When your Language Model cannot even do Determiners right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle

Judith Sieker and **Sina Zarrieß**
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{j.sieker, sina.zarriess}@uni-bielefeld.de

Abstract

The increasing interest in probing the linguistic capabilities of large language models (LLMs) has long reached the area of semantics and pragmatics, including the phenomenon of presuppositions. In this study, we investigate a phenomenon that, however, has not yet been investigated, i.e., the phenomenon of *anti-presupposition* and the principle that accounts for it, the *Maximize Presupposition!* principle (MP!). Through an experimental investigation using psycholinguistic data and four open-source BERT model variants, we explore how language models handle different anti-presuppositions and whether they apply the MP! principle in their predictions. Further, we examine whether fine-tuning with Natural Language Inference data impacts adherence to the MP! principle. Our findings reveal that LLMs tend to replicate context-based n-grams rather than follow the MP! principle, with fine-tuning not enhancing their adherence. Notably, our results further indicate a striking difficulty of LLMs to correctly predict determiners, in relatively simple linguistic contexts.

1 Introduction

Presuppositions have held a significant position in semantic and pragmatic studies over the past decades (e.g., Beaver et al. (2021)). Also in the research on linguistic knowledge represented in large language models (LLMs) (Belinkov and Glass, 2019; Ettinger, 2020; Schuster and Linzen, 2022), focus has shifted more and more towards exploring LLMs' capabilities in semantic and pragmatic discourse processing (Ruis et al., 2022; Hu et al., 2023; Sieker et al., 2023), encompassing the exploration of presuppositions (Jiang and de Marneffe, 2019; Jeretic et al., 2020). In this context, the primary emphasis has largely been on classification tasks, examining whether language models capture inferences triggered by presuppositions, using datasets that were developed in particular for Natural Language Inference (NLI) tasks; and it became apparent that LLMs mostly acquire surface-level patterns rather than deeply engaging with underlying knowledge representations (Jeretic et al., 2020; Kabbara and Cheung, 2022; Cong, 2022).

Our goal in this paper is to explore whether presuppositions and, in particular, pragmatic principles underlying the phenomenon of presuppositions, are captured in LLMs that are pretrained on large text corpora. Rather than relying on NLI sentence classification tasks, which typically involve various aspects of semantic and pragmatic linguistic knowledge, we aim to introduce a controlled linguistic diagnostic following the idea of minimal pair analysis, which has been well established in work on diagnosing syntactic knowledge in LLMs (Warstadt et al., 2020). More specifically, in this study, our focus lies on a special kind of presupposition that, despite its comparable prevalence in language, has not yet received equivalent attention in research, the so-called *anti-presupposition* (Percus, 2006) and, furthermore, the pragmatic principle that accounts for it, Heim (1991)'s Maximize Presupposition! principle (MP!).

Given that anti-presupposition triggers are simple instances found in minimal pairs that exhibit a clear preference for one trigger over the other (see Section 2), they present an optimal opportunity to explore the degree to which language models incorporate linguistic competencies, particularly pragmatic principles like the MP! principle. Still, to the best of our knowledge, it has remained unexplored how LLMs deal with anti-presuppositions and, beyond, whether they consider the MP! principle as part of their predictive processes. In this paper, we use a simple minimal pair analysis that tests the LLMs' ability to predict determiners, when being presented with contexts featuring anti-presupposition triggers, utilizing data sourced from the field of psycholinguistics. Given that determiners rank among the most frequent words in many languages, it may be

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 180–198
December 7, 2023. ©2023 Association for Computational Linguistics

expected that LLMs are able to not only pick up their surface-level occurrence patterns, but learn deeper generalizations about their discourse-level functions and principles of use. Yet, as we show in this paper, LLMs show a striking difficulty to correctly predict determiners in our diagnostic setup, challenging the idea that determiners pose only minimal challenges for LLMs (Yang et al., 2023) and supporting the hypothesis that they struggle with fundamental aspects of linguistic reasoning.

In the following, Section 2 presents background on anti-presuppositions and the Maximize Presupposition! principle as well as prior research in the domain, Section 3 describes the setup of our experimental investigation and Section 4 describes the results.

2 Background

(Anti-)presuppositions. Presuppositions are background information or information that interlocutors assume to be part of the common ground (Stalnaker, 1973). Generally, presuppositions are introduced by particular words or syntactic constructions, so-called presupposition triggers (Beaver et al., 2021). For instance, in Example (1), taken from Schneider et al. (2019), the indefinite determiner *a* triggers the presupposition that there is more than one (unique) pen.

(1) Please hand me a pen.

What is more, the indefinite determiner raises an intriguing linguistic question, leading us to the central focus of this paper: Its classification is uncertain in terms of whether it triggers a presupposition or is more fittingly categorized as a so-called *anti-presupposition* (Percus, 2006). That is, certain expressions containing presupposition triggers might appear inappropriate in situations where the truth of a presuppositionally stronger element is implied – in other words, where the presuppositionally stronger element is part of the common ground (cf., e.g., Percus (2006), Schneider et al. (2019), Blunier (2022)). For a clearer illustration, see (2), where the presupposition that there is precisely one sun is fulfilled in the common ground, therefore, making the sentence appear odd.

(2) A sun is shining.

More precisely: The determiner "the" carries a stronger presupposition compared to the indefinite determiner "a" (i.e., it is *presuppositionally stronger*), as "the" implies both existence and uniqueness. Therefore, in (2), using the indefinite determiner becomes inappropriate, as it is established that only one entity of the mentioned type exists. Similarly, uttering (1) is appropriate in a context with three pens and a single pencil, yet it is unsuitable in a context featuring only one pen and three pencils. In the latter, due to uniqueness, a definite determiner would be anticipated.

Like presuppositions, anti-presuppositions are associated with specific words or constructions. Apart from the (in-)definite determiners, other triggers include *both* and *all*, as well as the verbs *know* and *believe*, see (3) and (4) from Percus (2006).

(3) John assigned the same exercise to all of Mary's students. –> *anti-presupposes* that Mary has exactly two students.

(4) Mary believes that Jane is pregnant. –> *anti-presupposes* that Jane is pregnant.

Maximize presupposition! principle (MP!). The phenomenon of anti-presuppositions can be attributed to a broader pragmatic principle proposed by Heim (1991), adding to Grice (1975)'s conversational maxims. This principle is known as the *Maximize Presupposition!* principle (MP!), and it states: *Presuppose as much as possible!* Essentially, MP! accounts for anti-presuppositions by proposing that sentences will be blocked in contexts where other sentences with stronger presuppositions (while being identical in all other respects) would convey the same meaning. In other words, MP! mandates the speaker to always use the strongest presupposition among a set of alternatives, provided that these presuppositions are fulfilled (Percus, 2006; Schneider et al., 2019; Blunier, 2022; Panzeri and Foppolo, 2021). For example, in the case of (4), MP! predicts that Jane is not pregnant. This is because if she were indeed pregnant, the speaker would have chosen the alternative "know" (which would introduce the presupposition of Jane's pregnancy). Put differently: when hearing (4), the hearer assumes that the presupposition of the stronger alternative is false, that is, the presupposition of the stronger alternative is anti-presupposed. Similarly, we can explain the anti-presupposition in (3): Here, "both" blocks the use of a sentence containing "all" since "both" represents the presuppositionally stronger alternative. However, it's essential to note that (anti-)presuppositions are context-dependent, meaning that they need to be evaluated in relation to the knowledge or beliefs shared between the conversational participants (e.g., Beaver et al. (2021)). For instance, revisiting (3) in a specific context where "Mary is a new teacher at a small, private school, and John is her colleague who knows that Mary has only two students," the interpretation of (3) changes. I.e., in this context, (3) no longer gives rise to the same anti-presupposition as it would without this additional information.

