

Findings of the Fourth BabyLM Challenge & Workshop: Going Beyond English

BabyLM Team

Leshem Choshen
IBM Research, MIT

Ryan Cotterell
ETH Zürich

Mustafa Omer Gul
Cornell University

Jaap Jumelet
University of Groningen

Tal Linzen
NYU

Aaron Mueller
Boston University

Suchir Salhan
University of Cambridge

Raj Sanjay Shah
Georgia Tech

Alex Warstadt
UC San Diego

Ethan Gotlieb Wilcox
Georgetown University

Abstract

This report summarizes the findings from the 4th BabyLM Challenge. The BabyLM Challenge is a shared task aimed at closing the data-efficiency gap between human and machine language learners. This year, the challenge was held as part of the 2nd BabyLM Workshop. We retained the text-only evaluation tracks, *Strict* and *Strict-Small*, while allowing multimodal data and interactive training within these tracks. We also introduced a new *Multilingual* track, where language models were required to perform well on BabyLM evaluations in English, Dutch, and Mandarin Chinese. We observe that GPT-BERT remains a strong architectural foundation for sample-efficient pre-training, and that more training epochs, recursion/looping, and data preprocessing generally improve performance. This year’s winning submissions to the new multilingual track also show that LLM-generated parallel data can be effective at improving multilingual capabilities in low-data settings.

1 Introduction

In recent years, the trend has been toward ever-larger-scale language model (LM) training. This leads to LM methods research often being considered the domain of industry research. Further, LM methods research is also often considered irrelevant to the study of language learning and processing in humans, as the number of words required to train a state-of-the-art model is orders of magnitude greater than the number of words a human would hear in their lifetime.

To advance the science of language modeling at the academic scale and create more cognitively

plausible LMs, the BabyLM Challenge encourages researchers to train language models with the amount of language typical of human language acquisition. This paper presents and analyzes the main findings from the fourth iteration of the BabyLM Challenge.¹ We also describe participating systems and some key takeaways from the BabyLM Workshop, to which participants could submit papers without submitting to the challenge.

The objective of the BabyLM Challenge is to train a model with 100M words or fewer. For entrants wishing to work at an even smaller scale, we also organized a 10M-word track. The challenge explicitly refrains from restrictions on anything other than the word count. In doing so, we hope to encourage new approaches that improve LMs’ sample efficiency and reveal why standard LLMs are so data hungry. The previous year’s interaction track has been folded into these tracks. This year, we additionally introduce a multilingual track inspired by BabyBabelLM (Jumelet et al., 2026a). The multilingual track encourages authors to train efficient models that perform well in three languages (English, Dutch, and Mandarin Chinese) within the same fixed word budget as the monolingual tracks.

Summary of takeaways. BabyLMs continue to improve at cognitive-alignment and NLP tasks in the *Strict* and *Strict-Small* tracks. Statistical analyses of challenge contributions show that models with longer sequences, more training epochs, and more layers tend to perform better. Data preprocessing and specific model architectures (e.g.,

¹For findings from previous years, see Warstadt et al. (2023); Hu et al. (2024); Charpentier et al. (2025).

GPT-BERT) also boost overall performance. In the multilingual track, we observe overall higher scores in English and Dutch, with models’ performance on Chinese slightly underpowered based on their overall capabilities. The BabyLM workshop contributions explored a range of topics, with particular emphasis on multilingual analysis and data-efficient learning via architectural and training innovations. Overall, we were pleased to see an increase in the number of BabyLM participants, indicating growing interest in the intersection of cognitive modeling and data-efficient pretraining.

2 Competition Details

2.1 Data

Text-only Corpora. We provided updated versions of the BabyLM corpora used in the previous challenges, with 10M-word and 100M-word datasets for the *Strict-Small* and *Strict* tracks (Choshen et al., 2026). These corpora included child-directed speech from CHILDES; dialogue from the spoken portion of the British National Corpus and Switchboard; children’s stories from Project Gutenberg; movie and television subtitles from OpenSubtitles; and articles from Simple English Wikipedia.

Dataset Revision. The main change from previous years was detoxification of the training data. This revision followed prior work identifying potentially toxic content in the BabyLM corpus and toxic generations during interactive training (Salhan et al., 2025; Trhlík et al., 2026).

Multimodal and Multilingual Data. Participants could use multimodal training data within the relevant track’s word budget. The new Multilingual track focused on English, Dutch, and Chinese, using data from BabyBabelLM (Jumelet et al., 2026a). Participants could choose their own mixture of these languages, subject to the total data budget and language-specific Byte Premium adjustments.

2.2 Strict & Strict-Small Tracks

Track Overview. The *Strict-Small* and *Strict* tracks required submissions to be trained on datasets of at most 10M and 100M words, respectively. Participants could use either the provided BabyLM corpus or construct their own training datasets (provided they adhered to the word

limit). Participants could use any model architecture that could assign log-likelihoods or pseudo-log-likelihoods to strings of text. We evaluated both tracks on language-only tasks. This year, we also merged the *Multimodal* and *Interaction* tracks from previous BabyLM offerings into the *Strict* and *Strict-Small* tracks.

Interactive Training. Participants could use an external model in their submission model’s training pipeline. Like last year’s *Interaction* track, the external model could provide scalar rewards or natural language feedback, or generate training examples conditioned on the submission model’s outputs. External models were restricted to a pre-defined list, but their training and fine-tuning were otherwise unrestricted. External model training text counted toward the word budgets for the respective tracks. The constraints also required that the submission model could not access the external model’s tokenizer, weights, hidden states, or output distribution.

Training Duration Limitations. We restricted competition submissions to at most 100M words of cumulative training exposure for *Strict-Small* and 1B words for *Strict*, counting repeated passes over the same input. For models that generated text during training, the exposure budget also includes their generated output. These limits corresponded to ten epochs over a dataset at the full word budget.

Intermediate Checkpoints. Following last year’s setup, we asked participants to submit intermediate model checkpoints at increasing word-exposure intervals. Specifically, we required checkpoints every 1M words through 10M words and every 10M words through 100M words; for *Strict*, we also required checkpoints every 100M words through 1B words.

