

BabyLM Challenge 2023

Held at the 27th Conference on Computational Natural
Language Learning

Singapore
6 – 7 December 2023

ISBN: 978-1-7138-8586-3

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora</i>	
Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen and Ryan Cotterell	1
<i>GPT-wee: How Small Can a Small Language Model Really Get?</i>	
Bastian Bunzeck and Sina Zarriß	35
<i>Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks</i>	
Clayton Fields, Osama Natouf, Andrew McMains, Catherine Henry and Casey Kennington	47
<i>Mini Minds: Exploring Bebeshka and Zlata Baby Models</i>	
Irina Proskurina, Guillaume Metzler and Julien Velcin	58
<i>Grammar induction pretraining for language modeling in low resource contexts</i>	
Xuanda Chen and Eva Portelance	69
<i>ChapGTP, ILLC’s Attempt at Raising a BabyLM: Improving Data Efficiency by Automatic Task Formation</i>	
Jaap Jumelet, Michael Hanna, Marianne de Heer Kloots, Anna Langedijk, Charlotte Pouw and Oskar van der Wal	74
<i>Penn & BGU BabyBERTa+ for Strict-Small BabyLM Challenge</i>	
Yahan Yang, Elior Sulem, Insup Lee and Dan Roth	86
<i>Too Much Information: Keeping Training Simple for BabyLMs</i>	
Lukas Edman and Lisa Bylinina	89
<i>Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior?</i>	
Aryaman Chobey, Oliver Smith, Anzi Wang and Grusha Prasad	98
<i>CLIMB – Curriculum Learning for Infant-inspired Model Building</i>	
Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery and Lisa Beinborn	112
<i>Acquiring Linguistic Knowledge from Multimodal Input</i>	
Theodor Amariuca and Alexander Scott Warstadt	128
<i>Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures</i>	
Julius Steuer, Marius Mosbach and Dietrich Klakow	142
<i>Baby’s CoThought: Leveraging Large Language Models for Enhanced Reasoning in Compact Models</i>	
Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer and Ercong Nie	158
<i>ToddlerBERTa: Exploiting BabyBERTa for Grammar Learning and Language Understanding</i>	
Ömer Veysel Çağatan	171
<i>CogMemLM: Human-Like Memory Mechanisms Improve Performance and Cognitive Plausibility of LLMs</i>	
Lukas Thoma, Ivonne Weyers, Erion Çano, Stefan Schweter, Jutta L Mueller and Benjamin Roth	180

<i>BabyStories: Can Reinforcement Learning Teach Baby Language Models to Write Better Stories?</i> Xingmeng Zhao, Tongnian Wang, Sheri Osborn and Anthony Rios	186
<i>Byte-ranked Curriculum Learning for BabyLM Strict-small Shared Task 2023</i> Justin DeBenedetto	198
<i>McGill BabyLM Shared Task Submission: The Effects of Data Formatting and Structural Biases</i> Ziling Cheng, Rahul Aralikkatte, Ian Porada, Cesare Spinoso-Di Piano and Jackie CK Cheung	207
<i>Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings</i> David Samuel	221
<i>Not all layers are equally as important: Every Layer Counts BERT</i> Lucas Georges Gabriel Charpentier and David Samuel	238
<i>WhisBERT: Multimodal Text-Audio Language Modeling on 100M Words</i> Lukas Wolf, Klemen Kotar, Greta Tuckute, Eghbal Hosseini, Tamar I. Regev, Ethan Gotlieb Wilcox and Alexander Scott Warstadt	253
<i>A surprisal oracle for active curriculum language modeling</i> Xudong Hong, Sharid Loáiciga and Asad Sayeed	259
<i>Mmi01 at The BabyLM Challenge: Linguistically Motivated Curriculum Learning for Pretraining in Low-Resource Settings</i> Maggie Mi	269
<i>Baby Llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty</i> Inar Timiryasov and Jean-Loup Tastet	279
<i>BabyLM Challenge: Curriculum learning based on sentence complexity approximating language acquisition</i> Miyu Oba, Akari Haga, Akiyo Fukatsu and Yohei Oseki	290
<i>Better Together: Jointly Using Masked Latent Semantic Modeling and Masked Language Modeling for Sample Efficient Pre-training</i> Gábor Berend	298
<i>Lil-Bevo: Explorations of Strategies for Training Language Models in More Humanlike Ways</i> Venkata S Govindarajan, Juan Diego Rodriguez, Kaj Bostrom and Kyle Mahowald	308
<i>Towards more Human-like Language Models based on Contextualizer Pretraining Strategy</i> Chenghao Xiao, G Thomas Hudson and Noura Al Moubayed	317
<i>Increasing The Performance of Cognitively Inspired Data-Efficient Language Models via Implicit Structure Building</i> Omar Momen, David Arps and Laura Kallmeyer	327
<i>Pre-training LLMs using human-like development data corpus</i> Khushi Bhardwaj, Raj Sanjay Shah and Sashank Varma	339
<i>On the effect of curriculum learning with developmental data for grammar acquisition</i> Mattia Oppè, J. Morrison and N. Siddharth	346
<i>Optimizing GPT-2 Pretraining on BabyLM Corpus with Difficulty-based Sentence Reordering</i> Nasim Borazjanizadeh	356