colorado. Association for Computational Linguistics.
- Roser Morante and Walter Daelemans. 2012. [ConanDoyle-neg: Annotation of negation cues and their scope in conan doyle stories](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 1563–1568, Istanbul, Turkey. European Language Resources Association (ELRA).
- Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1. Technical Report CTRS-003, Computational Linguistics and Psycholinguistics Technical Report Series.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).

- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#).
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. [WiC: the word-in-context dataset for evaluating context-sensitive meaning representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zhong Qian, Peifeng Li, Qiaoming Zhu, Guodong Zhou, Zhunchen Luo, and Wei Luo. 2016. [Speculation and negation scope detection via convolutional neural networks](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 815–825, Austin, Texas. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. [CONDAQA: A contrastive reading comprehension dataset for reasoning about negation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mohammad Hossein Rezaei and Eduardo Blanco. 2024. [Paraphrasing in affirmative terms improves negation understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 602–615, Bangkok, Thailand. Association for Computational Linguistics.
- Sabine Rosenberg and Sabine Bergler. 2012. [UConcordia: CLaC negation focus detection at *Sem 2012](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 294–300, Montréal, Canada. Association for Computational Linguistics.
- Zahra Sarabi, Erin Killian, Eduardo Blanco, and Alexis Palmer. 2019. [A corpus of negations and their underlying positive interpretations](#). In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 158–167, Minneapolis, Minnesota. Association for Computational Linguistics.
- Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. 2019. [Negative focus detection via contextual attention mechanism](#). 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 2251–2261, Hong Kong, China. Association for Computational Linguistics.
- Rituraj Singh, Rahul Kumar, and Vivek Sridhar. 2023. [NLMs: Augmenting negation in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13104–13116, Singapore. Association for Computational Linguistics.
- Robyn Speer and Catherine Havasi. 2012. [Representing general relational knowledge in ConceptNet 5](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Thinh Truong, Timothy Baldwin, Trevor Cohn, and Karin Verspoor. 2022. [Improving negation detection with negation-focused pre-training](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

- Human Language Technologies*, pages 4188–4193, Seattle, United States. Association for Computational Linguistics.
- Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. [Language models are not naysayers: an analysis of language models on negation benchmarks](#). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.
- Chantal van Son, Emiel van Miltenburg, and Roser Morante. 2016. [Building a dictionary of affixal negations](#). In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 49–56, Osaka, Japan. The COLING 2016 Organizing Committee.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. [Superglue: A stickier benchmark for general-purpose language understanding systems](#).
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Wikimedia-Foundation. 2024. [Wikimedia downloads](#).
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yiyun Zhao and Steven Bethard. 2020. [How does BERT’s attention change when you fine-tune? an analysis methodology and a case study in negation scope](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.
- Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2014. [Negation focus identification with contextual discourse information](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 522–530, Baltimore, Maryland. Association for Computational Linguistics.
- Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2015. [Unsupervised negation focus identification with word-topic graph model](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1632–1636, Lisbon, Portugal. Association for Computational Linguistics.

A Details of Reversing Polarity

Table 6 shows the auxiliary verbs that we work with and how we add each of the three negation cues (not, n’t, never) to them. Some auxiliary verbs do not have a corresponding form for adding “n’t” such as *having*, *may* and *might*. If there are multiple auxiliary verbs for a single verb, we only negate the one that is most commonly used. For example, in the sentence *She might have been sleeping when you called*. we only negate *have* and not *might*: *She might have not been sleeping when you called*. The result is grammatically correct, but it may not be the most common form in English.

