Findings of the Third BabyLM Challenge: Accelerating Language Modeling Research with Cognitively Plausible Data

BabyLM Team

Lucas Charpentier[1], **Leshem Choshen**[2,3], **Ryan Cotterell**[4], **Mustafa Omer Gul**[5], **Michael Y. Hu**[6], **Jing Liu**[7], **Jaap Jumelet**[8], **Tal Linzen**[6], **Aaron Mueller**[9], **Candace Ross**[10], **Raj Sanjay Shah**[11], **Alex Warstadt**[12], **Ethan Gotlieb Wilcox**[13], **Adina Williams**[10]

[1]LTG, University of Oslo [2]IBM Research [3]MIT [4]ETH Zürich [5]Cornell University [6]NYU [7]ENS-PSL [8]University of Groningen [9]Boston University [10]Meta AI [11]Georgia Tech [12]UC San Diego [13]Georgetown University

Abstract

This report summarizes the findings from the 3rd 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 an expanded BabyLM Workshop that invited paper submissions on topics relevant to the BabyLM effort, including sample-efficient pretraining and cognitive modeling for LMs. For the challenge, we kept the text-only and text-image tracks from previous years, but also introduced a new *interaction* track, where student models are allowed to learn from feedback from larger teacher models. Furthermore, we introduce a new set of evaluation tasks to assess the "human likeness" of models on a cognitive and linguistic level, limit the total amount of training compute allowed, and measure performance on intermediate checkpoints. We observe that new training objectives and architectures tend to produce the best-performing approaches, and that interaction with teacher models can yield high-quality language models. The strict-small and interaction tracks saw submissions that outperformed the baselines. We do not observe a complete correlation between training FLOPs and performance. This year's BabyLM Challenge shows that there is still room to innovate in a data-constrained setting, and that community-driven research can yield actionable insights for language modeling.

1 Introduction

Language modeling (LM) has become increasingly compute-intensive in the past decade, and is thus often cast as the preserve of tech giants. LM research is also often dismissed as irrelevant to the study of language and mind, 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 Large Language Models (LLMs) with the amount of language typical of human language acquisition. This paper presents and analyzes the main findings from the third iteration of the BabyLM Challenge.[1] We also present the winning submissions and some key takeaways from the BabyLM workshop, to which participants could submit papers without needing to submit 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 LLMs' sample efficiency and reveal why standard LLMs are so data hungry. Previous BabyLM iterations (Hu et al., 2024) included a multi-modal text-image track. While we keep this track, this year, we also introduced an interactive track. The interactive track enables training on direct interaction with a teacher model via the student model's generated outputs, rather than passive exposure to human-generated texts. Inspired by interactions

[1]For findings from previous years, see Warstadt et al. (2023); Hu et al. (2024). For a write-up focused on implications for psycholinguists, see Wilcox et al. (2025).

Proceedings of the First BabyLM Workshop, pages 399–420

during human language acquisition, we hoped to encourage researchers to investigate the benefits of adapting the text seen to the model's needs (see §3.3).

Summary of takeaways. As in the previous two iterations of the BabyLM Challenge, curriculum learning was a common approach. However, the most effective approaches were those that proposed architectural innovations or modifications to the training objective or procedure. Winners included a diffusion language model (Kosmopoulou et al., 2025), a mixture-of-experts model (Tapaninaho, 2025), and a reinforcement learning-based interactive approach (Martins et al., 2025).

2 Competition Details

Track Overview. The third BabyLM Challenge included four competition tracks: the returning *Strict*, *Strict-Small*, and *Multimodal* tracks and the newly added *Interaction* track.

The *Strict* and *Strict-Small* tracks require submissions to be trained on datasets of 100M and 10M words or less, respectively. Participants were free to use the provided BabyLM corpus or construct their own training datasets, provided that they adhered to the track's word limitations. Models in this track were evaluated on language-only evaluation tasks.

