Context Matters: A Pragmatic Study of PLMs' Negation Understanding

Reto Gubelmann
University of St.Gallen
Rosenbergstrasse 30
9012 St.Gallen
reto.gubelmann@unisg.ch

Siegfried Handschuh
University of St.Gallen
Rosenbergstrasse 30
9012 St.Gallen
siegfried.handschuh@unisg.ch

Abstract

In linguistics, there are two main perspectives on negation: a semantic and a pragmatic view. So far, research in NLP on negation has almost exclusively adhered to the semantic view. In this article, we adopt the pragmatic paradigm to conduct a study of negation understanding focusing on transformer-based PLMs. Our results differ from previous, semantics-based studies and therefore help to contribute a more comprehensive – and, given the results, much more optimistic – picture of the PLMs' negation understanding.

1 Introduction

Transformer-Based pre-trained language models (PLMs) have become the *de facto* standard in a variety of natural language processing tasks. Based on the original transformer architecture (Vaswani et al., 2017), researchers have proposed a number of extraordinarily successful architectures, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and smaller versions such as DistilBERT (Sanh et al., 2019). Such transformer-based models perform impressively at standard natural language understanding (NLU) benchmarks, often outperforming the human benchmark, as evinced by the GLUE and SuperGLUE Leaderboards (Wang et al., 2018 and Wang et al., 2019).

While it is impossible to deny the performance of these models at such benchmarks, it is another, particularly challenging question whether this performance is driven by simple shallow heuristics or by any real understanding of the languages that they are processing.[1] This study contributes to answering this question with a focus on negation.

Answering the question is important for both theoretical and practical reasons. On the theoretical side, the study contributes to a more accurate understanding of the driving forces behind the predictions issued by PLMs. The phenomenon of negation, being both highly semantically relevant and having a small footprint on the syntactic surface of a sentence, is ideally suited for this purpose. For real-world applications, it is crucial to know whether the model predicts based on simple contextual clues or on a real understanding of negation. For instance, for NLU applications, it makes a substantial difference whether a certain microblog recommends that one should "(not) get vaccinated against covid-19".

The question is challenging for two main reasons. First, because the models' performance is typically strong. However, it is wrong, not correct predictions that potentially unveil underlying heuristics. Second, as we will show, it requires careful, linguistically conscientious construction of the datasets to be able to draw sound conclusions even from wrong predictions. For example, if a model predicts "fly" in example (1), this has been taken by most researchers as a clear indication that the model does not understand negation, especially if its confidence in this prediction is similar to the confidence with which it predicts "fly" in example (2).

(1) Birds cannot **fly**.

(2) Birds can **fly**.

In contrast, our hypothesis is that this behavior of the models is not due to lack of negation understanding, but to a failure to, as it were, resolve the context in samples such as (1) and (2). While human beings automatically read sentences such as (1) within a default context, perhaps something like a biology class in primary school, where it is clear that birds can fly, this does not mean that it is never sensible or appropriate to say that they

[1] Compare the appendix, section A, for a more elaborate sketch of the theoretical background of our concept of real understanding.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4602 - 4621

cannot.[2] Our hypothesis is grounded in the linguistic research tradition called "pragmatics", while virtually all research in NLP focusing on negation understanding is based on the competing tradition called "semantics" (for a case in point, see Kassner and Schütze, 2020, who use samples very similar to (1) and (2)). Semanticists tend to assume that truth values and appropriateness of propositions are unambiguous and context-independent. Pragmatics, in contrast, emphasizes the importance of context and of syntactic, prosodic and other details to judge the appropriateness of a sentence. See below, section 3, for more details and references.

To test our hypothesis, we construct datasets that, while being designed to be challenging to the models, provide a micro-context that allow us to rule out failure to resolve context as a cause of wrong predictions by the models.

In detail, we contribute to the investigation of transformer-based PLMs in three ways. **First**, following the pragmatic tradition in linguistics, we develop a novel testing approach for negation understanding. Rather than using isolated sentence-pairs such as (2) and (1), our approach builds on automatically creating pragmatically and stylistically sound micro-contexts. **Second**, by tailoring our datasets to the individual models to be tested, and by varying a number of possibly influencing factors, we are able to pin down precisely the true driving forces behind the models' predictions. **Finally**, by fine-tuning the most successful models, we gain a view towards the potential for improving the performance with such fine-tuning.

2 Previous NLP Research on Negation

We emphasize that our study is not directly connected to research on negation clue and scope detection (for an overview on that research, see Khandelwal and Sawant, 2019). Rather, our experiments test whether the models are able to process the information contained in a negated sentence to predict a semantically admissible token in a following sentence.

Hence, our research is connected to other work that examines various aspects of transformer-based PLMs (the field is often called "BERTology", testifying to the dominance of BERT-focused studies in this area). Rogers et al. (2020) provides an overview. There are studies examining the inner functional differentiation of the system's parts, such as Voita et al. (2019), whose findings suggest that many of the attention heads of the original transformer are superfluous; Kovaleva et al. (2019) examine the functions of BERT's attention heads and find none that are specifically dedicated to negations. Wiedemann et al. (2019) report state of the art performance in word-sense disambiguation using BERT's contextualized word embeddings. Forbes et al. (2019) study the models' abilities to learn so-called "physical commonsense". Zhang et al. (2021) find that large models fare better in particular regarding common-sense reasoning (and show little improvement over smaller models with regard to semantic or syntactic tasks).

As we are trying to get the models to commit mistakes that reveal underlying shallow heuristics, our research is connected to so-called adversarial attack or probing studies. These studies are trying to go beyond the NLU benchmarks such as GLUE and SuperGLUE to see whether the models achieve their impressive performance using shallow heuristics or real understanding. There are such studies in the field of argument reasoning (Niven and Kao, 2019) and natural language inference (McCoy et al., 2019). Geirhos et al. (2020) have proposed a general diagnosis of the problem of shallow heuristics, and Ribeiro et al. (2020) have urged a more comprehensive, multi-dimensional approach to testing the abilities of these models instead of simply submitting them to automated benchmarks.

Furthermore, as we are studying negation by testing whether the models are able to draw very simple inferences, research in natural language inference (NLI) is also relevant for our work. In this regard, Gururangan et al. (2018) show that, in the main datasets used in NLI, negated sentences are biased towards contradiction, Wallace et al. (2019) show that certain triggers can be inserted context-independently and lead to a stark decline in NLI accuracy, and Hossain et al. (2020) show that simply ignoring negation does not substantially decrease model performance in many NLI datasets. Notably, Jeretic et al. (2020) is among the rare NLP studies that presuppose a pragmatic background, studying the ability of PLMs to cope with implicatures.

Of particular importance for our study are contri-

[2] For instance, when observing an ostrich, somebody might use (1) as a shorthand for "These are some birds that cannot fly". Furthermore, one can imagine example (1) to pop up in multiple-choice examinations, many dialogues ("Birds cannot fly? – Of course they can!"; "Barcelona cannot win the Champions league!" – "Oh yes, and birds cannot fly."), fictional literature, etc.

butions by Warstadt et al. (2020), Ettinger (2020), and Kassner and Schütze (2020). All three studies develop minimal pairs to examine the ability of PLMs to correctly categorize a number of linguistic phenomena, including negation. Hence, all of these studies are squarely based on the semantic side of the ongoing debate in linguistics between semantics and pragmatics (see above, section 1, examples (1) and (2)), and all of them find that the models largely ignore negation, as for each minimal pair, the predictions differ little between positive and negated sentence; Ettinger (2020) finds slightly better performance for more natural examples, hinting at the relevance of pragmatics in this context.

3 Linguistics & Philosophy of Language: Understanding Negation Between Semantics and Pragmatics

Negation is a multi-faceted phenomenon that can be realized in a number of ways. In the study of negation, it is common to distinguish two very different approaches in linguistics and philosophy of language: semantics and pragmatics (for a recent discussion of the distinction, see Preyer (2018)). The tradition of semantics has been initiated in its modern, formal-logical form by Frege (1892). Arguably the most important analysis of negation in this semantic tradition is Russell (1905). For a recent contribution in this tradition with a focus on computability, see Moot and Retoré (2019). As mentioned above, (section 1), semanticists often try to context-independently assess the truth of a proposition.