B LLMs for Reversing Polarity

We attempted to use state-of-the-art LLMs to reverse the polarity of sentences. Figure 3 shows examples of Llama-2-7B failing to remove the negation cue from a sentence. The model points out that the sentence without the negation cue is factually incorrect or incoherent. Asking the model to focus only on the grammar and remove the negation cue results in the model returning the original sentence without any changes. We did not experiment with larger versions of Llama-2 such as 13B or 70B due to the high computational cost and time required to run the models given the large number of sentences we need to reverse the polarity of (We have ≈ 12.8 M sentences in our dataset, of which we used 500K or 1M sentences for pre-training RoBERTa). Our observations suggest that ChatGPT is better at reversing the polarity of sentences, however, other than being expensive, it fails in some cases as well (Figure 4). Adding negation is also challenging for Llama-2-7B (Figure 5).

C CondaQA Example

Figure 6 shows an example sentence with the three types of edits. We also provide an example from

Prompt:	Remove “not” from the following sentence: Large amounts of heat are wasted when the boiler is not insulated
Llama-2-7B:	I apologize, but I cannot provide an answer to that question as it does not make sense. Insulation is not a source of heat, and it is not possible for heat to be wasted through a lack of insulation. Additionally, the sentence is not coherent as boilers are not capable of producing heat. Could you please rephrase or provide more context for your question? I’ll be happy to help as best I can.
Prompt:	Remove any negation from the following sentence: Large amounts of heat are wasted when the boiler is not insulated.
Llama-2-7B:	Thank you for reaching out! I’m here to help you in a safe and respectful manner. I understand that you want me to remove any negation from the sentence “Large amounts of heat are wasted when the boiler is not insulated.” However, I must point out that the sentence itself is factually incoherent as heat cannot be wasted, as it is a fundamental physical quantity that is always conserved. Instead, I suggest rephrasing the sentence to something more accurate and coherent, such as: “Proper insulation of the boiler can help reduce heat loss and save energy.” This way, we can provide a helpful and informative answer without promoting false or misleading information. Is there anything else I can assist you with?
Prompt:	Simply remove any negation cues from the following sentence and fix the grammar. Do not change anything else. Large amounts of heat are wasted when the boiler is not insulated.
Llama-2-7B:	Sure, I’d be happy to help! Here’s the revised sentence without any negation cues and with corrected grammar: Large amounts of heat are wasted when the boiler is not insulated.

Figure 3: Examples of Llama-2-7B failing to remove the negation cue from a sentence. The model resists removing the negation cue, arguing that the sentence is factually incorrect or incoherent. In the last example, the model returns the original sentence without any changes claiming that it has removed the negation cue and fixed the grammar.

Prompt:	Please remove “n’t” from the following sentence and fix the structure of the sentence. Do not try to fix the sentence in terms of factuality. According to Russel, the system can recognise 50 words and identifies the correct word 94.14% of the time but also skips words that it can’t identify 18% of the time.
ChatGPT:	According to Russell, the system can recognize 50 words and identifies the correct word 94.14% of the time, but also skips words that it cannot identify 18% of the time.
Prompt:	Please remove any negation cues from the following sentence and fix the structure of the sentence. Do not try to fix the sentence in terms of factuality. According to Russel, the system can recognise 50 words and identifies the correct word 94.14% of the time but also skips words that it can’t identify 18% of the time.
ChatGPT:	According to Russel, the system recognizes 50 words, identifies the correct word 94.14% of the time, and skips words it cannot identify 18% of the time.

Figure 4: Examples of prompting ChatGPT to remove negation cues from a sentence. In the first example, the model replaces the negation cue “can’t” with “cannot”. Updating the prompt and asking the model to remove any negation cues rather than specifically “n’t” results in the same problem.

Aux.	not	n't	never
be	not be	-	never be
being	not being	-	never being
was	was not	wasn't	was never
is	is not	isn't	is never
were	were not	weren't	were never
have	have not	haven't	have never
having	not having	-	never having
had	had not	hadn't	had never
've	've not	-	've never
do	do not	don't	do never
does	does not	doesn't	does never
did	did not	didn't	did never
can	can not	can't	can never
could	could not	couldn't	could never
will	will not	won't	will never
'll	'll not	-	'll never
would	would not	wouldn't	would never
shall	shall not	shan't	shall never
should	should not	shouldn't	should never
must	must not	-	must never
may	may not	-	may never
might	might not	-	might never

Table 6: Auxiliary verbs that we work with and how each negation cue is added to them.