2.3 Multilingual Track

The Multilingual track was introduced this year, in which participants trained the BabyLM model on a mixture of languages. We limited the number of languages to keep the competition format computationally feasible on smaller hardware. With a similar compute budget as the *Strict* track of 100M tokens, participants were free to define a mixture of English, Chinese, and Dutch data. Models were scored based on the average performance across these three languages.

2.4 Evaluation

Evaluation for the *Strict* and *Strict-Small* tracks remained similar to prior editions. We divided tasks into two groups: **NLP tasks** that could be measured against ground truth, such as GLUE and BLiMP, and **Human-like tasks**, which were correlated with human judgments. From previous edition’s results, it became clear that these human-like correlations are quite unstable and caused some models to score highly on the leaderboard by pure chance. Leaderboard winners are now decided only based on NLP task performance. Evaluation of the **multilingual** track was leveraged from BabyBabelLM, which contains both zero-shot and fine-tuning evaluation tasks for the three languages of the track. Tasks include language-specific linguistic evaluation (Warstadt et al., 2020; Suijkerbuijk et al., 2025; Liu et al., 2026; Jumelet et al., 2026b), classification tasks (Adelani et al., 2024), and question-answering (Romanou et al., 2024).

Hidden tasks. We supplemented the released evaluation suite with hidden tasks to assess generalization beyond the public evaluations. Across all three tracks, we introduced Global-PIQA (Chang et al., 2026), which tested physical commonsense reasoning using parallel and non-parallel evaluation sets. We used the English tasks for *Strict* and *Strict-Small*, and the English, Dutch, and Chinese tasks for the Multilingual track. For the Multilingual track, we also introduced **part-of-speech probing**, released two weeks before the submission deadline. Linear probes trained on Universal Dependencies annotations assessed the part-of-speech information available in model representations across languages (Nivre et al., 2020; Ravishankar et al., 2019). For Chinese, we added the Hanzi-structure and Hanzi-pinyin tasks, which use minimal pairs to assess models’ knowledge of character structure and phonology, respectively (Song et al., 2026).

Multilingual reading-time evaluation. We introduced new evaluations on eye-tracking data using the Multilingual Eye-Tracking Corpus (MECO) (Siegelman et al., 2022). Models were evaluated on how their surprisals improve prediction of human first-pass reading times, controlling for word length, frequency, and sentence position. Models were evaluated on first language (L1) MECO data for Dutch, Chinese, and English, and second language (L2) English.

2.5 Baselines

Similar to the evaluation, baselines for the *Strict* and *Strict-Small* tracks are mostly the same as in the last iteration, except for removing GPT-BERT as a baseline. We have retrained the naive, autoregressive GPT-2 Small baseline (Radford et al., 2019) using the same recipe, but with the revised dataset. As a result of the merging of the Interaction track with the *Strict* and *Strict-Small* tracks this year, we have also retrained the Interaction baselines from the Third BabyLM Challenge (Charpentier et al., 2025) on the updated dataset.

For the Multilingual track, we re-use the autoregressive GPT-2 Small training recipe but vary the dataset construction process. Our primary baseline for the track mixes the BabyLM corpora for English, Dutch, and Chinese (Jumelet et al., 2026a) such that each language occupies a similar proportion after Byte-level Premium adjustments. For each language dataset, we subsample the necessary number of words while preserving the proportion of tokens contributed by each constituent data source for the language. We then train a BPE tokenizer on this data mixture and train the model without any further interventions in terms of language ordering. This is a naive baseline intended to provide participants with the resources and utilities to process and mix the datasets.

3 Submissions

We received 44 Challenge submissions, 8 ARR submissions, and 58 workshop submissions. Out of these, we accepted 36 Challenge submissions (81.8%), 7 ARR submissions (87.5%), and 35 workshop papers (60.3%). This number of submissions marks a stark increase compared to prior editions, which we take as a sign of the community growing and there being a strong interest in the Challenge as a way to test data-efficient language modeling in a controlled environment, and the Workshop as a place to explore broader scientific questions related to this topic.

4 Analysis of Challenge Results

The BabyLM Challenge submissions explore a wide range of strategies for improving language modeling under strict data and compute constraints. We survey the challenge submissions, followed by a quantitative analysis of leaderboard scores.

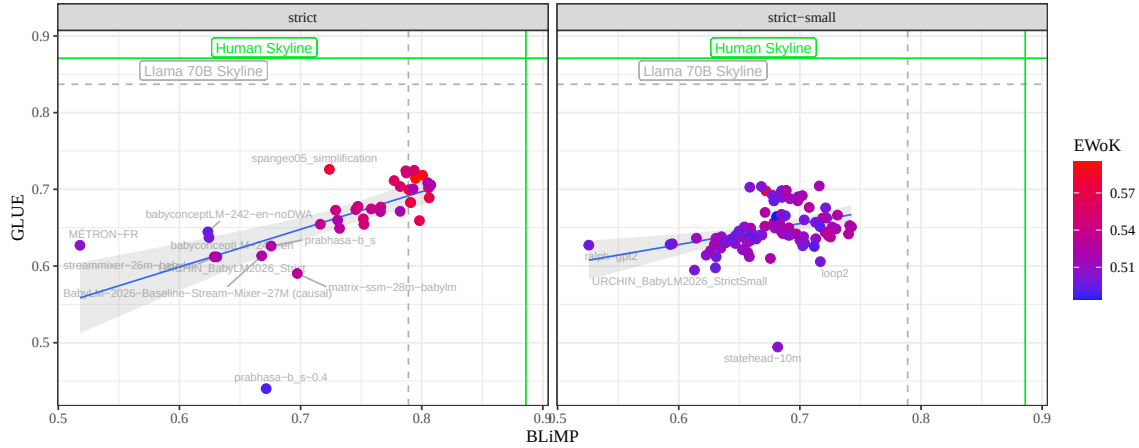


Figure 1: Comparison between models’ scores on BLiMP and GLUE vs. Human scores reported for these tasks. Scores are shown for the *Strict* and *Strict-Small* tracks. We visualize these two tasks, as they have been used consistently through all iterations of the BabyLM challenge. For overall scores, see fig. 3. All leaderboard submissions are considered here.

