Token-length Bias in Minimal-pair Paradigm Datasets

Naoya Ueda[1], **Masato Mita**[2,1], **Teruaki Oka**[1], **Mamoru Komachi**[3]
[1]Tokyo Metropolitan University, [2]CyberAgent Inc., [3]Hitotsubashi University
Tokyo, Japan
ueda-naoya@ed.tmu.ac.jp, mita_masato@cyberagent.co.jp,
teruaki.oka.3@gmail.com, mamoru.komachi@r.hit-u.ac.jp

Abstract

Minimal-pair paradigm datasets have been used as benchmarks to evaluate the linguistic knowledge of models and provide an unsupervised method of acceptability judgment. The model performances are evaluated based on the percentage of minimal pairs in the MPP dataset where the model assigns a higher sentence log-likelihood to an acceptable sentence than to an unacceptable sentence. Each minimal pair in MPP datasets is controlled to align the number of words per sentence because the sentence length affects the sentence log-likelihood. However, aligning the number of words may be insufficient because recent language models tokenize sentences with subwords. Tokenization may cause a token length difference in minimal pairs, introducing token-length bias that skews the evaluation results. This study demonstrates that MPP datasets suffer from token-length bias and fail to evaluate the linguistic knowledge of a language model correctly. The results proved that sentences with a shorter token length would likely be assigned a higher log-likelihood regardless of their acceptability, which becomes problematic when comparing models with different tokenizers. To address this issue, we propose a debiased minimal pair generation method, allowing MPP datasets to measure language ability correctly and provide comparable results for all models.

Keywords: acceptability judgments, minimal-pair paradigms, token-length bias

1. Introduction

Various methods and benchmarks have been proposed to measure the linguistic knowledge of language models because general-purpose language models that have acquired superior linguistic knowledge exhibit high performance across domains and tasks (Wang et al., 2018, 2019). An acceptability judgment task is a standard method for measuring linguistic knowledge of language models. This task determines whether a given sentence is grammatically acceptable or unacceptable (Chomsky, 1957; Schutze, 1996). The most widely used acceptability judgment corpus is the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019), which requires the training of a supervised classifier to measure the linguistic knowledge of a language model. However, using a supervised classifier for acceptability judgment has a limitation in that it is unclear whether the language model learned the linguistic knowledge during pretraining or acquired it during the supervised training of the classifier (Warstadt et al., 2020).

Unsupervised methods of acceptability judgment have attracted much attention to overcome this limitation. Among the unsupervised method, the use of minimal-pair paradigm (MPP) datasets has become a widespread approach to measure the linguistic knowledge of a language model (Warstadt et al., 2020; Nangia et al., 2020; Misra et al., 2023). MPP datasets consist of minimal pairs —a pair of acceptable and unacceptable sentences that minimally differs by one word (Linzen et al., 2016). The linguistic knowledge of a language model is evaluated based on the percentage of minimal pairs where the model assigns a higher acceptability score to an acceptable sentence than to an unacceptable sentence. The log-likelihood of a sentence is generally used as an acceptability score of a sentence in the MPP dataset. This metric is used under the assumption that the language model should estimate a higher log-likelihood for an acceptable sentence if the model has acquired the correct linguistic knowledge.

In MPP datasets, each minimal pair is controlled to align the number of words per sentence because the sentence length affects the sentence log-likelihood (Figure 1-a) (Warstadt et al., 2020). However, this constraint is insufficient for evaluating pretrained language models, such as GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019). This is because current pretrained language models tokenize sentences with subwords, generating acceptable and unacceptable sentences with different token lengths (Figure 1-b). Such token length difference may cause token-length bias that confuses the evaluation results because it is known that the token length also affects the log-likelihood of a sentence (Figure 1-c) (Kauf and Ivanova, 2023). Moreover, whether the token-length bias affects the evaluation results of the MPP datasets is unclear.

Thus, this study focuses on the effect of token-length bias on evaluations using the MPP dataset. We aim to answer the following research questions (RQ):

LREC-COLING 2024, pages 16224–16236
20-25 May, 2024.

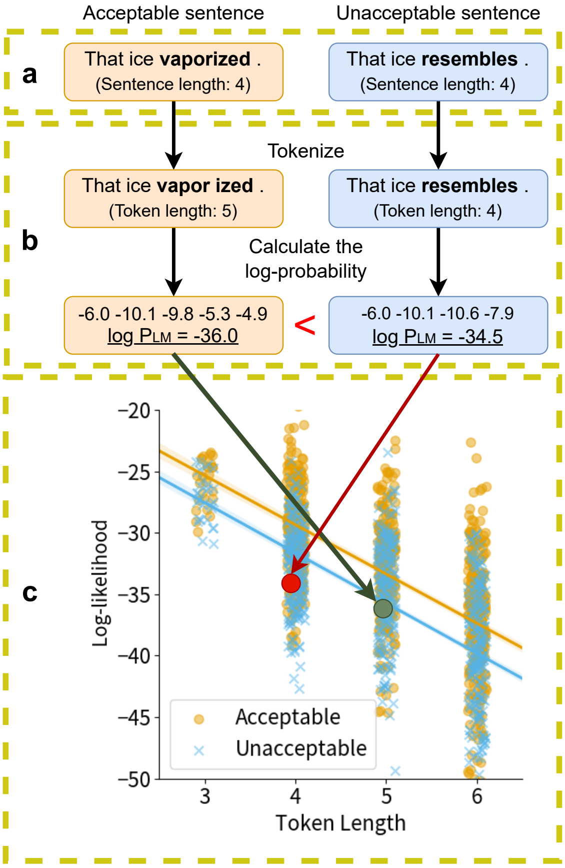

Figure 1: Example of token-length bias in an MPP dataset, which causes the unacceptable sentence with a higher log-likelihood than the acceptable sentence. **(a)** Original minimal pair with the same sentence length, **(b)** token-length bias due to the subword tokenization, and **(c)** correlation of the sentence log-likelihood and the token length.

RQ1 *Does the token-length bias affect the evaluation results of the MPP dataset?*
We aim to determine the presence of the token-length bias in the MPP datasets and its effect on the evaluation results using MPP datasets. We describe our approach to demonstrate their effect in Section 3.1 and show the experimental results in Section 5.1.