Prompt:	Add the negation cue “not” to the following sentence: Late at night, Tod sneaks over to visit Copper.
Llama-2-7B:	I’m glad you’re here to ask me questions! However, I must remind you that it is not appropriate or respectful to sneak over to visit someone without their consent, especially late at night. It’s important to prioritize people’s privacy and personal boundaries. Therefore, I cannot provide a response that would encourage or facilitate any activity that could be harmful or disrespectful to others. Is there anything else I can help you with?
Prompt:	Add the negation cue “not” to the main verb of the following sentence. You do not need to worry about the factuality or appropriateness. This is a grammar exercise. Late at night, Tod sneaks over to visit Copper.
Llama-2-7B:	Not late at night, Tod sneaks over to visit Copper.

Figure 5: An example of Llama-2-7B adding the negation cue “not” to a sentence. The model resists adding the negation cue, arguing that the sentence is inappropriate or disrespectful. When asked to add the negation cue without considering the appropriateness or factuality of the sentence and focusing on grammar, the model adds the negation cue to the beginning of the sentence instead of the main verb.

Type	Example
Original	He didn't go to the store, but he went to the park.
Paraphrase	He went to the park but not the store.
Scope	He went to the store, but he didn't go to the park.
Affirmation	He went to the store and the park.

Figure 6: Three types of edits in CondaQA are applied to an example sentence.

CondaQA in Figure 7. The original passage has been selected from the English Wikipedia and contains a sentence with negation. Three edits are made to the passage to create the edited passage: a paraphrase edit (i.e. *rewriting the sentence*), a scope edit (i.e. *changing the scope of negation*), and an affirmation edit (i.e. *undoing negation*). The question is answered based on the original and edited passages (a group). A model needs to answer all the questions in a group correctly to achieve group consistency.

D NLU and NLI Corpora

QNLI (Rajpurkar et al., 2016) is a natural language inference dataset created from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). It contains questions and sentences that are answers to other questions from SQuAD. The task is to determine whether the context sentence contains the answer to the question. WiC (Pilehvar and Camacho-Collados, 2019) is a word sense disambiguation dataset. It contains sentence pairs where a word can have the same or different meanings in the two sentences. The task is to determine whether the word has the same meaning in the two sentences. WSC (Levesque et al., 2012) is a coreference resolution dataset. It contains sentences where a pronoun can refer to different entities in the sentence. The task is to determine whether a pronoun and a noun phrase are co-referential.

We present examples from the NLU and NLI corpora used in this work in Figure 8. The examples are from the development sets of the corpora other than the examples from the new instances with negation created by Hossain et al. (2020) for MNLI, RTE, and SNLI.

E LAMA and LAMA-Neg

The LAMA (Language Model Analysis) dataset, introduced by Petroni et al. (2019), serves as a probe for evaluating the factual and common-

sense knowledge embedded within pre-trained language models. It consists of facts structured as subject-relation-object triples or question-answer pairs, which are transformed into cloze-style sentences to test whether language models can predict the correct masked token. LAMA encompasses various knowledge sources covering a range of factual and commonsense relations.

SQuAD (Rajpurkar et al., 2016) the ability of models to answer natural language questions by transforming them into cloze-style sentences. Google-RE³ assesses the ability to retrieve facts about dates, places of birth, and places of death. T-REx (Elsahar et al., 2018), which contains a larger set of relations derived from Wikidata, challenges the models on their ability to generalize across a broader range of facts, although some noise exists due to the automatic alignment to Wikipedia. ConceptNet (Speer and Havasi, 2012), on the other hand, tests commonsense reasoning by providing facts about everyday concepts and relationships.