Pragmatic studies of negation understanding have traditionally had a focus on the readings of ambiguous negated sentences, and on systematic ways in which conversational contexts and other non-semantic features systematically disambiguate such sentences (e.g., identifying types of contexts in which example (1) is read as containing a negated universal quantifier, as opposed to the types of contexts in which it is read as containing a negated existential quantifier). Noveck (2009) provides an overview on recent experimental-pragmatic research on negation understanding. For a study with a focus on the influence of context on human's understanding of negation, see Kaup (2009). In his seminal study of negation, Horn (2001, 368f.) also discusses contextual factors, Davis (2016) continues in Horn's footsteps.

Furthermore, according to the orthodox Gricean version of conversational implicatures (see Davis, 2019 for an introduction, for canonical texts by Grice see Grice, 1975, Grice and Strawson, 1956, and Grice, 1978), one can assume that participants follow conversational maxims, including the one of relation. This maxim urges the participants to a conversation to only contribute statements that are relevant. Accordingly, merely repeating the same assertion in a discourse would be seen as apparently violating this maxim, and hence as calling for a non-standard interpretation according to which the statement is, *pace* first appearances, in agreement with that maxim.

Based on this pragmatic perspective on meaning, and on its emphasis on the importance of context in particular, testing anybody's negation understanding abilities with minimal pairs such as examples (1) and (2) is questionable: There are many contexts in which it is appropriate to say that birds cannot fly, and these contexts might be more common than others where, say, it is appropriate to say that birds cannot breastfeed – even if the latter, but not the former would be considered true from a zoological point of view. This pragmatic perspective then grounds our hypothesis that it is context resolution, not negation understanding, that explains the models' performance on minimal-pairs such as (1) and (2), which is the main evaluation method in current research.

4 Dataset

Following our pragmatic outlook and our hypothesis, we construct our datasets always using micro-contexts to guide the models and to avoid confounding inability to determine context with inability to understand negation. Furthermore, we pay careful attention to grammatical details that might influence prediction, and we construct our positive examples (those not containing a negation) such that they respect the maxim of relation.

We here give the generic way how we create our datasets. In the following section 5, we discuss the experiment-specific details of the templates used in each of the experiments. As we are tailoring our datasets to each individual model, we actually create some 12 million of potentially different test sentences in total. We have discussed each of the templates (43 in total) with a native speaker and philosopher of language, Dr. David Dolby. We detail how his review has influenced the dataset in the appendix, section C.

4.1 Step 1: Hand-Craft Templates

For each experiment, the first step consists in hand-engineering suitable templates. A simple example for the kind of template that we want to test the models on is given by (3), together with possible replacements for the placeholders in curly brackets.

(3) FNAME{Petra} is PROF{an architect} who doesn't like to ACT{sail}. However, she does like to MASK.

In template (3), "FNAME" (its male counterpart being "MNAME") is a placeholder for entries in a list of female first names to be used in the following step, "PROF" is a placeholder for a profession, also to be used in the next step. The ACT-placeholder will be replaced with a verb specific to the respective name and profession as well as to the respective PLM under scrutiny. A model that understands negation is not going to predict "sail" (or, more generally speaking, any verb taking the position of the ACT-placeholder, what we call ACT-replacement) to replace the MASK token. We would call such a prediction an *exactly wrong prediction*.

4.2 Step 2: Fill in First Names and Professions

Once the templates such as the one given in (3) are available, we expand each of them into 9.1k unsaturated sentences by replacing the F/MNAME- and PROF-Placeholders with pre-set lists of male and female names and professions. [3]

Having completed step 2, our template (3) might have been developed into the unsaturated sentence shown in (4). The name and profession placeholders have been replaced by real names and professions; the only remaining placeholder is "ACT".

(4) Jessica is a printer who doesn't like to ACT. However, Jessica does like to [MASK].

4.3 Step 3: Extract Tailored ACT-Tokens of Specific Probability Ranks

In this third step, we replace the ACT-placeholders with verbs that are specific not only to the given unsaturated sentence (that is, specific to the given sentential context and to the specific combination of name and profession), but also to the model under scrutiny. We achieve this by individually extracting each model's predictions of the desired probability rank for the MASK in (5) and use it to replace the ACT-placeholder(s) in the unsaturated sentence at issue.

(5) Jessica is a printer and she likes to [MASK].

We run our first experiment with probability ranks 0, 50, 100, and 200. The goal of this procedure is to control for the overall probability of the ACT-token that the model might be tempted to repeat.

We expect that the models are more likely to wrongly repeat a negated ACT-token if it has a low probability rank: In this case, the model is inclined to predict it in contexts involving the given gender, first name, and profession; therefore, absent any understanding of negation, using an ACT-token of low probability rank makes it likely that the model will predict it to fill the MASK in patterns such as (3). In contrast, the model is very unlikely to predict an ACT-token of very high probability rank in these contexts. As a consequence, we expect the models to be less inclined to predict such ACT-tokens of high probability rank. However, the very occurrence even of such ACT-tokens might incline some models to predict them, despite both the overall low probability rank and the presence of negation.

For instance, assuming that we want to use probability rank 0 to fully specify unsaturated sentence (4) with regard to `roberta-large`, our method proceeds as follows. For `roberta-large`, the top three predictions to fill the MASK in (5) are: draw (prob. 0.21), write (prob. 0.16), and travel (prob. 0.15). Hence, we replace the ACT-placeholder in (4) with "draw", yielding (6). If we were running the experiment with probability rank 200, we would adapt the probability rank of the token used to replace the ACT-placeholder accordingly.

(6) Jessica is a printer who doesn't like to draw. However, she does like to <mask>.

Here, we have a fully-fledged, grammatical sentence with a MASK token. While "Jessica" and "printer" have been inserted using the lists, "draw" has been dynamically selected specific to both "Jessica" and "printer" as well as to the model under scrutiny by letting the model predict the MASK token in (5).

Unlike minimal pairs such as (1) and (2), this example (6) gives a minimal context: The situation in which the prediction of the MASK-token is to be made is one where Jessica does not like to draw.

[3] We use the top 100 male and female names in the USA between 1920 and 2019 according to the US social security administration. See here. For the professions, we use a list of 91 common professions.

As a consequence, if the model predicts "draw", it cannot be because of lack of understanding of context.

5 Experiments

For our experiments, we use the models provided by Huggingface (Wolf et al., 2019). We fine-tune the most promising models using a dataset that has been filtered from English Wikipedia with a rather simple regular expression, yielding 315 thousand sentences. For details, see the appendix, section D.

We have conducted three different experiments using the dataset creation method spelled out in the previous section. All of our experiments obey the basic pragmatic requirement to use micro-contexts, we pay attention to syntactic details that might matter for prediction, and we observe the maxim of relation for positive samples (requiring that we don't expect the models to merely repeat information). The first experiment forms the basis, it examines the ability of the models to correctly use information contained in negated sentences for prediction in a later sentence. In the second experiment, we use three kinds of misprimes to see whether this confuses the models, and in the final experiment, we test whether the models are sensitive to changes of referents. For a full list of the templates used in all experiments, see the appendix, section B; the scripts as well as these templates are also available on github.[4]

Experiment 1 In the first experiment, we used sentences of the form (7) and (8) to test the models' sensitivity to negation.

(7) MNAME is PROF who doesn't like to ACT. However, he does like to MASK.

(8) FNAME is PROF who tries to ACT as often as possible. So, she really does like to MASK.

Note that, with these templates, we made sure that the negated versions have a higher subsequence overlap than the corresponding positive versions, as the PLMs have a reputation for reacting strongly to such subsequences ("subsequence overlap" here refers to the overlap in tokens between the context where the ACT-token occurs and the context where the MASK occurs, see McCoy et al. 2019). This means that, if the results show that the models predict an ACT-token to fill the MASK more often with positive than with negated sentences, this cannot be because of the subsequence heuristic, as following this heuristic would pull the models ' prediction in the opposite direction. From a linguistic perspective, examples such as (8) obey Grice's maxim of relation by introducing new information in the second sentence.

Furthermore, we varied gender and the number of ACT-tokens in the first sentence (ranging from 1 to 3), and we also varied the extent to which we syntactically express the contrast in the negated version, motivated by pragmatic attention to syntactic detail. Starting from example (7), we first removed the conjunction ("However"), then we also removed the "does", which also marks a contrast. In the positive version, we added templates that do not have a conjunction signaling implication ("So"). Overall, this yields 30 different templates, which we expanded into sentences as described in the previous section. Then, we let the models predict the tokens to fill the MASK.