Related Work. Research from psychology and psycholinguistics has explored anti-presuppositions and the MP! Principle, albeit with much lesser emphasis compared to other pragmatic principles or phenomena. For example, both Yatsushiro (2008) and Panzeri and Foppolo (2021) examined whether children are sensitive to the MP! principle. Both studies both found an evolutionary trend, wherein sensitivity to the principle increased with age. Furthermore, Bade and Schwarz (2021) conducted four experiments to investigate the derivation of inferences triggered by different anti-presupposition triggers, finding results in support of the MP! principle. Also, Schneider et al. (2019) investigated the processing efforts of English definite and indefinite determiners. They employed a mouse-tracking study and found that processing of the indefinite determiner is more difficult than processing the definite determiner, as well providing evidence for the MP! principle.

Now, when it comes to investigating anti-presuppositions and the MP! principle within the realm of language models, there seems to be a gap in research. However, in general, there is an increasing interest in examining the linguistic capabilities captured in LLMs, often facilitated through the utilization of linguistic test suites and experimental datasets (e.g., Belinkov and Glass (2019); Ettinger (2020)). And, while much previous research has concentrated on dissecting the syntactic competence of LLMs (e.g., Hu et al. (2020); Marvin and Linzen (2018)), recent investigations have extended to exploring the prowess of LLMs in semantic and pragmatic discourse processing (e.g. Ruis et al. (2022); Hu et al. (2023); Sieker et al. (2023)). Overall, as illustrated by, e.g., Chang and Bergen (2023), it appears that LLMs are capable of performing basic logical reasoning tasks; yet, they still face challenges when confronted with complex reasoning.

In the context of exploring presuppositions within the field of LLMs, one case study, for example, is carried out by Jeretic et al. (2020). Their investigation centered around the extend to which NLI models are able toy capture the inferences triggered by presuppositions (and implictures), leading to mixed results. For example, their findings suggest that the BERT model rejects presuppositions involving numeracy (e.g., those containing the trigger "both") and that, in general, language models occasionally lack knowledge of basic word meanings. Furthermore, Kabbara and Cheung (2022)'s study, which consisted of fine-tuning LLMs on Jeretic et al. (2020)'s ImpPres dataset to assess their performance on tasks involving presuppositions, indicated that the models predominantly relied on surface-level lexical and structural cues, rather than engaging in any form of pragmatic reasoning. Cong (2022) conducted a minimal pair analysis of presuppositions (and scalar implicaturs), also by fine-tuning LLMs on Jeretic et al. (2020)'s ImpPres dataset, testing the language models on a cloze task. The results yieled a mixed picture, for example, revealing that GPT-3's performance was mostly at chance, whereas DistillBERT displayed some understanding of the implications.

The preceding studies concerning the pragmatic knowledge captured in LLMs (not only, but also in the area of presuppositions) highlight that the models face notable challenges in this domain, especially when dealing with more complex forms of reasoning. Nonetheless, these discoveries also emphasize the valuable role of psycholinguistic datasets when evaluating the (pragmatic) linguistic capabilities of language models. Given that the study of anti-presuppositions and the MP! principle has yet to be examined within the context of language models, i.e., representing a domain that awaits exploration, we approach this task by incorporating carefully controlled psycholinguistic data, which we will outline in the next section.

3 Experimental Setup

We investigate whether language models follow the MP! principle by analyzing their predictions involving two different anti-presupposition triggers. To carry out this investigation, we conduct minimal pair analyses. This paradigm involves contrasting two linguistic items that are nearly identical except for a single aspect and is, therefore, particularly well-suited for investigating anti-presuppositions as these tend to appear in pairs. Furthermore, the technique of minimal pair analysis is not only commonly employed in linguistic experiments, but it has also been shown by several studies that it can furthermore be a productive approach for investigating the linguistic properties captured in language models (e.g., Marvin and Linzen (2018); Warstadt et al. (2020); Cong (2022); Hu and Levy (2023)).[1]

Data. We ground our study on German data from Schneider et al. (2019), who investigated the MP! principle with regards to definite and indefinite determiners in the context of visualized stories.[2] The data from Schneider et al. (2019) offer a valuable avenue to investigate anti-presuppositions, given that (anti-)presuppositions are sensitive to context, i.e., their interpretation very much depends on the shared knowledge or beliefs among the conversational participants (cf. Section 2).

Concretely, in their study, Schneider et al. (2019) utilized mouse-tracking to examine how definite and indefinite determiners are processed, both when used felicitously and infelicitously. For this, participants were asked to judge the appropriateness of sentences in the context of a visualized story. Each experimental trial began with a context that featured a shopping basket with three pieces of fruit, accompanied with a context sentence, such as "*Jan's mother was shopping. She bought one banana and two pears*". Then, participants were presented with the next part of the story, where Jan received one of the initially introduced fruit (e.g., "*Of these, Jan received the banana.*"), and asked to judge this stimulus sentence against the provided context by selecting "true" or "false" response boxes. Schneider et al. (2019)'s study had six conditions, resulting from combining two determiners (definite vs. indefinite) with three types of sentences (false vs. felicitous vs. infelicitous). While in infelicitous conditions the (anti-)uniqueness presupposition of the determiner used in the stimulus sentence was violated, in felicitous conditions, the context satisfied this presupposition: That is, for definite determiners, Jan received the unique fruit, while for indefinite determiners, Jan received one of the non-unique fruits. Cf. (5) for sentences examples of the felicitous conditions.[3]

(5) **Context**: Jan's mother was shopping. She bought one banana and two pears.

a. **Felicitous definite:** Of these, Jan received **the** banana.

b. **Felicitous indefinite:** Of these, Jan received **a** pear.

Schneider et al. (2019) find, among other things, that, for both felicitous and infelicitous conditions, the mouse cursor's deviation towards the final response location commenced later for sentences with indefinite determiners, and in general, that there was a delay in response for infelicitous sentences. Their findings, thus, indicate that processing indefinite determiners is more challenging than processing definite determiners, providing evidence in support of the MP! principle.

Prompts. We utilize Schneider et al. (2019)'s experimental items to construct prompts, in which we mask the position of the definite and indefinite determiners, respectively, allowing us to test the language models' predictions for these words (and with it to investigate whether determiners really are "easy" words for LLMs (Yang et al., 2023)). Prompts always consist of a context sentence, followed by the sentence that contains the masked determiner, i.e. they follow the format in (6).