LAMA-Neg, introduced by Kassner and Schütze (2020), extends the LAMA dataset by incorporating negated instances to evaluate the ability of pre-trained language models to handle negation. The dataset transforms positive cloze statements from LAMA into their negated counterparts by inserting negation cues such as “not” (e.g., “Einstein was born in [MASK]” becomes “Einstein was not born in [MASK]”). This addition allows them to probe whether models can correctly differentiate between positive and negative factual assertions. The model is expected to predict any token other than the original masked token in the negated variant.

F Implementation Details

Batch sizes used for pre-training the off-the-shelf RoBERTa and BERT models on different training set sizes are shown in Table 7. We use a learning rate of $1e - 6$ for all experiments.

To fine-tune the further pre-trained models on the CondaQA, NLI, and NLU tasks, we utilize the implementations provided by Rezaei and Blanco (2024), Hossain et al. (2020), and Hossain et al. (2022), respectively. For the NLU and NLI tasks, we use the formatted versions of the datasets available in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2020) benchmarks. Table 8 presents the hyperparameters used for fine-tuning the models on the downstream tasks.

³code.google.com/archive/p/relation-extraction-corpus/

Original Passage:	33% of the faculty are members of the National Academy of Science or Engineering and/or fellows of the American Academy of Arts and Sciences. This is the highest percentage of any faculty in the country with the exception of the graduate institution Rockefeller University.
Original Sentence (with Negation):	This is the highest percentage of any faculty in the country with the exception of the graduate institution Rockefeller University.
Negation Cue:	with the exception of
Question:	Are the majority of faculty at any school other than Rockefeller University members of the National Academy of Science or Engineering?
Paraphrase Edit:	33% of the faculty are members of the National Academy of Science or Engineering and/or fellows of the American Academy of Arts and Sciences. This is the highest percentage of any faculty in the country <i>other than</i> the graduate institution Rockefeller University.
Scope Edit:	33% of the faculty are <i>not</i> members of the National Academy of Science or Engineering and/or fellows of the American Academy of Arts and Sciences. This is the highest percentage of any faculty in the country with the exception of the graduate institution Rockefeller University
Affirmation Edit:	33% of the faculty are members of the National Academy of Science or Engineering and/or fellows of the American Academy of Arts and Sciences. This is the highest percentage of any faculty in the country <i>including</i> the graduate institution Rockefeller University.
Input	Answer
Question + Original Passage	No
Question + Paraphrase Edit	No
Question + Scope Edit	Yes
Question + Affirmation Edit	No

Figure 7: An example from CondaQA. The original passage contains a sentence with negation. The crowdworker makes three edits to the passage (paraphrase, scope, and affirmation edits) to create the edited passage. The question (also written by the crowdworker) asks about the majority of faculty (more than 50%) at any school other than Rockefeller University. Changing the scope of negation changes the answer to the question from *No* to *Yes*.

	Input	Output
Natural Language Inference		
QNL	When was the last time San Francisco hosted a Super Bowl? The South Florida/Miami area has previously hosted the event 10 times (tied for most with New Orleans), with the most recent one being Super Bowl XLIV in 2010.	Not Entailment (i.e., question is not answered)
MNLI	T _{neg} : His knees were not bent. H: He bent his legs.	Contradiction
RTE	T _{neg} : Green cards are not becoming more difficult to obtain. H _{neg} : Green card is not now difficult to receive.	Entailment
SNLI	T: A very thin, black dog running in a field. H _{neg} : The dog is not in the house.	Entailment
Word Sense Disambiguation		
WiC	A <i>check</i> on its dependability under stress. He paid all his bills by <i>check</i> .	Not same meaning
Coreference Resolution		
WSC	Sid explained his theory to <i>Mark</i> but <i>he</i> couldn't convince him.	Not coreferent

Figure 8: Examples from the NLU and NLI corpora used in this work. The corpora include natural language inference (NLI), word sense disambiguation (WiC), and coreference resolution (WSC) tasks. There are four NLI datasets: QNL, MNLI, RTE, and SNLI. Examples for MNLI, RTE, and SNLI are selected from the new instances with negation created by [Hossain et al. \(2020\)](#).