Experiment 2 In the second experiment, we wanted to further probe the robustness of the models' negation understanding by adding specific misprimes. Examples (9), (10), and (11) illustrate the patterns used here.

(9) MNAME is PROF who doesn't like to ACT. Of course, many people like to ACT. MNAME, in contrast, likes to MASK.

(10) MNAME is PROF who doesn't like to ACT. Many people, but not MNAME, like to ACT. MNAME likes to MASK.

(11) MNAME is PROF who doesn't like to ACT. Today is Tuesday and the Sun is shining. MNAME likes to MASK.

By varying gender as well as the presence or absence of a contrastive conjunction ("in contrast"), we obtained 10 templates, which we expanded and saturated as described above.

The basic idea behind templates of the kind of (9) and (10) is to bring to light shallow heuristics that are based on occurrence of the verbs in the context of the MASK to be filled. The expectation is that, if the models do not represent any logical structures involving negation, then the logically more explicit patterns of the form (10) are going to be more misleading to them than the less explicit ones of the form (9): in the former sentences, the proper

[4] https://github.com/retoj/transnegpaper_acl2022pub.

name appears again in the context of the activity that is not supposed to be predicted as a filler of the MASK. The random sentence inserted in (11), finally, is intended to test whether the models can transfer information contained in a negation across an unconnected sentence (one, notably, that might appear odd to humans as well, but which would not lead them to forget about the negated sentence).

Experiment 3 This third experiment, finally, is entirely dedicated to assessing whether the models are sensitive to changes in referents. By using templates of the form (12) (3 in total), we wanted to see whether the models are sensitive to obvious changes in referents.

(12) MNAME is PROF who doesn't like to ACT. Unlike MNAME, Cleopatra does like to MASK.

In templates such as (12), we included gender-incongruity and different proper names to clearly signal that the person of the first sentence that does not like to ACT is not the same as the Cleopatra of the second sentence.

6 Results

In the following, we report on the results of the experiments conducted as described in the previous section. The key figure that we are reporting is the percentage at which the model predicted an ACT-replacement token (in the following: "ACT-token"), for instance, "draw" in (6)) to fill the MASK. We call this figure "%-ACT-Repetition". Generally, a high percentage of ACT-repetitions implies poor performance with negated first sentences: these are the exactly wrong predictions. In contrast, a high percentage of ACT-repetitions is ok with positive first sentences; indeed, it is often a natural completion of the sentence, given the information in the first sentence. Furthermore, we report on results obtained by using ACT-tokens of probability rank 0 and 50 (for information on probability ranks, see above, section 4.3).

Finally, `xlnet-large-cased` performed so poorly that it has been excluded in the display of results in the present section. Since the predictions of this model were often ungrammatical, listing them along with the others would have given a false impression of equivalence; for instance, predicting ")" as an ACT-token in example (5) and "," to fill the MASK in example (6) would not count as an exactly wrong prediction, but it would of course be mistaken. In this sense, our approach requires that the predictions by the models be more or less sensible and grammatical, which all models except for this one fulfilled. However, `xlnet-large-cased`'s performance is given in detail in the appendix, section G, figure 12.

Experiment 1 Figure 1 displays the effect of negation on prediction. The values shown are percentages of cases where the model predicted (one of) the ACT-token(s) to fill the MASK in the final sentence.

As mentioned, in general, repeating such an activity token is correct if the activities are not in the scope of a negation, but exactly wrong if they are in such a scope, as in example (6). In figure 1, the red-to-reddish bars give the percentage at which the models wrongly predicted an ACT-token to fill the MASK, even though it was excluded by a negation in the first sentence. The green bars show the percentage at which the models predicted an ACT-token that has not previously been excluded via negation (which is perfectly fine).

In the first column of figure 1, the results of `neg-roberta-large` are displayed. This model wrongly predicts one of the ACT-tokens to fill the MASK in roughly 8% of cases if the first sentence is negated. The presence or absence of contrastive signals matter little. If the first sentence is not negated, the percentage goes up to some 52% on average, despite the fact that the lexical overlap is much smaller with positive templates. This yields a delta of 44% between negated and positive templates. Such a high delta indicates a sensitivity of the model for negation. In contrast, if both scores are low, they might be low simply because the model is unable to retain any information and hence predicts something completely unrelated to the context, or even something ungrammatical.

It is notable that fine-tuning does show a significant, albeit slightly unstable effect both regarding `roberta-large` and regarding `bert-large-cased`. Furthermore, as mentioned previously, the effect is much stronger with `xlnet-large`, as the vanilla version was simply predicting gibberish (closing brackets, for instance), while the fine-tuned version shows very strong performance.

Figure 2 shows further results from experiment one. We are here only showing the performance on negated templates. This means that, gener-

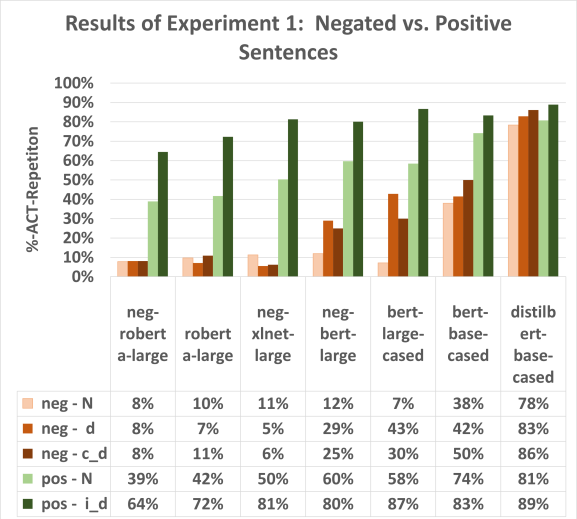


Figure 1: Percentages of prediction of ACT-token(s) ("neg": negated sentences, "pos": positive sentences, "N": no additional markers, "d": contrastive "does", "c_d" contrastive does with contrastive conjunction, "i_d": implicative conjunction plus "does" for emphasis). These predictions are ok if there is no negation in the first sentence (greenish bars) and exactly wrong if the first sentence is negated (red-reddish bars).

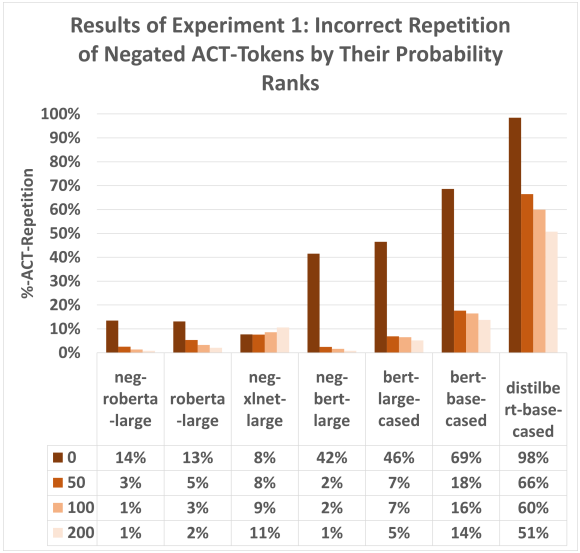


Figure 2: Percentages of erroneous predictions of negated ACT-tokens by probability rank of ACT-tokens.

ally, an ACT-repetition is wrong here, as it has been explicitly excluded by negation (for example, compare (7)). Furthermore, we are showing the results categorized by probability rank of ACT-token chosen. This means that, for instance, `neg-roberta-large` wrongly predicts an ACT-token to fill the MASK in almost 14% of all cases if the probability rank of the ACT-token in question is 0, while it does so in less than 3% of cases for the lower probability ranks.

Figure 2 clearly shows that the BERT models (both large and base) react very strongly to high probability ranks: `neg-bert-large`'s error rate drops from almost 40% to about 2% if the probability rank of the act token is lowered from 0 to 50, 100, or 200. XLNET, in contrast (and quite surprisingly), has a lower error rate with probability rank 0 tokens than with tokens of lower probability ranks.

Experiment 1 has also shown some differences in performance of PLMs depending on the gender of the templates. In particular, the RoBERTas perform worse with male gender than with female gender. For details, see section F.