In the *Multimodal* track, participants trained multimodal vision-language models. Participants were allowed to use any model and training procedure, provided that the model could assign (pseudo) log-likelihoods to strings of text, conditioned on input images. Submissions could be trained on any arbitrary dataset of 100M words or less, including our provided corpus for the *Multimodal* track, which is split evenly between text-only and paired image-text data. Models in this track were evaluated on both language-only and multimodal tasks.

New to this year, the *Interaction* track enabled participants to explore how feedback and interaction could assist with sample-efficient modeling. Here, an external model different from the participants' submission model could be incorporated into the training pipeline. Participants were prohibited from exposing the external model's weights, hidden states, or output distribution to the submission model, but were otherwise unrestricted in how they instantiated "interactions." The external model could, for instance, give scalar or natural language feedback to the submission model or produce training data conditioned on the submission model's outputs. Similar to previous tracks, the submission model could be exposed to at most 100M external words, which could come either from regular datasets or the external model. Furthermore, the submission model could not generate more than 100M words of its own. Finally, we restricted the external model to a pre-determined list of models (namely Llama3.1-8B-Instruct, Llama3.2-3B-Instruct, Llama3.1-1B-Instruct (Dubey et al., 2024), and any language model below 1B parameters). Participants were allowed to fine-tune these models without any restriction. Models in this track were evaluated on language-only evaluation tasks.

The data composition of the corpora for each competition track is described in full in Table 1.

Training Duration Limitations. This year, we restricted models to a fixed amount of training data exposure, counting repeated passes over the same input, specifically at most 100M words for the *Strict-Small* track and at most 1B words for the other tracks. This decision was motivated by two goals of BabyLM. Firstly, BabyLM aims towards developmentally plausible training. While memories of inputs could have an impact on learning beyond the initial exposure, dozens or hundreds of repeated exposures are developmentally implausible. Secondly, BabyLM aims towards democratizing pretraining research. We observed in the 2024 BabyLM Challenge that larger numbers of training epochs improved model performance, which gives groups with greater computational resources a significant advantage if no limitations exist. Although the new limitation does not eliminate all advantages of greater compute, such as for hyperparameter tuning, it helps ensure that successful training procedures are more reproducible and accessible to teams with modest resources.

Intermediate Checkpoints. We additionally required participants to submit intermediate model checkpoints corresponding to different word exposure amounts. We specifically ask for checkpoints for every 1M words until 10M words are seen, every 10M words until 100M words are seen, and every 100M words until 1B words are seen. Each checkpoint would then be evaluated on a subset of less compute-intensive tasks. The motivation behind this is that the training dynamics of LMs can be compared to the learning trajectories of children, which is valuable from the cognitive modeling per-

spective. Results from this analysis are presented in Figure 5.

3 Baselines

In this section, we detail the baselines and their associated training procedures for each competition track. When possible, we set winning entries from the past competition year as the baseline for a given track. Each baseline is meant to encourage participants to innovate and improve beyond existing models and approaches.

3.1 Strict and Strict-Small Tracks

For the *Strict* and *Strict-Small* tracks, we used last year's winning submission, GPT-BERT (Charpentier and Samuel, 2024), and the GPT-2 Small (Radford et al., 2019) architecture naively trained with an auto-regressive language modeling loss as baselines. GPT-BERT is based on the architecture of LTG-BERT (Samuel et al., 2023), a BERT-style model developed to work with low amounts of data. It uses disentangled attention from DeBERTa (He et al., 2021), both pre- and post-layer normalization as in NormFormer (Shleifer et al., 2021), span masking, and GEGLU activation functions in the feed-forward layers. In addition to using LTG-BERT as a base, GPT-BERT uses both the masked and auto-regressive language modeling objectives to train the models. To achieve this, the authors used a variation of standard masked language modeling called masked next token prediction, where the outputs are shifted in the same way as in the auto-regressive training. By training with both objectives, the models can be used both as an encoder and a decoder.

As GPT-BERT is trained with both masked and autoregressive language modeling losses, we train three variants for it: one focused on the autoregressive loss, another focused on the masked loss, and finally, another with equal focus on both losses. Baselines for each track were trained using the corresponding BabyLM corpus.

