

BabyLM 2025

**The First BabyLM Workshop: Accelerating Language
Modeling Research with Cognitively Plausible Data**

Proceedings of the Workshop

November 8, 2025

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN TODO

Introduction

We are excited to welcome attendees to the First BabyLM Workshop! This follows two years of the BabyLM Challenge (now in its third iteration). The workshop will be co-located with the 2025 Conference on Empirical Methods in Natural Language Processing on November 8, 2025 in Suzhou, China.

This year, the program includes an oral session for the winning shared task papers, two oral sessions for the award-winning workshop papers, and a poster session for all accepted submissions. There is also an introductory presentation from the organizers summarizing the challenge, this year’s winning submissions, and trends across submissions. In addition, we have two invited talks—one from a language modeling expert and another from a cognitive modeling expert.

We received 32 workshop submissions (many of which also included system submissions to the BabyLM Challenge) and 12 direct challenge submissions. We are grateful to the challenge participants, whether in the challenge or workshop tracks, for advancing the science of language modeling. The participants’ efforts are essential to advancing the state of cognitively plausible and sample-efficient language modeling.

We also extend our thanks to the organizers of EMNLP for their significant efforts in sustaining a conference of its scale, and in providing an environment for the BabyLM community.

Finally, we thank our program committee members—largely sampled from the participants of the challenge—for committing their time to help us curate an excellent program.

—**The BabyLM Organizing Committee:** Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, Adina Williams

Organizing Committee

Program Chairs

Lucas Charpentier, University of Oslo
Leshem Choshen, Massachusetts Institute of Technology
Ryan Cotterell, ETH Zürich
Mustafa Omer Gul, Cornell University
Michael Y. Hu, New York University
Jing Liu, École Normale Supérieure - PSL
Jaap Jumelet, University of Groningen
Tal Linzen, New York University
Aaron Mueller, Boston University
Candace Ross, Meta AI
Raj Sanjay Shah, Georgia Tech
Alex Warstadt, UC San Diego
Ethan Wilcox, Georgetown University
Adina Williams, Meta AI

Program Committee

BabyLM Challenge Reviewers

Lisa Beinborn, Alessandro Bondielli

Luca Capone

Lukas Edman

Achille Fusco

Nalin Kumar

Mateusz Lango, Sharid Loáiciga

Francesca Padovani, Alexandros Potamianos

Asad B. Sayeed, Jun Suzuki

Joonas Tapaninaho

BabyLM Workshop Reviewers

Xuanda Chen

Justin DeBenedetto

Steven Y. Feng

Kata Gabor, Zebulon Goriely, Leon Guertler

Patrick Haller, Michael Y. Hu

Mayank Jobanputra

Marianne De Heer Kloots, Andrey Kutuzov

Mateusz Lango, Yukyung Lee, Jing Liu, Sharid Loáiciga

Bhavivyva Malik

Ercong Nie

Miyu Oba, Yohei Oseki

Omkar Pangarkar

Anthony Rios

David Samuel, Raj Sanjay Shah, Ekta Sood, Shikhar Srivastava, Julius Steuer, Michal Štefánik

