The Self-Contained Negation Test Set

David Kletz[1,2] and **Pascal Amsili**[2] and **Marie Candito**[1]
(1) Université Paris Cité & LLF (CNRS/UPC)
(2) Université Sorbonne Nouvelle & Lattice (CNRS/ENS-PSL/USN)
david.kletz@sorbonne-nouvelle.fr, marie.candito@u-paris.fr, pascal.amsili@ens.fr

Abstract

Several methodologies have recently been proposed to evaluate the ability of Pretrained Language Models (PLMs) to interpret negation. In this article, we build on Gubelmann and Handschuh (2022), which studies the modification of PLMs' predictions as a function of the polarity of inputs, in English. Crucially, this test uses "self-contained" inputs ending with a masked position: depending on the polarity of a verb in the input, a particular token is either semantically ruled out or allowed at the masked position. By replicating Gubelmann and Handschuh (2022) experiments, we have uncovered flaws that weaken the conclusions that can be drawn from this test. We thus propose an improved version, the *Self-Contained Neg Test*, which is more controlled, more systematic, and entirely based on examples forming minimal pairs varying only in the presence or absence of verbal negation in English.

When applying our test to the `roberta` and `bert` base and large models, we show that only `roberta-large` shows trends that match the expectations, while `bert-base` is mostly insensitive to negation. For all the tested models though, in a significant number of test instances the top-1 prediction remains the token that is semantically forbidden by the context, which shows how much room for improvement remains for a proper treatment of the negation phenomenon.

1 Introduction

The treatment of negation by PLMs has recently been the subject of various works whose conclusions are fairly contradictory.

On the one hand, Kassner and Schütze (2020) and Ettinger (2020) compare the predictions of Transformer-based language models (Vaswani et al., 2017) in minimal pairs varying in polarity (1).

(1) a. A robin is a [MASK].
b. A robin is not a [MASK].

Noting that changes of polarity in the model's inputs result in little or no change for both top-1 predictions and the entire vocabulary distribution, these authors conclude that the models are insensitive to negation.

However, it has been established that the presence of negation can be detected in contextual representations. Celikkanat et al. (2020) thus find "traces" of negation on the negated verb, its subject, its object. Moreover, the extent to which negation is diffused in contextual representations follows syntactic constraints: Kletz et al. (2023) show that the presence of negation in contextual representations is stronger for tokens within the scope of negation, this effect being visible even when controlling for the distance between the token and negation.

As pointed out by Gubelmann and Handschuh (2022), this apparent contradiction can be explained by the fact that Kassner and Schütze (2020) study the factual knowledge of models, and therefore use contexts involving world knowledge, such as (1). The inability of models *not* to predict *bird* in the negated case could be explained by stored factual knowledge taking precedence over the ability to capture that negation reverses the truth value of a proposition. Especially as there is an asymmetry in the number of acceptable words to replace the mask: only a few are possible for the positive version, but a huge number are for the negative one, which can't be a favorable situation when only the top-1 prediction is studied.

Gubelmann and Handschuh (2022) have thus proposed a test where the inputs supplied to the models are self-contained (in our terminology): a context sentence is followed by a target sentence containing a masked position. The context sentence is either negative or affirmative. In the negative case, it renders semantically impossible a certain token at the masked position (*sail* in example (2)), which is itself plausible in the positive case (see section 2.2).

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 212–221

(2) Jessica is an architect who doesn't like to sail. However, she does like to [MASK].

Gubelmann and Handschuh (2022) observe a variable sensitivity to negation depending on the models tested, suggesting that the truth-value inversion effect of negation is more or less captured depending on the model.

In this article, we build on (Gubelmann and Handschuh, 2022) (hereafter **GH22**), taking up the idea of self-contained inputs, allowing us to target understanding of the semantics of negation independently of world knowledge. Our contributions are the following:

- a finer-grained analysis of GH22 experiments, uncovering a much more contrasted picture. In particular, averaged results for different input patterns mask significant sensitivity to factors other than negation (e.g., an intensifier *really* or *does*).
- the development of a more controlled test[1], using self-contained inputs organized in minimal pairs differing only in polarity, as well as the introduction of control tests (double negation, use of a non-negative adverb instead of *not*, and variations on coreference between NPs in the context sentence and the target sentence).