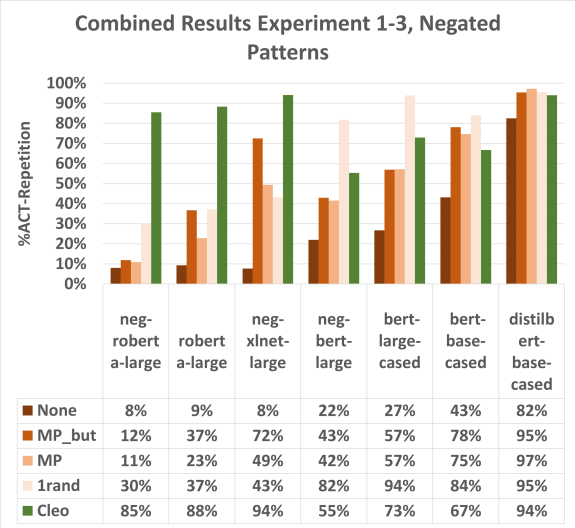
Experiment 2 Figure 3 shows percentages of ACT-repetition depending on the misprime or additional sentence inserted (for the interpretation of the Cleo-row, see below, experiment 3). Overall, the results show that all models are confused by the misprimes, but notably, `neg-roberta-large` only really loses performance with the one random sentence inserted. Furthermore, the results resemble the one presented above in figure 1 as far as the smaller models are performing worse, with distilbert being clearly at a loss. Fine-tuned XLNET loses much precision with the misprime with a "but", indicating a lack of representation of logical structure (see above, section 5).

Experiment 3 With "Cleo", the templates used are such that it should be maximally clear that the person that does not like to ACT according to the first sentence is clearly different from the person that does like to MASK (see (12)). And the models are very sensitive to that, as figure 3 shows. The RoBERTas, for instance, repeat the ACT-token with some 86% probability, regardless of the presence of a contrastive conjunction or a contrastive "does" in the second sentence, while this figure has not exceeded 30% for any of the misprimes with `neg-roberta-large`.

7 Discussion

In the following, we first conduct a brief analysis of the predicted tokens, then we discuss four insights that flow from our results.

Analysis of Predicted Tokens Overall, per probability rank, each model issued 400k predictions of activities that persons might like to do; as it



	neg-robert a-large	robert a-large	neg-xlnet-large	neg-bert-large	bert-large-cased	bert-base-cased	distilbert-base-cased
None	8%	9%	8%	22%	27%	43%	82%
MP_but	12%	37%	72%	43%	57%	78%	95%
MP	11%	23%	49%	42%	57%	75%	97%
1rand	30%	37%	43%	82%	94%	84%	95%
Cleo	85%	88%	94%	55%	73%	67%	94%

Figure 3: Percentages of exactly wrong predictions by inserted element (probability ranks 0 and 50): Either a misprime (MP, see example (9)), a misprime with a but (MP_but, see example (10)), or a random sentence (1rand, see example (11)). Finally, "Cleo" refers to the clearly different referent in the second sentence ("Cleopatra", see example (12)).

were, they made a guess as to what hobby a person with a specific gender, first name, and profession might have. Extensive inspection shows that in the clear majority of cases, with the stark exception of `xlnet-large-cased`, these predictions resulted in grammatical, sensible sentences, with some decline in grammaticality with higher probability ranks. Notably, the insertion of a random sentence ("Today is Tuesday and the sun is shining") has not had any observable impact on the semantics of predictions. For instance, it did not lead to a higher rate of predictions of activities involving the outdoors, or generally requiring good weather.

For the runs conducted with `neg-roberta-large` and tokens of probability rank 0, the predicted tokens belong to just 108 verb types, five of which account for 57% of all predictions: read, write, cook, dance, and sing (read accounts for 22% alone).

Clear Sensitivity to Negation The first basic insight provided by experiment one is thoroughly positive: With the exception of `distilbert-base-cased`, all models show sensitivity to negation (see figure 1): they are much less inclined to ACT-repetition if the token has been negated, even if the ACT-token is highly probable, given the model, the gender, the first name and the profession, and even though there is much less subsequence overlap with the positive (not negated) templates than with the negated ones. Hence, almost all models showed sensitivity to negation despite the contrary pull exerted by several shallow heuristics.

Note also that the models are generally not overly sensitive to contrast-highlighting elements in the negated sentences, but very sensitive to implication-indicating elements in the positive sentences, which, considering examples such as (8), seems pragmatically sound: In the absence of such an element, one might think that the first and the second sentence in these templates have little connection except for the common referent. Furthermore, note that the BERT-models perform significantly worse with patterns containing contrast-highlighting elements. This is surprising as one would expect that such elements would make it easier for the models to realize that a prediction of an ACT-token to replace the MASK is inaccurate.

These results put the pioneering findings by Ettinger (2020), Kassner and Schütze (2020), Warstadt et al. (2020), and Ribeiro et al. (2020) in perspective: our study clearly shows that pre-trained transformer-based language models *do* show sensitivity to negation. In figure 1, `neg-roberta-large`'s tendency to repeat verbs that replace ACT-placeholders drops by some 44% if these replacements are in the scope of a negation. Clearly, this model does not simply ignore negation.

There are two explanations for this contrast with earlier research. The first one consists simply in a reminder that these earlier studies did not test models based on the RoBERTa or XLNET architectures. Furthermore, Ribeiro et al. (2020) only use base-sizes of the models they test. Still, even with `bert-large-cased` and `bert-base-cased`, the difference between previous findings and our result are stark.

This means that a second explanation is needed. We suggest that the best candidate for such an explanation is precisely our hypothesis that, in these early studies, the models are struggling less with negation and more with the contextualization of the tasks: they are not unable to represent negation; rather, they are unable to identify a default context that rules out certain predictions *ab initio*.

Inferences to the best explanations are always fallible (for the standard study of abductive inferences,

see Lipton 2004): There could be another explanation that we have failed to consider that explains the difference in performance. However, given that, we are using highly controlled, synthetic, and relatively simple sentences, that we have extensively varied syntactic structure, gender, profession, and first name, and that we have tested a number of misprimes, all resulting in the same basic outcome, namely a sensibility to negation, we feel confident that we can rule out other explanations to the degree to which this is possible given current methods in BERTology. We therefore do take our results as providing strong support for our main hypothesis.

The Influence of Probability Ranks Furthermore, experiment one also shows that the model- and context-specific probability rank, which was controlled for in this experiment by a new method, is highly relevant for prediction (see figure 2). This effect is particularly pronounced for the BERT models, and much less for the RoBERTas and XLNET.

Another interesting aspect of the influence of probability ranks is that fine-tuning seems to reduce error rate less for highly probable tokens. Finally, the figures also show that the big drop in error rate occurs between probability ranks 0 and 50. Between 50 and 200, little further reduction occurs.

Robustness Against Misprimes & Random Insertions We have been testing the models' robustness against the insertion of two different misprimes as well as a random sentence in experiment 2. For results, see figure 3. These results show a very nuanced picture. First, `neg-roberta-large` is hardly disturbed by the misprimes, with the exception of the random sentence: its error rate doesn't surpass 12%. In stark contrast, `neg-xlnet-large`'s error rate skyrockets with all of the misprimes inserted: it shows an increase in error rate from 8 to 43%, and well beyond that. The other models are somewhere in between; notably, the BERT-models are struggling most with the random sentence inserted.

As robustness against misprimes indicates dependence on real understanding rather than shallow heuristics, these results further corroborate the finding of experiment 1 that the best performing models, `neg-roberta-large` in particular, understand negation.

Changes in reference The results of this experiment are surprising when the RoBERTas and fine-tuned XLNET are considered. All three models repeat the ACT-token with a probability of more than 85% if it is clear that the subject of the activity in question is clearly distinct from the subject of which the same activity has been negated in the first sentence. These results indicate that the models are clearly sensible to such changes in referents.

8 Conclusion

In this paper, we have examined the extent to which contemporary transformer-based PLMs understand negation. We have done so by presenting the models with tailored masked language modelling tasks that are structured in a way that is pragmatically sound and that ensures that the known shallow heuristics are of no help. We have found that all but two models are clearly sensitive to negation. It seems justified to say that the best-performing model understands negation, as it erroneously repeats a negated token in only 12% of cases even when strong misprimes are used (with the exception of the questionable insertion of a random sentence, where the figure is 30%), and it shows clear sensitivity to changes in reference. Our results complement and partly contrast earlier, semantics-based studies of PLMs' negation understanding.

