Word-Order Error Detection Helps Data-Efficient Language Models Learn Syntax

Alexander Fung[1]*, Chengxu Zhuang[1]*, Steven T. Piantadosi[2], Jacob Andreas[1], Evelina Fedorenko[1]

[1]Massachusetts Institute of Technology, [2]University of California, Berkeley

Contact: alexfung@mit.edu

Research Question

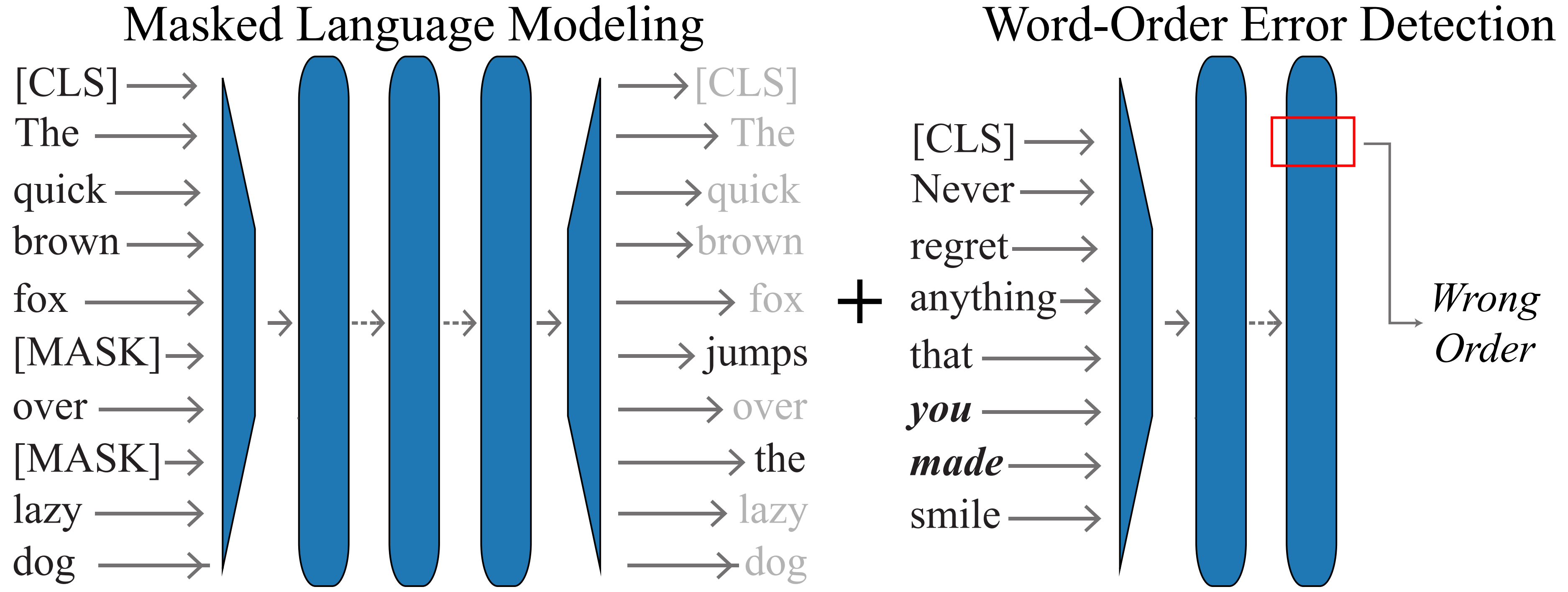
Can language models learn word-order sensitivity with developmentally plausible amounts of data, and can this improve syntax understanding generally?

Word-Order Objective

While grammatically ill-formed sentences are generally out-of-distribution for language models, they are common in human speech that children are exposed to.

We augment masked language models with an additional word-order error detection loss:

Training Objectives



After training models on 5–50M tokens of data, we evaluate them on our self-constructed word-order and function words benchmarks, as well as BLiMP.