4.1 Survey

Architecture. Submissions investigated mixture-of-experts models (Mustafa et al., 2026; Jain et al., 2026; Bohn et al., 2026), recurrent and weighted architectures (Furr, 2026; Miles, 2026; Gülbahar et al., 2026; Kamzela et al., 2026; Fan et al., 2026), higher-order attention (Chan, 2026), and parameter-efficient embedding and projection layers (Mosquera-Gómez et al., 2026). Other approaches included a horizontal spiking architecture that operates without attention (Chiang, 2026) and input-dependent routing over earlier-layer representations (Wang, 2026). Furr (2026) demonstrated grammatical-discrimination capabilities with about 10M parameters and a fixed-size recurrent state, although its gains did not consistently extend to semantic or entity-tracking tasks. Gülbahar et al. (2026) explored recursive Transformers that increase computational depth without proportionally increasing the number of parameters.

Several papers examined whether additional recurrent computation could compensate for limited parameters or data (Hotta et al., 2026; Haller et al., 2026; Fan et al., 2026). The results depended on the task and comparison budget. For example, Miles (2026) reported lower validation bits per byte with looped depth, but the downstream gain was small relative to its uncertainty, and training took longer. Hotta et al. (2026) found a text-only BLiMP advantage for recurrence under matched inference compute, but recurrence did not clearly mitigate the performance decline associated with their multimodal training condition. These findings motivate

distinguishing parameter efficiency, data efficiency, and computational cost.

Training Objective and Optimization. Submissions explored auxiliary objectives (Palomino, 2026; Benedikt et al., 2026), preference optimization (Štěpánková et al., 2026; Mishra and Roy, 2026), adaptive teacher feedback based on corrections to student-generated text (Islam et al., 2026), and optimizer and learning-rate choices (Satheesh et al., 2026). Other work combined masked and causal language modeling in a multilingual OLMo2 model (Wei, 2026), introduced objectives for syntagmatic and paradigmatic relations (Yang et al., 2026b), or varied the balance between replaced-token detection and masked language modeling (Mikat and Fahim, 2026). Štěpánková et al. (2026) investigated how preference-data selection affected downstream performance, while Mishra and Roy (2026) applied preference post-training after downstream performance had saturated. The evidence did not uniformly favor additional objectives or feedback; the data2vec auxiliary did not yield significant task-level gains in the controlled comparison of Benedikt et al. (2026), and the official-corpus baseline retained the highest overall score in Islam et al. (2026).

Data composition and curricula. Submissions examined multilingual language allocation and targeted data design (Mori et al., 2026), tokenization, packing, and curriculum learning (Yang et al., 2026a), code-switching curricula (Rooryck et al., 2026; Huang et al., 2026), meaning-preserving

rewrites (Wu et al., 2026), and concise parallel text generated by an LLM (Edman and Fraser, 2026). Hwang et al. (2026) compared mixed multilingual training with a curriculum that began with monolingual exposure. Mixed training performed better in all 15 matched configurations, and several language-dominant mixtures exceeded the balanced baseline; these comparisons used one run per configuration. Rodriguez et al. (2026) combined curriculum pretraining with layer repetition at inference.

Tokenization and representation. Submissions studied subword segmentation learned during language-model training (Meyer, 2026), semantically informed morphological tokenization for English (Rodriguez, 2026), and character-level tokenization in Qwenfant (Edman and Fraser, 2026). Other papers investigate how models can acquire particular capabilities or representations, including dynamic concept computation (yizhe xu et al., 2026), linguistic parameters (Chesi et al., 2026), reducible-loss spans and self-generation (Kalfus and Goyal, 2026), and practical life skills (Štěpánková et al., 2026). Finally, several submissions examine broader questions about data-efficient learning, including whether looped computation can substitute for parameters (Fan et al., 2026), whether recurrent depth is beneficial under data limitations (Haller et al., 2026), whether layer duplication can improve small models (Rodriguez et al., 2026), how relational inductive biases affect learning (Yang et al., 2026b), and how dense supervision and curriculum-guided rebalancing can improve small-data pretraining (Mikat and Fahim, 2026).

Together, these submissions span architectural, data, objective, tokenization, and training-procedure interventions for studying how much language learning can be achieved from limited exposure. The highest-scoring Challenge submissions span several approaches to improving learning under limited data. These papers illustrate the breadth of the Challenge: improving small language models can involve changing the model’s computational structure, the representation of its input, the composition and ordering of its experience, or the optimization procedure used to learn from that experience.

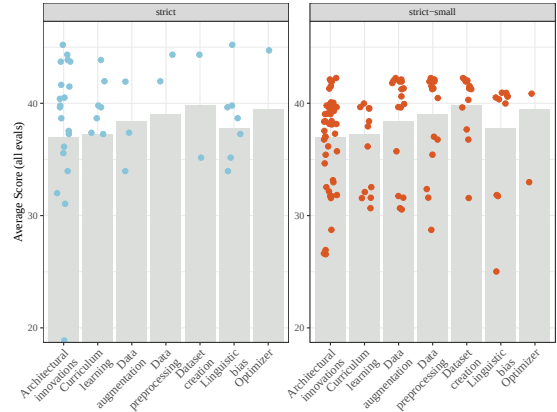


Figure 2: Approaches to the challenge broken down by score. Submissions that use more than one approach are counted multiple times.

4.2 Quantitative Analysis

We conduct various ‘meta-analyses’ on the models submitted to the BabyLM Leaderboard to gain an understanding of the types of submissions people have made this year, and which model properties drive strong performance. Participants had to enter various model parameters and experimental design choices, which we can leverage to plot and fit correlations between these properties and leaderboard scores. We hope this will inspire next year’s participants to set up their experimental design.

Figure 1 shows the GLUE performance set out against BLiMP performance for the Strict and Strict-Small models. Models are set out against a Human and LLM ‘skyline’. While strong GLUE performance remains challenging for models of this size, this result does show that their *linguistic* abilities are on par with Llama3-70B, a model several orders of magnitude larger.

In Figure 2 we show a breakdown of model scores by Contribution Type, using contribution type categories from previous BabyLM findings papers. The majority of model submissions introduced architectural innovations, resulting in a wide range of model scores. Curriculum learning remains popular, while data augmentation and data preprocessing both see a strong set of model contributions. Contribution Type alone, however, is not sufficient to explain model results in detail, so we also conduct a broader linear modeling analysis.