References

Emily Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics*, pages 5185–5198.

Wayne Davis. 2019. Implicature. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Wayne A Davis. 2016. *Irregular negatives, implicatures, and idioms*. Springer.

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.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Hans-Johann Glock. 2019. Agency, intelligence and reasons in animals. *Philosophy*, pages 1–27.

H. P. Grice and P. F. Strawson. 1956. In defense of a dogma. 65(2):141–150.

H Paul Grice. 1978. Further notes on logic and conversation. In *Pragmatics*, pages 113–127. Brill.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Laurence R. Horn. 2001. *A natural history of negation*. CSLI.

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.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. 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.

Barbara Kaup. 2009. How are pragmatic differences between positive and negative sentences captured in the processes and representations in language comprehension. In *Semantics and Pragmatics: From Experiment to Theory*. Palgrave Macmillan.

Aditya Khandelwal and Suraj Sawant. 2019. Negbert: A transfer learning approach for negation detection and scope resolution. *arXiv preprint arXiv:1911.04211*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Peter Lipton. 2004. *Inference to the Best Explanation*, 2 edition. Routledge.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David Marr. 2010 [1982]. *Vision*. The MIT Press.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Richard Moot and Christian Retoré. 2019. Natural language semantics and computability. *Journal of Logic, Language and Information*, 28(2):287–307.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Ira A Noveck. 2009. Meaning and inference linked to negation: An experimental pragmatic approach. In *Semantics and Pragmatics: From Experiment to Theory*. Palgrave MacMillan.

Gerhard Preyer. 2018. *Beyond Semantics and Pragmatics*. Oxford University Press.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

John Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing 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.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv:1909.10430*.

Ludwig Wittgenstein. 2006/1953. Philosophische untersuchungen. In *Werkausgabe Band 1*. Suhrkamp.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

A Our Concept of Real Understanding

It is a matter of dispute which conditions are sufficient to credit any being with "real" linguistic understanding. In philosophy of mind as well as in cognitive psychology there are two very broad camps. On the one hand, there is a representationalist one that emphasizes what is going on inside the mind or brain of the being in question (compare Searle, 1980, 417, who argues that "causal powers equal to those of the brain" are necessary, or Marr, 2010 [1982], whose pioneering monograph on vision science became one of the founding documents of cognitive science). On the other hand, there are neo-behaviorist researchers who emphasize the importance of the being's behavior for any judgment on its understanding (see Glock, 2019). This study sides with the neo-behaviorists in focusing on the behavior of the models in question rather than on their internal going-ons. Hence, to really understand negation, a being must be able to react to negated sentences competently in a variety of circumstances with a performance that is comparable to what humans are capable of.

Note that the representationalists generally agree that the neo-behaviorists requirements are necessary; they just doubt that they are sufficient for real understanding. Hence, if it should turn out that current models do not satisfy the neo-behaviorist's requirement, the representationalist would agree that, as a consequence, the models do not really understand negation.

Furthermore, note that this question is related to another one that Bender and Koller (2020) have raised forcefully: What kind of relation (if any) is expressed by meaning? Following Searle (1980), they argue that it is a relationship between a linguistic item, say a word, and something extralinguistic. On this view, understanding negation would require understanding the meaning of negation which, supposedly, involves a word-world-relation. Continuing our loosely Wittgensteinian neo-behaviorism, we disagree: the meaning of a word is given by its use in language (Wittgenstein, 2006/1953, § 43). Unfortunately, at this point, we can only point to this disagreement without properly engaging the arguments by Bender and Koller (2020).

B Full list of Templates

In tables 1, 2, and 3, we give the full list of templates with their features by experiment. "N-ACTs" refers to the number of activities used (that is, one, two or three verbs that the person in question is supposed to like or not to like doing), "Negtype" expresses whether the template is positive or contains a negation, "Conj" refers to the kind of conjunction used (if any), "Add. El." specifies whether an additional sentence between the first and the final sentence is inserted, or whether a shift in reference occurs ("Cleo"), and "MorF" specifies th gender of the first names and pronouns used.

C Analysis of Stylistic Proposals

As mentioned above, section 4, we have discussed all of the templates with a native speaker and philosopher of language, Dr. David Dolby. He has made the following suggestions, which we adopted for our main experiments. In addition, we carried out all of the experiments with the original templates as well. We here list his suggestions as well as the effect on model performance with regard to the six templates in total which were affected.

Proposal 1 (3 templates affected) In sentences as the following, it was proposed to replace "to swim, nor does she like to fish, and she also doesn't like to surf" with "to swim, fish, or surf": "Petra is an architect who doesn't like to swim, nor does she like to fish, and she also doesn't like to surf. She does like to MASK."

Proposal 2 (2 templates affected) In sentences as the following, it was proposed to replace "to swim, to fish, and to surf" with "to swim, fish, and surf": "Petra is an architect who tries to swim, to fish, and to surf as often as possible. So, she really does like to MASK."

Proposal 3 (1 template affected) In sentences as the following, it was proposed to replace "In contrast with Peter," with "Unlike Peter,": "Peter is an architect who doesn't like to swim. In contrast with Peter, Cleopatra does like to MASK."

We have run all the experiments with both of these variants, and the differences in performance are given in table 4.

In general, it seems that slight variation in performance between these variants is an indicator of a more profound understanding of the sentences in question (similar to a model's robustness against misprimes). A human evaluator would be able to understand that predicting either swim, fish, or surf in sentences such as the one quoted in proposal 1

N-ACTs	Negtype	Conj	Add. El.	MorF	Template
3	not	contr-does	None	m	MNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. However, he does like to MASK.
3	not	does	None	m	MNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. He does like to MASK.
3	not	None	None	m	MNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. He likes to MASK.
3	None	implic.-does	None	m	MNAME is PROF who tries to ACT3, ACT2, and ACT1 as often as possible. So, he really does like to MASK.
3	None	None	None	m	MNAME is PROF who tries to ACT3, ACT2, and ACT1 as often as possible. He really likes to MASK.
2	not	contr-does	None	m	MNAME is PROF who doesn't like to ACT2, nor does he like to ACT1. However, he does like to MASK.
2	not	does	None	m	MNAME is PROF who doesn't like to ACT2, nor does he like to ACT1. He does like to MASK.
2	not	None	None	m	MNAME is PROF who doesn't like to ACT2, nor does he like to ACT1. He likes to MASK.
2	None	implic.-does	None	m	MNAME is PROF who tries to ACT2, and to ACT1 as often as possible. So, he really does like to MASK.
2	None	None	None	m	MNAME is PROF who tries to ACT2, and to ACT1 as often as possible. He really likes to MASK.
1	not	contr	None	m	MNAME is PROF who doesn't like to ACT. However, he does like to MASK.
1	not	does	None	m	MNAME is PROF who doesn't like to ACT. He does like to MASK.
1	not	None	None	m	MNAME is PROF who doesn't like to ACT. He likes to MASK.
1	None	implic.-does	None	m	MNAME is PROF who tries to ACT as often as possible. So, he really does like to MASK.
1	None	None	None	m	MNAME is PROF who tries to ACT as often as possible. He really likes to MASK.
3	not	contr-does	None	f	FNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. However, she does like to MASK.
3	not	does	None	f	FNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. She does like to MASK.
3	not	None	None	f	FNAME is PROF who doesn't like to ACT3, ACT2, or ACT1. She likes to MASK.
3	None	implic.-does	None	f	FNAME is PROF who tries to ACT3, ACT2, and ACT1 as often as possible. So, she really does like to MASK.
3	None	None	None	f	FNAME is PROF who tries to ACT3, ACT2, and ACT1 as often as possible. She really likes to MASK.
2	not	contr-does	None	f	FNAME is PROF who doesn't like to ACT2, nor does she like to ACT1. However, she does like to MASK.
2	not	does	None	f	FNAME is PROF who doesn't like to ACT2, nor does she like to ACT1. She does like to MASK.
2	not	None	None	f	FNAME is PROF who doesn't like to ACT2, nor does she like to ACT1. She likes to MASK.
2	None	implic.-does	None	f	FNAME is PROF who tries to ACT2, and to ACT1 as often as possible. So, she really does like to MASK.
2	None	None	None	f	FNAME is PROF who tries to ACT2, and to ACT1 as often as possible. She really likes to MASK.
1	not	contr	None	f	FNAME is PROF who doesn't like to ACT. However, she does like to MASK.
1	not	does	None	f	FNAME is PROF who doesn't like to ACT. She does like to MASK.
1	not	None	None	f	FNAME is PROF who doesn't like to ACT. She likes to MASK.
1	None	implic.-does	None	f	FNAME is PROF who tries to ACT as often as possible. So, she really does like to MASK.
1	None	None	None	f	FNAME is PROF who tries to ACT as often as possible. She really likes to MASK.

