



UNIVERSITY OF
ILLINOIS
URBANA - CHAMPAIGN



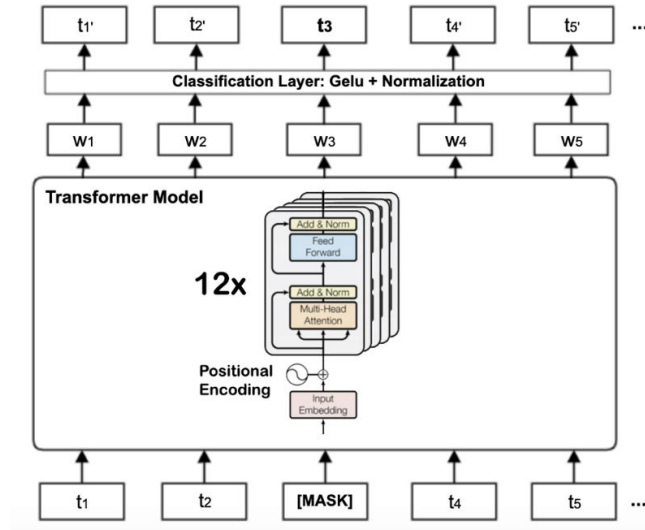
Penn
UNIVERSITY of PENNSYLVANIA

BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language

Philip A Huebner, Elicor Sulem, Cynthia Fisher, Dan Roth

CoNLL 2021

Transformer language models (TLMs) revolutionized NLP



Transformer language models (TLMs) revolutionized NLP

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	69.5	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	81.5	99.6	78.3	80.1	80.5	93.3	86.6	81.3	84.1	70.6	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Table 3: Percentage accuracy of four baseline models and raw human performance on BLiMP using a forced-choice task. A random guessing baseline would achieve an accuracy of 50%.

Bridging Language Acquisition Research with NLP

- Language Acquisition Research
 - How do children acquire the grammar of their native language?
 - What is the contribution of language exposure and conceptual knowledge?

- NLP
 - How to build systems for learning and using natural language data?
 - How much supervision is necessary?



**UPenn Cognitive
Computation
Group**

Bridging Language Acquisition Research with NLP

- The challenge:
 - TLMs are trained on billions of words
 - Existing corpora and benchmarks are unsuited for questions in acquisition research

- Specific Questions:
 - Do TLMs scale-down to psychologically plausible corpus sizes ?
 - How much grammar can TLMs learn given only input to children aged 1-6 years ?

Bridging Language Acquisition Research with NLP

- We made available NLP tools to researchers outside NLP
 - a lightweight TLM trained on a small corpus of child-directed input
 - **BabyBERTa** based on RoBERTa (Liu et al., 20219)
 - a test suite for evaluating grammatical knowledge of masked language models
 - **Zorro** inspired by BLiMP (Warstadt et al., 2020)




<https://github.com/phueb/BabyBERTa>
<https://huggingface.co/phueb/BabyBERTa>



<https://github.com/phueb/Zorro>

Language Data: From 3 Domains

- Child-directed transcribed speech
 - AO-CHILDES
 - children aged 1-6 years
- Adolescent-directed written News articles
 - AO-Newsela
 - targeted to K5-10 students
- Adult-directed written Wikipedia articles
 - Wikipedia-1, Wikipedia-2, Wikipedia-3



language input changes
with age

Language Data: From 3 Domains

Corpus	Sentences	Avg sentence length		Questions (proportion)
		Sub-tokens	Words	
AO-CHILDES	723,524	7.33	6.38	0.42
AO-Newsela	442,571	22.37	15.97	0.01
Wikipedia-1	525,917	31.71	24.77	0.00
Wikipedia-2	525,903	31.71	24.78	0.00
Wikipedia-3	525,352	31.74	24.80	0.00

Table 6: Descriptive statistics for each of our corpora. The reported number of sentences was computed after excluding sentences that contain more than 128 sub-word tokens. The precise number of sentences is irrelevant, because we control for data quantity by stopping training at a pre-defined number of steps. The proportion of questions was determined based on counting question marks.