GPT-BERT. In line with the challenge requirements, we train the *Strict* and *Strict-Small* models for 10 epochs. Our *Strict* models have around 120M parameters with 12 layers and 12 attention heads. We use a batch size of 131 072 tokens and train for 12 330 steps. Our *Strict-Small* models have around 31M parameters with 12 layers and 6 attention heads. We use a batch size of 16 384 tokens and train for 9 914 steps. For both tracks, we use a warmup-cosine-cooldown learning rate scheduler with a maximum learning rate of 7×10^{-3}. The first 1.6% of steps are used for linear warmup, and the final 1.6% of steps are used for linear cooldown. For the masked objective, we start the masking ratio at 0.3 and linearly decay it to 0.15. We use a sequence length of 128 tokens for the first 60% of training steps, we then increase the sequence length to 256 tokens for the next 20%, and for the final 20% we use a sequence length of 512.

We train three variants of GPT-BERT. The auto-regressive focus uses a 93.75-6.25 mix of auto-regressive to masked ratio. The mixed focus uses a balanced 50-50 mix of auto-regressive to masked ratio. Finally, the masked focus uses a 6.25-93.75 mix of auto-regressive to masked ratio. All three models in each track are evaluated both in the masked next-token prediction (MNTP) and auto-regressive styles. A complete list of hyperparameters can be found in the HuggingFace Model Hub; the HuggingFace names of the models can be found in Appendix B.

GPT-2. We additionally train the GPT-2 Small (Radford et al., 2019) with a purely auto-regressive loss as a naive baseline. We first chunk the BabyLM corpus into datapoints of 512 tokens each. The model is trained for 10 epochs with a batch size of 16 (containing 8192 tokens per step). We use a learning rate of 5×10^{-5} with a cosine-decay scheduler that warms up the learning rate in the initial 1% of training. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer.

3.2 Multimodal Track

As no submissions outperformed our *Multimodal* baselines in the 2024 BabyLM Challenge, we re-released them for this year. We train the GIT (Wang et al., 2022) and Flamingo (Alayrac et al., 2022) architectures on the BabyLM corpus for the *Multimodal* track, and use a frozen DINO model with the ViT-B/16 architecture as the image encoder (Caron et al., 2021).

We perform training on the BabyLM corpus for the *Multimodal* track. We train the models for 4 epochs, where each epoch consists of one pass over the text-only half of the corpus and four passes over the remaining image-text paired data (resulting in 250M word exposures per epoch). We use a learning rate of 10^{-4}, with a linear learning

rate scheduler, and train with the AdamW optimizer (Loshchilov and Hutter, 2019).

3.3 Interaction Track

For the *Interaction* track, we provide a baseline that explores how corrections in natural language can be incorporated into language model training. We split training into 20 rounds of interaction. At each round, the student model, initialized with the GPT-2 Small architecture (Radford et al., 2019), is given incomplete data points sampled from the BabyLM training corpus. For each data point, the student samples a completion. The teacher model, chosen to be Llama-3.1-8B-Instruct (Dubey et al., 2024), is then prompted to revise the student model's completion based on grammaticality, coherence, and relevance to the input. The student model is then first trained with the language modeling loss on the full teacher-corrected datapoint and is then further finetuned with SimPO (Meng et al., 2024), a preference optimization algorithm, where the teacher and student completions are the winning and losing responses, respectively.

We split each constituent dataset of the BabyLM corpus into 20 equally sized chunks prior to training. At each round, a chunk is sampled at random from each constituent dataset without replacement. Each chunk is then split into data points consisting of 512 tokens. The student is provided the first 256 tokens of each data point as context for generation. We then sample student completions with nucleus sampling (Holtzman et al., 2020) where $p = 0.8$. Teacher corrections are similarly sampled using nucleus sampling with $p = 0.8$. The prompt can be found in the Appendix.