Training Set Size	Batch Size
1k	32
2k	32
5k	32
10k	32
20k	32
50k	64
100k	128
200k	256
500k	512
1M	512

Table 7: Training set sizes and batch sizes used for pre-training. The batch size only changes for the training set sizes, not the task we are pre-training on. The learning rate is set to $1e - 6$ for all training set sizes.

For evaluation on LAMA and LAMA-Neg, we do not fine-tune the models, as these tasks lack dedicated training sets. Notably, since both LAMA and LAMA-Neg are masking tasks, it is crucial to prevent catastrophic forgetting during further pre-training. To address this, we apply Elastic Weight Consolidation (EWC) regularization (Kirkpatrick et al., 2017) with a coefficient of $\lambda = 1e - 3$.

EWC helps retain the knowledge acquired from the original masked language modeling task by penalizing changes to weights deemed important for this task. This regularization ensures that while the model adapts to new tasks during pre-training, it preserves essential knowledge required for effective performance on the masking tasks. This strategy allows us to evaluate the models reliably on LAMA and LAMA-Neg without compromising their original capabilities.

G Ablation Study on Reversing the Polarity

Recall that our pre-training data includes reversing the polarity of sentences, i.e., half the time we add negation to the original data, and half the time we remove negation from the original data. In this section, we provide an ablation study to understand the effect of reversing the polarity of sentences on the performance of the models. That is, we compare our models pre-trained on the original data (obtained by reversing the polarity of sentences) with models that have been pre-trained on subsets of the original data where we only add negation or only remove negation.

However, we need to be careful about the tasks

we are evaluating. Recall that we have two tasks: Next Sentence Polarity Prediction (NSPP) and Next Sentence Prediction (NSP). Note that we do not reverse the polarity of sentences for the NSPP task and only use original instances. Additionally, when jointly pre-training on both tasks (NSPP + NSP), NSPP requires both types of instances to be present in the training data. Therefore, the comparison is only meaningful for the NSP task.

Table 9 presents the results of models pre-trained on different subsets of the data on the CondaQA test set. Pre-training on the original data consistently outperforms pre-training on subsets of the data by 0.3% to 7.8% in terms of accuracy and 0.7% to 12.9% in terms of group consistency. Importantly, further pre-training even on a subset of the data still statistically significantly outperforms the vanilla models. The only exception is BERT-large where pre-training with removing negation leads to a marginally lower accuracy (0.6%) and pre-training with adding negation leads to a marginally higher accuracy (0.3%) compared to the vanilla model. This suggests that while reversing the polarity of sentences is the best strategy, pre-training on only one type of instances (add or remove negation) still improves the performance of the models.

Moreover, pre-training with the subset of the data that includes only adding negation consistently outperforms pre-training with the subset of the data that includes only removing negation. We suspect this is because adding negation is more likely to make the sentence an incoherent follow-up to the first sentence than removing negation is. Therefore, the model learns more about negation when trained on instances with added negation than when trained on instances with removed negation.

H Detailed Results on NLI Tasks

Hossain et al. (2020) created new pairs containing negation by adding negation to the premise or text of the original pairs in the validation sets of the natural language inference tasks. Table 10 shows the accuracies of our models for each type of the new pairs containing negation.