BabyBERTa



<https://github.com/phueb/BabyBERTa>

- Design Considerations:
 - base-model is state-of-the art TLM (RoBERTa trained with MLM objective)
 - accessible to researchers without access to high-performance computing resources
 - optimized for grammatical knowledge acquisition
 - no unmasking
- Hyper-parameters
 - identified by tuning MLM performance on a held-out portion of AO-CHILDES

Test Suite for Grammatical Knowledge



<https://github.com/phueb/Zorro>

- based on BLiMP (Benchmark of Linguistic Minimal Pairs) (Warstadt et al., 2019)
- content words are counterbalanced across domains
 - each content word has approx. equal probability of occurring in each of our corpora
 - eliminates evaluation bias

Test Suite for Grammatical Knowledge



<https://github.com/phueb/Zorro>

Phenomenon	Paradigm	Examples	
		Well-formed	Not well-formed
Det-subject agreement	across_1_adjective between_neighbors	<i>look at this happy piece .</i> <i>this color must be commercial .</i>	<i>look at this happy pieces .</i> <i>this colors must be commercial .</i>
Subject-verb agreement	across_prepositional_phrase across_relative_clause in_question_with_aux in_simple_question	<i>the brother by the lion is red .</i> <i>the pages that i like were dirty .</i> <i>where does the bird go ?</i> <i>what color was the piece ?</i>	<i>the brothers by the lion is red .</i> <i>the page that i like were dirty .</i> <i>where does the birds go ?</i> <i>what color was the pieces ?</i>
Anaphor agreement	pronoun_gender	<i>she will give herself the wire .</i>	<i>she will give himself the wire .</i>
Argument structure	dropped_argument swapped_argument transitive	<i>my brother moves fast .</i> <i>they built the mouse that farm .</i> <i>will robert eat ?</i>	<i>my brother moves to .</i> <i>the mouse built that farm they .</i> <i>will robert force ?</i>
Binding	principle_a	<i>sarah thinks about herself making a tree .</i>	<i>sarah thinks about herself makes a tree .</i>
Case	subjective_pronoun	<i>they gave the person the tour .</i>	<i>the person gave they the tour .</i>
Ellipsis	n_bar	<i>allen got one roman brain and chris got two .</i>	<i>allen got one brain and chris got two roman .</i>
Filler-gap	question_object wh_question_subject	<i>laura got the suit that the bird cut .</i> <i>chris reached the bear that is washing trains .</i>	<i>laura got what the suit cut the bird .</i> <i>chris reached who the bear is washing trains .</i>
Irregular	verb	<i>sarah spoke without thinking last night .</i>	<i>sarah spoken without thinking last night .</i>
Island effects	adjunct_island coord_struct_constraint	<i>what did robert eat while facing the kiss ?</i> <i>what did sarah and the person work for ?</i>	<i>what did robert eat the kiss while facing ?</i> <i>what did sarah work for and the person ?</i>
Local attractor	in_question_with_aux	<i>can the husband change ?</i>	<i>can the husband changes ?</i>
NPI licensing	matrix_question only_npi_licensor	<i>would william ever keep the movie ?</i> <i>only his rabbit will ever be in her magic .</i>	<i>william would ever keep the movie ?</i> <i>even his rabbit will ever be in her magic .</i>
Quantifiers	existential_there superlative	<i>there was a leg that anne made .</i> <i>no bird could catch more than six plants .</i>	<i>there was most leg that anne made .</i> <i>no bird could catch at least six plants .</i>

Table 5: Examples of well-formed and not well-formed sentences for each paradigm in our grammar test suite. Each paradigm consists of 4,000 sentences (2,000 minimal pairs).

Test Suite for Grammatical Knowledge



<https://github.com/phueb/Zorro>

Phenomenon	Paradigm	Examples	
		Well-formed	Not well-formed
Det-subject agreement	across_1_adjective between_neighbors	<i>look at this happy piece .</i> <i>this color must be commercial .</i>	<i>look at this happy pieces .</i> <i>this colors must be commercial .</i>