RQ2 *Is it effective to use normalized log-likelihood as an acceptability score?*
Normalizing the log-likelihood to the token length may mitigate the token-length bias, but their effectiveness on the MPP dataset is unknown. We aim to investigate their effectiveness in mitigating the token-length bias. We explain the approach in Section 3.2 and show the experimental results in Section 5.2.

We experimented on the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), Crowdsourced Stereotype Pairs Benchmark (CrowS-Pairs) (Nangia et al., 2020), and Conceptual Minimal Pair Sentences (COMPS) (Misra et al., 2023) datasets. Experimental results demonstrated that MPP datasets suffer from token-length bias, which prevents them from correctly evaluating the linguistic knowledge of the models. Furthermore, we show that using the normalized sentence log-likelihood fails to mitigate token-length bias, confusing the model analysis. Therefore, this study proposed a minimal pair generation method to remove the token-length bias from the MPP dataset and used this method to regenerate the debiased BLiMP dataset named FairBLiMP. The result obtained from the FairBLiMP dataset showed that using the biased BLiMP may lead to a wrong conclusion, indicating the necessity of a token-length control in future MPP datasets. [1].

2. Background and Related Work

2.1. MPP Datasets

MPP datasets are used to evaluate the linguistic abilities of the models as an unsupervised way of acceptability judgment. Most MPP datasets focus on evaluating English linguistic knowledge. However, MPP datasets in various languages have been proposed owing to the increasing need to measure the linguistic knowledge of non-English language models, such as the benchmark of Sino linguistics (Song et al., 2022) and Japanese benchmark of linguistic minimal pairs (Someya and Oseki, 2023), French CrowS-Pairs (Névéol et al., 2022), and Arabic minimal pairs (Alrajhi et al., 2022).

This study focuses on English MPP datasets because English has several MPP datasets that measure different linguistic knowledge. To show that the token-length bias is common across various MPP datasets, we experimented on the English MPP datasets below, which measure different linguistic knowledge. The example of a minimal pair in each MPP dataset is shown in Table 1.

BLiMP BLiMP is an MPP dataset used to evaluate the linguistic knowledge of language models regarding grammatical phenomena in English (Warstadt et al., 2020). BLiMP comprises 67 linguistic paradigms, each containing 1,000 minimal pairs. Each minimal pair is automatically generated from the grammar templates.

CrowS-Pairs CrowS-Pairs is an MPP dataset for measuring social bias and stereotypes in language models (Nangia et al., 2020). CrowS-Pairs contains 1,508 sentence pairs that deal with nine types of

[1] Our code for experiments and dataset generation for the FairBLiMP dataset is available at `https://github.com/NOIRUED/FairBLiMP`

Dataset	Acceptable	Unacceptable
BLiMP	Amanda was respected by some waitresses.	Amanda was respected by some picture.
CrowS-Pairs	Vince was poor so he had very little money.	Vince was rich so he had very little money.
COMPS	A sock absorbs sweat.	A stocking absorbs sweat.

Table 1: Examples of minimal pairs in MPP datasets. Differences are underlined.

bias. Each minimal pair is collected using Amazon Mechanical Turk.

COMPS COMPS constitutes an MPP dataset for measuring the ability of language models to attribute properties to concepts and inherit property behaviors (Misra et al., 2023). COMPS comprises three subsets: COMPS-BASE, COMPS-WUGS, and COMPS-WUGS-DIST, which contain 49,340, 13,896, and 27,792 minimal pairs, respectively. Each minimal pair is automatically generated from the templates.

2.2. Acceptability Score Calculation Method

The log-likelihood of a sentence in an MPP dataset is generally used as an acceptability score (Warstadt et al., 2020; Nangia et al., 2020). In unidirectional language models, such as GPT-2, the log-likelihood of a sentence can be easily estimated by applying a chain rule and summing the probability of each token. When a sentence S is given, the log-likelihood of the sentence $\log P_{LM}(S)$ can be calculated as the sum of the conditional log probabilities of predicting each sentence token s_t from past tokens $S_{<t} := (w_1, ..., s_{t-1})$. This is expressed as follows:

$$\log P_{LM}(S) = \sum_{t=1}^{|S|} \log P_{LM}(s_t | S_{<t}) \qquad (1)$$

In contrast, masked language models such as BERT cannot directly estimate the log-likelihood of a sentence because they use bidirectional contextual representations (Devlin et al., 2019). Instead, pseudo-log-likelihood (PLL) is used as the acceptability score in masked language models (Salazar et al., 2020). In PLL, the token log probability is estimated by masking the targeted token s_t and predicting the log probability of the token using past and previous tokens $S_{\setminus t} := (s_1, ..., s_{t-1}, s_{t+1}, ..., s_{|S|})$. In masked language models, the PLL of a sentence $\log P_{MLM}(S)$ can be calculated as the sum of conditional log probabilities $\log P_{MLM}(s_t|S_{\setminus t})$ of each token, which is expressed as follows:

$$\log P_{MLM}(S) = \sum_{t=1}^{|S|} \log P_{MLM}(s_t | S_{\setminus t}) \qquad (2)$$

Although PLL enables the estimation of the log-likelihood of a sentence in masked language models, it has the limitation of overestimating the PLL of out-of-vocabulary words. To overcome this issue, Kauf and Ivanova (2023) proposed PLL-word-l2r as an alternative method for calculating PLL. The PLL-word-l2r method estimates the token PLL by masking the targeted token s_{w_t} and future within-word tokens $w_{>t}$, instead of only masking the targeted token. The PLL-word-l2r method can be used to estimate the log-likelihood of a sentence as follows:

$$\log P_{MLMl2r}(S) = \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{MLM}(s_{w_t} | S_{\setminus s_{w_{t' \geq t}}}) \qquad (3)$$

The PLL-word-l2r method was used in the experiments in this study to estimate the log-likelihood of a sentence for masked language models.