In Table 1 we show the β coefficients of a linear model fitted on the normalized average NLP task scores of the 118 BabyLM submissions in the Strict and Strict-Small tracks. Significant variables

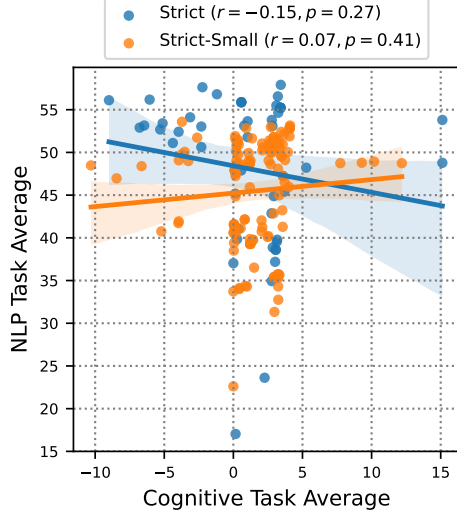


Figure 3: Connection between NLP task performance and “Cognitive tasks”. Cognitive tasks are correlated with human scores and include reading time prediction and age-of-acquisition.

are highlighted, indicating various model properties driving strong performance. *Max Sequence Length* is positively correlated with performance, as is the number of training epochs and the number of layers (more so than the number of parameters). Out of the Contribution Types, *Data Preprocessing* is the only significant factor. Model Type again shows, as in similar years, that the hybrid GPT-BERT model (Charpentier and Samuel, 2024), which combines the strengths of decoder and encoder models, remains a very strong model architecture. Encoder models outperform decoder models; encoder-decoder models perform poorly. Most contributions used the AdamW optimizer, but none of the optimizer types had a significant effect.

Finally, we look at the connection between NLP task performance and the Human-like (Cognitive) tasks in Figure 3. Measuring ‘human likeness’ remains a challenge for evaluation; many of the tasks we considered for this did not work well at the BabyLM scale, and the remaining few tasks we do have turned out to be brittle and prone to noisy artifacts. Neither Strict nor Strict-Small models show a significant effect of NLP Task performance on cognitive task performance. Improving Cognitive evaluation remains an open question that we look forward to addressing in next year’s edition.

4.3 Multilingual Track

For the Multilingual Track, we compare model performance cross-lingually. Figure 4 shows language

Variable	Value	n	β	p
Batch Size			-0.12	0.16
Learning Rate			0.19	0.17
Max Seq Length			0.22	0.02
PFLOPS			-0.03	0.72
Token Set Size			-0.07	0.44
# Training Epochs			0.27	<0.01
# of Attention Heads			0.16	0.20
# of Layers			0.20	0.03
# of Parameters (M)			0.05	0.66
Main Contribution	Architectural	62	-0.24	0.20
	Controlled experiments	37	-0.08	0.70
	Curriculum learning	19	0.34	0.21
	Data augmentation	13	-0.18	0.50
	Data preprocessing	11	0.72	0.03
	Dataset creation	7	0.01	0.97
	Hyperparameter tuning	6	0.16	0.67
	Linguistic bias	17	-0.21	0.37
	Teacher models	4	-0.08	0.89
Model Type	Training objective	27	-0.02	0.95
	Decoder	68	-0.07	0.70
	Encoder	36	0.57	<0.01
	Encoder-Decoder	8	-1.48	<0.01
Optimizer	GPT-BERT	6	0.97	<0.01
	AdamW	71	0.24	0.30
	LAMB	26	-0.11	0.64
	Muon	10	-0.12	0.65
	MuonAdamW	8	-0.40	0.19
	Other	3	0.39	0.39
Track	strict	35	0.27	<0.01
	strict-small	83	-0.27	<0.01

Table 1: Linear model predicting z -scored NLP Average for 118 BabyLM Strict and Strict-Small submissions ($R^2 = 0.62$). Continuous predictors are z -scored; categorical coefficients are deviations from the mean across levels.

averages against overall multilingual averages. It can be seen in this figure that across all models, performance on Dutch and English is slightly above average, whereas Chinese is below average. Whether this is driven by easier language transfer between English and Dutch requires a broader experimental setup, though. In next year’s edition, we will consider ways to incorporate a broader set of languages while keeping the track computationally light enough to make it easy to participate.

4.4 Challenge: Winning Submissions

Strict-Small Track Winner. Satheesh et al. (2026) achieved the highest score on the Strict-Small track with the BabySteps model. These models are based on GPT-BERT. Via various ablation studies, they find that a well-tuned baseline model can perform exceptionally well without any out-of-the-ordinary interventions to the training data or method. Muon-family optimizers (AdaMuon for Strict-Small, Muon for Strict) and careful learning rate tuning yielded the largest and most reliable gains, whereas curriculum learning, synthetic data, distillation, and data mixing did not reliably improve the overall score.

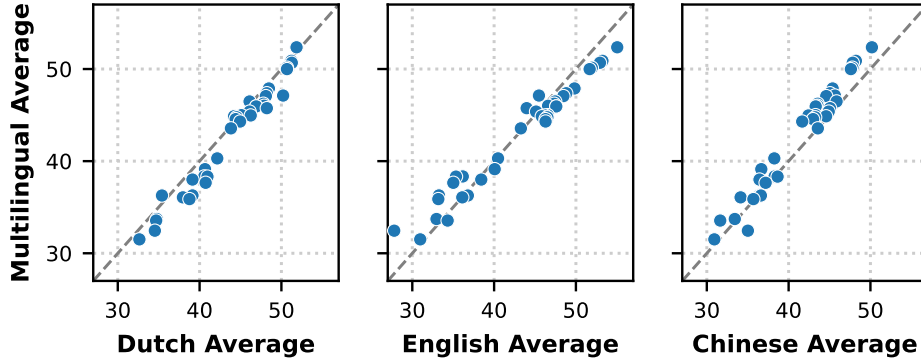


Figure 4: Model performance of the multilingual track submissions, with average language performance set out against overall multilingual average performance. Scores below the diagonal have skewed performance towards that language.

Strict Track Winner. [Satheesh et al. \(2026\)](#) also achieved the highest score on the Strict track. The architecture and methods for their submission to this track were largely identical to those of the strict-small track. We note that RecGPT ([Gülbahar et al., 2026](#)) came a close second. It is a decoder-only model that reuses one Transformer block recurrently to spend more computation per token without adding parameters. It also factorizes the vocabulary embeddings and uses a loss term that predicts the next hidden state.