Table 1: Templates used in experiment 1.

N-ACTs	Negtype	Conj	Add. El.	MorF	Template
1	not	contr-does	MP-but	m	MNAME is PROF who doesn't like to ACT. Many people, but not MNAME, like to ACT. MNAME, in contrast, likes to MASK.
1	not	none	MP-but	m	MNAME is PROF who doesn't like to ACT. Many people, but not MNAME, like to ACT. MNAME likes to MASK.
1	not	contr-does	MP	m	MNAME is PROF who doesn't like to ACT. Of course, many people like to ACT. MNAME, in contrast, likes to MASK.
1	not	none	MP	m	MNAME is PROF who doesn't like to ACT. Of course, many people like to ACT. MNAME likes to MASK.
1	not	contr-does	1rand	m	MNAME is PROF who doesn't like to ACT. Today is Tuesday and the Sun is shining. MNAME likes to MASK.
1	not	contr-does	MP-but	f	FNAME is PROF who doesn't like to ACT. Many people, but not FNAME, like to ACT. FNAME, in contrast, likes to MASK.
1	not	none	MP-but	f	FNAME is PROF who doesn't like to ACT. Many people, but not FNAME, like to ACT. FNAME likes to MASK.
1	not	contr-does	MP	f	FNAME is PROF who doesn't like to ACT. Of course, many people like to ACT. FNAME, in contrast, likes to MASK.
1	not	none	MP	f	FNAME is PROF who doesn't like to ACT. Of course, many people like to ACT. FNAME likes to MASK.
1	not	contr-does	1rand	f	FNAME is PROF who doesn't like to ACT. Today is Tuesday and the Sun is shining. FNAME likes to MASK.

Table 2: Templates used in experiment 2.

N-ACTs	Negtype	Conj	Add. El.	MorF	Template
1	not	contr	Cleo	m	MNAME is PROF who doesn't like to ACT. Unlike MNAME, Cleopatra does like to MASK.
1	not	does	Cleo	m	MNAME is PROF who doesn't like to ACT. Cleopatra does like to MASK.
1	not	None	Cleo	m	MNAME is PROF who doesn't like to ACT. Cleopatra likes to MASK.

Table 3: Templates used in experiment 3.

Model	**Delta Prop. 1 (3 templ. aff.)**	**Delta Prop. 2 (2 templ. aff.)**	**Delta Prop. 3 (1 templ. aff.)**
neg-roberta-large	1.40%	1.0%	14.14%
roberta-large	1.33%	1.3%	11.54%
neg-xlnet-large	5.19%	-0.4%	0.86%
neg-bert-large	10.51%	4.1%	22.31%
bert-large-cased	15.21%	6.6%	7.23%
bert-base-cased	1.08%	5.0%	3.40%
distilbert-base-cased	3.61%	5.0%	4.00%

Table 4: Difference in percentage of ACT-repetition for the templates affected, depending on whether or not Dolby's proposals were adopted (positive value means higher repetition without adopting Dolby's proposals). As usual, the average of probability ranks 0 and 50 was used. For instance, adopting proposal 1 leads to a decrease of ACT-repetition of 1.4% for neg-roberta-large with regard to the three templates affected.

is inadmissible, regardless of whether the sentence is phrased in the slightly clumsier phrasing before adopting Dolby's first proposal: The difference is in style, not in logical form.

In the specific cases at hand, it must be noted that only proposal 1 concerns templates that logically exclude certain predictions, while proposals 2 and 3 concern templates where certain predictions are suggested, e.g., swim, fish and surf in the example listed in the description of proposal 2, or swim in the example listed in the description of proposal 3. So, the substantial differences in performance caused by proposal 3 do not indicate that the models performed substantially worse without adopting the proposal. It merely means that they were less inclined to repeat the ACT-token in the sentence.

Furthermore, the results have minor impact on the results of the experiment as a whole, as even for the largest variation, found with `bert-large-cased`, the difference concerns only 3 out of 18 negated templates in total, implying that overall performance of the model improves by only 2.5%.

Finally, the figures fit with the findings from experiment 2 (see figure 3): the RoBERTas, in particular the fine-tuned version, show almost no susceptibility to these surface phenomena, whereas `bert-large-cased` reacts strongly, suggesting a shallower processing and less understanding of logical structure.

D Details on Fine-Tuning

To fine-tune the models to sentences involving negation, we use a regular expression to extract sentences containing a negation token together with a contrastive conjunction from English Wikipedia[5]; the goal was to filter for sentences that, in addition to containing a negation, also contain parts that can only be predicted correctly if the negation is taken into account. For instance, in sentence (13), taken from the fine-tuning corpus, a model that has no understanding whatsoever of negation is not going to predict a sensible token in the penultimate position ("play", right before ".").

(13) He was also selected by Zimbabwe for the 2014 African Nations Championship but didn't play.

By filtering for such patterns, we extract 351k sentences from English Wikipedia.

[5]See this Wikipedia entry for details.

Based on some initial exploratory tests, we fine-tuned `roberta-large`, `xlnet-large-cased` as well as `bert-large-cased`. In view of the fine-tuning corpora used, we label the resulting models **neg-roberta-large**, **neg-bert-large** and **neg-xlnet-large** respectively. For fine-tuning, the scripts provided by Huggingface (Wolf et al., 2019) were adapted. Fine-tuning took place on 4 GPUs of a DGX-2. The fine-tuned models where then tested together with the vanilla ones in the experiments.

The python regular expression used to extract the sentences is the following (note that we are filtering out sentences with character length less than 65, as they are usually not full sentences):

```
re.search('( no )|( none* )|(n't)',sentence) and
re.search('rather|even|but',sentence) and len(sentence) >64
```

Fine-tuning occurred with the following settings (again taken from Huggingface):

- num_train_epochs=1,
- per_gpu_train_batch_size=64,

E Details on Experiments

The experiments were conducted on four GPUs of a DGX-2, processing time of one template (i.e. 9.1k sentences) per model varied widely depending on the model, from about 300 seconds for distilbert to 1500 seconds for the fine-tuned version of xlnet. The script used for these experiments as well as all necessary input-files is available on github.[6] The script builds on the standard scripts provided by Huggingface, see Wolf et al. (2019). For an illustration of the algorithm used, see algorithm 1.

F Results of Experiment 1 by Gender

Table 4 shows the results of experiments 1 and 2, restricted to negated templates, by gender. The RoBERTas perform slightly worse with male gender than with female, as they wrongly predict an ACT-token more often with male than with female templates. For the remaining models, there is no clear trend.

G Detailed Results of all models

In figures 5-12, we are giving all the results of all models, including vanilla `xlnet-large-cased`, which has been

[6]https://github.com/retoj/transnegpaper_acl2022pub

```
for MODEL in model-array do
    for TEMPLATE in template-array do
        for NAME in name-Array do
            for PROF in prof-array do
                Obtain ACT1-3 by letting
                  MODEL predict [MASK]
                  in sentences like (5)
                Let MODEL predict the
                  [MASK] in the sentence
                  built from TEMPLATE,
                  NAME, PROF, ACT1-3.
            end
        end
    end
end
```

Algorithm 1: The method used to evaluate the models.