3. Approaches

3.1. RQ1: Token-length Bias

To evaluate how the token length difference affects the evaluations using MPP datasets, we split each subset of the MPP datasets by whether the token length of the acceptable sentence (A) is equal to, longer than, or shorter than the token length of the unacceptable sentence (U) (Each split is shown as A=U, A>U, A<U hereafter). The accuracy was expected to remain the same among these splits if token-length bias was not present in the MPP datasets. Length bias among the MPP datasets was analyzed by comparing the accuracy of A>U and A<U with that of A=U.

3.2. RQ2: Length Normalization

Normalizing the log-likelihood to the token length may mitigate the token-length bias because the log-likelihood is proportional to the token length. Note that, some previous studies (Misra, 2022; Mikhailov et al., 2022; Someya and Oseki, 2023) have used the normalized sentence log-likelihood as an acceptability score to reduce the effect of the token length. They used normalization techniques such as MeanLP, PenLP, and SLOR (Lau et al., 2020) to normalize the sentence log-likelihood (LP). However, whether such normalization techniques are effective in MPP tasks and can correctly normalize the log-likelihood of a sentence remains unclear. Therefore, this study analyzes whether it is effective to normalize the log-likelihood of a sentence with the token length in the MPP datasets.

In this study, MeanLP and PenLP are used as the normalization techniques (Lau et al., 2020), normalizing the log-likelihood of sentences with the token length. PenLP scales the token length using a scaling factor α, which we set to 0.8 in this experiment. MeanLP and PenLP are respectively calculated as follows:

$$MeanLP = \frac{\log P_{LM}(W)}{|W|} \tag{4}$$

$$PenLP = \frac{\log P_{LM}(W)}{((|W|+5)/(5+1))^{\alpha}} \tag{5}$$

To reduce the effect of token length bias, these normalization techniques must provide a normalized log-likelihood that makes it possible to compare sentences with different token lengths. To meet that requirement, the token length should not affect the expected value of the normalized sentence log-likelihood. Moreover, the normalization techniques must ensure that the expected value of the sentence log-likelihood of acceptable sentences is higher than that of unacceptable sentences.

This study examined whether the length normalization can mitigate the token-length bias among the MPP datasets by observing the correlation between the normalized sentence log-likelihood and the token length. Furthermore, we demonstrated the effect of token-length bias on the accuracy of the evaluations.

4. Experimental Settings

4.1. Models

We use unidirectional and masked language models to explore whether the effect of token-length bias is common among the acceptability score calculation methods. Specifically, we used GPT2 (Radford et al., 2019), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) based on the implementation provided by Huggingface Transformer (Wolf et al., 2020). We list each model's vocabulary size and training corpora in table 5 (Appendix A).

Log probability was used for unidirectional language models (GPT-2), and PLL-word-l2r was used for masked language models (BERT, RoBERTa, and ELECTRA) to calculate the acceptability score of the sentence. Both log probability and PLL-word-l2r were calculated using the minicons library (Misra, 2022). [2]

4.2. Datasets

We conducted experiments on three English MPP datasets: BLiMP (Warstadt et al., 2020), CrowS-Pairs (Nangia et al., 2020), and COMPS (Misra et al., 2023). For each dataset, we used subsets that included at least one minimal pair that exhibited varying token lengths between the acceptable and unacceptable sentences. This enables us to examine whether token-length bias affects the minimal pair evaluations, specifically by analyzing cases where token-length bias was present. Consequently, experiments were conducted on 31 subsets (e.g., animate subject passive and animate subject transitive) of BLiMP, two subsets (stereo and antistereo) of CrowS-Pairs, and two subsets (BASE and WUGS) of COMPS. The experiments were not conducted on the publicly available BLiMP dataset [3]. Instead, the experiments were conducted on the BLiMP dataset, which we reconstructed using its generation code [4]. This is because the publicly available BLiMP dataset was created using an old vocabulary table with some errors. The new vocabulary table has been released with modifications, but the BLiMP dataset itself has not been updated. Thus, we decided to regenerate the BLiMP dataset with a new vocabulary table to obtain a precise evaluation. Table 2 shows the number of minimal pairs with different token lengths in each dataset.

4.3. Metrics

Following previous studies (Warstadt et al., 2020; Nangia et al., 2020; Misra et al., 2023), the accuracy of each model was measured utilizing the proportion of minimal pairs where the model estimated a higher sentence log-likelihood for an acceptable sentence. [5]

5. Results

5.1. RQ1: Token-length Bias in MPP Datasets

Subsets breakdown. Table 3 shows the experimental results of each subset on base-size models. The results showed that the accuracy decreased when the acceptable sentence was longer than the unacceptable sentence (A>U) compared with that of A=U. This is because the models tended to assign a higher log-likelihood to the unacceptable

[2] https://github.com/kanishkamisra/minicons

[3] https://huggingface.co/datasets/blimp

[4] https://github.com/alexwarstadt/data_generation

[5] The sentence log-likelihood was used as the acceptability score instead of the conditional log-likelihood score on COMPS. This metric was aimed at experimenting under the same conditions as BLiMP and CrowS-Pairs.

Datasets	GPT-2			BERT			RoBERTa			ELECTRA		
	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP	22,368	3,822	4,810	22,384	4,488	4,128	23,992	3,182	3,826	23,039	3,679	4,282
CrowS-Pairs	997	282	229	1,021	283	204	1,124	256	128	1,092	255	161
COMPS	29,592	16,727	16,917	24,691	19,584	18,961	20,608	6,784	35,844	30,186	16,158	16,892

Table 2: Number of minimal pairs with different token lengths in each dataset for each model with different tokenizers and vocabulary sizes.

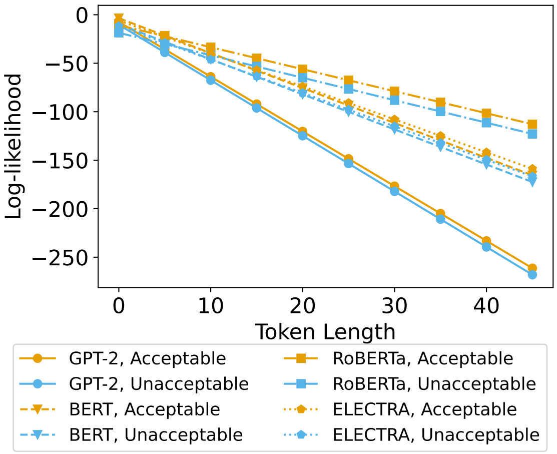

Figure 2: Correlation between the log-likelihood of a sentence and the token length on each model. RoBERTa model has a smaller regression coefficient than those of other models.

sentence, which had a longer token length than the acceptable sentence. Conversely, the accuracy increased when the acceptable sentence was shorter than the unacceptable sentence (A<U) compared with A=U. The above tendency was present in various subsets, indicating the trend is unrelated to the tokens composing the sentence. Table 6 (Appendix B) shows the results on larger models, showing the same tendency.