(6) Jan's mother was shopping. She bought one {unique_fruit} and two {non-unique fruits}. Of these, Jan received [MASK] {unique_fruit | non-unique fruit}.

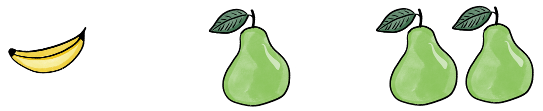
Just like in Schneider et al. (2019)'s study, the second sentence of a prompt starts with the word "Davon" ("Of these") to underscore that Jan received a fruit from the three fruits introduced in the context sentence. To evaluate if the choice of the proper name "Jan" influenced the language models' performance, we experimented with alternative names, including potentially more internationally recognized ones like "Tom" or "Peter." As it turned out, the choice of names did not yield any discernible differences in the results. Therefore, we opted to retain Schneider et al. (2019)'s original prompts. And, as well similar to Schneider et al. (2019), we employ seven fruit types (Banane (*banana*), Zitrone (*lemon*), Orange (*orange*), Birne (*pear*), Ananas (*pineapple*), Pflaume (*pear*), Erdbeere (*strawberry*)), each paired with the feminine determiner to prevent early disambiguation and ensure consistency in sentence structure. The absence

[1] All source code for replicating the experimental investigations can be found here: https://github.com/clause-bielefeld/antipresuppositions.

[2] Their data is publicly available here: https://osf.io/w5yr4/.

[3] As in our study we only make use of the felicitous conditions to investigate MP! in LLMs, we will not further elaborate on the other conditions. Please refer to the original paper for more details, also regarding the experimental setup.

Context: Jan's mother was shopping. She bought one banana and two pears.



(a) **Unique fruit**: Of these, Jan received [the | a] banana.

(b) **Non-unique fruit**: Of these, Jan received [a | the] pear.

(c) **Pair of fruits**: Of these, Jan received [both | all] pears.

Figure 1: Exemplified conditions of our study (images included for illustration purposes only).

of any fruit starting with a vowel further offers the advantage of obviating the need to elicit the probability of English "an".

Conditions. In our study, we make use of the *felicitous definite* and *felicitous indefinite* conditions (cf. (5)), which we refer to as **unique fruit** and **non-unique fruit** condition, respectively (cf. Figure 1). Furthermore, we go beyond Schneider et al. (2019)'s original design and introduce supplementary conditions to enhance the robustness of our design and derive more insightful conclusions about the potential existence of the MP! principle in language models. That is, on the one hand, we extend the original German data with its corresponding English translations, following the pattern as seen in (6), for instance. Firstly, this allows to explore possible crosslingual differences. Secondly, as in English we introduce the unique fruit with the numeral "one" (e.g., "She bought one banana"), this allows us to investigate the inability to interpret the indefinite determiner as a numeral, a contrast to the German "eine" where such an interpretation is feasible (cf. Schneider et al. (2019)).

On the other hand, we include another minimal pair in our investigation, namely the anti-presupposition triggers "beide" (*both*) and "alle" (*all*), referred to with the **pair of fruits** condition (cf. Figure 1). These triggers are not only well compatible with the experimental items, they further allow us to investigate potential differences between anti-presupposition triggers and to compare our investigation to other studies that included these triggers, e.g., Jeretic et al. (2020) (see Section 2). To investigate *both* and *all*, we retain the sentence structure and keep the masked token at the same position as in (6), and simply change the number of the received fruit to the plural form. Here, as *both* is "presuppositionally stronger" than *all* (see Section 2), language models should predict "both" rather than "all" if they follow the MP! principle in their predictions.

All together, for both English and German, we investigate whether LLMs adhere to the MP! principle when making predictions with the conditions summarized and exemplified in Figure 1.[4] In accordance with MP!, the language models should predict the word highlighted in green with a higher score compared to the word highlighted in red.

Models. We use the Hugging Face framework for reproducibility, employing their *Fill-Mask*-pipeline and the models listed below:

1. bert-base-german-cased (for German)
2. bert-base-cased (for English)
3. bert-base-multilingual-cased (for German and English)
4. xlm-roberta-base (for German and English)

We focus our investigation on BERT and BERT model variants. As we are interested in evaluating whether language models adhere to the MP! principle, and fine-tuned models do not necessarily reflect the linguistic properties of language models in general (cf. Ettinger (2020) or Chang and Bergen (2023)), we center our investigation on the predictions of these base models. However, due to less favorable outcomes regarding the models' compliance with the MP! principle, we also investigate whether there could be an effect from fine-tuning these models on exisiting NLI datasets. Furthermore, even though recent LLM evaluations strongly rely on ChatGPT or GPT-4 (e.g., Cai et al. (2023), Kocoń et al. (2023)), we omit such models from our analysis. This is motivated by the absence of token probability access through the OpenAI API. Furthermore, in line with Hu and Levy (2023), we are of the view that that restricting our interactions with LLMs to high-level prompting might result in missing the opportunity to measure and understand their linguistic capabilities more comprehensively.

For fine-tuning the models, we make use of Wang et al. (2019)'s SuperGLUE benchmark.[5] We

[4] For simplicity, Figure 1 shows only the English version.

[5] That is, of its BoolQ (Boolean Questions, Clark et al. (2019)), COPA (Choice of Plausible Alternatives, Roemmele et al. (2011)), MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al. (2018)), CB (CommitmentBank, de Marneffe et al. (2019)), and RTE (Recognizing Textual Entailment) datasets.

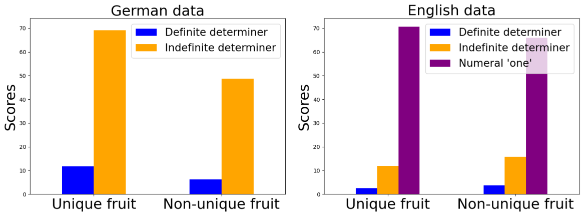


Figure 2: Mean probability scores of definite and indefinite determiners for the *unique fruit* and *non-unique fruit* conditions, as exemplified in Figure 1. German on the left, English on the right. Results are aggregated over the individual LLMs.

preprocess the data by concatenating the pertinent input sequences (similar to Raffel et al. (2019)) and then masking the specific minimal pairs of interest (i.e., the words "the," "a," "all," and "both"). We fine-tune all models on approximately 10.000 datapoints, using three epochs, a learning rate of 2e-5 and a weight decay of 0.01. As the SuperGLUE benchmark is available in English only, we investigate fine-tuning only for English data.[6]

Evaluation. To assess whether the LLMs adhere to the MP! principle in their predictions, we prompt the models with our sentences and get their 150 predicted words and respective probability scores for the masked token. Our assumption is that a language model adheres to the MP! principle in its predictions when the score for the *felicitous* token (e.g., "the"/"die" in the *unique fruit* condition) surpasses the score for the *infelicitous* token (e.g., "a"/"eine" in the *unique fruit* condition). We calculate Ettinger (2020)' metrics of Completion Sensitivity and Prediction Accuracy. With Completion Sensitivity, we measure the percentage of items for which the models assign a higher score to the felicitous token than to the infelicitous token. With Prediction Accuracy, we calculate the percentage of items for which the felicitous token is among the model's top k predictions. Further, in the Tables in Appendix A.2, we collect all mean probability scores. The efficacy of using such direct probability measurements for assessing language model generalization capabilities has been shown by, for example, Hu and Levy (2023) and Hu et al. (2023).