Alex Warstadt

Han Yang

Li Zhou, Richard Zhu

Table of Contents

<i>Rethinking the Role of Text Complexity in Language Model Pretraining</i> Dan John Velasco and Matthew Theodore Roque	1
<i>Contrastive Decoding for Synthetic Data Generation in Low-Resource Language Modeling</i> Jannek Ulm, Kevin Du and Vésteinn Snæbjarnarson	29
<i>Unifying Mixture of Experts and Multi-Head Latent Attention for Efficient Language Models</i> Sushant Mehta, Raj Dandekar, Rajat Dandekar and Sreedath Panat	42
<i>Are BabyLMs Deaf to Gricean Maxims? A Pragmatic Evaluation of Sample-efficient Language Models</i> Raha Askari, Sina Zarrieß, Özge Alacam and Judith Sieker	52
<i>Model Merging to Maintain Language-Only Performance in Developmentally Plausible Multimodal Models</i> Ece Takmaz, Lisa Bylinina and Jakub Dotlacil	66
<i>TafBERTa: Learning Grammatical Rules from Small-Scale Language Acquisition Data in Hebrew</i> Anita Gelboim and Elior Sulem	76
<i>FORGETTER with forgetful hyperparameters and recurring sleeps can continue to learn beyond normal overfitting limits</i> Yamamoto Rui and Keiji Miura	91
<i>Large Language Models and Children Have Different Learning Trajectories in Determiner Acquisition</i> Olivia La Fiandra, Nathalie Fernandez Echeverri, Patrick Shafto and Naomi H. Feldman	100
<i>Design and Analysis of few Million Parameter Transformer-based Language Models trained over a few Million Tokens Dataset</i> Yen-Che Hsiao and Abhishek Dutta	109
<i>What is the Best Sequence Length for BabyLM?</i> Suchir Salhan, Richard Diehl Martinez, Zebulun Goriely and Paula Buttery	130
<i>BitMar: Low-Bit Multimodal Fusion with Episodic Memory for Edge Devices</i> Euhid Aman, Esteban Carlin, Hsing-Kuo Kenneth Pao, Giovanni Beltrame, Ghaluh Indah Permata Sari and Yie-Tarng Chen	147
<i>Exploring smaller batch sizes for a high-performing BabyLM model architecture</i> Sharid Loáiciga, Eleni Fysikoudi and Asad B. Sayeed	155
<i>BLiSS: Evaluating Bilingual Learner Competence in Second Language Small Language Models</i> Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery and Weiwei Sun	160
<i>Sample-Efficient Language Modeling with Linear Attention and Lightweight Enhancements</i> Patrick Haller, Jonas Golde and Alan Akbik	175
<i>Looking to Learn: Token-wise Dynamic Gating for Low-Resource Vision-Language Modelling</i> Bianca-Mihaela Ganescu, Suchir Salhan, Andrew Caines and Paula Buttery	192
<i>A Comparison of Elementary Baselines for BabyLM</i> Rareş Păpuşoi and Sergiu Nisioi	218
<i>Two ways into the hall of mirrors: Language exposure and lossy memory drive cross-linguistic grammaticality illusions in language models</i> Kate McCurdy, Katharina Christian, Amelie Seyfried and Mikhail Sonkin	226

<i>What did you say? Generating Child-Directed Speech Questions to Train LLMs</i>	
Whitney Poh, Michael Tombolini and Libby Barak	237
<i>Beyond Repetition: Text Simplification and Curriculum Learning for Data-Constrained Pretraining</i>	
Matthew Theodore Roque and Dan John Velasco	246
<i>CurLL: A Developmental Framework to Evaluate Continual Learning in Language Models</i>	
Pavan Kalyan Tankala, Shubhra Mishra, Satya Lokam and Navin Goyal	256
<i>A Morpheme-Aware Child-Inspired Language Model</i>	
Necva Bölücü and Burcu Can	279
<i>Do Syntactic Categories Help in Developmentally Motivated Curriculum Learning for Language Models?</i>	
Arzu Burcu Güven, Anna Rogers and Rob Van Der Goot	288
<i>SlovakBabyLM: Replication of the BabyLM and Sample-efficient Pretraining for a Low-Resource Language</i>	
Ľuboš Kriš and Marek Suppa	301
<i>Single layer tiny Co4 outpaces GPT-2 and GPT-BERT</i>	
Noor Ul Zain, Mohsin Raza Naseem and Ahsan Adeel	313
<i>Teacher Demonstrations in a BabyLM's Zone of Proximal Development for Contingent Multi-Turn Interaction</i>	
Suchir Salhan, Hongyi gu, Donya Rooein, Diana Galvan-Sosa, Gabrielle Gaudeau, Andrew Caines, Zheng Yuan and Paula Buttery	323
<i>Influence-driven Curriculum Learning for Pre-training on Limited Data</i>	
Loris Schoenegger, Lukas Thoma, Terra Blevins and Benjamin Roth	356
<i>Understanding and Enhancing Mamba-Transformer Hybrids for Memory Recall and Language Modeling</i>	
Hyunji Lee, Wenhao Yu, Hongming Zhang, Kaixin Ma, Jiyeon Kim, Dong Yu and Minjoon Seo	380
<i>Findings of the Third BabyLM Challenge: Accelerating Language Modeling Research with Cognitively Plausible Data</i>	
Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox and Adina Williams	399
<i>Dialogue Is Not Enough to Make a Communicative BabyLM (But Neither Is Developmentally Inspired Reinforcement Learning)</i>	
Francesca Padovani, Bastian Bunzeck, Manar Ali, Omar Momen, Arianna Bisazza, Hendrik Buschmeier and Sina Zarrieß	421
<i>CLASS-IT: Conversational and Lecture-Aligned Small-Scale Instruction Tuning for BabyLMs</i>	
Luca Capone, Alessandro Bondielli and Alessandro Lenci	436
<i>Mask and You Shall Receive: Optimizing Masked Language Modeling For Pretraining BabyLMs</i>	
Lukas Edman and Alexander Fraser	445
<i>Once Upon a Time: Interactive Learning for Storytelling with Small Language Models</i>	
Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid and Lisa Beinborn	454