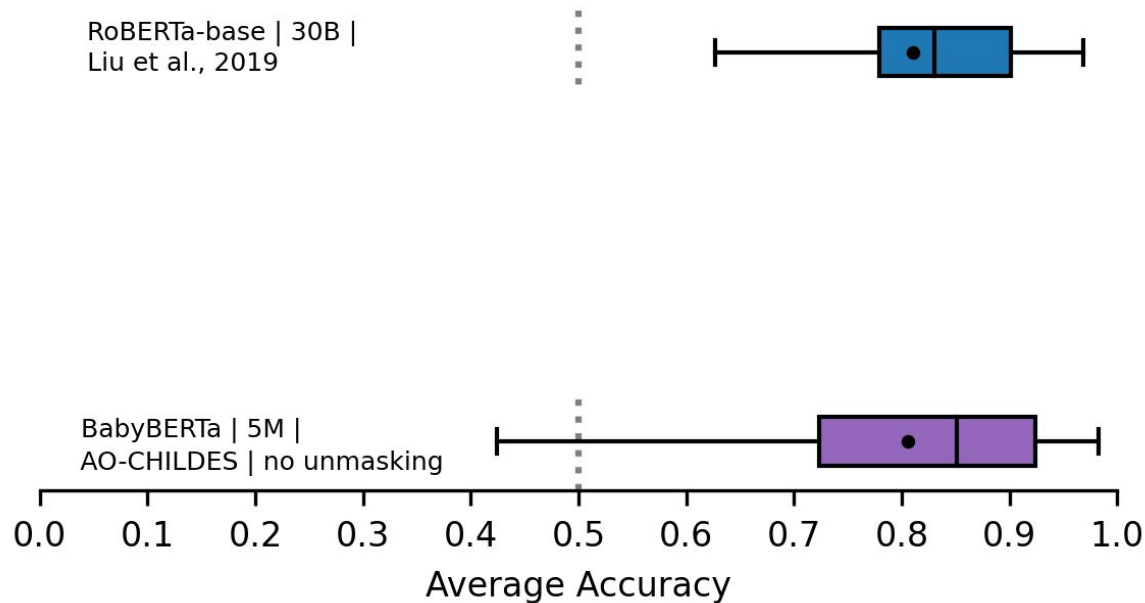
Test Suite for Grammatical Knowledge



<https://github.com/phueb/Zorro>

- Evaluation Procedure
 - forced choice task
 - each sentence in a pair is scored using “holistic scoring” Zaczynska et al. (2020)
 - random guessing baseline would achieve an accuracy of 50%
 - a word-frequency baseline results in an accuracy of 50%

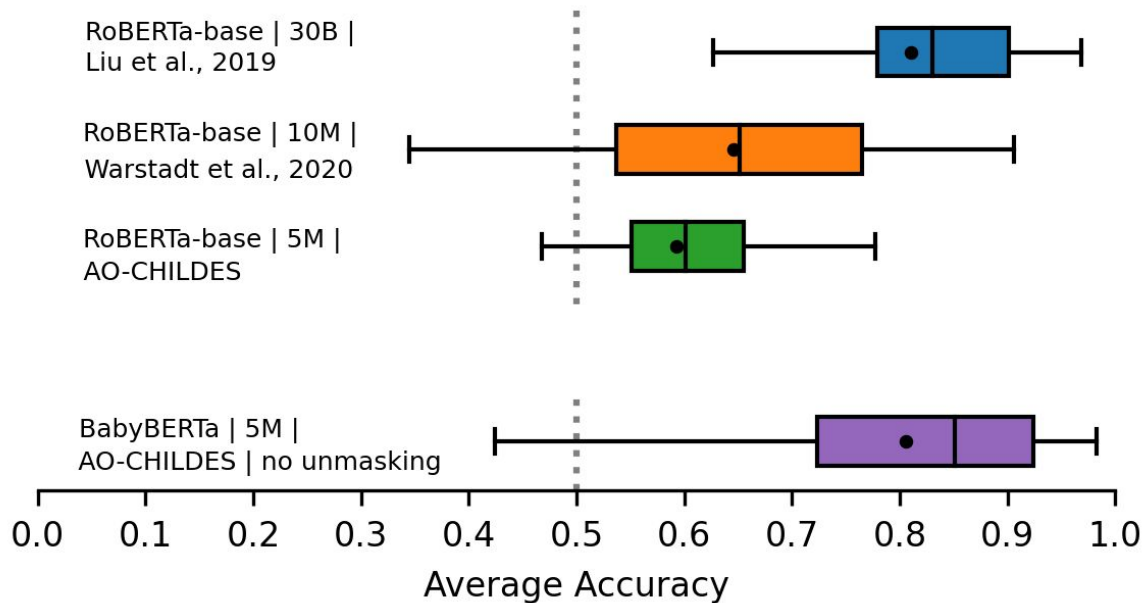
BabyBERTa achieves near RoBERTa performance



Average across paradigms.