4 Results and Discussion

Determiners. Figure 2 depicts the probability scores for the definite and indefinite determiners, for German and English data, aggregated over the individual models (but see Table 3 in the Appendix for results separated for models). Strikingly, comparing the two determiners across languages and conditions (and also models, cf. Table 3), the indefinite determiner is predominantly predicted with the highest score. Thus, the definite determiner is predicted with exceedingly low scores, even in the *unique fruit* condition; this is particularly noticeable for German. In English, the indefinite determiner also obtains highest scores across the conditions, however, the scores for the indefinite determiner are notably lower. Remarkably, in German, the indefinite determiner receives a higher score even in the *unique fruit* condition compared to the *non-unique fruit* condition. Furthermore, the Completion Sensitivity scores displayed in Table 1 highlight that while all models attain 100% or nearly 100% in the *non-unique fruit* condition, the highest score of the *unique fruit* condition is at only 14.29%, achieved by the bert-base-multilingual-cased model. Still, as visible in the Prediction Accuracy scores in Table 2, the definite determiner, despite not being favored over the indefinite determiner as expected by the MP! principle, appears within the top three predictions in the *unique fruit* condition for most models.

With this distribution it may seem that the models adhere to the MP! principle in their predictions at least for the *non-unique fruit* condition. However, we posit that this observed distribution, instead, is a result of the LLMs merely repeating the singular determiner or numeral present in the context of the prompt. This interpretation is supported by the crosslingual comparison: in German, the determiner "eine" is a homonym for the indefinite determiner "a" and the numeral "one" in English, and the German LLMs repeat this word in all conditions. Plus, in the English prompts, where the unique fruit is introduced with the numeral "one", the LLMs clearly favour to predict this numeral in the continuation, across both conditions, as shown in Figure 2.

Turning our attention now to the results of the fine-tuned models which are presented in Appendix A.2.1. First, in Table 7, we can observe consistently higher values for the two determiners and notably lower values for the number "one", across all models. Thus, fine-tuning on NLI data indeed seems to have a substantial effect on the prediction of the models. However, the models continue to exhibit

[6] Cf. Appendix A.1 for preprocessed examples of each dataset. Also, see the github (https://github.com/clause-bielefeld/antipresuppositions) for more specifics on the data preprocessing and the fine-tuning approach.

model	lang.	unique	n-unique	pair
bert-base-german-cased	DE	0	100	85.71
bert-base-cased	EN	14.29	85.71	90.48
bert-base-multilingual-cased	DE	7.14	92.86	7.14
	EN	14.29	85.71	0
xlm-roberta-base	DE	7.14	92.86	80.95
	EN	0	100	100

Table 1: Completion Sensitivity scores for the *unique fruit* (here: "unique"), *non-unique fruit* (here: "n-unique") and *pair of fruits* (here: "pair") conditions, cf. Figure 1.

model	lang.	unique	n-unique	pair
bert-base-german-cased	DE	95.24	100	59.52
bert-base-cased	EN	0	85.71	4.76
bert-base-multilingual-cased	DE	100	85.71	85.71
	EN	100	85.71	0
xlm-roberta-base	DE	80.95	78.57	83.33
	EN	16.67	100	85.71

Table 2: Prediction Accuracy scores for for the *unique fruit* (here: "unique"), *non-unique fruit* (here: "n-unique") and *pair of fruits* (here: "pair") conditions, cf. Figure 1. For *unique fruit* and *non-unique fruit* conditions $k = 3$, for *pair of fruits* condition $k = 40$.

a clear inclination toward predicting the indefinite determiner more strongly than the definite determiner, again also within the *unique fruit* condition. This is also supported by the Completion Sensitivity scores in Table 5. Further, although Completion Sensitivity in the *unique fruit* condition for the fine-tuned xlm-roberta model slightly improves, all other values remain unchanged. However, the Prediction Accuracy scores in Table 6 show that the determiners are now almost always among the top three predictions, which is an improvement, for example, for the bert-base-cased and xlm-roberta-base models in the *unique fruit* condition. Still, overall, fine-tuning on NLI data does not seem to have resulted in the models adhering more closely to the MP! principle or decreasing their reliance on replicating bigrams from the provided context. Rather, the results indicate that masking the determiners during fine-tuning has caused the models to overly predict these words without effectively capturing underlying linguistic patterns.

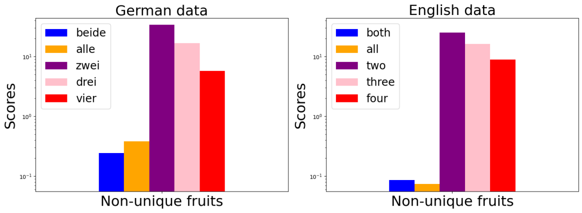


Figure 3: Mean probability scores of "beide"/"both" and "alle"/"all" for the *pair of fruits* condition, as exemplified in Figure 1. As scores were so small, we use a logarithmic scale instead of a linear scale. German on the left, English on the right. Results are aggregated over the individual LLMs.

"Both" and "all". The outcomes concerning our supplementary condition *pair of fruits* involving the presupposition triggers "beide"/"both" and "alle"/"all" are depicted in Figure 3 (see Table 4 in the Appendix for results separated for models). First, it is noteworthy that the scores assigned to these anti-presupposition triggers in both German and English are notably low. Thus, in general, the models rarely predict "beide"/"both" or "alle"/"all" when prompted with our sentences. This is also substantiated by the fact that we had to adjust the parameter k to a value of 40 when calculating the Prediction Accuracy for this condition, in order to obtain comparable values at all (cf. Table 2). Further, we can observe that the German triggers "beide" and "alle" are predicted with a higher probability compared to their English counterparts. However, only for English, "both" is predicted with a higher probablity than "all", as expected according to the MP! principle.

Intriguingly, we find noticeable differences between the models, as, for example, shown by the Completion Sensitivity in Table 1. For instance, while the results for the models bert-base-german-cased and xlm-roberta-base may suggest that these two models, to some extent, adhere to the MP! principle in their predictions, the results of the bert-base-multilingual-cased model speak otherwise (also visible in Table 4). Given the overall remarkably low values for the words "both" and "all", we further examined the predicted words and scores for the numerals "zwei"/"two", "drei"/ "three" and "vier"/"four" (cf., e.g., Figure 3 again). Clearly, we find that the models predominantly predict numbers for the masked tokens and, particularly the number "zwei"/"two", which, in this context, appears to be sensible. Intriguingly, however, the models also tend to predict those numbers that do not align with the given context at all. That is, if the context indicates that the mother only bought two bananas, Jan should not be receiving three or more of them. Interestingly, these results are reminiscent of the outcomes observed by Jeretic et al. (2020). On the one hand, they suggest a potential deficiency in the models' comprehension of basic word meanings, such as that of the term "two". On the other hand, these authors similarly encountered challenges for

LLMs when dealing with presuppositions involving numeracy (cf. Section 2). We posit that the general inclination to predict numerals in this condition, and especially the number "zwei"/"two", might again be attributed to the models relying on (and replicating) words detected in the prompt rather than considering pragmatic principles such as the MP! principle.

