Erratum: ‘‘BLiMP: The Benchmark of Linguistic Minimal Pairs for English’’

Alex Warstadt[1], Alicia Parrish[1], Haokun Liu[2], Anhad Mohananey[2]
Wei Peng[2], Sheng-Fu Wang[1], Samuel R. Bowman[1,2,3]

[1]Dept. of Linguistics
New York University

[2]Dept. of Computer Science
New York University

[3]Center for Data Science
New York University

warstadt@nyu.edu

Abstract

We correct wrongly reported results on BLiMP.

We wrongly reported some results on BLiMP. The error does not change any of our conclusions in the paper. We made the following changes:

1. We corrected 12 values reported in Table 3 of the paper, shown in Table 1 of this statement.

2. In Section 5.1 of the paper, we wrote, ‘‘Only GPT-2 performs well above chance, and it remains 20 points below humans.’’ The corrected version reads ‘‘Only GPT-2 performs well above chance, and it remains 14 points below humans.’’

3. We updated Figure 2 in the paper to reflect the updated results.

These errors resulted from the erroneous inclusion in our calculations of the final results of two paradigms that we had eliminated from the dataset. These paradigms had been eliminated due to low human agreement with the labels. They belonged to the FILLER-GAP and ISLAND EFFECTS phenomena, respectively. Thus, only results for these phenomena and the overall results are affected.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 867–868, 2021. https://doi.org/10.1162/tacl_x_00375
Action Editor: Mark Steedman. Published 2/2021.

	Overall		Filler Gap		Island	
model	*Original*	*Corrected*	*Original*	*Corrected*	*Original*	*Corrected*
5-gram	~~60.5~~	61.2	~~58.1~~	60.2	~~53.7~~	57.2
LSTM	~~68.9~~	69.8	~~72.5~~	73.9	~~42.9~~	46.6
TXL	~~68.7~~	69.6	~~64.9~~	66.6	~~45.8~~	48.4
GPT-2	~~80.1~~	81.5	~~79.0~~	81.3	~~63.1~~	70.6

Table 1: Original and corrected values in Table 3 of the original paper.