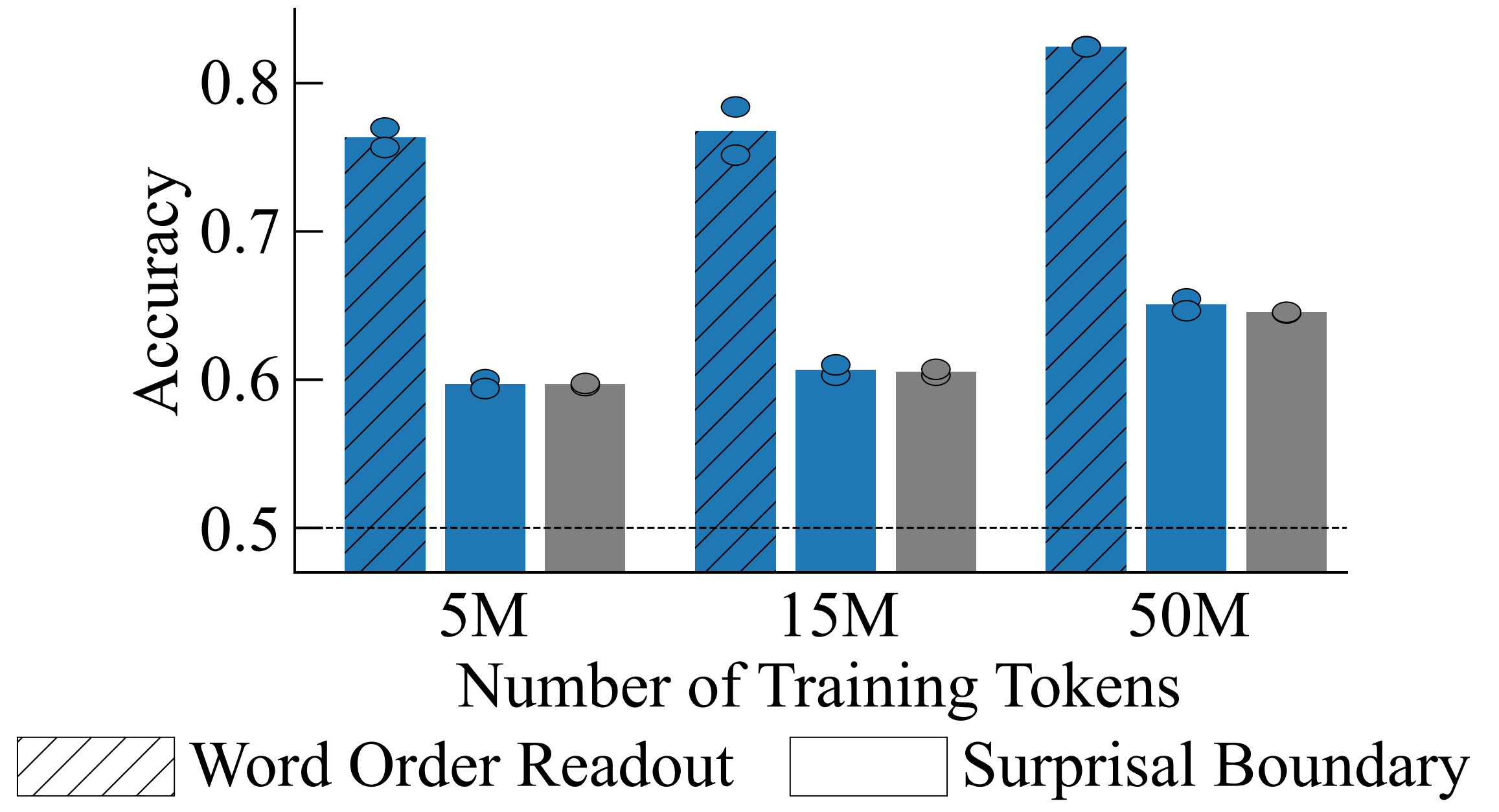
Results

Word Order

Networks trained with word-order objective achieve significantly **better results on classifying incorrectly ordered sentences**.