Average accuracy is best out of 1 (RoBERTa-base | 30B), 3 (RoBERTa-base | 10M), and 10 for all others.

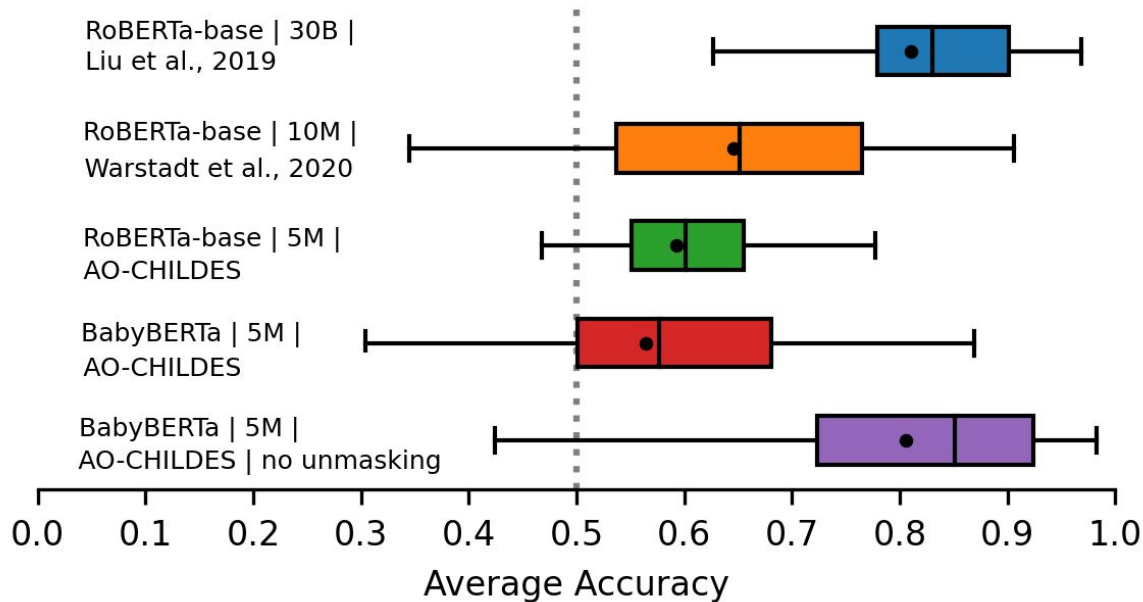
BabyBERTa achieves near RoBERTa performance



Average across paradigms.

Average accuracy is best out of 1 (RoBERTa-base | 30B), 3 (RoBERTa-base | 10M), and 10 for all others.

BabyBERTa achieves near RoBERTa performance



Average across paradigms.

Average accuracy is best out of 1 (RoBERTa-base | 30B), 3 (RoBERTa-base | 10M), and 10 for all others.