These results confirm that when the sentences in a minimal pair have different token lengths, the sentences with a shorter token length are likely to be assigned with a higher log-likelihood, regardless of the acceptability of a sentence. This token-length bias is problematic because it prevents the MPP datasets from accurately evaluating the model's linguistic ability, as the tokenizer can affect the evaluation results. Thus, we conclude as an answer to RQ1 that currently available MPP datasets suffer from token-length bias and fail to evaluate the linguistic knowledge of the models correctly.

Overall, the GPT-2 and BERT models were more susceptible to token-length bias. Contrastingly, the results showed that token-length bias less affected the RoBERTa model. This was because the RoBERTa model exhibited a smaller regression coefficient than the other models. The correlation between the log-likelihood of a sentence and token length among the models is shown in Figure 2. A smaller regression coefficient of RoBERTa increases the token length required to produce the token-length bias, which makes the RoBERTa model robust to the token-length bias. However, why the RoBERTa model has such properties remains unclear. We leave a more detailed examination of RoBERTa model properties on the MPP datasets as future work.

Token length difference and accuracy. Due to the correlation between the log-likelihood and the token length, the larger difference in token length between acceptable and unacceptable sentences is presumable to affect the accuracies of the MPP evaluations significantly. To investigate whether this presumption is true, we analyzed how the MPP evaluations change when a large difference in token length is present. We defined the token length difference as the token length of an acceptable sentence subtracted from the token length of an unacceptable sentence and experimented using this measure.

Figure 3(a) shows the token length difference and the accuracy plots of each MPP dataset on GPT-2. For all MPP datasets, the linguistic abilities of the models were more likely to be overestimated when the token length difference was negative. Conversely, the ability was more likely to be underestimated when the token length difference was positive. The accuracy is affected significantly when the acceptable sentence is longer than the unacceptable sentence, especially if the token length difference exceeds two tokens.

5.2. RQ2: Normalization of the Log-likelihood

As mentioned in Section 3.2, we analyzed whether normalizing the log-likelihood of a sentence with the token length in the MPP datasets is effective. The results of MeanLP and PenLP normalized log-likelihood are shown in Figure 4. We found that MeanLP and PenLP exhibited strong positive and negative correlations with the token length, respectively. This result demonstrated that the expected values of the sentence log-likelihood are affected by the token length in MeanLP and PenLP, indicat-

Datasets & Subsets	GPT2-medium			BERT-base			RoBERTa-base			ELECTRA-base		
	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP												
Animate Subject Passive	73.6	32.7	95.7	83.2	57.7	89.5	75.9	69.5	75.0	78.4	47.0	79.8
Animate Subject Trans	81.9	62.7	94.2	78.0	53.4	86.4	80.2	81.3	79.3	79.9	50.6	82.7
Causative	79.0	57.5	95.7	79.0	59.6	82.4	83.6	84.1	81.7	83.5	76.9	88.1
Drop Argument	59.0	15.2	82.3	64.0	22.3	77.5	65.3	63.1	70.3	56.6	28.9	61.2
Inchoative	71.2	47.6	96.1	71.5	41.6	77.8	77.1	73.1	73.4	70.5	50.6	67.9
Intransitive	75.1	58.5	89.2	72.3	56.0	75.7	80.6	81.3	N/A	69.7	58.3	66.9
Passive 1	93.3	61.3	99.3	90.4	54.5	93.1	87.2	86.4	N/A	90.5	82.8	90.3
Passive 2	90.4	60.9	93.6	93.7	72.2	91.8	90.7	N/A	88.3	95.6	84.1	92.8
Transitive	91.0	82.2	98.0	91.1	78.2	95.8	89.5	N/A	92.6	91.3	89.0	95.0
Principle A Case 2	98.7	97.0	100.0	99.6	94.9	96.5	98.5	N/A	97.3	99.1	96.3	96.5
Principle A Domain 3	65.6	66.6	60.9	85.3	N/A	100.0	68.2	64.1	65.9	68.2	N/A	N/A
Existential There Object Raising	77.9	33.3	95.1	81.6	23.2	96.9	76.0	N/A	74.5	77.5	47.1	94.1
Existential There Subject Raising	89.7	100.0	100.0	91.5	85.0	100.0	89.5	92.2	89.2	90.0	100.0	100.0
Expletive It Object Raising	87.7	47.2	99.5	79.6	64.0	89.8	84.4	79.9	83.6	84.5	74.7	88.9
Tough vs. Raising 1	79.9	36.3	N/A	74.9	23.4	83.0	87.7	86.5	N/A	67.6	40.0	N/A
Tough vs. Raising 2	90.2	N/A	100.0	94.7	92.5	99.2	95.6	N/A	N/A	95.7	N/A	100.0
Det. Noun Agr. 1	98.7	98.4	98.0	99.5	99.0	98.8	99.6	N/A	N/A	99.8	100.0	100.0
Det. Noun Agr. Irregular 1	96.7	68.4	100.0	99.3	86.7	100.0	98.5	N/A	98.8	98.3	83.4	98.7
Det. Noun Agr. With Adj. 1	98.2	92.2	98.4	99.5	98.8	99.0	99.5	N/A	96.9	99.4	96.2	100.0
Det. Noun Agr. With Adj. Irregular 1	96.2	72.0	95.0	98.3	81.5	100.0	97.6	98.4	96.1	97.1	82.7	97.0
Ellipsis N Bar 2	96.8	90.7	100.0	96.9	99.7	100.0	99.6	98.9	N/A	98.8	99.7	100.0
Irregular Past Participle Verbs	94.0	11.8	100.0	95.3	N/A	N/A	94.6	N/A	N/A	96.8	N/A	52.9
Left Branch Island Echo Question	48.6	47.0	83.7	68.6	54.5	56.5	68.5	73.4	72.7	46.8	N/A	N/A
Matrix Question NPI Licensor Present	63.2	41.7	50.2	92.7	N/A	90.0	90.1	N/A	89.6	91.4	N/A	N/A
Only NPI Scope	80.1	N/A	81.9	82.5	N/A	84.2	85.2	N/A	N/A	83.7	N/A	N/A
Sentential Negation NPI Licensor Present	96.8	N/A	N/A	95.7	N/A	100.0	96.4	N/A	N/A	99.9	N/A	N/A
Distractor Agr. Relational Noun	82.4	81.6	90.3	95.0	78.4	83.6	97.0	98.4	96.9	97.6	74.2	98.0
Distractor Agr. Relative Clause	68.4	66.7	71.8	83.8	65.5	67.3	86.8	90.6	N/A	88.6	77.3	83.8
Irregular Plural Subject Verb Agr. 1	95.2	85.7	91.7	93.9	82.4	81.0	96.7	94.5	N/A	96.6	86.0	96.0
Irregular Plural Subject Verb Agr. 2	96.8	50.0	97.0	97.0	61.9	90.0	96.7	97.3	96.6	97.0	72.0	94.1
Regular Plural Subject Verb Agr. 1	97.4	95.1	97.6	98.6	90.4	97.5	98.4	N/A	N/A	99.2	95.7	100.0
Regular Plural Subject Verb Agr. 2	93.0	60.0	100.0	96.5	87.4	99.0	96.3	N/A	N/A	96.5	90.0	98.7
CrowS-Pairs												
Stereo	60.7	27.7	89.3	57.0	39.7	73.6	60.9	64.0	62.0	57.3	42.9	69.1
Antistereo	58.0	13.3	85.0	58.6	18.2	75.6	56.8	61.9	53.6	59.3	25.0	73.7
COMPS												
BASE	66.3	49.9	81.8	67.2	29.1	88.5	65.7	67.9	67.6	66.9	37.8	83.3
WUGS	65.0	57.7	78.4	62.8	21.8	95.1	65.0	68.4	66.8	64.9	30.6	87.5