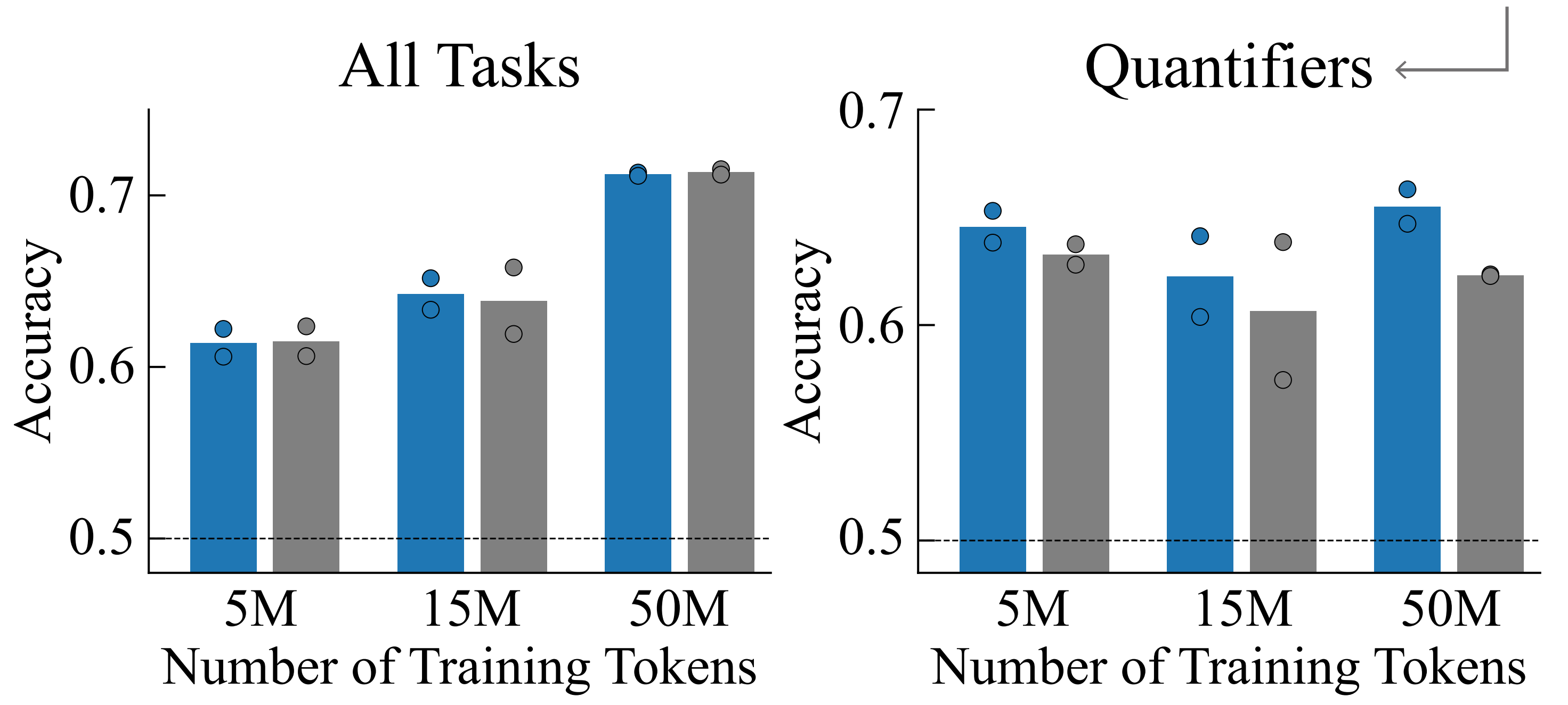
This improvement is only evident when using the trained word-order readout.

She resisted the urge to shiver. — *Correct*
She resisted the to urge shiver. — *Wrong*



BLiMP

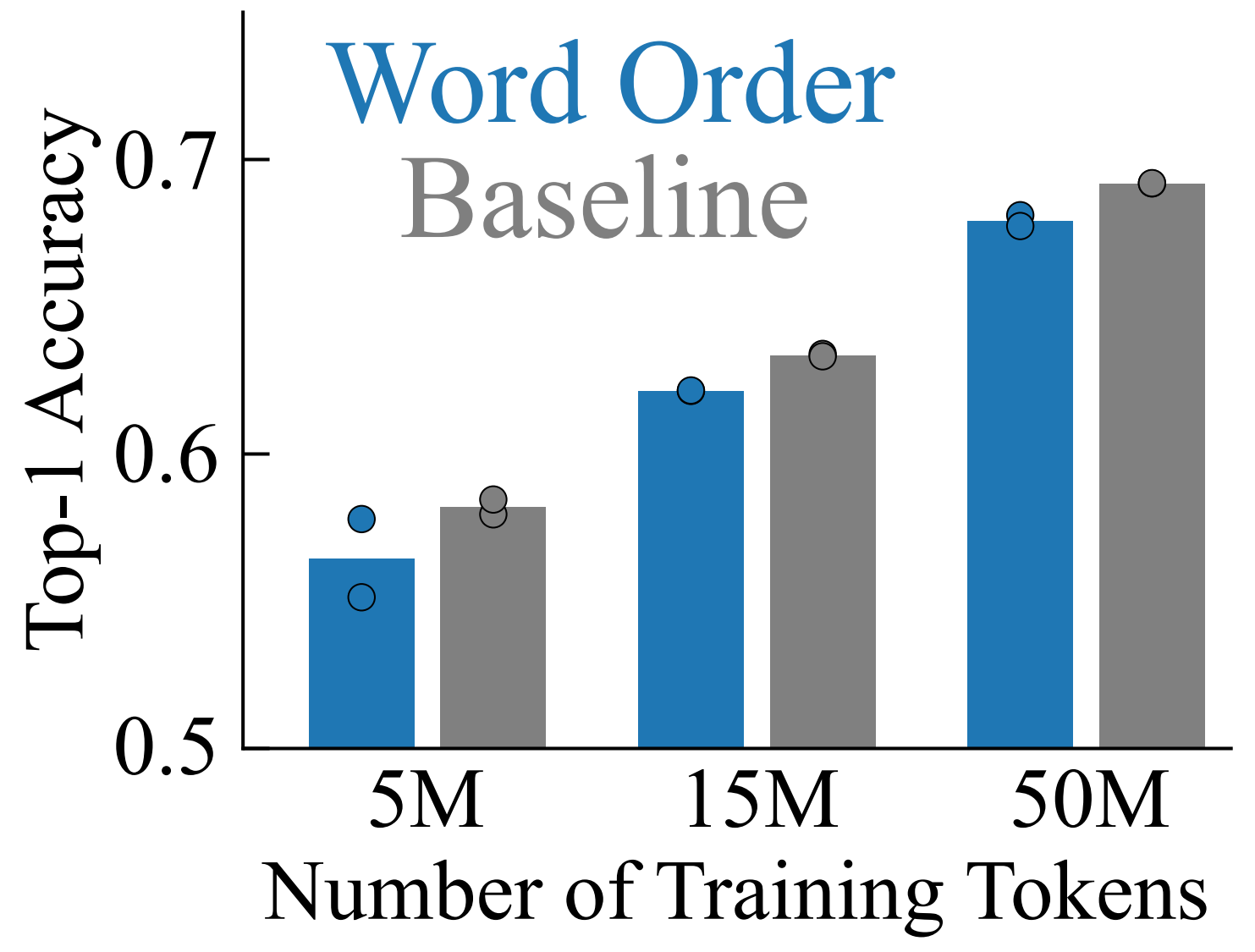
Every fish was there astounding Sally. — *Correct*
There was every fish astounding Sally. — *Wrong*