We optimize the student model with AdamW (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-5} and set $\beta = 2$ and $\gamma = 1$ for SimPO. We add the language modeling loss on the winning completion, with a scaling coefficient of 0.2, as a regularizer during preference optimization training, following Dubey et al. (2024). For each round of interaction, we perform 7 epochs of training with the regular language modeling loss on full teacher-corrected datapoints, followed by 2 epochs with SimPO.

4 Evaluation

For evaluation, we kept the tasks from previous year's edition. For the *Strict* and *Strict-Small* these are the (Super)GLUE suite of NLP tasks (Wang et al., 2018, 2019), the linguistic minimal pairs of BLiMP (Warstadt et al., 2020), and the Elements of World Knowledge (EWoK) dataset (Ivanova et al., 2024), which measures pragmatic, commonsense, and discourse knowledge. For evaluation of the *Multimodal* track, we test again on Visual Question Answering (VQA, Agrawal et al., 2015; Goyal et al., 2017), WinoGround (Thrush et al., 2022), and DevBench (Tan et al., 2024).

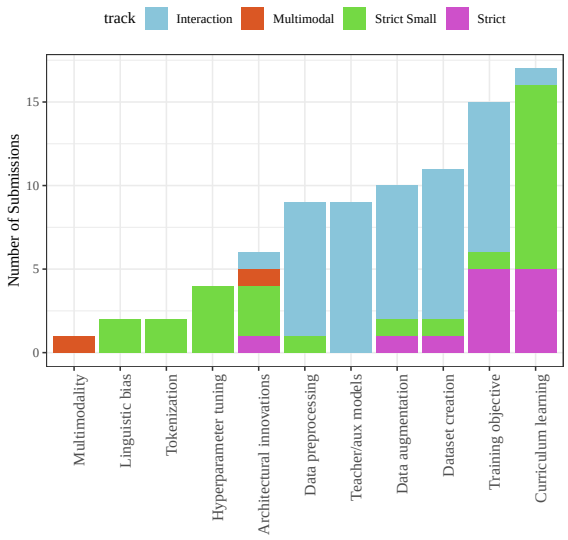

Figure 1: **Number of submissions by approach**. Curriculum learning was again the most popular approach. This year, we encouraged more teacher/auxiliary-model approaches in the interaction track.

4.1 New Tasks

This year, we additionally included tasks that measure *psychometric fit* to human language learners and linguistic abilities of aspects not covered by BLiMP. For selecting these tasks, we focused on the following two aspects of a model being 'human-like': i) connecting model behavior and internals to cognitive aspects of human language processing, such as reading time prediction, and ii) assessing how human-like a model's generalizations are on various tasks related to reasoning and morphology. We excluded tasks that could not be reasonably acquired from the BabyLM training data. Below, we describe the tasks in more detail.

Morphological Generalization Weissweiler et al. (2023) introduce a task for testing morphological generalization, based on a past tense formation task of nonce ("*wug*") words: e.g. *veed* $\rightarrow$ *ved*/*veeded*/*vode*. Similar to this task, we also include the task of Hofmann et al. (2025), in which nonce adjectives are *nominalized* as

Dataset	Description	# Words (multimodal)	# Words (strict)	# Images
Localized Narratives[a]	Image Caption	27M	–	0.6M
Conceptual Captions 3M[b]	Image Caption	23M	–	2.3M
CHILDES[c]	Child-directed speech	14.5M	29M	–
British National Corpus (BNC), dialogue portion[d]	Dialogue	4M	8M	–
Project Gutenberg (children's stories)[e]	Written English	13M	26M	–
OpenSubtitles[f]	Movie subtitles	10M	20M	–
Simple English Wikipedia[g]	Written Simple English	7.5M	15M	–
Switchboard Dialog Act Corpus[h]	Dialogue	0.5M	1M	–
Total	–	100M	100M	2.9M

Table 1: Datasets for the *Multimodal* and *Strict* tracks of the 3rd BabyLM Challenge. Word counts are approximate and subject to slight changes. [a]Pont-Tuset et al. (2020) [b]Sharma et al. (2018) [c]MacWhinney (2000) [d]Consortium (2007) [e]Gerlach and Font-Clos (2020) [f]Lison and Tiedemann (2016) [g]`https://dumps.wikimedia.org/simplewiki/` [h]Stolcke et al. (2000)

either an -ity or -ness noun: e.g. *cormasive* → *cormasiveness*/*cormasivity*. We evaluate these tasks against human predictions: from the participant responses included in each of the above papers, we derive a distribution over human-preferred inflections. Our score for this task is then a correlation between the model's probability for each inflection against the human preference distribution.