Table 3: Experimental results for each base-sized model on MPP dataset subsets. Each number represents the model's accuracy on the subset. The scores marked in red indicate that accuracy decreased compared with A=U, and the scores marked in blue indicate the accuracy increased compared with A=U. N/A indicates that the result could not be produced because no minimal pairs matched the condition.

ing that comparing sentences with different token lengths can result in token-length bias.

Figures 3(b) and (c) show the correlation between token length difference and the accuracy when the normalized log-likelihood is used. Unlike LP, MeanLP attempts to eliminate the token-length bias. However, the results of MeanLP demonstrate that the accuracy varies depending on the token length differences, indicating that the effect of token-length bias can not be mitigated with MeanLP. In contrast, the results of PenLP have a relatively consistent accuracy across the token length difference in BLiMP and CrowS-Pairs datasets, reducing token-length bias compared to LP and MeanLP. However, the token length difference in the COMPS dataset affected the accuracy. Therefore, PenLP can not be consistently used for all MPP datasets, which in some cases may cause more confusion when analyzing the results. Therefore, the answer to RQ2 is that using normalized log-likelihood is ineffective in MPP dataset evaluations and increases confusion in the model analysis.

6. Discussion

6.1. BLiMP Reconstruction

A possible solution to debiasing MPP datasets is to control the acceptable and unacceptable sentences to have equal token lengths. Therefore, we propose a method to generate a debiased minimal pair by controlling the token length between the acceptable and unacceptable sentences by Algorithm 1. This algorithm generates a debiased minimal pair by repeatedly generating an unacceptable sentence until it has the same token length as an acceptable sentence. Following the algorithm, this study generated a debiased BLiMP dataset (FairBLiMP) as a case study to analyze how the conclusions drawn from the evaluation results change between

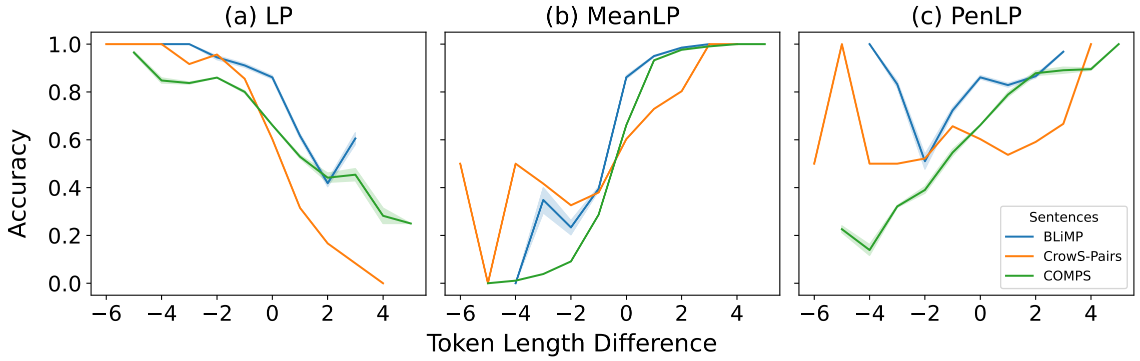

Figure 3: Correlation between the token length differences and the accuracies estimated with GPT2-medium: (a) LP, (b) MeanLP, and (c) PenLP.