The benefit of co-training with word-order objective on **BLiMP** is mainly on the **Quantifiers** task.

Function Words

on of
↑ ↑
I'm [MASK] top [MASK] it.



Word-order objective does not help predict the missing function words.

Conclusions

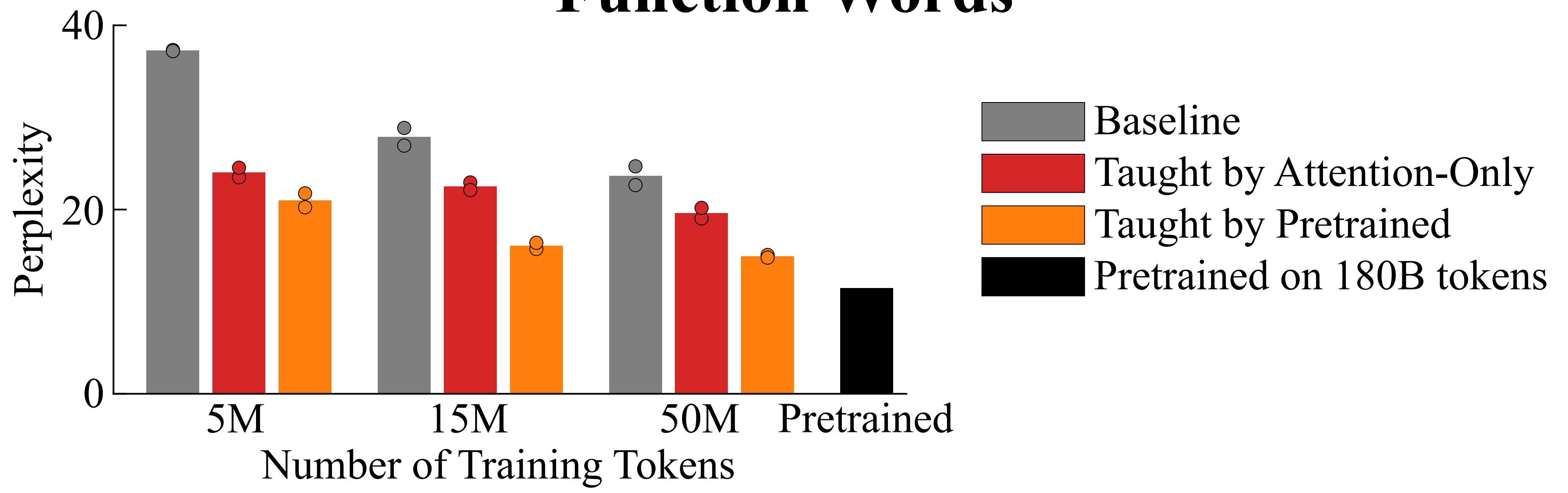
Data-efficient language models can learn to perform the word-order task, but this does not translate to general improvements in syntax understanding.

Future Directions

There may be too many aspects of syntax to capture effectively by adding task-specific denoising losses.

Alternatively, general syntactic representations can be intuitively understood through attention patterns. Promisingly, distilling attention patterns from a pretrained next word prediction model leads to **better function words understanding**:

Function Words



Here, an attention-only model is trained with only an attention objective, on only 5M tokens. Models then copy either the attention-only model's or the pretrained model's attentions and learn the next word prediction task.