Multilingual Track Winner. [Edman and Fraser \(2026\)](#) win the Multilingual track with Qwenfant-0.4B. Their performance appears to come primarily from data preprocessing: they pack training examples, use an LLM to summarise them into more concise text, and translate the summaries to produce parallel English, Dutch, and Chinese data. In addition, they use a relatively large Qwen-3 architecture and a character-level vocabulary. They also use LoRA-style fine-tuning, rather than the more typical full-rank fine-tuning.

5 Analysis of Workshop Papers

The BabyLM Workshop extended the Challenge’s focus on data-efficient language learning to cognitive modeling and related methodological questions, including studies outside the competition’s training constraints. The papers examined how multilingual experience, model structure, and training signals affect learning from limited data, and how standard benchmark results relate to more targeted measures of linguistic and developmental behavior.

5.1 Highlighted Workshop Papers

Multimodality. TinyLibrary used illustrated children’s books to investigate age-based curricula for small vision-language models ([Varghese, 2026](#)). Ordering training material by its intended reader age did not show a reliable advantage over random ordering, suggesting that reader age alone is an insufficient guide to model learning difficulty.

Multilinguality. The release of BabyBabelLM ([Jumelet et al., 2026a](#)) and the introduction of the multilingual challenge ([Choshen et al., 2026](#)) provided a setting for studying cross-linguistic learning. Workshop papers also addressed language-specific questions of corpus construction. [Haider and Can \(2026\)](#) examined how mixing Modern Standard Arabic and Dialectal Arabic affected learning under different tokenization schemes. [Pulipaka and Barak \(2026\)](#) used IndicTrans2 to translate the English BabyLM corpus into Hindi and Telugu, then trained language models on the resulting corpora, testing translation as a means of extending BabyLM-style training to additional languages.

Tokenization. [Badawy et al. \(2026\)](#) compared BPE with morphology-aware tokenization for Arabic. Their experiments did not show consistent downstream improvements over BPE, even when a segmentation constraint substantially reduced perplexity. Similarly, [Li et al. \(2026\)](#) found that reweighting BPE merges using association measures changed segmentation without reliably improving held-out bits per byte or grammatical evaluation.

Architectures and adaptive training. Relational BabyLM separated object-level and relational information using a Dual Attention Transformer and investigated an auxiliary next-latent prediction objective (Brasoveanu et al., 2026). The architectural change contributed more strongly to grammatical generalization than the auxiliary objective. Another work, ARMOR-DP, adjusted language sampling using per-language validation loss as the weakest language changed during training. Its dynamic sampling achieved a better trade-off between overall performance and balance across languages than a fixed sampling strategy that favored one language (Liao et al., 2026). The Arabic studies of Haider and Can (2026) and Badawy et al. (2026) also show how BabyLM’s multilingual extension supports research beyond the languages included in the main Challenge.

5.2 Creative Approaches

Cross-linguistic learning. Theodoropoulos et al. (2026) examined cross-linguistic effects in bilingual models trained on phonemic input. Ren (2026) tested transfer from a previously learned language, finding no uniform advantage for typologically closer language pairs. Mita et al. (2026) introduced typological flexibility into symbolic pretraining, retaining average grammatical gains while avoiding some construction-specific degradations associated with fixed structural priors.

Tokenization and representation. Li et al. (2026) found that reweighting BPE merges using association measures changed segmentation without reliably improving held-out bits per byte or grammatical evaluation. ? found that prosodically annotated phonemic input produced the strongest taxonomic organization among the input representations they tested. Finegan-Dollak (2026) found that representation standardization improved lexical similarity for models trained on large corpora but reduced performance for models trained at cognitively plausible data scales.

Data composition. Perdijk et al. (2026) introduced StoriesFTW, a synthetic Dutch children’s storytelling corpus evaluated using human and model judgments. Vaira et al. (2026) compared BabyLM and encyclopedia corpora at matched word budgets, finding that BabyLM data favored linguistic acceptability while encyclopedia text often favored entity tracking.

Training mechanisms. Related Challenge work by Hwang et al. (2026) examined how language exposure distribution and order shaped multilingual learning. Mixed multilingual training outperformed a curriculum beginning with monolingual exposure in all 15 matched configurations, although each configuration was evaluated with a single run. Veitsman et al. (2026) investigated pretraining on symbolic sequences, including music, probabilistic grammars, and cellular automata. Several sources improved subsequent language-modeling loss, but these improvements did not consistently translate into better downstream linguistic performance. Maheshwari (2026) found that reducing the number of causal-language-modeling updates mitigated late-training declines in grammatical evaluation for RWKV-7 models. Because update count and the learning-rate trajectory remained coupled, the experiments did not isolate the mechanism responsible for this benefit. Related Challenge work also examined practical-life-skills post-training through preference-data selection (Štěpánková et al., 2026).

Architectures and resource efficiency. Workshop papers compared causal, hybrid, and simulated ternary Transformers (Srivastav, 2026), and studied how token mixers and generative objectives interact under restricted data budgets (Laitenberger et al., 2026). Recursive approaches included shared-weight computation for non-autoregressive generation (Oytaç, 2026) and additional tied depth in hybrid recurrent–attention models (Sasidhar, 2026). Related Challenge work investigated whether looped computation could substitute for additional parameters (Fan et al., 2026). HydraLLM varied active model width through nested subnetworks that shared parameters (Kemme et al., 2026). Parameter Golf analyzed techniques from a separate community competition constrained by model-artifact size and training time (Paudel and Sheshappanavar, 2026). Taylor (2026) examined copy mechanisms for data-efficient language modeling. These studies addressed distinct constraints on data, computation, and model size, requiring comparisons that account for their different training and evaluation regimes.

Learning signals and context. Ploner and Akbik (2026) randomized token prediction order to expose models to different conditioning contexts during training. Öztürk et al. (2026) added sentence-

level learning signals, but found that a learned next-sentence representation had limited influence on generation in their controlled tests. Kim et al. (2026) likewise distinguished access to context from its effective use: their small translation models could become strongly dependent on prompt examples without gaining a consistent additional benefit from relevant retrieval. Wasserman and Beauchemin (2026) showed that single-token scoring for a French-only BabyLM was sensitive to tokenizer and answer-template choices, motivating native-language evaluation and controls for scoring artifacts.