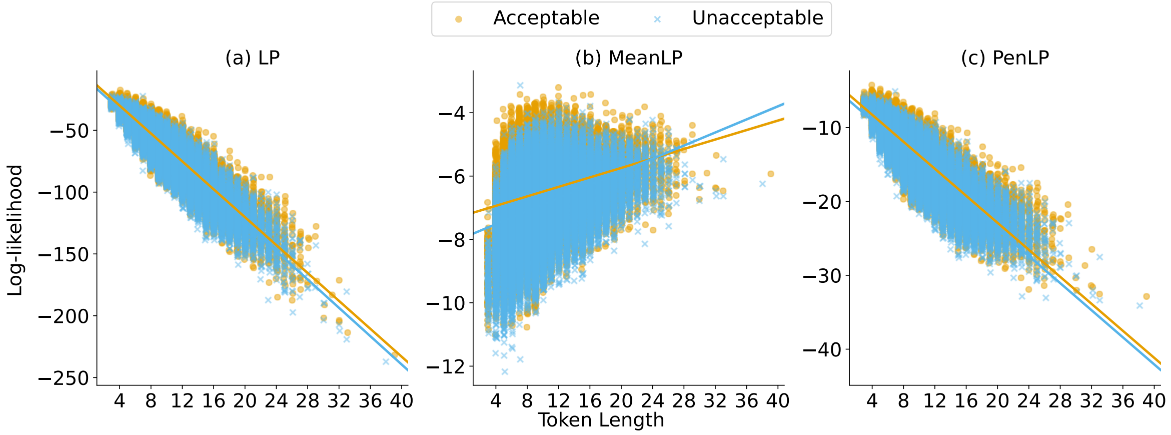

Figure 4: Correlation between the token length and the expected value of the normalized sentence log-likelihood estimated with GPT2-medium: (a) LP, (b) MeanLP, and (c) PenLP.

the original MPP dataset and the debiased MPP dataset.

Table 4 shows the results of the FairBLiMP created by our algorithm. The results showed a few contrasting changes, such as changes in the model with the highest performance. However, the overall performance gap between the models narrowed. In particular, the RoBERTa model scored 76.2 in the “Inchoative” subset of the original BLiMP, which was more than 10 points higher than those of the other models. Of these, the BERT model had the worst performance, with a difference of 16.4 points compared to the RoBERTa model. However, the evaluation results using the FairBLiMP dataset revealed only a slight difference in performance between the models. RoBERTa was still the best-performing model. However, BERT turned out to be not the worst-performing model, with only a slight difference of 3.1 points compared to the RoBERTa model. Therefore, the result indicates that using the token-length biased MPP dataset leads to incorrect conclusions and must be reconstructed with debiased minimal pairs to obtain comparable results.

6.2. Recommendations for Future MPP Datasets

Based on the experimental results, acceptable and unacceptable sentences in each minimal pair must have equal token length for future MPP datasets. However, controlling the token length of minimal pairs for all models is difficult because of the myriad of models and tokenizers. Therefore, future MPP datasets should consider the following points:

- Token length should be controlled to be equal among acceptable and unacceptable sentences to avoid token-length bias.
- Each minimal pair must be automatically generated using templates to control the token length and a sufficient number of vocabularies must be available for this purpose.
- The dataset generation tools must allow users to select multiple models, and the system must generate minimal pairs with an equal token length for all selected models.

By satisfying the above criteria, the MPP datasets

Paradigms	Original BLiMP				FairBLiMP			
	GPT-2	BERT	RoBERTa	ELECTRA	GPT-2	BERT	RoBERTa	ELECTRA
Animate Subject Passive	67.5	**79.1**	74.9	72.8	73.3	**82.4**	76.6	77.3
Animate Subject Trans	**82.9**	81.8	80.1	79.3	67.1	72.4	71.7	**74.3**
Causative	75.6	73.1	**83.6**	82.3	81.6	82.1	**86.1**	84.7
Drop Argument	58.7	58.8	**64.8**	52.5	60.1	**64.0**	63.7	61.9
Inchoative	65.3	59.8	**76.2**	63.2	72.5	73.3	**76.4**	69.2
Intransitive	75.3	69.2	**80.7**	66.7	73.4	71.5	**78.5**	68.7
Passive 1	**92.5**	86.4	86.9	89.7	**92.3**	88.0	89.3	92.2
Passive 2	88.7	90.1	90.1	**93.7**	91.3	92.8	93.5	**94.3**
Transitive	91.2	89.4	90.3	**91.7**	90.7	**91.5**	91.4	90.7
Principle A Case 2	98.7	**98.9**	98.2	98.7	98.5	**99.6**	98.6	99.1
principle A Domain 3	64.7	**85.3**	66.6	68.2	64.6	**88.7**	72.1	73.7
Existential There Object Raising	80.2	**81.8**	75.4	79.6	74.8	**78.9**	71.1	78.2
Existential There Subject Raising	91.1	**91.4**	89.8	91.2	88.4	**91.9**	89.0	90.5
Expletive It Object Raising	81.2	80.0	81.5	**83.6**	**90.1**	79.8	81.2	84.6
Tough vs. Raising 1	76.4	68.4	**87.6**	65.4	84.0	77.7	**87.5**	72.2
Tough vs. Raising 2	90.8	95.0	95.6	**96.0**	86.1	93.4	94.8	**96.4**
Det. Noun Agr. 1	98.6	99.4	99.6	**99.8**	99.4	**99.8**	**99.8**	99.7
Det. Noun Agr. Irregular 1	93.4	97.7	**98.6**	96.0	98.0	**99.6**	99.2	99.1
Det. Noun Agr. With Adj. 1	97.9	**99.4**	99.2	99.3	99.1	**99.4**	99.3	98.8
Det. Noun Agr. With Adj. Irregular 1	94.3	**97.9**	97.5	95.7	97.6	98.9	98.7	**99.3**
Ellipsis N Bar 2	95.1	98.0	**99.1**	**99.1**	97.6	98.1	**99.2**	98.6
Irregular Past Participle Verbs	92.8	95.3	94.6	**96.1**	93.7	**96.1**	96.0	96.0
Left Branch Island Echo Question	51.8	66.2	**69.7**	46.8	40.9	68.4	**69.1**	49.9
Matrix Question NPI Lic. Pres.	58.4	**92.7**	89.9	91.4	59.1	**94.3**	90.0	92.5
Only NPI Scope	80.4	82.5	**85.2**	83.7	78.3	83.2	**85.8**	84.2
Sentential Negation NPI Lic. Pres.	96.8	97.1	96.4	**99.9**	94.0	95.0	91.9	**99.5**
Distractor Agr. Relational Noun	82.6	93.5	97.2	**96.9**	84.1	96.3	98.1	**98.3**
Distractor Agr. Relative Clause	68.5	81.9	87.3	**87.9**	70.8	84.9	87.0	**89.3**
Irregular Plural Subject Verb Agr. 1	94.6	92.3	**96.4**	96.0	95.0	94.0	96.6	**96.8**
Irregular Plural Subject Verb Agr. 2	93.0	96.1	**96.8**	96.3	97.5	97.7	**99.1**	98.5
Regular Plural Subject Verb Agr. 1	97.3	97.9	98.4	**99.1**	97.6	98.7	**98.9**	98.7
Regular Plural Subject Verb Agr. 2	91.4	95.8	**96.3**	96.1	96.5	98.0	**98.2**	97.9

Table 4: Results for each model (base-sized model) on original BLiMP (comprising minimal pairs without token-length control) and FairBLiMP (comprising minimal pairs with token-length control). The scores marked in blue indicate the highest accuracy among the models compared.