Finally, this test enables us to make a detailed assessment of four models, and to conclude that among these, only `roberta-large` reasonably meets the defined criteria. Crucially, a number of models like `bert-large` seemed reasonably sensitive to negation in GH22, do pass our baseline test, but don't pass the control tests at all, calling into question the positive interpretation of the baseline test. This highlights the many limitations to PLMs' understanding of negation for English, and the need for highly controlled tests to reach solid conclusions.

2 Replication of (Gubelmann and Handschuh, 2022)

2.1 Presentation of the test

The GH22 test aims at studying the tokens predicted at a masked position, within an input consisting of two sentences, a C(ontext) sentence followed by a T(arget) sentence. The actual examples provided as input to PLMs are obtained by instantiating variables within patterns: the context sentence C contains a variable ACT, to be instantiated with a verb (called **ACT-token**), like *sail* in (3), embedded in a negative (*doesn't like to ACT*) or affirmative (*tries to ACT as often as possible*) phrase. We refer to **Cn** and **Cp** as the negative and affirmative contexts. The sentence T contains a masked position, and is defined in such a way that repetition of the ACT-token is acceptable with an affirmative context (Cp), while semantically impossible with a negative context (Cn): for example, repeating *sail* in the masked position is plausible in (4) while semantically impossible in (3).

(3) NAME(Jessica) is PROF(an architect) who **doesn't like** to ACT(sail). However, PRON(she) does like to [MASK].

(4) NAME(Jessica) is PROF(an architect) who **tries** to ACT(sail) **as often as possible**. So, PRON(she) really **likes to** [MASK].

The metric proposed by GH22 is the rate of repetition of the ACT-token (**%-ACT-repetition**), *i.e.* the percentage of instantiated examples for which the top-1 at position MASK is the ACT-token itself. In the Cp case, a high repetition rate is acceptable, as the ACT-token is not mandatory at this position, but plausible. In the Cn case, a high %-ACT-repetition is clearly a sign of a failure of the model, by construction of the input examples. Note, however, that a weak %-ACT-repetition may be due to a good "understanding" of negation by the model, but may also stem from inconsistencies, if the model predicts agrammatical tokens in top-1 in this context.

GH22 also varied other parameters, such as the presence or absence of ACT-coordinated verbs in C[2], intensifiers *does* and/or *really* in T, and a discourse connective in T (contrastive *however* if C is negative, mplicative *so* if C is affirmative). Details of the combinations tested by GH22 are given Table 1.

2.2 Pattern instantiation

The authors generated the input examples by instantiating first the variables NAME (with typically feminine or masculine first names), PROF (with a profession), and PRON (third person pronoun, same gender as NAME). Then, for any instantiated triplet (NAME, PROF, PRON), the ACT-token is

[1] *https://github.com/davkletz/self-neg-test*

[2] As in *NAME is a PROF who doesn't like to ACT, ACT1 or ACT2.*

pol.	main v in C	aux	adv	conn.
N	doesn't like to	✓	-	✓
N	doesn't like to	✓	-	-
N	doesn't like to	-	-	-
P	tries to	✓	✓	✓
P	tries to	-	✓	-

Table 1: Details of the parameter combinations tested by GH22. Columns: *pol.* is the polarity of the context (negative or positive); *main v in C* indicates which verb was used in the C sentence; *aux* (resp. *adv*) indicates whether *does* (resp. *really*) is used in T; *conn.* indicates whether T begins with a connective (contrastive *however* for Cn, and implicative *so* for Cp). In addition, in GH22, all these configurations are combined with the gender of the subject proper noun (fem/masc) and with 0, 1 or 2 verbs coordinated to ACT.

.

chosen by considering the tokens predicted at the masked position in the sentence (5): either the first one, the 50th, 100th or 200th in the predicted distribution. As a consequence, the examples that will be given as input vary from one model to another, the ACT-token being adapted for each triplet and model).