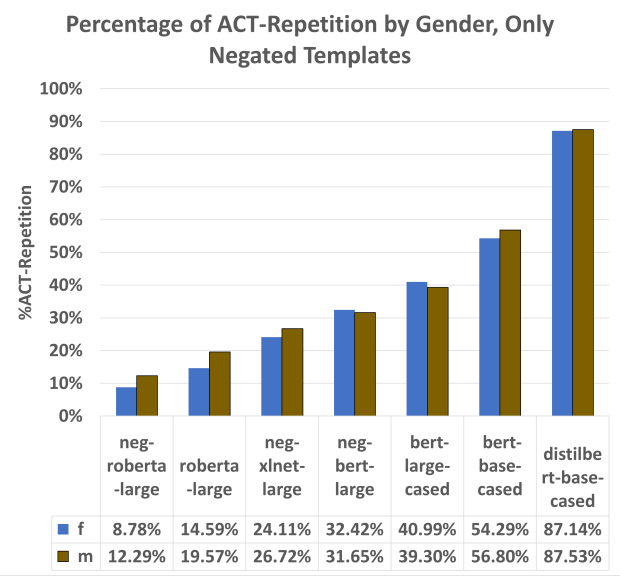


Figure 4: Performance in experiment 1 and 2, only negated templates and probability ranks 0 and 50, by gender.

excluded from the results in section 6 due to its extremely poor performance. As the final chart shows, the model repeats an ACT-token between 2% and 57% of cases. As a look at the prediction shows, this is because `xlnet-large-cased` often predicts ungrammatical and nonsensical tokens such as "that" or "." to fill the MASK both in (5), which is used to extract the ACT-tokens of suitable rank, and in fully specified templates such as (6). Listing these nonsensical predictions on a par with the grammatical predictions by the other models would have given a false impression of equivalence.

To increase readability, we have replaced "contr_does" with "contr" and "implic._does" with "implic".

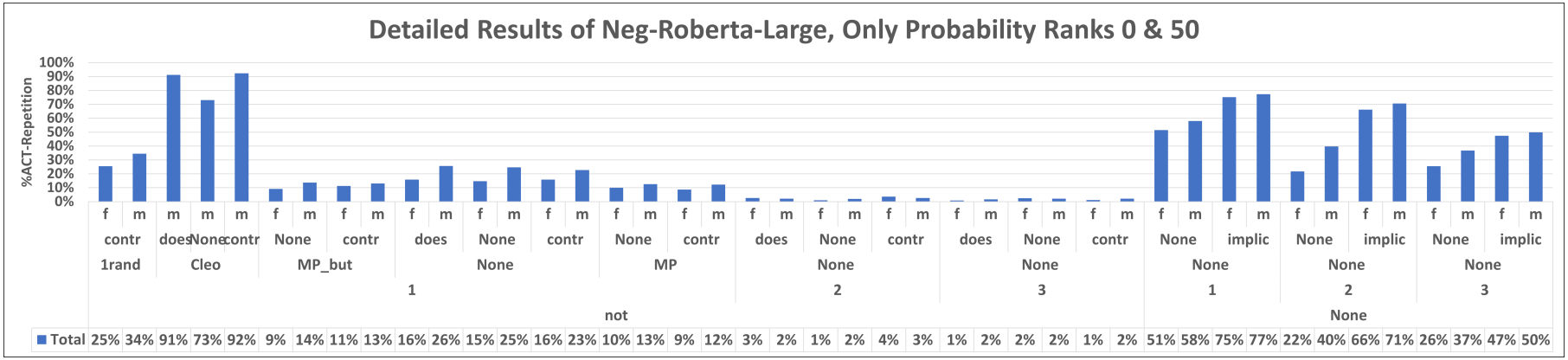


Figure 5: Performance of neg-roberta-large. Performance is shown only for probability ranks 0 and 50.

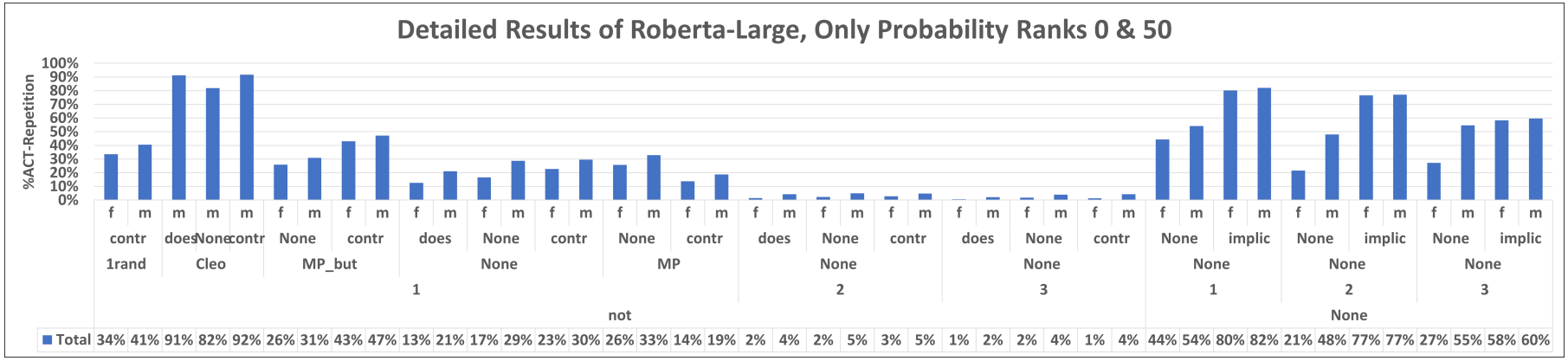


Figure 6: Performance of roberta-large. Performance is shown only for probability ranks 0 and 50.

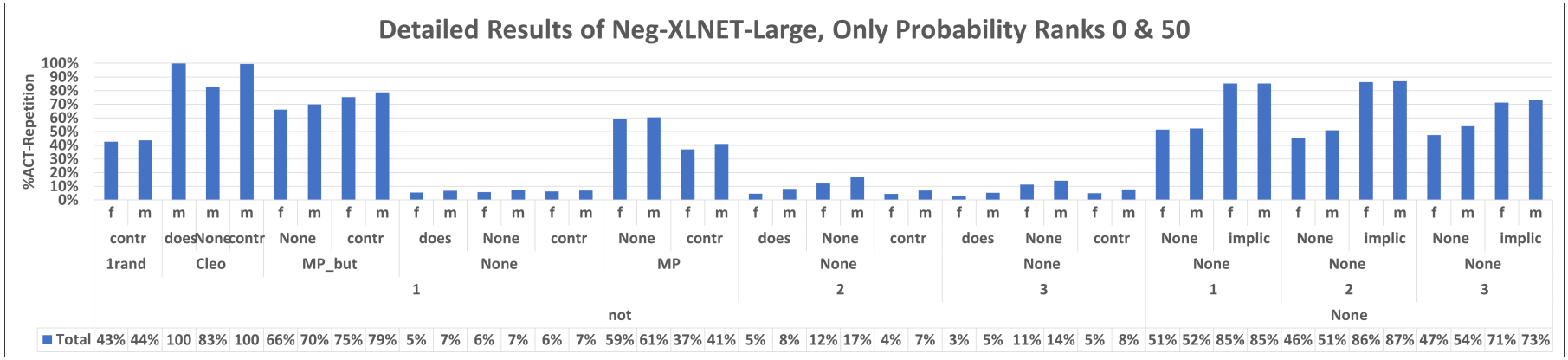


Figure 7: Performance of neg-xlnet-large. Performance is shown only for probability ranks 0 and 50.

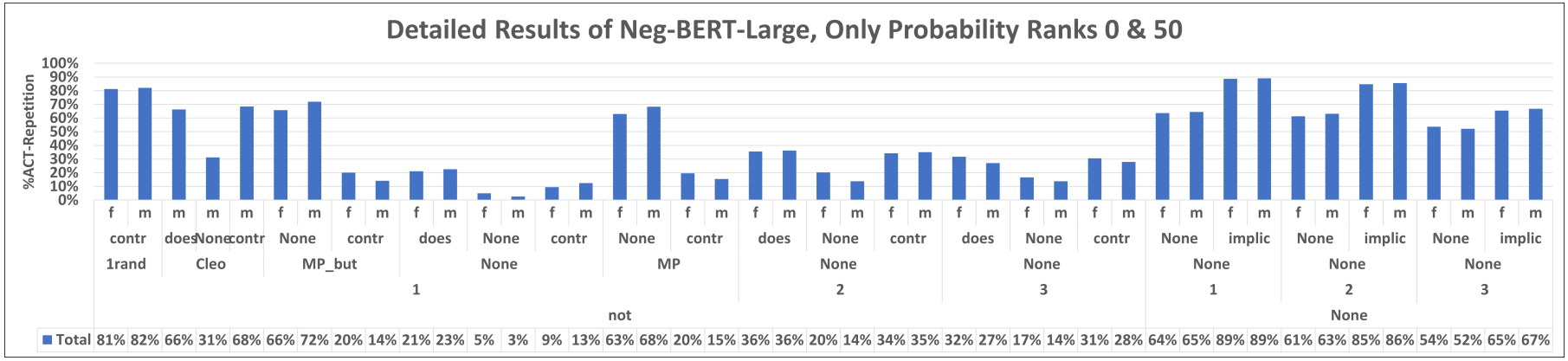


Figure 8: Performance of neg-bert-large. Performance is shown only for probability ranks 0 and 50.