<i>You are an LLM teaching a smaller model everything you know: Multi-task pretraining of language models with LLM-designed study plans</i>	
Wiktor Kamzela, Mateusz Lango and Ondrej Dusek	469
<i>Active Curriculum Language Modeling over a Hybrid Pre-training Method</i>	
Eleni Fysikoudi, Sharid Loáiciga and Asad B. Sayeed	488
<i>Linguistic Units as Tokens: Intrinsic and Extrinsic Evaluation with BabyLM</i>	
Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo and Cristiano Chesi	496
<i>Batch-wise Convergent Pre-training: Step-by-Step Learning Inspired by Child Language Development</i>	
Ko Yoshida, Daiki Shiono, Kai Sato, Toko Miura, Momoka Furuhashi and Jun Suzuki	508
<i>Pretraining Language Models with LoRA and Artificial Languages</i>	
Nalin Kumar, Mateusz Lango and Ondrej Dusek	525
<i>Masked Diffusion Language Models with Frequency-Informed Training</i>	
Despoina Kosmopoulou, Efthymios Georgiou, Vaggelis Dorovatas, Georgios Paraskevopoulos and Alexandros Potamianos	531
<i>MoEP: Modular Expert Paths for Sample-Efficient Language Modeling</i>	
Joonas Tapaninaho	540
<i>RecombiText: Compositional Data Augmentation for Enhancing LLM Pre-Training Datasets in Low-Resource Scenarios</i>	
Alexander Tampier, Lukas Thoma, Loris Schoenegger and Benjamin Roth	548

Program

Saturday, November 8, 2025

09:00 - 09:15 *Opening Remarks*

09:15 - 10:15 *BabyLM Challenge Orals*

CLASS-IT: Conversational and Lecture-Aligned Small-Scale Instruction Tuning for BabyLMs

Luca Capone, Alessandro Bondielli and Alessandro Lenci

Masked Diffusion Language Models with Frequency-Informed Training

Despoina Kosmopoulou, Efthymios Georgiou, Vaggelis Dorovatas, Georgios Paraskevopoulos and Alexandros Potamianos

MoEP: Modular Expert Paths for Sample-Efficient Language Modeling

Joonas Tapaninaho

Mask and You Shall Receive: Optimizing Masked Language Modeling For Pre-training BabyLMs

Lukas Edman and Alexander Fraser

Once Upon a Time: Interactive Learning for Storytelling with Small Language Models

Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid and Lisa Beinborn

10:15 - 11:00 *Break*

11:00 - 12:00 *Invited Talk 1*

12:00 - 14:00 *Lunch*

14:00 - 15:30 *Poster Session*

15:30 - 16:00 *Break*

16:00 - 17:00 *Invited Talk 2*

17:00 - 17:35 *BabyLM Workshop Orals*

Saturday, November 8, 2025 (continued)

Teacher Demonstrations in a BabyLM's Zone of Proximal Development for Contingent Multi-Turn Interaction

Suchir Salhan, Hongyi gu, Donya Rooein, Diana Galvan-Sosa, Gabrielle Gau-
deau, Andrew Caines, Zheng Yuan and Paula Buttery

Are BabyLMs Deaf to Gricean Maxims? A Pragmatic Evaluation of Sample-efficient Language Models

Raha Askari, Sina Zarrieß, Özge Alacam and Judith Sieker

Looking to Learn: Token-wise Dynamic Gating for Low-Resource Vision-Language Modelling

Bianca-Mihaela Ganesu, Suchir Salhan, Andrew Caines and Paula Buttery

17:35 - 17:45 *Awards and Closing Remarks*