(5) *NAME is PROF and PRON likes to [MASK]*

2.3 Critical analysis

The reason we give these details is that a careful examination of the data set and a replication of the GH22 experiments reveal variance in the results and certain asymmetries, making it difficult to draw firm conclusions.

Firstly, the GH22 test is not organized on the basis of minimal pairs varying only in polarity (like the pair (1) above, from Kassner and Schütze (2020)). Such minimal pairs cannot be formed, firstly because the parameter combinations are not exactly the same for cases with positive context (Cp) and those with negated context (Cn), and secondly because the embedding verb is different in Cp and Cn (*tries to ACT as often as possible* versus *doesn't like to* ACT). So it's hard to tell whether the variations in %-ACT-repetition are due to negation sensitivity or to other parameters.

These parameters (connectives and intensifiers) have a major impact on the interpretation of the examples, and more specifically on the discourse link between the context and target sentences. The test is based on:

- examples with an affirmative context, for which a repetition of the ACT is expected and corresponds to a discourse relation 'elaboration' (for instance, in the context "*Jessica is an architect who tries to dance as often as possible*", the second sentence "*She likes to dance*" goes in the same direction);
- examples with a negative context, for which a **non**-repetition of ACT is expected, corresponding to a 'contrast' between C and T.

In GH22, the interpretation of the discourse relation between C and T is supported by various clues in addition to the absence or presence of negation in C: (i) the possible co-reference between NAME and PRON, (ii) the semantic link between the main predicates in C and in T (e.g. for Cp cases, the link between *try to ACT as often as possible* and *like to ACT*) and (iii) the intensifiers *does* and *really* and the discourse connectives. Because they make the elaboration or contrast relation explicit, connectives make the test easier, and weaken the possibility of analyzing the models' "understanding" of negation. Intensifiers strengthen the elaboration relation in the positive case, but the effect is more ambiguous in the negative case. This excessive number of parameters weakens the interpretation that can be made of this test.

And indeed, while overall the PLMs tested show a sensitivity to negation in the GH22 results (*i.e.* the repetition rate is lower for Cn cases than for Cp), in replicating their experiments we observed significant variance depending on the various parameters cited above, notably the presence of coordinated verbs in C, and the rank for the choice of ACT-token (GH22 give results aggregating ranks 1 and 50). In the next sub-section, we present our replication of GH22 in detail, before moving on in the next section to our proposal for a more controlled test, based on minimal pairs varying only in polarity, and on additional control tests to ensure a correct interpretation of the results.

2.4 Partial replication

In this sub-section, we present our replication results for GH22. To focus on the ways models deal with negation, we have ignored a number of parameters, and systematized the combination of retained parameters. More specifically, we have limited ourselves to (i) patterns with no coordination in context sentences and (ii) instantiations using rank-1 ACT-token (we observed significant variations

with respect to these parameters). We also discard patterns with connectives, which allows us to reconcile the affirmative and negative versions of the tested inputs (since the connectives differ along with the polarity of C), and above all to remove cues favoring or hindering the repetition of the ACT.

To sum up, the parameters we have kept for this replication are the presence/absence of negation in C, the presence/absence of the intensifiers *does* and *really* in T, and we test all 8 combinations.

We apply this reduced test to the models `bert-large-cased` (Devlin et al., 2019) and `roberta-large` (Liu et al., 2019). The results are summarized in Table 2.

n°	pol.	aux	adv	roberta-l	bert-l
1	P	-	-	44.2	24.6
	N			27.3	3.3
2	P	✓	-	31.8	91.8
	N			25.1	58.3
3	P	-	✓	94.1	99.6
	N			25.3	73.6
4	P	✓	✓	99.8	100
	N			55.9	96.3

Table 2: %-ACT-repetition rates, for the two models `roberta-large` & `bert-large`, using GH22 patterns without coordination in the context sentence C nor discourse connectives. As in GH22, the ACT-token is chosen as the top-1 prediction for *NAME is a PROF and PRON likes to [MASK]*. Columns: *pol*: polarity in C; *aux*: presence of *does* in T; *adv*: presence of *really* in T.