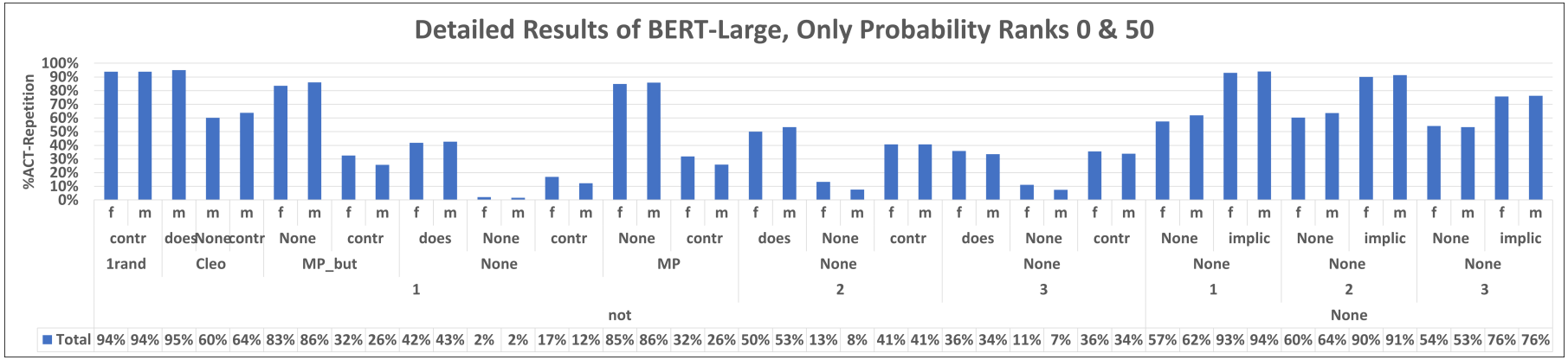


Figure 9: Performance of bert-large-cased. Performance is shown only for probability ranks 0 and 50.

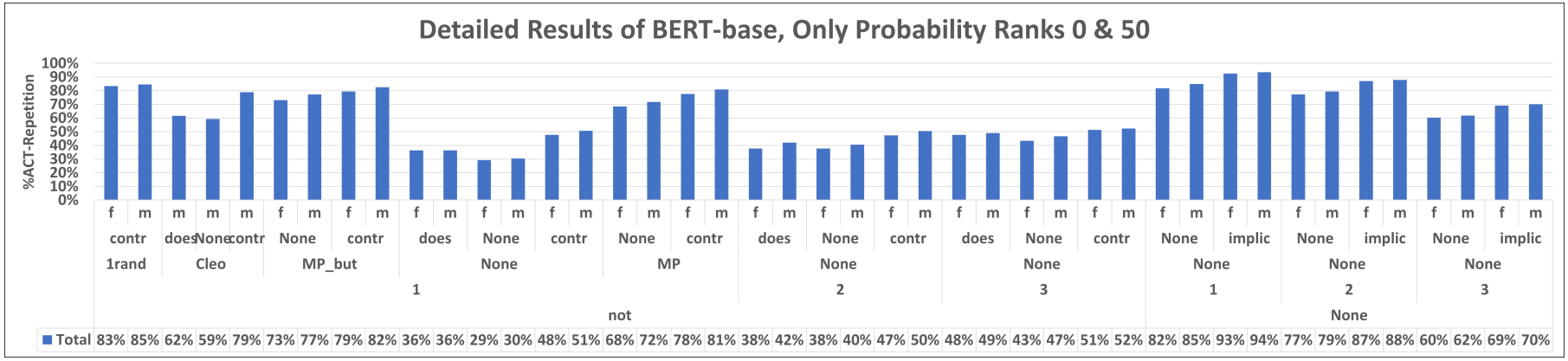


Figure 10: Performance of bert-base-cased. Performance is shown only for probability ranks 0 and 50.

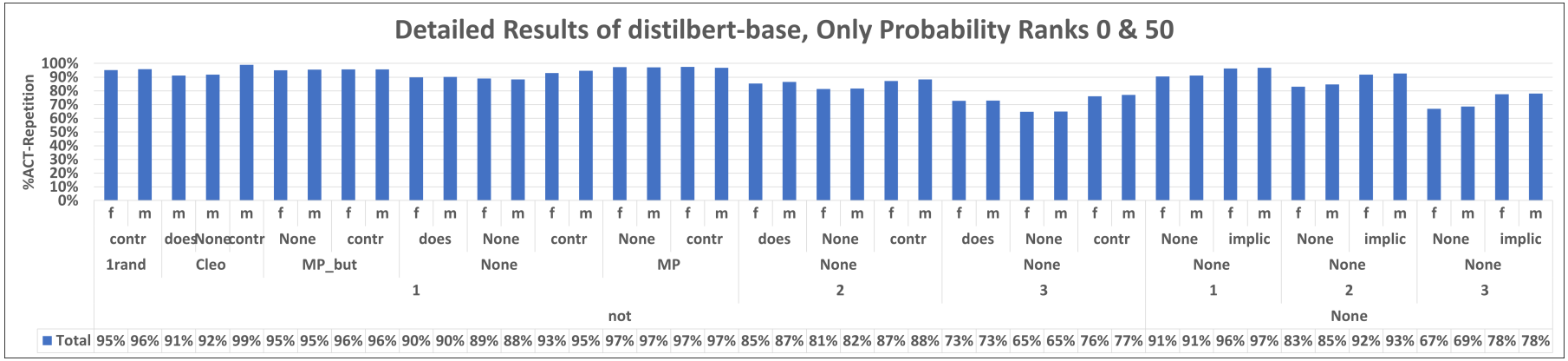


Figure 11: Performance of distilbert-base-cased. Performance is shown only for probability ranks 0 and 50.

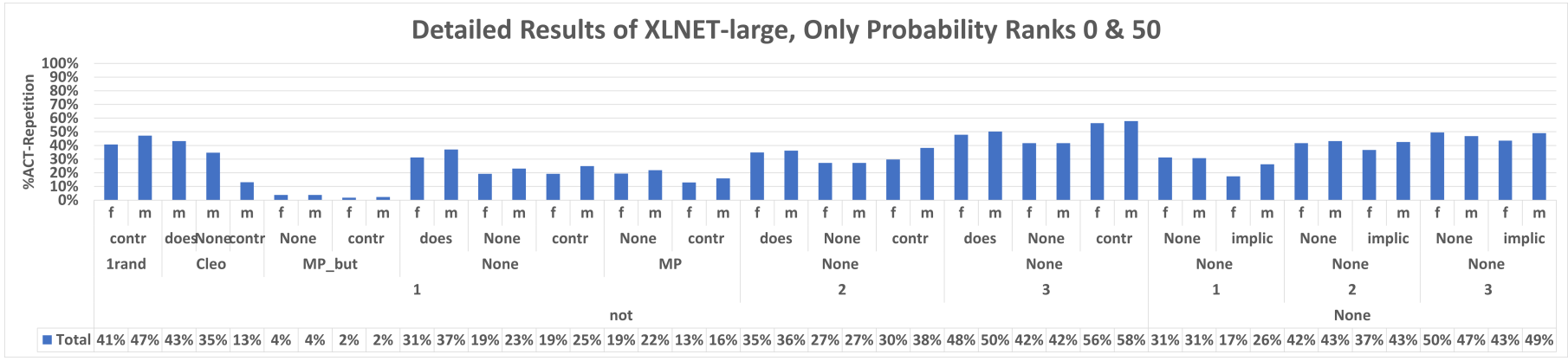


Figure 12: Performance of xlnet-large; as can be seen, the percentages of ACT-repetitions hardly differ between negated and positive sentences. Furthermore, note again that the model's predictions are very poor, often predicting clearly wrong token such as closing brackets ")" or other non-alphabetic characters.