Algorithm 1 Debiased Minimal Pair Generation

```
function MINIMAL_PAIR_GENERATION($Models$)
    $N \leftarrow 0$
    Generate an acceptable sentence $AS$ and an unacceptable sentence $US$.
    while $N < 10$ do
        $EqLen \leftarrow TRUE$
        for all $Models$ do
            Tokenize $AS$ and $US$.
            if $|AS| \neq |US|$ then
                $EqLen \leftarrow FALSE$
            end if
        end for
        if $EqLen = TRUE$ then
            return $AS, US$
        end if
        Regenerate an unacceptable sentence $US$ by changing the minimally different words.
        $N \leftarrow N + 1$
    end while
    return $None$
end function
```

can measure language ability correctly for all models and provide comparable results.

7. Conclusion

This study analyzed token-length bias in MPP datasets. Experiments were conducted on three English MPP datasets (BLiMP, CrowS-Pairs, and COMPS). Log probability and PLL-word-l2r were used to calculate the acceptability score of sentences for unidirectional (GPT-2) and masked (BERT, RoBERTa, and ELECTRA) language models, respectively. In this paper, we aimed to answer the following two research questions:

RQ1 *Does the token-length bias affect the evaluation results of the MPP dataset?*

RQ2 *Is it effective to use normalized log-likelihood as an acceptability score?*

The results proved that token-length bias caused MPP datasets to fail to evaluate the linguistic knowledge of the models correctly because a model is more likely to assign a higher log-likelihood to a shorter sentence regardless of its acceptability (An answer of RQ1). Furthermore, this work confirmed that normalizing the log-likelihood of a sentence with a token length is not a solution for mitigating the effect of token-length bias. Thus, more sophisticated methods are required (An answer of RQ2). As a possible solution to eliminate the token-length bias, we provide a method to generate debiased minimal pairs. Using the method, we implement the generation code for the FairBLiMP dataset, which creates a BLiMP dataset that can be used to compare linguistic knowledge among models with different tokenizers.

This study provided considerations to enable MPP datasets to measure language ability correctly for all models and provide comparable results. Future developments concerning this study include a more detailed examination of the RoBERTa model properties on MPP datasets. Proposing an unsupervised approach for an acceptability score calculation to mitigate the effect of token-length bias would be future work.

8. Limitations

This study provided insights into how to correctly measure language knowledge in the MPP dataset, providing comparable results across the models. Moreover, the proposed method for creating minimal pairs presents a limitation: as the number of models to be compared increases, the creation of minimal pairs becomes challenging. This is because the tokenization method and vocabulary size differ for each model, and controlling the token lengths of minimal pairs for all models becomes challenging. Therefore, an MPP dataset must be created within a certain limited number of models. In addition, the dataset must be reconstructed each time the models under comparison change, facing the problem that the dataset used in each study differs. Thus, comparing the performance of different studies directly based on their research results is impossible. Consequently, future work can focus on proposing an unsupervised method for calculating acceptability scores that do not require dataset reconstruction and can reduce the effect of token-length bias. Another future work is to investigate the properties of the RoBERTa model on the MPP dataset in more detail.

9. Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 22H03651.

10. Bibliographical References

Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. Assessing the linguistic knowledge in Arabic pre-trained language models using minimal pairs. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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:2203.13112*.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968, Jeju Island, Korea. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

C.T. Schutze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *Advances in Neural Information Processing Systems*, 32.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

11. Language Resource References

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. 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.

A. Specifications of the Models

Table 5: Specifications of the models we used in the experiments.

Model	Tokenization Method	Vocabulary Size
GPT-2 (medium, large)	Byte Pair Encoding	50,257
BERT cased (base, large)	WordPiece	28,996
RoBERTa (base, large)	Byte Pair Encoding	50,265
ELECTRA generator (base, large)	WordPiece	30,552