Entity Tracking Kim and Schuster (2023) tests entity state tracking in LMs, by describing a sequence of actions placing and removing items to and from various numbered boxes and evaluating a model's understanding of the contents of each box at a given moment. We revised the evaluation of this task to evaluate LMs' ability to assign the highest probability to the correct continuation (akin to BLiMP and EWoK) rather than requiring the model to generate the correct completion as in the original operationalization. This was done to enable simpler, zero-shot evaluation. We construct five candidate continuations, one of which is the ground-truth. Distractor continuations were constructed by copying prior contents of a given box, contents of an adjacent box, or the result of the most recent action. They were also synthetically generated by randomly swapping, adding, and removing objects from the box state.

Concept Knowledge Misra et al. (2023) introduce a task for testing the property knowledge of language models and whether they can infer that properties of superordinate concepts are inherited by subordinate concepts, each represented by nonce words. The dataset is composed of minimal pair sentences, and models are evaluated by whether they assign a higher probability to the correct sentence.

Reading Time Prediction de Varda et al. (2023) Connects LM predictions to human reading times, allowing us to assess to what extent LM processing is aligned with human language processing. To measure this, we do a correlation between the surprisal score (defined as the negative log probability of a word) of a word for a language model and either the time it took for a human to read the word or the time spent looking at the word. The more correlated the two metrics are, the more human-like a model is, following the previously established relationship between surprisal and reading time, wherein words that take longer to read are associated with higher surprisal scores (Wilcox et al., 2020, 2023).

Word Learning Chang and Bergen (2022) present a benchmark for tracking word surprisal across training checkpoints to extract learning curves and compute ages of acquisition for vocabulary items. We compute surprisal scores as the negative log probability of target words given their contexts in the C4-en-10k test set (a shuffled subset of the first 10,000 records from the English portion of the C4 corpus) across training steps. We then fit sigmoid functions to each word's learning trajectory. In the end, the benchmark enables direct comparison between the language model and child language development by computing correlation scores between model-derived and human Age-of-Acquisition data from the WordBank repository (Frank et al., 2016).

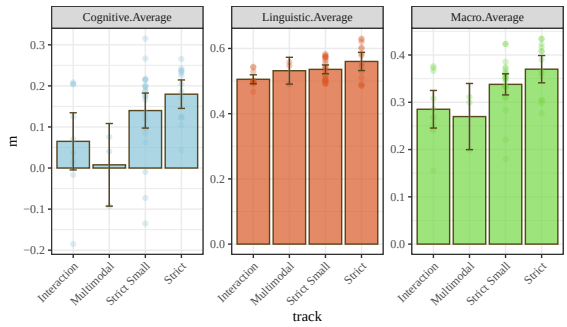

Figure 2: **Scores by track**. Despite the greater flexibility and tailored supervision allowed in the interaction track, performance was generally higher in the strict track. Multimodal models remain difficult to train, perhaps due to the track design (as discussed in Ganescu et al. (2025)).

4.2 Evaluation Pipeline

As in previous years, we distributed an open-source evaluation pipeline that could be run by all participants.[2] We rewrote the evaluation pipeline from scratch so as to make the structure of the repository significantly simpler than in previous years. This allowed participants to adapt it to their needs or unique architectures and debug any potential issues, as improving the computation efficiency. We provided a HuggingFace version that could be re-written to use only PyTorch modules.