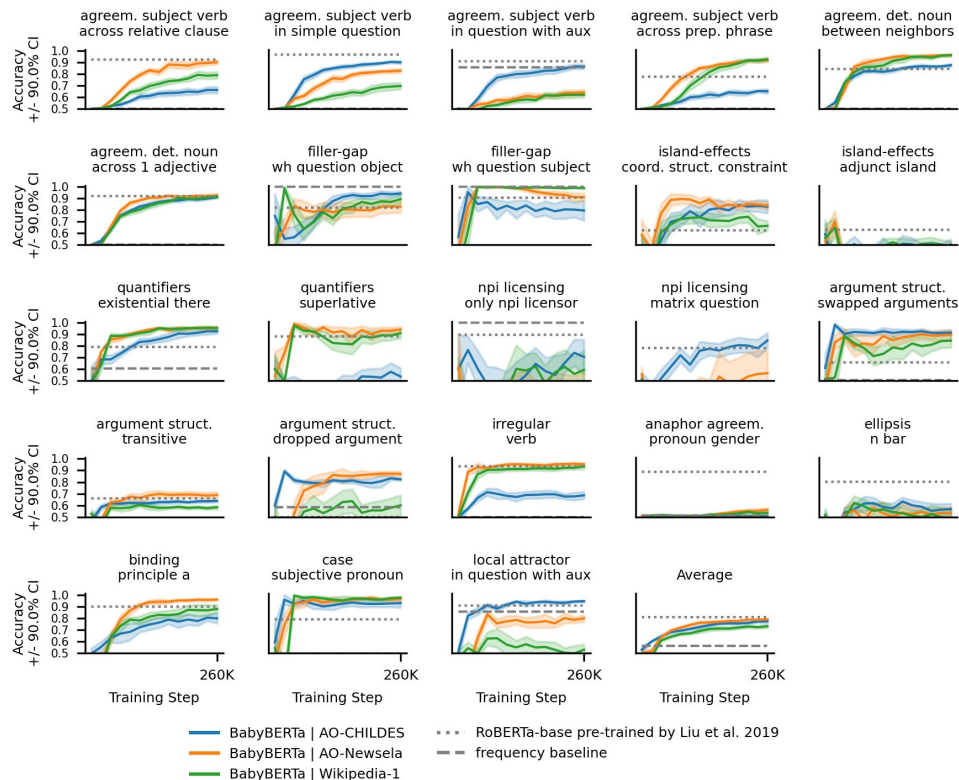
BabyBERTa achieves near RoBERTa performance

Model (Data Size)	Average Accuracy (across Paradigms)
RoBERTa-base – Liu et al., 2019 (30B)	81.1
RoBERTa-base - Warstadt et al., 2020 (10M)	64.5
RoBERTa-base on CHILDES (5M)	59.2
BabyBERTa with unmasking (5M)	56.4
BabyBERTa (5M)	80.5

BabyBERTa achieves near RoBERTa performance

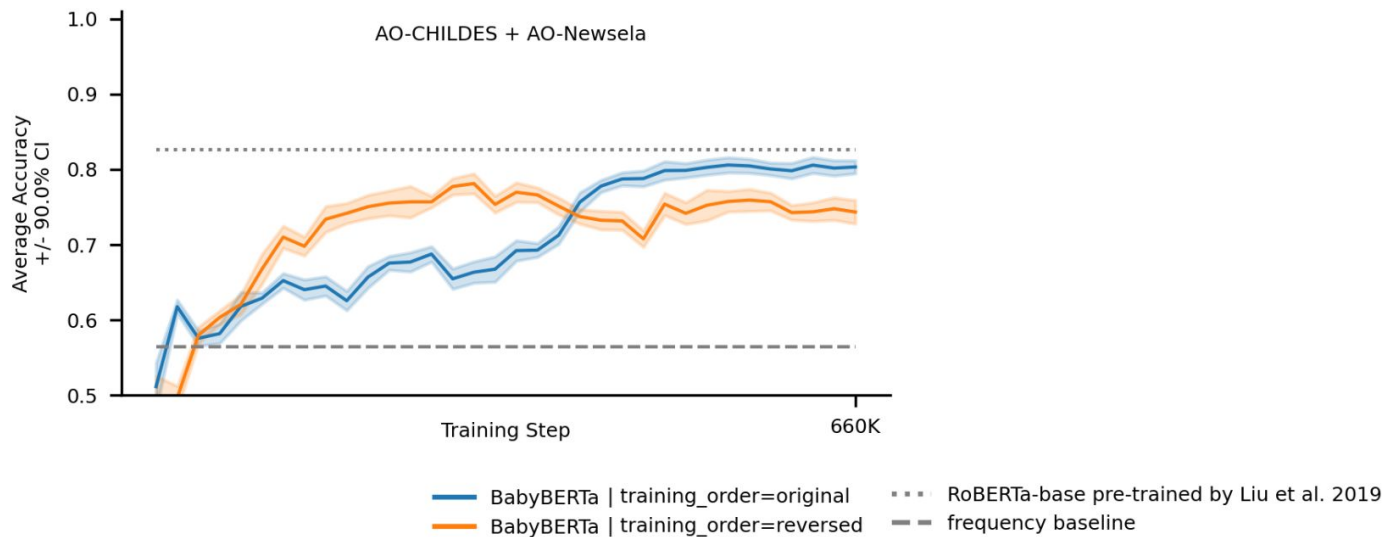
	RoBERTa-base	BabyBERTa
Parameters	125M	8M
Words in data	30B	5M
Hardware (GPU)	1024x V100	1x GTX1080
Training Time	24 hours	2 hours
Vocabulary Size	50265	8192
Average Accuracy	81.0	80.5