Cognitively inspired learning and evaluation. Huda (2026) identified differences in noun frequencies between the training corpora and recorded infant input, examining the developmental plausibility of the training distribution. Ding et al. (2026) probed acquisition-order information in model representations, finding that strong Challenge performance did not necessarily correspond to better recovery of children’s word acquisition order.

Other studies tested specific learning behaviors that broad benchmark averages can obscure. Cacioli (2026) observed repetition priming without the emergence of mutual exclusivity in the tested text-only models. Bylinina (2026) used visually grounded embedding initialization to test ostensive word learning, finding improvements in property knowledge that were not reflected in most grammatical benchmarks. Zhang (2026) showed that high aggregate accuracy on Japanese past-tense inflection concealed substantial errors in a rare irregular verb class.

6 Conclusion

We were pleased to see an increase in the overall number of BabyLM participants for this cycle, indicating a growing interest in cognitive modeling and data-efficient pretraining. We decided to include a multilingual track after consulting the BabyLM community, which made it clear that multilingualism was a top priority. We anticipate including the multilingual track in future versions of the challenge, and hope to grow the number of languages offered. In a similar vein, discussions with the community have highlighted model evaluation—particularly generation—as an additional area of interest. We anticipate addressing this in a future iteration of the challenge.

References

- David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. **SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.** In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian’s, Malta. Association for Computational Linguistics.
- Mohamed Badawy, Marion Di Marco, and Alexander Fraser. 2026. Bpe vs. morphology-aware tokenization: A case study on arabic language modeling at babyLM scale. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Martin Benedikt, Ivan Dostál, and Robin Klubarski. 2026. Does a data2vec auxiliary help small masked language models? a babyLM-2026 strict-small study. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Jeremias Bohn, Sebastian Feiertag, Nicolas Raube, Emil Schwenger, and Georg Groh. 2026. Domain-based routing in mixture-of-experts models for babyLMs. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Adrian Brasoveanu, Ece Takmaz, and Jakub Dotlacil. 2026. Relational attention for data-efficient language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Lisa Bylinina. 2026. Augustinian babyLM: What ostensive definition can and cannot teach a small language model. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Jon-Paul Cacioli. 2026. Repetition without exclusivity: Scale sensitivity of referential mechanisms in child-scale language models. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Joseph Yew Choe Chan. 2026. Can higher order attention help small language models? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akriti Kuri, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Carolina Condez, Anamaria-Roberta Hartl, Andiswa

Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anjali Chaudhary, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Anupam Nayak, Ashvanth S, Ayodeji Lana, Ayu Purwarianti, Bashar Alhafni, Benedict Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska Đurić, Blessing Ogundipe, Bolaotan Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Çinar, Cade Kane, Can Udomcharoenchaikit, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Nwokolo, Christopher Homan, Clément Sampebo, Cristina España-Bonet, Cynthia Amol, Daeyoep Lee, Dan Saattrup Smart, Dana Arad, Daniil Dzenhaliou, Dasol Choi, David Liu, David Semedo, David Anugraha, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dennis Owusu, Dhyuthy Krishna Kumar, Diogo Tavares, Diogo Glória-Silva, Divyanshu Goyal, DongGeon Lee, E. Kelly Buchanan, Elebe Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elias Herranen, Eliza Acharya, Eman Nisar, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Itiola, Faran Taimoor Butt, Fareeha Fayyaz Sheikh, Fathima Thekkekkara, Fatima Haouari, Faustin Nsengiyumva, Fenal Ashokbhai Ilasariya, Filbert Aurelian Tjjaranata, Firas Laakom, Francesca Grasso, Francesco Periti, Francesco Orabona, Gbenga Kayode Solomon, Genta Indra Winata, Gia Nghia Ngo, Gloria Udhedhe-oze, Gonçalves Vinagre, Gopi Naga Sai Ram Challagolla, Gorka Urbizu-Garmendia, Gouthami Vadithya, Guijin Son, Gulnaz Abdykadyrova, Gyan Swaroop Mohapatra, Hafeez Ullah, Hasteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Hitesh Laxmichand Patel, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, İbrahim Ethem Deveci, Ikhlasul Akmal Hanif, Ikhyun Cho, Inês Vieira, Inês Calvo, Isaac Manzi, Ismael Illa Salifou, Ismail Daud, Ismail Yusuf, Itay Itzhak, Ivan Zhelyazkov, Ivan Belashkin, Ivan Spada, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Kocoń, Jan Cuhel, Jauza Krito, Jebish Purbey, Jennifer Za, Jennifer Mickel, Jenny Kunz, Jessica Ratovondranto, Jeyarajalingam Varsha, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Seon Keun Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Julia Alekseenko, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Tshinu, Kättriin Kukk, Kaustubh Ponkshe, Kavsar Huseynova, Ke He, Kenneth Enevoldsen, Kent Joshua Alvarez, Kerem Zaman, Khalil Mrini, Kian Kyars, Komal Gour, Krishnakumar Lainitha, Krister Kruusmaa, Kunal Mukherjee, Kusum Chouhan, Laura Castro, Laura M. Porrino-Moscato, Lenny Sivi Za Nzambi, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Loretta Oma Jones, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik, Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos Garcia,

Marek Šuppa, Mariam D'Ciofalo, Marii Ojastu, Marium Attaullah, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu, Menan Velayuthan, Mhasilenuo Vizo, Michael Leventhal, Michał Marcińczuk, Mina Almasi, Mirna Potočnjak, Mithil Bangera, Mohammadamin Shafiei, Mohiba Ansari, Mridul Sharma, Mrityunjaya Indoria, Mughees Ur Rehman, Muhammad Ravi Shulthan Habibi, Murat Kolić, Murat Barkın Kinay, Nada Galant, Naina Singh Rathore, Naphat Permpredanun, Narada Maugin, Nathalie Norman, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Odunayo Ogundepo, Odunayo Kareemat Buliaminu, Oghojafor Godswill Fejiro, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun, Olasoji Akindejoye, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kinay, Osman Tursun, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Rodríguez, Pablo Gamallo, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Prakhara Agarwal, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Rahul Raja, Ram Mohan Rao Kadiyala, Raphael Merx, Rasmus Larsen, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zavar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, Sai Sandeep Kantareddy, Saleha Muzammil, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Sarveswaran Kengatharaiyer, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, Segun Afolayan, Senthilnathan M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibeji, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A, Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti More, Spyros Dragazis, Stefan Milosavljević, Stephan P. Kaufhold, Suba S, Sultan Alrashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Idoko, Thu Mai Do, Tilek Chubakov, Tina Munda, Tobiloba Owoe, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varvara Arzt, Vasily Konovalov, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoriya Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yash Bagla, Yeshil Bangera, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Zaid Alyafei, Zhargal Batozargalova, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and