The results are analyzed by comparing the P lines with their corresponding N lines, and considering the drop in repetition (drop = P rate − N rate). GH22 consider that the greater the drop, the more sensitive the model is to negation. For both models, a drop is indeed observed for all 4 pairs of P/N patterns, so we can say that the test is effective. Note, however, that for `bert-large-cased`, the drop is small in patterns with *really*. The model seems to interpret intensifiers as elaborations, and doesn't seem to be able to interpret a contrast despite the negation in the context sentence (*NAME is a PROF who doesn't like to ACT. PRON really does like to [MASK]*).

These results underline the strong interference of intensifiers on ACT repetition rates and drops, yet only a pattern without intensifiers (or connectives) comes close to a minimal pair targeting negation.

In addition, we believe that the method has a significant shortcoming. For patterns with positive polarity, the ACT repetition rate can be far from 100% (*e.g.*, 31.8 for `roberta-large`, pattern 2P). For this pattern, therefore, $100 - 31.8 = 68.2\%$ of the affirmative examples are such that the top-1 prediction is not the ACT. This corresponds to cases like (6)).

(6) Maria is a doctor who tries to **pad** as often as possible.
She does like to [MASK]top-1=**teach**.

However, in this case, GH22 count a non-repetition in the corresponding negative example (as in example (7)) as a proper handling of negation by the model.

(7) Maria is a doctor who doesn't like to **pad**.
She does like to [MASK]top-1=**follow**.

But such a non-repetition can no longer be taken as an evidence for an understanding of negation: it is not so striking not to repeat the single forbidden token (*pad*), since it was not repeated in the affirmative case (as in (6)).

In order to circumvent the shortcoming, we propose to consider by construction only examples leading to a repetition of the ACT-token in the affirmative case.

3 Our proposal: the *Self-Contained Neg Test*

3.1 Patterns

On the basis of the above findings, we propose to create a test that is strongly inspired by GH22, but that allows us to draw a more reliable conclusion: we want to be able to attribute any drop in the ACT-token repetition rate solely to negation, and thus judge whether a model has mastered the semantics of verbal negation.

We keep the principle of "self-contained" inputs, composed of a context sentence (C), and a target sentence (T) ending with a masked position, syntactically calling an infinitive verb. But we propose a single pattern for C and T, each sentence being either affirmed or negated, so that variation in C and T is limited to the presence or absence of negation. We give the two variants Cp and Cn and the two variants Tp and Tn in table 3. By combining these variants, we obtain four patterns (CpTp, CpTn, but also CnTp and CnTn).

Context	Target
Cp: NAME is a PROF who likes to ACT.	**Tp**: PRON is happy to [MASK].
Cn: NAME is a PROF who doesn't like to ACT.	**Tn**: PRON isn't happy to [MASK].

Table 3: Context and target sentences, either positive and negative, used for the base *Self-Contained Neg Test*.

Note that in CpTp, and to a lesser extent CnTn, although the repetition of the ACT-token is not mandatory, it leads to a pragmatically felicitous discourse. In contrast, the repetition is semantically forbidden in CpTn and CnTp.

3.2 Instantiation of examples

As in GH22, final examples are obtained by instantiating NAME, PROF and ACT (PRON is *she* or *he* depending on the gender of the proper noun instantiating NAME), but we modify the way ACT is instantiated, to resolve the shortcoming described in section 2.4: we only consider by construction positive examples (pattern CpTp) leading to a top-1 repetition. To do this, instead of using a different sentence, external to the test (GH22 used (5)), we take the CpTp pattern (*NAME is a PROF who likes to ACT. PRON is happy to [MASK].*), and for each pair [NAME, PROF], for each model, we instantiate ACT with an English intransitive verb such that the top-1 prediction at the masked position is that very same verb.