If we now briefly examine the scores of the fine-tuned models for this condition too, as shown in the Tables in Appendix A.2.1, it is evident that, even though there seems to be some enhancement in predictions concerning the adherence to the MP!, across all models, numbers (including those that do not make sense in the given context) are still predicted with the highest probability. These results, thus, again suggest that fine-tuning on NLI data does not enhance the models' capacity to adhere to the MP! principle in their predictions. Instead, they underscore that the models continue to replicate patterns from the context of the prompts, much like the outcomes observed in the other conditions. The rather limited difference in values between fine-tuned and base models for "all" and "both," unlike the distinct contrast observed for "the" and "a", could potentially stem from the less frequent occurrence of the words "all" and "both" in the datasets compared to the determiners.[7] See Section A.2.1 for a more detailled discussion of the results of the fine-tuned models in the *pair of fruits* condition.

Additional conditions. In order to gain further support for our finding that the LLMs primarily repeat patterns from the provided context, we examined additional conditions (with models and evaluation metrics remaining the same) and provide them in the Appendix: the conditions are documented in examples (7) to (11) in Appendix A.2.2, the results can be found in the subsequent Tables 9 to 14. We will briefly summarize these here.

The conditions (7), (8) and (9) were constructed to further examine the notion that the models predominantly reproduce words from the prompt when making predictions. This was achieved by incorporating the target words into the context of the prompts. The results provided in Tables 9, 10 and 11, respectively, make the impression that all models in all conditions (and languages) now follow the MP! principle. However this alignment, instead, implies a departure from the principle itself: The tendency to now predict the words "die/"the", "beide"/"both" as well as "a" for English with a higher score can be traced back to these words being introduced in the context sentence. In other words, these results support our proposition that LLMs primarily adhere to context-based bigrams and replicate them in their predictions, rather than operating in accordance with the MP! principle.

Next, see the conditions (10) and (11), aimed at investigating the possibility of "manipulating" the language models – not through fine-tuning, but via an alternative form of prompting – in order to encourage them to align more closely with the MP! principle. The results depicted in Tables 12 to 11 show that neither of the two strategies has succeeded in shifting the models' predictions towards the MP! principle. That is, all the Tables exhibit a striking resemblance in their distribution to their corresponding counterparts from the initial experiment. Furthermore, it appears that particularly the scenario where the prompt is enriched with additional information ((11)) has no (or even a negative) influence on the models' predictions – which could once again imply that language models occasionally struggle with basic word meanings (cf. Jeretic et al. (2020)).

Recap and comparative analysis. Summing up, our investigation shows that the LLMs investigated here do not follow the MP! principle when predicting masked tokens. That is, the results across all our conditions (i.e., those presented in Figure 1 and in (7) to (11)) strongly support the interpretation that the models' alternative strategy is based on replicating words detected in the given input prompt. With these results, we find analogies to existing research. For example, similar to some of Ettinger (2020)'s results on the CPRAG-102 test, our findings suggest that LLMs do not appropriately take into account the context provided by the preceding sentence. Also, our findings point to a potential deficiency in the models' comprehension of basic word meanings and challenges when dealing with presuppositions involving numeracy, a shortcoming also noted by Jeretic et al. (2020). And while Parrish et al. (2021) found that within their NOPE dataset, which consists of natural language examples, numerical determiners presented only minor challenges for NLI models, they too encountered similar challenges with newer models when testing them on Jeretic et al. (2020)' ImpPres dataset. Additionally, our findings point to an inadequacy of BERT model variants in effectively

[7] Which is, e.g., visible in the examples in Appendix A.1.

acquiring representations of entities – a factor that has been demonstrated to enhance the quality of generated text – aligning with observations made by, for example, Févry et al. (2020) and Yamada et al. (2020). And, what is more, our results show great resemblance to those of Kim et al. (2019) who investigated language models' understanding of function words. While their study wasn't focused on testing for MP!, it revealed that a probing classifier could not differentiate between correct and incorrect uses of definite or indefinite articles any better than chance, too countering the notion that determiners only present minimal challenges for LLMs (Yang et al., 2023).

5 Conclusion

This paper investigated whether LLMs follow the MP! principle by analyzing their predictions involving two different minimal pairs of anti-presupposition triggers, i.e., expanding prior work on probing the semantic and pragmatic discourse knowledge captured in LLMs. Our findings reveal that the language models we studied do not follow the MP! principle when predicting masked determiners but, instead, tend to repeat words from the given input prompt. Furthermore, we find that fine-tuning language models on NLI datasets does not enhance their capacity to adhere to the MP! principle; rather, it appears to lead to an excessive prediction of the triggers of interest, devoid of capturing underlying linguistic patterns. The observed deviation from the MP! principle contrasts with the assumption that determiners only pose minimal challenges for LLMs (Yang et al., 2023) and lends further support for the hypothesis that these models cannot truly grasp and reflect fundamental aspects of language use that govern presupposition and pragmatics. However, it's important to acknowledge that our conclusions are based on our choice of models and a relatively limited range of examples, and further research is needed to confirm and qualify this observation.

Considering that (anti-)presuppositions are omnipresent in everyday language, it is imperative for language models to effectively capture them, for example, in tasks such as Question Answering (Kim et al., 2021; Yu et al., 2023). Therefore, future research directions might involve exploring the potential utility of Visual Question Answering tasks to gain deeper insights into the models' understanding of prompts and whether they are able to appropriately represent discourse entities within them.

Limitations

While our study provides valuable insights into the behavior of LLMs regarding the MP! principle, we acknowledge the limitations of our approach that may restrict the generalizability of our results. One major limitation of our work is its confinement to a singular domain, i.e., to the context related to fruits, which we adopted from Schneider et al. (2019). While this delimited context was sufficient to illustrate the lack of MP! adherence in the studied language models and their difficulty to predict determiners, it would be beneficial to encompass a broader set of contexts and various types of prompts (or anti-presupposition triggers), in order to ascertain the generalizability of our findings. Another limitation of our study pertains to the limited exploration of differences between individual LLMs and to employing one fine-tuning-approach only, owing to space limitations. Additionally, we did not investigate language models from alternate families, for example, including those based on the GPT architecture. This limitation stemmed in part from the challenge of devising suitable prompts for such models. Therefore, we recognize the need for further research that encompasses a broader range of models of different sizes and training objectives, a more diverse set of templates, and an increased dataset size to achieve a more comprehensive understanding of how LLMs interact with the MP! principle.