- David Ifeoluwa Adelani. 2026. [Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures](#).
- Lucas Georges Gabriel Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, et al. 2025. Findings of the third babyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: why not both?](#) In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- Cristiano Chesi, Achille Fusco, Guido Formichi, Sarah Rossi, Hanna B. Cygan, Tommaso Sgrizzi, and Asya Zanollo. 2026. Exploring linguistic parameters in lms. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Po-Han Chiang. 2026. Urchin: A horizontal spiking language model for data-constrained pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. [BabyLM turns 4 and goes multilingual: Call for papers for the 2026 babyLM workshop](#).
- Chenyi Ding, Zachary Nicholas Houghton, and Masoud Jasbi. 2026. Reading acquisition order out of babyLM representations: Child-scale training produces developmentally aligned semantic spaces. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Lukas Edman and Alexander Fraser. 2026. Qwenfant: Train big character-level babyLMs on concise parallel text. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Tingshuo Fan, Hongtao Mu, Tianyu Zhou, Hansen Liu, and Tao Ji. 2026. Looped gpt-bert: Trading parameters for computation in small language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Catherine Finegan-Dollak. 2026. What helped with huge data hurt with small data: Standardization, representational geometry, and data scale. In *Proceedings of the Second BabyLM Workshop*, pages 379–393. Association for Computational Linguistics.
- Sam Furr. 2026. Four kilobytes of memory: A 10m-parameter fixed-state recurrent lm for babyLM strict-small. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Serdar Gülbahar, Alexander Fraser, and Lukas Edman. 2026. Recursive transformers for babyLM 2026. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Mozafar Haider and Burcu Can. 2026. Arabic babyLM: Effects of diglossia and tokenisation. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Patrick Haller, Filipe Laitenberger, Jonas Golde, and Alan Akbik. 2026. From puzzle-solver to language model: Recurrent inductive bias, data scale, and benchmark validity in babyLM. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Daichi Hotta, Yohei Kobashi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. Loop before you leap: What recurrent depth can and cannot buy in data-limited pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. [Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.
- Zhuojing Huang, Luise Pohlmann, and Lisa Beinborn. 2026. Introducing code-switched contexts to cognitively-inspired bilingual model training. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Imrul Huda. 2026. How far is the training corpus from real infant input? auditing babyLM against day-long home recordings, and what an age-of-acquisition curriculum does not buy. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- JiWoo Hwang, Lukas Edman, and Alexander Fraser. 2026. Language exposure matters: Distribution and order in multilingual babyLM pre-training. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Sadri Islam, Tagore Rao Kosireddy, Kevin Li, and Dylan Gaines. 2026. Learning from correction: Data-efficient language model training through adaptive teacher-model feedback. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Soham Jain, Harsh Singh, Divija Dewan, and Atul Dev. 2026. Xpertgpt: Multi-scale sparse expert routing for data-constrained language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.

- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrana, Francesca Padovani, Francois Meyer, Hai Hu, Julien Etxaniz, Laurent Prevot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, Suchir Salhan, Susana Zhou, Yurii Paniv, Ziyin Zhang, Arianna Bisazza, Alex Warstadt, and Leshem Choshen. 2026a. [BabyBabelLM: A multilingual benchmark of developmentally plausible training data](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3297–3329, Rabat, Morocco. Association for Computational Linguistics.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026b. [MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Jeremy Kalfus and Sharvin Goyal. 2026. Babblerm: Targeting closed-system self-generation at a 10m-word model’s own reducible-loss spans. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Wiktór Kamzela, Ondrej Dusek, and Mateusz Lango. 2026. Scaling compute instead of parameters: Exploring iterative transformers for data-efficient language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Oscar Kemme, Janek Haberer, Ali Hojjat, and Olaf Landsiedel. 2026. Hydrallm: Adaptive width slicing for elastic text and vision-language models. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Heesu Kim, Shota Sato, and Mamoru Komachi. 2026. Retrieval without relevance: Prompt dependence in small machine translation models. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Filipe Laitenberger, Patrick Haller, Jonas Golde, and Alan Akbik. 2026. What survives the babylm budget? a controlled study of token mixers and generative objectives. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Zy Li, Phu Han, Trinh Tran, Lam Duong, and Andrew Fang. 2026. Do association-reweighted bpe merge rules help data-efficient language modeling? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Hanyu Liao, Yijing Chen, and Tao Ji. 2026. When the weakest language keeps moving: Armor-dp for multilingual babylm. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2026. [A systematic assessment of language models with linguistic minimal pairs in Chinese](#). *Transactions of the Association for Computational Linguistics*, 14:755–771.
- Yash Maheshwari. 2026. Halved clm exposure mitigates late-training degradation in small recurrent language models. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Francois Meyer. 2026. Subword segmental babylms: Learning to tokenise for sample-efficient pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Aftab Hossain Mikat and Tamjid Hasan Fahim. 2026. Dense supervision, gentle decay: Curriculum-guided rebalancing of electra-style rtd and mlm for small-data pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Sebastian Miles. 2026. Weight-tied looped depth for data-constrained pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Ujjwal Mishra and Sudipta Roy. 2026. Spend the wasted epochs: Preference post-training from the saturation checkpoint for sample-efficient language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Masato Mita, Max Müller-Eberstein, Misaki Yamaguchi, and Yohei Oseki. 2026. Linguistically inspired multilingual pre-pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Naoki Mori, Seiji Kyo, and Goda kohki. 2026. Language allocation and targeted data design for a multilingual babylm. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Rafael Mosquera-Gómez, Juan Felipe Rodriguez, Mateo Restrepo, Rodrigo Meireles de Oliveira, and Andrés Felipe González Giraldo. 2026. Parameter-efficient architectures for developmentally plausible pretraining: Optimizing embedding and projection layers under babylm strict-small constraints. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Noor Eidirhn Mustafa, Valdemar Melkersson, and Pierre Nugues. 2026. Moep+: Enhancing modular expert paths for data-efficient language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. [Universal Dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