More precisely, the instantiation procedure is as follows: we have four lists (100 female proper nouns, 100 male proper nouns, 91 professions, and a number of monotokenized intransitive verbs, the number varying according to the tokenizer of the models). For proper nouns and professions, we re-use the GH22 lists. For verbs, we use monotokenized infinitives among English verbs that may have an intransitive usage, by cross-referencing the list on this wiktionary page `https://en.wiktionary.org/wiki/Category:English_intransitive_verbs` and the verbs present in Verbnet (Schuler, 2006). We apply this procedure to two bert models (`bert-base-cased`, `bert-large-cased`) and two roberta models (`roberta-base` and `roberta-large`), and we obtain 597 and 106 verbs for the `bert` and `roberta` models respectively.

For each PLM and for each of the 2*100*91=18200 [NAME, PROF] pairs, we compute the subset of verbs in the list that lead to a top-1 repetition. The number of such [NAME, PROF, verb] triplets is shown in row number 4 of table 5. We then randomly select at most 20 verbs for each model and each [NAME,PROF] pair (row 6 of table 5). Note that these subsets hence depend on the tested model.

Table 4 illustrates the process for the pair [*Jessica, dancer*].

Instantiated NAME/PROF: *Jessica, dancer*		
Tested verb: *smoke*		
Tested example: *Jessica is a dancer who likes to smoke. She is happy to [MASK].*		
model	top 1 pred.	retained?
`bert-base-cased`	*smoke*	✓
`bert-large-cased`	*smoke*	✓
`roberta-base`	*dance*	no
`roberta-large`	*chat*	no

Table 4: Example of selection of [NAME, PROF, ACT] triplets, for a given instantiated [NAME=*Jessica*, PROF=*dancer*] pair. When instantiating ACT with *smoke*, the top-1 at the MASK position is *smoke* (repetition) for the models `bert-base-cased` and `bert-large-cased`, and will eventually be selected when retaining 20 random such verbs for the given input pair.

For each model, the triplets thus obtained to instantiate NAME, PROF, ACT are then used to form the saturated examples for each of the patterns.

3.3 Test interpretation

For each pattern, we can detail how a decrease or stability in %-ACT-repetition should be interpreted in relation to the rate of 100% repetition for CpTp. As we'll always be comparing %-ACT-repetition with the 100% rate obtained by construction for CpTp, we prefer to consider a measure of rate decrease: **drop** = 100 − %-ACT-repetition.

- **CnTp**: this pattern is an evolution of patterns proposed by GH22, but designed here to form a true minimal pair with CpTp. By construction, ACT-token is semantically impossible at the masked position, so a small drop would mean that the model doesn't interpret negation in C. On the contrary, the larger the drop (the maximum being 100), the more likely it is that the model interprets correctly the negation in C. Note that any other verb is semantically

model	bert-b-c	bert-l-c	roberta-b	roberta-l
1. Available verbs	597	597	106	106
2. Available NAME,PROF pairs	18200	18200	18200	18200
3. Tested triplets (row 1 $\times$ row 2) (*10^6)	10.9	10.9	1.9	1.9
4. $\hookrightarrow$ leading to ACT repetition (*10^6)	2.4	2.0	1.2	0.4
5. Ratio (row 4/row 3, %)	21.7	18.4	61.9	18.3
6. Selected triplets	364000	363922	362027	107856

Table 5: Statistics for the selection stage of triplets instantiating NAME, PROF, ACT, for each model. **Row 3:** number of tested triplets (NAME, PROF, verb). **Row 4:** number of such triplets for which the instantiated CpTp example leads to a repetition (top-1 prediction is identical to the ACT-token). **Row 6:** number of selected triplets among those of row 4 (retaining at most twenty verbs for each [NAME,PROF] pair).

and discursively possible in the masked position, and corresponds to a contrast discourse relation between Cn and Tp.