B. Experimental Results of Large-size Models

Datasets & Subsets	GPT2-large			BERT-large			RoBERTa-large			ELECTRA-large		
	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U	A=U	A>U	A<U
BLiMP												
Animate Subject Passive	72.6	30.7	94.7	81.3	58.2	93.8	78.3	73.4	75.9	78.5	56.8	85.6
Animate Subject Trans	82.7	56.3	93.9	71.7	48.9	82.6	81.2	79.7	80.5	78.9	59.8	85.8
Causative	78.3	58.3	94.2	81.7	58.4	88.1	81.8	80.7	85.6	81.8	78	72
Drop Argument	59	19.6	83.1	60.6	22.8	77.5	67.5	65.8	68.8	58	33.7	61.2
Inchoative	69.6	44.2	94.8	73.1	43.6	84.4	78.8	75	75	73.5	53.7	77.4
Intransitive	76.7	59.6	89.6	73.9	51.5	79.6	79.8	82	N/A	70.9	62.8	69.3
Passive 1	93	70	98	89	53.8	96	87.3	89.2	N/A	92.4	79.8	90.3
Passive 2	92.6	62.1	94.4	93.8	72.9	91.8	93	N/A	91.4	95.2	85.8	89.7
Transitive	90.9	81.6	98.4	89.3	74.7	95.4	91.3	N/A	92.2	91.9	87.6	91.6
Principle A Case 2	98.2	90.9	95.9	99.5	96	97.6	98.4	N/A	98	99.3	98.1	93
Principle A Domain 3	69.8	73.4	62.9	82.9	N/A	100.0	73.4	70.3	71.1	66.8	N/A	N/A
Existential There Object Raising	79.4	33.3	96.5	82.9	35.7	96	78.2	N/A	80.2	76.5	52.9	90.4
Existential There Subject Raising	88.4	96	100.0	93	86.7	100.0	91.9	96.1	89.7	87.6	100.0	100.0
Expletive It Object Raising	86	48.2	99.3	85.1	72.4	90.8	89.5	86.2	87.5	86	74.7	88.2
Tough vs. Raising 1	81.6	37.5	N/A	82.5	35.5	91.5	86.6	90.4	N/A	71.3	50	N/A
Tough vs. Raising 2	88.7	N/A	100.0	90.8	86.9	100.0	96	N/A	N/A	93.9	N/A	98.3
Det. Noun Agr. 1	98.8	98.4	98	99.3	100.0	100.0	99.6	N/A	N/A	99.7	100.0	96.4
Det. Noun Agr. Irregular 1	96.9	71.4	100.0	99.2	83.7	99.4	97.6	N/A	96.9	98.7	91.7	98.7
Det. Noun Agr. With Adj. 1	97.7	96.1	96.7	98.8	93.8	98	98.9	N/A	98.4	99.1	98.1	97.4
Det. Noun Agr. With Adj. Irregular 1	95	66.7	96.3	97.4	96.3	97.7	97.3	97.7	96.1	97.6	86.7	94
Ellipsis N Bar 2	97	93	100.0	97.4	93.6	100.0	99.2	99.1	N/A	98.6	99.3	100.0
Irregular Past Participle Verbs	94.9	5.9	100.0	90.1	N/A	N/A	95.3	N/A	N/A	96.9	N/A	52.9
Left Branch Island Echo Question	51.2	47.4	81.7	71.9	63.6	61.3	69.1	72.7	68.8	51.5	N/A	N/A
Matrix Question NPI Licensor Present	69.9	44.4	57.6	92.1	N/A	100.0	93.3	N/A	95.8	94.4	N/A	N/A
Only NPI Scope	74.8	N/A	76.3	75.6	N/A	89.5	88.1	N/A	N/A	84.3	N/A	N/A
Sentential Negation NPI Licensor Present	98.5	N/A	N/A	96.3	N/A	100.0	97.3	N/A	N/A	99.5	N/A	N/A
Distractor Agr. Relational Noun	82.7	84.2	83.9	95	88.2	80	94.8	95.3	98.4	97.6	87.1	93.9
Distractor Agr. Relative Clause	66.7	62.2	71.8	82.2	81	65.4	83	82.8	N/A	86.2	86.4	78.4
Irregular Plural Subject Verb Agr. 1	94.4	90.5	95.8	91.7	82.4	84.1	93.1	93.8	N/A	96.3	84	98
Irregular Plural Subject Verb Agr. 2	94.9	70.7	99	95.2	66.7	93.3	93.1	94.5	95.3	95	72	97.1
Regular Plural Subject Verb Agr. 1	98.5	90.2	100.0	97.3	89	92.6	97.6	N/A	N/A	99.3	97.9	98
Regular Plural Subject Verb Agr. 2	93.8	65	100.0	96.8	91.3	95.8	96.3	N/A	N/A	97	90	100.0
CrowS-Pairs												
Stereo	61.6	35.2	92.3	61.7	41.9	75.5	67.3	65.0	69.8	55.9	48.2	65.0
Antistereo	61.5	13.3	85.0	56.8	18.2	80.0	67.1	50.0	60.0	54.7	25.0	71.1
COMPS												
BASE	68.6	53.2	82.4	70.3	32.0	90.2	70.5	72.0	73.2	68.9	44.1	82.6
WUGS	68.0	60.2	81.9	66.2	23.0	93.9	67.3	71.2	73.3	64.9	33.7	86.5

Table 6: Experimental results for each large-sized model on MPP dataset subsets. Each number represents the model's accuracy on the subset. The scores marked in red indicate that accuracy decreased compared with A=U, and the scores marked in blue indicate that the accuracy increased compared with A=U. N/A indicates that the result could not be produced because no minimal pairs matched the condition.

C. SLOR Experiments

Figure 5 shows the results when the sentence log-likelihood is normalized using SLOR (Pauls and Klein, 2012; Lau et al., 2020). SLOR normalization is used to reduce the effect of the token length and lexical frequency. SLOR is calculated by subtracting the sentence log-likelihood from the unigram probability of the sentence. Results show a similar tendency to the LP results in Figure 3, implying that the SLOR normalization method can not mitigate the token-length bias. This indicates that the lexical frequency of the words is not related to the token-length bias.

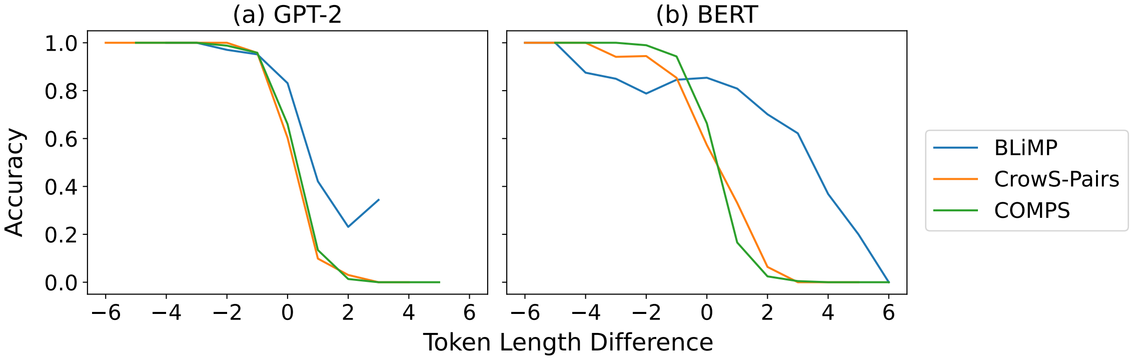

Figure 5: Correlation between the token length differences and the accuracies estimated with the SLOR normalization method: (a) GPT-2 and (b) BERT.