Ethics Statement

The data we used in this study was obtained either from psycholinguistic publications or was generated by the authors without the use of harmful content. Additionally, no experiments were conducted involving human participants, and no new models or datasets are being introduced. The primary objective of this paper is to provide insights into internal knowledge of modern LLMs and contribute to enhancing their interpretability. Therefore, while we do not foresee any ethical concerns specific to this paper, the broader ethical concerns pertaining to LLMs remain of relevance to our research (cf., e.g., Bender et al. (2021).

Acknowledgements

The authors acknowledge financial support by the project “SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia.

References

Nadine Bade and Florian Schwarz. 2021. New data on the competition between definites and indefinites. *Experiments in Linguistic Meaning*, 1(0):15–26.

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2021. Presupposition. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

David Blunier. 2022. Antipresuppositions, logophors and shifted indexicality. In *Proceedings of the 23rd Amsterdam Colloquium.*, pages 45–54.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does ChatGPT resemble humans in language use?

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *arXiv.org*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Yan Cong. 2022. Psycholinguistic diagnosis of language models' commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 17–22, Dublin, Ireland. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23.

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.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

H. P. Grice. 1975. *Logic and Conversation*, pages 41 – 58. Brill, Leiden, Niederlande.

Irene Heim. 1991. *Artikel und Definitheit*, pages 487–535. De Gruyter Mouton, Berlin • New York.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations.

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.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213, Florence, Italy. Association for Computational Linguistics.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Francesca Panzeri and Francesca Foppolo. 2021. Children's and adults' sensitivity to gricean maxims and to the maximize presupposition principle. *Front. Psychol.*, 12:624628.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Orin Percus. 2006. Antipresuppositions. *Theoretical and Empirical Studies of Reference and Anaphora: Toward the establishment of generative grammar as an empirical science,*, pages 52–73.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified Text-to-Text transformer.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Cosima Schneider, Carolin Schonard, Michael Franke, Gerhard Jäger, and Markus Janczyk. 2019. Pragmatic processing: An investigation of the (anti-)presuppositions of determiners using mouse-tracking. *Cognition*, 193:104024.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Judith Sieker, Oliver Bott, Torgrim Solstad, and Sina Zarrieß. 2023. Beyond the bias: Unveiling the quality of implicit causality prompt continuations in language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 206–220, Prague, Czechia. Association for Computational Linguistics.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online. Association for Computational Linguistics.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2023. Learning better masking for better language model pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7255–7267, Toronto, Canada. Association for Computational Linguistics.

Kazuko Yatsushiro. 2008. Quantifier acquisition: Presuppositions of "every". *SuB*, 12:663–677.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Data Preprocessing

Here, we present examples of our preprocessing approach for each of the datasets we consider from Wang et al. (2019)'s SuperGLUE benchmark.

BoolQ (Boolean Questions, Clark et al. (2019))

- **Original input:**
  - question – is there any meat in a chiko roll
  - passage – Chiko Roll – A Chiko Roll's filling is primarily cabbage and barley, as well as carrot, green beans, beef, beef tallow, wheat cereal, celery and onion. This filling is partially pulped and enclosed in a thick egg and flour pastry tube designed to survive handling at football matches. The roll is typically deep-fried in vegetable oil.
  - label – 1
- **Preprocessed input:**
  - Chiko Roll – A Chiko Roll's filling is primarily cabbage and barley, as well as carrot, green beans, beef, beef tallow, wheat cereal, celery and onion. This filling is partially pulped and enclosed in a thick egg and flour pastry tube designed to survive handling at football matches. The roll is typically deep-fried in vegetable oil. Is there any meat in a chiko roll? Yes.
- **Masked input:**
  - Chiko Roll – [MASK] Chiko Roll's filling is primarily cabbage and barley, as well as carrot, green beans, beef, beef tallow, wheat cereal, celery and onion. This filling is partially pulped and enclosed in [MASK] thick egg and flour pastry tube designed to survive handling at football matches. [MASK] roll is typically deep-fried in vegetable oil. Is there any meat in [MASK] chiko roll? Yes.

COPA (Choice of Plausible Alternatives, Roemmele et al. (2011))

- **Original input:**
  - premise – The girl had a phobia of dogs.
  - choice1 – She rescued an abandoned dog.
  - choice2 – She was bitten by a dog.
  - question – cause
  - label – 1
- **Preprocessed input:**
  - The girl had a phobia of dogs. What was the cause for this? She was bitten by a dog.
- **Masked input:**
  - [MASK] girl had [MASK] phobia of dogs. What was [MASK] cause for this? She was bitten by [MASK] dog.

MultiRC (Multi-Sentence Reading Comprehension, Khashabi et al. (2018))

- **Original input:**

  - paragraph – You know that friction also causes heat. Think about when you rub your hands together. It is friction that makes them warm. But why does this happen? Friction causes the molecules on rubbing surfaces to move faster. Faster moving particles have more heat energy. Heat from friction can be useful. Can you think of other places where you might find friction? Friction also lets you light a match. Heat from friction can also cause problems. It can cause a car to overheat. To reduce friction, oil is added to the engine. Oil coats the surfaces of moving parts. This coating of oil makes them slippery. When things are slippery there is less friction. Have you ever seen a sign that says, slippery when wet? This too has to do with friction. Water, like oil, can reduce friction. The wet surface may allow your shoes to slide more easily.
  - question – What can happen in a car when there is too much friction?
  - answer – It can overheat

- **Preprocessed input:**
  - You know that friction also causes heat. Think about when you rub your hands together. It is friction that makes them warm. But why does this happen? Friction causes the molecules on rubbing surfaces to move faster. Faster moving particles have more heat energy. Heat from friction can be useful. Can you think of other places where you might find friction? Friction also lets you light a match. Heat from friction can also cause problems. It can cause a car to overheat. To reduce friction, oil is added to the engine. Oil coats the surfaces of moving parts. This coating of oil makes them slippery. When things are slippery there is less friction. Have you ever seen a sign that says, slippery when wet? This too has to do with friction. Water, like oil, can reduce friction. The wet surface may allow your shoes to slide more easily. What can happen in a car when there is too much friction? It can overheat.

- **Masked input:**
  - You know that friction also causes heat. Think about when you rub your hands together. It is friction that makes them warm. But why does this happen? Friction causes [MASK] molecules on rubbing surfaces to move faster. Faster moving particles have more heat energy. Heat from friction can be useful. Can you think of other places where you might find friction? Friction also lets you light [MASK] match. Heat from friction can also cause problems. It can cause [MASK] car to overheat. To reduce friction, oil is added to [MASK] engine. Oil coats [MASK] surfaces of moving parts. This coating of oil makes them slippery. When things are slippery there is less friction. Have you ever seen [MASK] sign that says, slippery when wet? This too has to do with friction. Water, like oil, can reduce friction. [MASK] wet surface may allow your shoes to slide more easily. What can happen in [MASK] car when there is too much friction? It can overheat.