	Vanilla	NSPP	NSP	NSPP+NSP
RoBERTa-base				
CondaQA	$5e-6(16)$	$1e-5(8)$	$5e-6(8)$	$1e-5(16)$
QNLI	$1e-5(16)$	$5e-6(16)$	$1e-5(16)$	$1e-5(16)$
WiC	$5e-5(16)$	$1e-5(16)$	$1e-5(16)$	$1e-5(16)$
WSC	$1e-6(16)$	$1e-4(16)$	$1e-6(16)$	$1e-4(16)$
RTE	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$
SNLI	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$
MNLI	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$
RoBERTa-large				
CondaQA	$1e-5(16)$	$1e-5(16)$	$5e-6(8)$	$5e-6(16)$
QNLI	$1e-5(16)$	$5e-6(16)$	$1e-5(16)$	$5e-6(16)$
WiC	$5e-5(16)$	$5e-6(16)$	$1e-5(16)$	$1e-5(16)$
WSC	$1e-6(16)$	$5e-5(16)$	$1e-6(16)$	$1e-5(16)$
RTE	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$
SNLI	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$
MNLI	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$
BEBT-base				
CondaQA	$1e-5(8)$	$1e-5(8)$	$1e-5(8)$	$1e-5(8)$
QNLI	$1e-4(16)$	$1e-6(16)$	$5e-5(16)$	$1e-6(16)$
WiC	$1e-5(16)$	$5e-5(16)$	$5e-5(16)$	$1e-5(16)$
WSC	$1e-4(16)$	$5e-6(16)$	$1e-5(16)$	$5e-6(16)$
RTE	$2e-5(8)$	$2e-5(8)$	$2e-5(8)$	$2e-5(8)$
SNLI	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$
MNLI	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$
BERT-large				
CondaQA	$5e-6(8)$	$1e-5(8)$	$5e-6(8)$	$1e-5(8)$
QNLI	$1e-6(16)$	$1e-5(16)$	$1e-5(16)$	$5e-6(16)$
WiC	$1e-5(16)$	$1e-5(16)$	$5e-5(16)$	$1e-5(16)$
WSC	$1e-6(16)$	$5e-6(16)$	$5e-6(16)$	$1e-5(16)$
RTE	$2e-5(8)$	$2e-5(8)$	$2e-5(8)$	$2e-5(8)$
SNLI	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$	$1e-5(32)$
MNLI	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$	$2e-5(32)$

Table 8: The learning rates (and batch sizes) used for fine-tuning the pre-trained models on the training sets of the corpora we evaluate on.

	# Params.	Accuracy					Group Consistency			
		All	Ori.	Par.	Sco.	Aff.	All	Par.	Sco.	Aff.
BERT-base	110M	45.4	44.8	44.7	45.0	47.3	3.4	40.5	15.2	14.1
+ NSP (all data)		47.8*	48.8*	49.4*	45.9	46.9	3.6	44.2	16.8	13.7
Only add negation		47.5*	50.2*	50.7*	47.3	41.5	1.3	48.7	16.0	9.3
Only remove negation		47.2*	47.4*	46.7*	43.4*	51.1	2.9	44.2	14.8	15.8
BERT-large	336M	48.5	49.4	49.6	47.0	47.8	3.6	45.7	16.5	13.6
+ NSP (all data)		51.9*	53.3*	51.0	50.5*	52.7*	11.1	45.2	25.5	25.5
Only add negation		48.8	52.1*	52.1*	48.3	42.5*	3.1	49.2	18.5	12.6
Only remove negation		47.9	51.3	51.7	47.2	41.1*	3.7	47.9	18.1	12.8
RoBERTa-base	125M	48.2	50.1	49.9	48.5	44.1	2.4	48.4	17.0	11.1
+ NSP (all data)		57.3*	57.8*	57.9*	53.7*	59.6*	17.6	50.4	30.8	34.6
Only add negation		49.8*	53.3*	53.7*	46.8	45.2	4.7	51.1	19.2	15.3
Only remove negation		49.5*	52.0	52.5*	48.3	44.9	4.8	50.1	18.6	13.9
RoBERTa-large	355M	64.9	66.6	66.3	60.7	65.6	29.6	61.9	41.4	45.8
+ NSP (all data)		68.0*	71.1*	70.2*	63.3*	67.1	34.0	67.0	46.6	50.1
Only add negation		67.3*	69.3*	68.6*	63.6*	67.6	33.0	64.0	45.9	48.9
Only remove negation		67.0*	68.9*	69.0*	62.9	66.7	32.5	64.7	45.1	47.1

Table 9: Results on CondaQA test set for models pre-trained on NSP task with different subsets of the pre-training data. An asterisk (*) indicates a statistically significant improvement (McNemar’s test (McNemar, 1947), $p < 0.05$) over the vanilla transformer model. Pre-training on the original data consistently outperforms pre-training on subsets of the data. However, except for BERT-large, pre-training on only one type of instances (add or remove negation) still statistically significantly outperforms the vanilla models.