- **CpTn**: we also add the case where negation is in the target sentence, and therefore closer to MASK. Here again, ACT-token is semantically impossible at the masked position, and the drop interpretation is the same as for CnTp.
- **Control pattern CnTn**: here the repetition of the ACT-token is discursively natural. A high-performing pattern is expected to have only a marginal drop. The pattern is used to check that a negation in one sentence is correctly interpreted in relation to the polarity in the other sentence, and not just in isolation.
- **Control pattern CpTv**: we also add a control where the modification with respect to CpTp is not the addition of the negative adverb, but the addition of another adverb, *very*, in T. This pattern makes it possible to check whether a drop in CpTn is really attributable to the negation in T, and not simply to the addition of any adverb. More generally, as ACT has been instantiated to obtain 100% repetition in the CpTp pattern, this CpTv pattern makes it possible to check the stability of ACT-token repetition: if a model's predictions are often different depending on the presence or absence of *very* in *NAME is a PROF who likes to ACT. PRON is (very) happy to ACT*, then this would be a sign that any change could potentially have a lot of impact, and it would prevent any positive interpretation of a drop for this model.

3.4 Properties of the test

These patterns have been chosen to limit the factors that can be used to interpret the discourse relation between C and T. In our case, the interpretation is solely driven by (i) the coreference between NAME and PRON, (ii) the semantic link between *like to ACT* and *be happy to ACT* and (iii) the absence or presence of negation on these predicates: only (iii) varies within the test, (i) and (ii) remain stable, and no intensifier or discourse connector cues are added that would favor or hinder the repetition of ACT-token.

In this way, we can form true minimal pairs varying only by a negation, in C or in T: for each triplet instantiating NAME, PROF and ACT, we have four minimal pairs (CpTp / CpTn), (CpTp / CnTp), (CnTn / CnTp) and (CnTn / CpTn).

By forcing an ACT repetition rate of 100% for the CpTp pattern, we totally avoid positive examples that don't lead to a repetition, which render the corresponding negative examples unusable (cf. sub-section 2.4). What's more, the CpTp pattern now serves as a reference point, and decreases in %-ACT-repetition are more comparable one with another, whether for a comparison between models, for the same pattern, or a for comparison between patterns for the same model. Finally, we make sure to obtain instantiated examples where the discourse relation between *like to ACT* and *be happy to ACT* is "understood". In this way, a lower repetition rate in a negative context will be all the more significant.

Note that the procedure to select [NAME, PROF, verb] triplets yields a large number of ACT-repetitions in CpTp (cf. the ratios for each model provided at row 5 of table 5). This confirms that

repetition in the CpTp pattern is pragmatically felicitous, although not mandatory. We observe that this ratio is much higher for the `roberta-base` model compared to the other three models. We cannot state whether this stems from a higher tendency to repeat tokens or from a preference to interpret the discourse relation between the two sentences as an elaboration.

3.5 Models evaluation

We apply our test to the four above-mentioned models and provide the results in Table 6.

Recall that passing our test implies having strong drops for the CpTn and CnTp patterns, and that these drops be greater than in the control pattern CpTv, and to a lesser extent in the CnTn pattern.

The `bert-base-cased` model fails the test completely: the drop is almost non-existent for the CpTn and CnTp patterns. Moreover, the drop is much stronger for the CpTv control pattern: in the context of *NAME is a PROF who likes to ACT*, the model repeats ACT less often in *PRON is very happy to MASK* than in *PRON isn't happy to MASK*.

Although the drop of the `bert-large-cased` model is larger for the CpTn and CnTp configurations than those of `bert-base-cased`, its drop in the CpTv configuration is still too high to conclude that this model understands negation.

The `roberta-base` model shows drops closer to our expectations: its CnTn drop is smaller than those of CpTn and CnTp (20.7 and 46.7). But since the drop for the CpTv control is 21.3, only the 46.7 drop is significant.

Finally, the model that seems to have acquired the most robust understanding of negation is `roberta-large`, having both a drop of over 50% for CpTn and CnTp, and a small drop for the controls. The maximum drop is obtained for CnTp, i.e. with a negation in the context sentence. It remains to be investigated why this configuration is better handled than CpTn.

To sum up, none of the models reaches drops close to 100% for CpTn and CnTp: many examples lead to a repetition of a token that is semantically forbidden by the context sentence. Nevertheless, it seems that the `roberta` models, and in particular `roberta-large`, "understand" the semantic value of verbal negation in English better than `bert`. Moreover, within a family of models, the `large` version performs better.