CB (CommitmentBank, de Marneffe et al. (2019))

- **Original input:**
  - premise – Matthew rode on feeling a little more at peace with himself. He skirted the spruce plantation and supposed that at some point he should tell Sara about it. He could imagine that she might be interested in its money-making propensity at the end of the year.
  - hypothesis – Sara might be interested in its money-making propensity at the end of the year
  - label – 0

- **Preprocessed input:**
  - Sara might be interested in its money-making propensity at the end of the year: Matthew rode on feeling a little more at peace with himself. He skirted the spruce plantation and supposed that at some point he should tell Sara about it. He could imagine that she might be interested in its money-making propensity at the end of the year.

- **Masked input:**
  - Sara might be interested in its money-making propensity at [MASK] end of [MASK] year: Matthew rode on feeling [MASK] little more at peace with himself. He skirted [MASK] spruce plantation and supposed that at some point he should tell Sara about it. He could imagine that she might be interested in its money-making propensity at [MASK] end of [MASK] year.

RTE (Recognizing Textual Entailment) datasets

- **Original input:**
  - premise – The University has also apologized for the incident saying, "[we] are very sorry that the incident happened and the person will be dealt with according to law. The university is a place for discussion, debate and considered argument, not for shoe throwing". According to authorities, there was never any real threat to the prime minister. The man will appear before a judge on February 10.
  - hypothesis – A man threw a shoe at the Prime Minister.
  - label – 0
- **Preprocessed input:**
  - A man threw a shoe at the Prime Minister: The University has also apologized for the incident saying, "[we] are very sorry that the incident happened and the person will be dealt with according to law. The university is a place for discussion, debate and considered argument, not for shoe throwing". According to authorities, there was never any real threat to the prime minister. The man will appear before a judge on February 10.
- **Masked input:**
  - [MASK] man threw [MASK] shoe at [MASK] Prime Minister: [MASK] University has also apologized for [MASK] incident saying, "[we] are very sorry that [MASK] incident happened and [MASK] person will be dealt with according to law. [MASK] university is [MASK] place for discussion, debate and considered argument, not for shoe throwing". According to authorities, there was never any real threat to [MASK] prime minister. [MASK] man will appear before [MASK] judge on February 10.

A.2 Further Results

model	**language**	**condition**	**def. determiner**	**indef. determiner**	**English ’one’**
bert-base-german-cased	German	unique fruit	8.91	69.11	–
		non-unique fruit	8.10	**67.52**	–
bert-base-cased	English	unique fruit	1.38	8.65	78.95
		non-unique fruit	1.46	**10.46**	78.77
bert-base-multilingual-cased	German	unique fruit	19.16	64.96	–
		non-unique fruit	7.79	**28.34**	–
	English	unique fruit	5.99	20.78	42.65
		non-unique fruit	8.96	**25.72**	33.50
xlm-roberta-base	German	unique fruit	7.18	73.28	–
		non-unique fruit	2.76	**50.43**	–
	English	unique fruit	0.51	6.30	90.05
		non-unique fruit	0.46	**11.28**	85.42

Table 3: Model predictions for German and English data for the *unique fruit* and *non-unique fruit* conditions, as exemplified in Figure 1. Also, for English, predictions for the numeral "one". We report the mean values of each word to be the predicted masked token as the final scores (mutliplied by 100). Bold values indicate results that conform to the predictions of the MP!.

model	**language**	**condition**	**beide/both**	**alle/all**	**zwei/two**	**drei/three**	**vier/four**
bert-base-german-cased	German	pair of fruits	**0.15**	0.09	40.82	11.44	5.02
bert-base-cased	English		**0.11**	0.08	27.72	20.27	9.55
bert-base-multilingual-cased	German	pair of fruits	0.38	0.89	32.28	16.98	2.94
	English		0.04	0.11	6.52	8.19	4.67
xlm-roberta-base	German	pair of fruits	**0.20**	0.18	28.89	21.39	9.15
	English		**0.11**	0.04	40.90	19.94	12.45

Table 4: Model predictions for German and English data for the *pair of fruits* condition, as exemplified in Figure 1. We report the mean values of each word to be the predicted masked token as the final scores (mutliplied by 100). Bold values indicate results that conform to the predictions of the MP!.

A.2.1 Fine-tuned models

model	language	unique fruit	non-unique fruit	pair of fruits
bert-base-cased^{FT}	English	14.29	85.71	69.05
bert-base-multilingual-cased^{FT}	English	14.29	85.71	100
xlm-roberta-base^{FT}	English	1.19	98.81	100

Table 5: Completion Sensitivity Accuracy scores for the **fine-tuned** models for English and the *unique fruit*, *non-unique fruit* and *pair of fruits* conditions, as exemplified in Figure 1.

model	language	unique fruit	non-unique fruit	pair of fruits
bert-base-cased^{FT}	English	100	85.71	30.95
bert-base-multilingual-cased^{FT}	English	100	100	38.10
xlm-roberta-base^{FT}	English	100	100	54.76

Table 6: Prediction Accuracy scores for the **fine-tuned** models for English and the *unique fruit*, *non-unique fruit* and *pair of fruits* conditions, as exemplified in Figure, as exemplified in Figure 1. Here, for all conditions $k = 3$.

model	language	condition	def. determiner	indef. determiner	English 'one'
bert-base-cased^{FT}	English	unique fruit	17.46	80.43	1.39
		non-unique fruit	15.95	**81.47**	1.60
bert-base-multilingual-cased^{FT}	English	unique fruit	15.55	83.94	0.27
		non-unique fruit	15.01	**84.66**	0.18
xlm-roberta-base^{FT}	English	unique fruit	4.63	94.46	0.66
		non-unique fruit	4.41	**94.69**	0.54

Table 7: Predictions of the **fine-tuned** models for English data for the *unique fruit* and *non-unique fruit* conditions, as exemplified in Figure 1. Also, predictions for the numeral "one". Again, for each condition, we report the mean values of the determiners to be the predicted masked token as the final score, and we multiply the mean scores by 100. Bold values indicate conditions where the results conform to the predictions of MP!.

model	language	condition	beide/both	alle/all	zwei/two	drei/three	vier/four
bert-base-cased^{FT}	English	pair of fruits	**10.78**	7.33	15.88	12.24	3.61
bert-base-multilingual-cased^{FT}	English	pair of fruits	**3.88**	1.90	5.55	3.87	2.73
xlm-roberta-base^{FT}	English	pair of fruits	**12.63**	2.10	11.28	15.75	6.78

Table 8: Predictions of the **fine-tuned** models for English data for the *pair of fruits* condition, as exemplified in Figure 1. Again, we report the mean values of each word to be the predicted masked token as the final score, and we multiply the mean scores by 100. Bold values indicate conditions where the results conform to the predictions of MP!.

"Both" and "all". In Tables, 5, 6 and 8, it is evident that, in comparison to the base models, "both" is now being predicted with a higher probability than "all" for all models, i.e., including the bert-base-multilingual-cased model. Additionally, as visible in Table 8, the values for the numbers show a decrease. However, there is also an increase in the values for "all". Most importantly, across all models, numbers (including those that do not make sense in the given context) are still predicted with the highest probability. Interestingly, the Completion Sensitivity values in Table 5 once again exhibit variations among the models. That is, fine-tuning leads to a substantial enhancement in the bert-base-multilingual-cased model (from 0% to 100%), while the xlm-roberta-base model maintains its high performance and the bert-base-cased model, in contrast, even experiences a decline. Conversely, the Prediction Accuracy results in Table 6 demonstrate a marked enhancement across all models, as evident from our ability to set parameter $k = 3$. Remarkably, for all models, "both" now ranks within the top three predictions, a significant shift from its previous distribution within the top 40 predictions.