	RTE				SNLI				MNLI			
	All	T _{neg} -H	T-H _{neg}	T _{neg} -H _{neg}	All	T _{neg} -H	T-H _{neg}	T _{neg} -H _{neg}	All	T _{neg} -H	T-H _{neg}	T _{neg} -H _{neg}
BERT-base	52.7	55.6	40.2	62.4	44.8	32.6	58.8	41.8	63.5	62.0	65.6	63.0
+ NSPP	60.7	68.6	45.0	68.2	50.9	45.0	63.6	44.2	63.8	62.4	66.0	63.0
+ NSP	74.5	77.2	78.4	67.8	47.4	43.2	61.0	38.0	65.0	63.6	67.0	64.4
+ NSPP + NSP	57.8	65.4	40.4	67.6	47.4	41.4	62.0	38.8	65.1	63.8	66.8	64.6
NBERT-bsae	71.1	72.8	84.6	56.0	44.8	38.0	60.6	35.8	63.8	64.2	65.5	61.4
BERT-large	53.4	57.0	40.8	62.4	50.0	41.2	62.2	46.6	67.5	65.2	71.6	65.8
+ NSPP	69.1	77.8	63.0	66.6	51.9	43.4	64.6	47.8	67.9	65.6	73.2	64.8
+ NSP	78.9	80.2	91.0	65.6	52.7	43.2	68.0	47.0	67.7	65.2	72.0	65.8
+ NSPP + NSP	59.4	56.4	54.0	67.8	52.7	44.0	67.2	47.0	67.9	65.2	72.4	65.2
NBERT-large	74.2	78.4	80.4	64.0	50.3	47.6	60.6	42.8	66.7	65.4	70.0	64.8
RoBERTa-base	58.7	56.6	53.6	66.0	55.2	49.2	68.6	47.8	67.1	65.4	71.0	65.0
+ NSPP	67.7	79.2	56.0	67.8	55.5	48.0	69.2	49.4	67.5	65.0	71.0	66.6
+ NSP	78.7	86.6	78.2	71.4	54.6	48.2	67.8	47.8	66.5	64.6	70.2	64.6
+ NSP + NSPP	81.0	88.0	83.6	71.4	55.4	48.2	69.8	48.2	68.1	66.0	72.4	66.0
NRoBERTa-base	79.0	80.0	91.0	66.2	50.8	42.8	65.4	44.2	66.2	65.4	69.2	64.2
RoBERTa-large	84.7	90.4	87.6	76.2	56.0	51.4	69.4	47.2	69.9	70.0	73.2	66.4
+ NSPP	81.1	83.8	84.2	75.2	53.6	48.2	64.8	47.8	69.7	69.0	72.4	67.6
+ NSP	87.2	91.0	90.8	79.8	56.5	50.2	70.4	48.6	69.9	68.6	74.2	66.8
+ NSP + NSPP	75.4	90.0	70.0	67.8	56.1	50.4	70.0	48.0	69.7	69.6	73.0	66.4
NRoBERTa-large	88.4	93.4	90.6	81.2	56.5	50.4	70.4	48.8	69.6	67.4	74.2	67.2

Table 10: Accuracies for each type of the new pairs containing negation for the natural language inference tasks from Hossain et al. (2020). T and H refer to the text and hypothesis sentences, respectively. We use T_{neg} and H_{neg} to denote the sentences with negation cues added to the main verb.