4 Additional controls: forcing non-coreference

In the results analyzed in the previous section, the drop can only be interpreted as an understanding of negation if the model has resolved the co-reference between the proper noun in C and the pronoun in T. In the absence of such a resolution, a repetition of the ACT is neither forbidden nor required.

While the ability of `bert` to resolve coreference has been evidenced by Clark et al. (2019), we need to ensure that this resolution is effective in the case of our patterns. To do this, instead of directly testing coreference resolution, we build a set of alternative examples to the base examples, in which non-coreference is forced. In practice, we replace the pronoun in T by a proper noun other than the one used in C, with two variants, depending on whether or not these two proper nouns have the same gender (cf. examples 1 and 2 table 7). If the model does indeed resolve coreference in base examples, then we should observe a much smaller drop for examples with forced non-coreference: in the absence of coreference, the context sentence no longer gives information about the target sentence, and therefore no longer prohibits or favors the choice of a particular token. As the sequences have been selected to favor repetition of the ACT token, this repetition should however remain high.

We also consider cases where we help the model establish a co-reference, so as to test only the impact of negation, independently of the models' ability to establish the co-reference between the proper noun and the pronoun in the basic examples. To this end, we use the same proper noun in C and T (cf. example 3 table 7). The repetition gives a less natural example, but in which the coreference is forced.

Triplets are selected using the same procedure as in section 3.2, namely retaining only triplets leading to a top-1 repetition for the CpTp pattern, and at most 20 verbs for a given [NAME,PROF] pair.

Results for `roberta` models are provided in table 8 (those for `bert` models are in appendix A, table 9). A first observation is that the number of selected triplets (first row of tables 8 and 9) undergoes a severe decrease. This is consistent with the fact that in such configurations, repetitions are pragmatically less felicitous.

An efficient model is expected to obtain small drops for the Non-Coref columns, while retaining

Pattern	bert-b-c	bert-l-c	roberta-b	roberta-l
CpTn	3.6	44.7	27.7	64.7
CnTp	1.2	16.5	66.9	82.8
CnTn	1.5	9.7	12.1	43.5
CpTv	25.5	42.9	23.3	26

Table 6: Drops of the %-ACT-repetition with respect to 100% for CpTp, when applying the *Self-Contained Neg Test* to four PLMs. To pass the test, drops should be high for the first two lines, and low for the last 2.

id	Type	Context	Target
1	non coref, same gender	Joyce is a designer who likes to smoke.	Janet really likes to [MASK].
2	non coref, other gender	John is a dentist who likes to dance.	Anna really likes to [MASK].
3	forced coref by repetition	Judith is a diplomat who likes to drink.	Judith really likes to [MASK].

Table 7: Examples of the corefence control test. In 3, coreference is forced by using the same name in C and T. In 1 and 2, coreference is ruled out by using distinct names, of either same or different genders.

large drops in the Coref column, for the CpTn and CnTp cases. The drops in CnTn and CpTv controls should remain small.

Pattern	Coref	Non-Coref	
		Same-gend.	Other-gend.
# ($\times 10^3$)	118.7	39.6	44.7
CpTn	3.9	7.5	6.9
CnTp	12.6	3.3	3.4
CnTn	1.5	3.4	2.4
CpTv	8.7	4.5	4.8

(a) roberta-base

Pattern	Coref	Non-Coref	
		Same-gend.	Other-gend.
# ($\times 10^3$)	60.8	4.4	5.1
CpTn	28.9	11.7	12.1
CnTp	64.1	1.9	10.8
CnTn	14.3	6.2	7.1
CpTv	17.3	9.4	11.5

(b) roberta-large

Table 8: **Last 4 rows**: drops of the %-ACT-repetition for the roberta models, when forcing coreference by using the same name in C and T (**Coref**) or forcing non-coreference using different names (**Non-Coref**), either of same or different genders. **First row (#)**: number of selected [NAME, PROF, ACT] triplets, among those leading to a top-1 repetition in the CpTp pattern (still retaining at most twenty verbs for each [NAME,PROF] pair).