Hidden Tasks As in previous years, we released a set of *hidden* evaluation tasks to control for overfitting to the public evaluation tasks. The hidden tasks this year were COMPS (Misra et al., 2023), the past tense formation *wug* task (Weissweiler et al., 2023), and the word learning trajectory task (Chang and Bergen, 2022). We released these tasks two weeks before the submission deadline.

Zero-shot vs. Finetuning A criticism of the evaluation procedure in previous editions was that the finetuning tasks presented a considerable computational overhead. We investigated to what extent these tasks can be evaluated using zero-shot prompting instead, but unfortunately concluded that the limited data size does not allow for robust in-context learning to emerge.[3] Therefore, we kept the existing finetuning tasks in (Super)GLUE but made the finetuning more efficient in two ways: subsampling large tasks and eliminating highly correlated tasks.

First, we sub-sampled the finetuning tasks of (Super)GLUE larger than $10,000$ training samples down to $10,000$. In our tests, we found that randomly subsampling large datasets like MNLI down to $O(1e4)$ still reliably differentiated between existing open-source models on the HuggingFace Model Hub without significantly increasing the variance due to our subsampling procedure: different subsamples of size $O(1e4)$ still gave the same stable ranking across open-source models after finetuning. Second, if models' performances on two tasks were consistently highly correlated with each other, such as with MNLI and QNLI, we eliminated one of the two tasks from our evaluations. Ultimately, we kept the following tasks from (Super)GLUE: BoolQ, MultiRC, RTE, WSC, MRPC, QQP, and MNLI. For any of these tasks larger than $10,000$ training samples, we subsampled down to $10,000$.

Next to this, we also release the evaluation tasks in two ways: a *fast* and *full* version. The *fast* evaluation consists of 20% of the data of each task (including the zero-shot tasks). This lessens the computational overhead that comes with our introduction of the evaluation of intermediate checkpoints: we only require the *full* evaluation to be run on the final model checkpoint.

5 Submission

This year, we used HuggingFace Spaces, HuggingFace Model Hub, and OpenReview for the submissions to both the workshop and challenge.

Challenge Results Submission. The participants to the challenge had to submit their results, both for the final checkpoint and intermediate checkpoints, through a leaderboard found in a HuggingFace Spaces.[4] The participants were required to submit their predictions in a JSON format; for predictions of the final model, each example consisted of an ID and a value (a text completion for non-classification tasks, and a label for classification tasks). For the intermediate checkpoints, the participants submitted the subtask scores for each checkpoint.

[2] github.com/babylm/evaluation-pipeline-2025

[3] Olsson et al. (2022) show that the *induction heads* required for in-context learning develop only after exposure to 2.5–5 billion tokens. Developing sample-efficient methods that enable such mechanisms to emerge under much smaller data budgets remains an exciting prospect for BabyLM-related research.