- Doğa Oytaç. 2026. Trlm: Shared-weight recursive reasoning for non-autoregressive language generation. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Ahmet Rifat Öztürk, Yağız Ekrem Dalar, Nedim Mutlu Sezer, Feyzi Arda Salihoğlu, and Ömer Faruk Aksoy. 2026. Sentence-level foresight in small language models: A diagnostic study of learning signals and calibration. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Alonso Palomino. 2026. Exploring auxiliary objectives through multi-task learning for small-data language model pretraining: A babylm 2026 submission. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Prashanna Mani Paudel and Shivanand Venkanna Sheshpanavar. 2026. Parameter golf: What really works? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Sabijn Perdijk, Gijs Wijnholds, Suzan Verberne, and Max J. van Duijn. 2026. Stories for the win: Creating and evaluating a synthetic Dutch children’s storytelling dataset. In *Proceedings of the Second BabyLM Workshop*, pages 450–469. Association for Computational Linguistics.
- Max Ploner and Alan Akbik. 2026. Can randomly permuting the lm decoding order boost data-efficiency? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Venkata Naga Sai Vishnu Rohit Pulipaka and Libby Barak. 2026. Can translation bridge the data gap for low-resource languages? a case study of hindi and telugu. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).
- Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. [Multilingual probing of deep pre-trained contextual encoders](#). In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.
- Pengfei Ren. 2026. Does typological distance affect cross-lingual transfer? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Elizabeth Granda Rodriguez, David Escobar Ruiz, Adrián Rojas Fernandez, and Andres Felipe Gomez Rojas. 2026. Growing smarter, not bigger: Curriculum pretraining and layer duplication in small language models. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Juan Diego Rodriguez. 2026. Tokens of meaning: Semantically-informed morphological tokenization in low-resource language modeling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2024. [Include: Evaluating multilingual language understanding with regional knowledge](#).
- Dries Rooryck, Alexander Dazhen Cai, Yonatan Belinkov, David Alvarez-Melis, and Kianté Brantley. 2026. Feeding babylms macaroni: Code-switching curricula cause cross-lingual convergence. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Suchir Salhan, Hongyi gu, Donya Rooein, Diana Galvan-Sosa, Gabrielle Gaudeau, Andrew Caines, Zheng Yuan, and Paula Buttery. 2025. [Teacher Demonstrations in a BabyLM’s Zone of Proximal Development for Contingent Multi-Turn Interaction](#). In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Aditya Sasidhar. 2026. When does depth recursion pay? a parameter-matched study of weight-tied hybrid transformers on 100m unique words. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Siddhi Satheesh, Jorge C. Murillo, and Yubei Chen. 2026. Babysteps at babylm 2026: Stronger optimization and causal baselines for data-constrained pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Noam Siegelman, Sascha Schroeder, Cengiz Acarturk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo,

- and Victor Kuperman. 2022. [Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus \(meco\)](#). *Behavior Research Methods*, 54:2843–2863.
- Siyan Song, Zhiheng Qian, Yunhao Zhang, Linyang He, Xiaozhe Ji, Yingxin Lin, Hongao Zhu, Chongtian Shao, Chuhan Lang, Luan Li, et al. 2026. The first chinesebabyLM challenge: training data-efficient and cognitively plausible language models for chinese. *arXiv preprint arXiv:2607.10745*.
- Arnav Srivastav. 2026. Causal, hybrid, and ternary transformers under a restricted data budget. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, 51(4):1267–1301.
- Josh Taylor. 2026. Hard-wiring the induction head: The surprising effectiveness of copy mechanisms in data-efficient language modelling. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Nikitas Theodoropoulos, Maria Lymperaïou, and Giorgos Filandrianos. 2026. Cross-linguistic effects in bilingual phoneme babyLMs. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Filip Trhlík, Andrew Caines, and Paula Buttery. 2026. [Bias dynamics in BabyLMs: Towards a compute-efficient sandbox for democratising pre-training debiasing](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33356–33377, San Diego, California, United States. Association for Computational Linguistics.
- Edoardo Vaira, Georgios Chavales, and Pierre Nugues. 2026. Different corpora, different skills? a paired study of small language models. In *Proceedings of the Second BabyLM Workshop*, pages 404–416. Association for Computational Linguistics.
- Dheeraj Varghese. 2026. Tinylibrary: Do age-graded curricula help small vision-language models? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Yana Veitsman, Jonas Mayer Martins, Jonathan Lautenschlager, and Lisa Beinborn. 2026. Structural priors for data-efficient language learning. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Yupei Wang. 2026. What does input-dependent depth routing buy in data-limited language modeling? In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. [Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*.
- Adam Zachary Wasserman and David Beauchemin. 2026. Right tool, right job: Native-language evaluation, tokenizer sensitivity, and methodological findings from a French-only BabyLM. In *Proceedings of the Second BabyLM Workshop*, pages 360–378. Association for Computational Linguistics.
- Zichao Wei. 2026. Olmo2 with hybrid mlm-clm training for the babyLM 2026 multilingual challenge. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Shaoxiang Wu, Alexander Fraser, and Lukas Edman. 2026. Pairing original text with meaning-preserving rewrites for sample-efficient small-data language model pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Fuming Yang, KAI-HAO YANG, Hsuan Lien, Lukas Edman, and Alexander Fraser. 2026a. Optimizing multilingual babyLM: The roles of tokenization, data packing, and curriculum learning. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Yuwei Yang, Shihao Pu, Gefei Fung, and Yiming Yang. 2026b. Learning what can combine and what can substitute: Relational inductive biases for babyLM pretraining. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- yizhe xu, Ming Song, DANNI HE, Jianghao Liu, Yitong Wang, and Qing Cai. 2026. BabyconceptLM: Exploring dynamic concept computation under restricted language exposure. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Wen Zhang. 2026. Not all irregularity is equal: Causally isolating a rare failure mode in japanese morphological inflection. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.
- Barbora Štěpánková, Vojtěch Lanz, and Jakub Szymanik. 2026. The impact of preference optimization data selection: Teaching practical life skills to babyLM. In *Proceedings of the 2nd BabyLM Workshop at EMNLP 2026*, Budapest, Hungary.