A.2.2 Additional conditions

In comparison to our main conditions exemplified in Figure 1, we indicate the modified or newly included words in the prompts of these additional conditions by underlining.

(7) **[Definite determiner in the context sentence.]**
Context: Jan's mother was shopping. She bought the banana and two pears.
a. **Unique fruit** (the / a): Of these, Jan received [MASK] banana.
b. **Non-unique fruit** (a / the): Of these, Jan received [MASK] pear.

(8) **[Indefinite determiner in the context sentence for English.]**
Context: Jan's mother was shopping. She bought a banana and two pears.
a. **Unique fruit** (the / a): Of these, Jan received [MASK] banana.
b. **Non-unique fruit** (a / the): Of these, Jan received [MASK] pear.

(9) **["both" in the context sentence.]**
Context: Jan's mother was shopping. She bought one banana and both pears.
a. **Pair of fruits** (both / all): Of these, Jan received [MASK] pears.

(10) **[Adjectives "einzige"/"single" in the context sentence.]**
Context: Jan's mother was shopping. She bought a single banana and two pears.

a. **Unique fruit** (the / a): Of these, Jan received [MASK] banana.
b. **Non-unique fruit** (a / the): Of these, Jan received [MASK] pear.

(11) **[Extended information in the stimulus sentence.]**
Context: Jan's mother was shopping. She bought one banana and two pears.

a. **Unique fruit** (the / a): Of the items that Jan's mother bought, Jan received [MASK] banana."
b. **Non-unique fruit** (a / the): Of the items that Jan's mother bought, Jan received [MASK] pear."
c. **Pair of fruits** (both / all): Of the items that Jan's mother bought, Jan received [MASK] pears.

model	language	condition	def. determiner	indef. determiner	English 'one'
bert-base-german-cased	German	unique fruit	**48.01**	29.54	–
		non-unique fruit	23.05	**51.47**	–
bert-base-cased	English	unique fruit	**44.48**	12.97	32.88
		non-unique fruit	11.74	**15.05**	62.41
bert-base-multilingual-cased	German	unique fruit	**83.80**	4.74	–
		non-unique fruit	15.93	**17.47**	–
	English	unique fruit	**82.99**	9.19	0.64
		non-unique fruit	48.05	23.65	7.66
xlm-roberta-base	German	unique fruit	**41.28**	19.42	–
		non-unique fruit	9.65	**36.91**	–
	English	unique fruit	**20.13**	11.50	63.10
		non-unique fruit	12.86	**13.87**	69.18

Table 9: Model predictions for German and English data for the *unique fruit* and *non-unique fruit* conditions with the **definite determiner in the context sentence**, as exemplified in (7).

model	language	condition	def. determiner	indef. determiner	English 'one'
bert-base-cased	English	unique fruit	5.68	34.19	42.36
		non-unique fruit	2.89	**23.00**	61.07
bert-base-multilingual-cased	English	unique fruit	14.79	61.95	3.24
		non-unique fruit	14.21	**51.15**	9.62
xlm-roberta-base	English	unique fruit	2.04	25.08	68.90
		non-unique fruit	1.03	**23.16**	71.24

Table 10: Model predictions for the *unique fruit* and *non-unique fruit* conditions with the **indefinite determiner in the context sentence for English**, as exemplified in (8).

model	language	condition	beide/both	alle/all	zwei/two	drei/three	vier/four
bert-base-german-cased	German	pair of fruits	**5.98**	0.75	33.92	7.37	2.86
bert-base-cased	English	pair of fruits	**0.72**	0.17	38.60	18.16	7.04
bert-base-multilingual-cased	German	pair of fruits	**23.38**	10.42	13.99	10.86	0.65
	English	pair of fruits	**0.39**	0.31	10.82	9.89	5.29
xlm-roberta-base	German	pair of fruits	**3.94**	0.56	41.60	15.15	4.94
	English	pair of fruits	**0.73**	0.09	55.72	17.69	8.15

Table 11: Model predictions for German and English data for the *pair of fruits* condition with **"both" in the context sentence**, as exemplified in (9).

model	language	condition	def. determiner	indef. determiner	English ’one’
bert-base-german-cased	German	unique fruit	8.47	62.04	–
		non-unique fruit	7.56	**65.34**	–
bert-base-cased	English	unique fruit	1.74	9.18	75.46
		non-unique fruit	1.52	**10.44**	78.05
bert-base-multilingual-cased	German	unique fruit	9.95	42.49	–
		non-unique fruit	5.59	**24.69**	–
	English	unique fruit	7.34	20.31	32.19
		non-unique fruit	9.49	**26.70**	28.10
xlm-roberta-base	German	unique fruit	3.30	49.30	–
		non-unique fruit	1.79	**41.83**	–
	English	unique fruit	0.36	5.74	90.32
		non-unique fruit	0.48	**9.50**	87.39

Table 12: Model predictions for German and English data for the *unique fruit* and *non-unique fruit* conditions with the **adjectives "einzige"/"single" in the context sentence**, as exemplified in (10).

model	language	condition	def. determiner	indef. determiner	English ’one’
bert-base-german-cased	German	unique fruit	7.61	57.68	–
		non-unique fruit	7.25	**56.33**	–
bert-base-cased	English	unique fruit	1.82	7.05	79.67
		non-unique fruit	1.94	**8.25**	80.11
bert-base-multilingual-cased	German	unique fruit	14.76	65.44	–
		non-unique fruit	4.67	**29.82**	–
	English	unique fruit	2.44	17.01	68.75
		non-unique fruit	3.78	**29.35**	50.84
xlm-roberta-base	German	unique fruit	1.91	60.91	–
		non-unique fruit	0.73	**38.63**	–
	English	unique fruit	0.86	10.32	85.02
		non-unique fruit	0.85	**15.09**	80.75

Table 13: Model predictions for German and English data for the *unique fruit* and *non-unique fruit* conditions with **extended information in the stimulus sentence**, as exemplified in (11).

model	language	condition	beide/both	alle/all	zwei/two	drei/three	vier/four
bert-base-german-cased	German	pair of fruits	0.14	0.18	12.76	4.30	1.37
bert-base-cased	English	pair of fruits	0.10	0.18	21.76	18.59	9.26
bert-base-multilingual-cased	German	pair of fruits	0.37	1.10	41.81	14.04	2.20
	English	pair of fruits	**0.11**	0.05	34.21	18.36	7.09
xlm-roberta-base	German	pair of fruits	**0.13**	0.11	33.80	17.52	8.62
	English	pair of fruits	**0.17**	0.10	40.70	19.87	10.21

Table 14: Model predictions for German and English data for the *pair of fruits* condition with **extended information in the stimulus sentence**, as exemplified in (11).