We can see this is not the case for the roberta-base model: the drops in the upper left part of table 8a are smaller with respect to the "vanilla" examples (with a pronoun in T sentences). It is as if the model interpreted the same two names as non-coreferent. It is though impossible to conclude whether this is the case (in which case the smaller drops do not mean that negation is misunderstood), or whether the model interpreted the coreference correctly, but failed to interpret negation.

On the other hand, the trends observed for roberta-large (table 8b) do follow our expectations: the drops do remain large for the Coref case for CpTn and CnTp (and significantly larger than for the CnTn and CpTv controls) but they are small for the non-coreference patterns. This confirms the observations made for this model with the previous test, and thus further confirms the abilitiy of this model to capture the semantics of verbal negation.

5 Conclusion

In this paper we propose a methodology and a dataset to study PLMs' abilities to correctly interpret the semantics of negation, more precisely verbal negation in English. We were inspired by Gubelmann and Handschuh (2022), who proposed *self-contained* examples, consisting of two sentences, the first serving as a context that favors or hinders the repetition of a certain verb in the second sentence. After critically analyzing this test, we propose an improved version, which is more controlled, more systematic, and entirely based on examples forming minimal pairs varying only in the presence or absence of verbal negation. We have sought to minimize the interpretations that the models have to make in addition to the negation

interpretation, so that the observed results can be more reliably interpreted as the model's good or bad "understanding" of negation.

We applied our test to four pretrained Transformer-based language models. A detailed analysis of the results shows a continuum of situations: `bert-base` is globally unable to take verbal negation into account, `bert-large` is a little better at first sight, but the control tests we added show its limitations. `roberta-base` partially passes the basic test, but is disappointing when it comes to controlling co-reference resolution. Only the `roberta-large` model shows trends in line with expectations, for both base and control patterns, clearly showing some ability to capture the semantics of verbal negation in English.

However, for all the models we tested, a significant number of examples get a top-1 prediction that is exactly the token semantically forbidden by the context. This shows how much room for improvement remains for this type of models.

We chose to focus on verbal negation, being the most frequent form of negation in English, but we plan to extend our test to other forms of negation. Extension to other languages is also considered.

6 Limitations

The *Self-Contained Neg Test* only works on a Masked language modeling task. As such it is clearly designed for bidirectional models. Applying it to generative language models would require a complete rethinking of the test.

References

Hande Celikkanat, Sami Virpioja, Jörg Tiedemann, and Marianna Apidianaki. 2020. Controlling the Imprint of Passivization and Negation in Contextualized Representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Online. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

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.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of PLMs' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4602–4621, Dublin, Ireland. 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.

David Kletz, Marie Candito, and Pascal Amsili. 2023. Probing structural constraints of negation in pretrained language models. In *The 24rd Nordic Conference on 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. *CoRR*, abs/1907.11692.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

A Results non coreference for `bert` models

Pattern	Coref	Non-Coref	
		Same-gend.	Other-gend.
# ($\times 10^3$)	21.4	7.2	
CpTn	2.0	3.4	2.9
CnTp	2.2	2.3	2.3
CnTn	1.4	4.1	3.7
CpTv	10.3	13.1	15.3

(a) `bert-base`

Pattern	Coref	Non-Coref	
		Same-gend.	Other-gend.
# ($\times 10^3$)	47.1	14.3	16.5
CpTn	41.1	47.9	44.7
CnTp	5.8	6.4	5.7
CnTn	2.9	7.3	5.8
CpTv	38.5	44.8	45.7

(b) `bert-large`

Table 9: **Last 4 rows**: drops of the %-ACT-repetition for the `bert` models, when forcing coreference by using the same name in C and T (**Coref**) or forcing non-coreference using different names (**Non-Coref**), either of same or different genders. **First row (#)**: number of selected [NAME, PROF, ACT] triplets, among those leading to a top-1 repetition in the CpTp pattern (still retaining at most twenty verbs for each [NAME,PROF] pair).