BabyBERTa performance by phenomenon and corpus

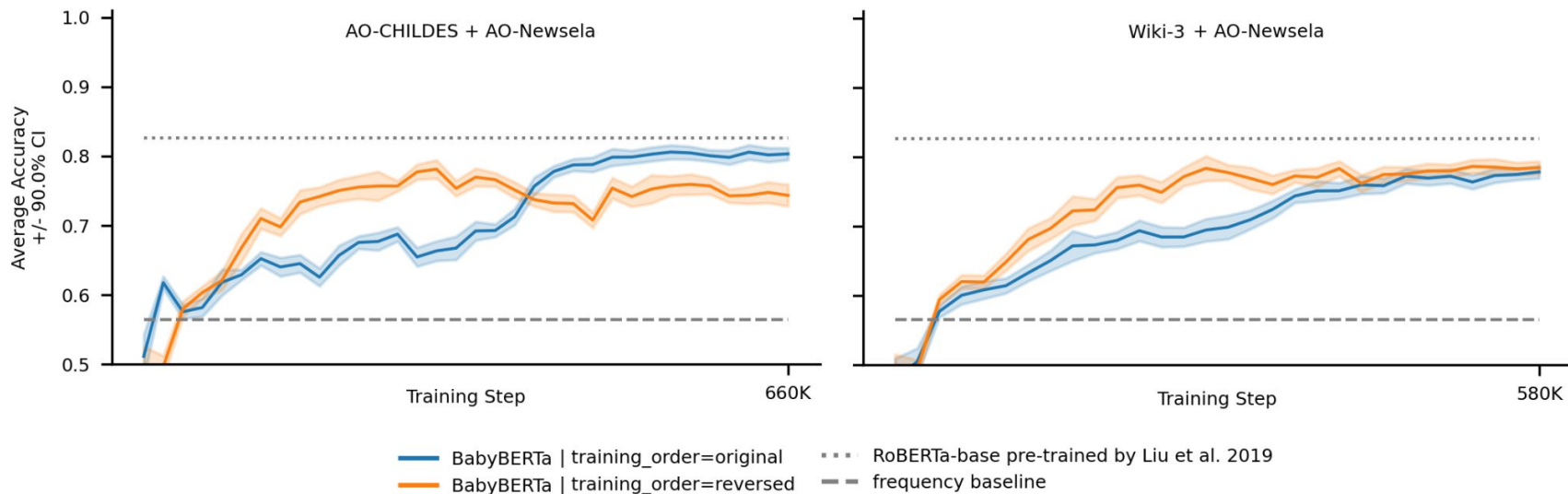


Corpus	Average Accuracy
AO-CHILDES	77.2
AO-Newsela	79.0
Wikipedia-1	73.0

Age-ordered training helps grammar learning



Age-ordered training helps grammar learning



Conclusions

- We provide new tools for using TLMs in language acquisition research.
- TLMs can achieve good performance on grammaticality tests when given an input of comparable quantity and quality to an average English-speaking six-years old.
 - Performance is comparable to that of RoBERTa-base trained on 30B words.
 - Child-directed language is a good starting point for training.

Discussion

- **Alternative evaluation: “MLM scoring”** (Salazar et al., 2020)
 - Does not affect BabyBERTa (without unmasking)
 - Models with unmasking achieve higher performance when evaluating with this measure
 - Our “holistic” evaluation better approximates the conditions under which humans produce acceptability judgments.
- **AO-CHILDES is transcribed speech as opposed to written language.**
 - Future Work: Experimenting with transcriptions of Adult Spoken Language.
- **Unmasking may be important for downstream tasks.**
 - Future Work: Experimenting on downstream tasks

Thanks to Co-authors



Elior Sulem**, Cynthia Fisher*, Dan Roth**

*Department of Psychology, University of Illinois at Urbana-Champaign

**Department of Computer and Information Science, University of Pennsylvania

Acknowledgments

This research was supported by a grant from the NICHD (HD-054448) and by Contracts FA8750-19-2-0201 and FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

I am currently looking for post-doctoral positions starting Fall 2022