Insights from the first BabyLM Challenge: Training sample-efficient language models on a developmentally plausible corpus

Alex Warstadt
ETH Zurich, Zurich, Switzerland

Aaron Mueller
Northeastern University, Boston, Massachusetts, United States

Leshem Choshen
IBM, Givataim, Israel

Ethan Wilcox
ETH Zurich, Zurich, Switzerland

Chengxu Zhuang
MIT, Cambridge, Massachusetts, United States

Adina Williams
Meta Platforms Inc., New York, New York, United States

Ryan Cotterell
Institute for Machine Learning, Zurich, Switzerland

Tal Linzen
New York University, New York, New York, United States

Abstract

Language models have great potential as cognitive models for studying human language acquisition, but current models are far less data-efficient than human learners. Children acquire language from 100 million words or less, but large language models are trained on trillions of words. We discuss the prospects for improving language models' developmental plausibility through a meta-analysis of results from the 2023 BabyLM Challenge. BabyLM was a competition that invited participants to train a language model on a 100 million-word corpus including transcribed speech and child-appropriate texts. Results from over 30 submissions showed that new machine learning techniques and increased training iterations yielded models that outperformed leading large language models in grammar, language understanding, and linguistic generalization, while cognitively plausible approaches such as curriculum learning were less effective. We discuss the implications of these and other findings for computational cognitive modeling and explore ideas to ensure future competitions' contributions to cognitive science.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*