[4] BabyLM-community/babylm-leaderboard-2025-all-tasks

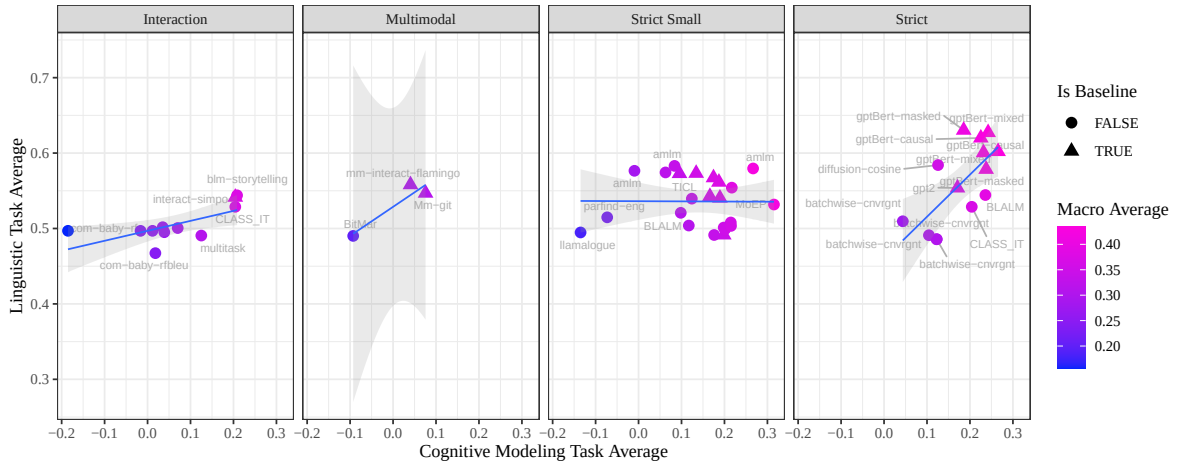

Figure 3: **Overview of the results.** We found a positive correlation between linguistic and cognitive modeling task performance, except for the *Strict-Small* track. The baselines (winning methods from previous years' challenges) remain strong, especially in the multimodal and strict tracks.

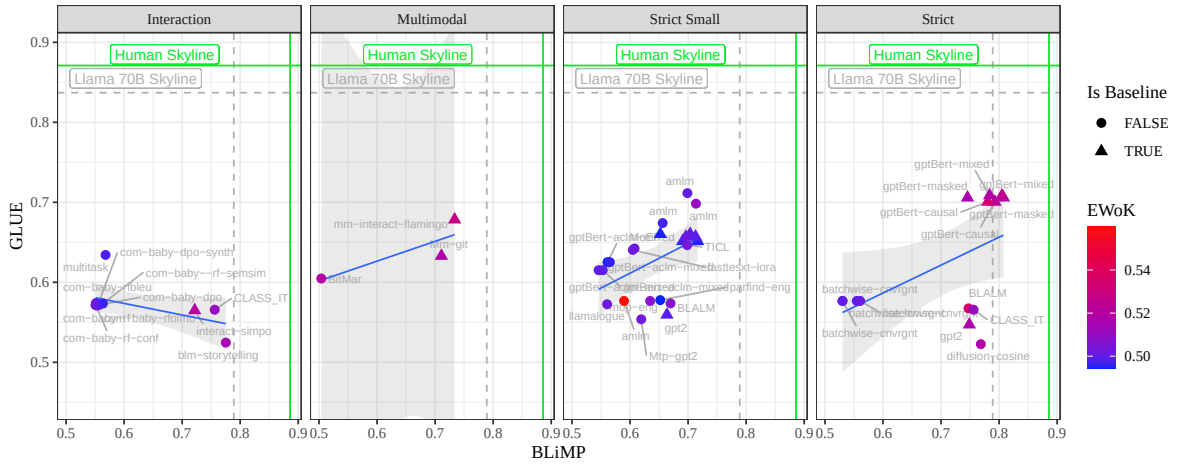

Figure 4: **Comparison of Results on BLIMP and GLUE to human scores.** Some models in the *Strict* and *Interaction* tracks are comparable to a Llama 70B parameter model on the BLiMP task. Models still fall short of skylines and human scores for GLUE.

Submission Form. In addition to submitting their results, the participants were required to fill in additional details about their training in the HuggingFace submission. These included: hyperparameters such as learning rate, scheduler, number of epochs, size of model, seed, and batch size, to name a few; information on the training dataset; number of FLOPS for both training and development; preprocessing or augmentation of data; and a short description of their model. The form can be found on the submit tab of the leaderboard.

Paper Submission. The participants were asked to submit their papers through OpenReview. Challenge participants were asked to submit papers detailing their methodology, research, and findings. Those participating in the associated BabyLM workshop were asked to submit papers thematically related to the goals of the challenge.

Artifact Submission. The participants of the challenge were also required to make their models and intermediate checkpoints available by submitting them to the HuggingFace Model Hub.

6 Competition Results

In this section, we describe the results of the competition, track winners and our selections for Outstanding Papers, which were chosen from both the challenge and workshop paper submissions. We received 32 papers to the workshop, 12 papers to the challenge, and 32 models to the challenge leaderboard. The submission counts per track are in Table 2. Similar to last year, we found low participa-

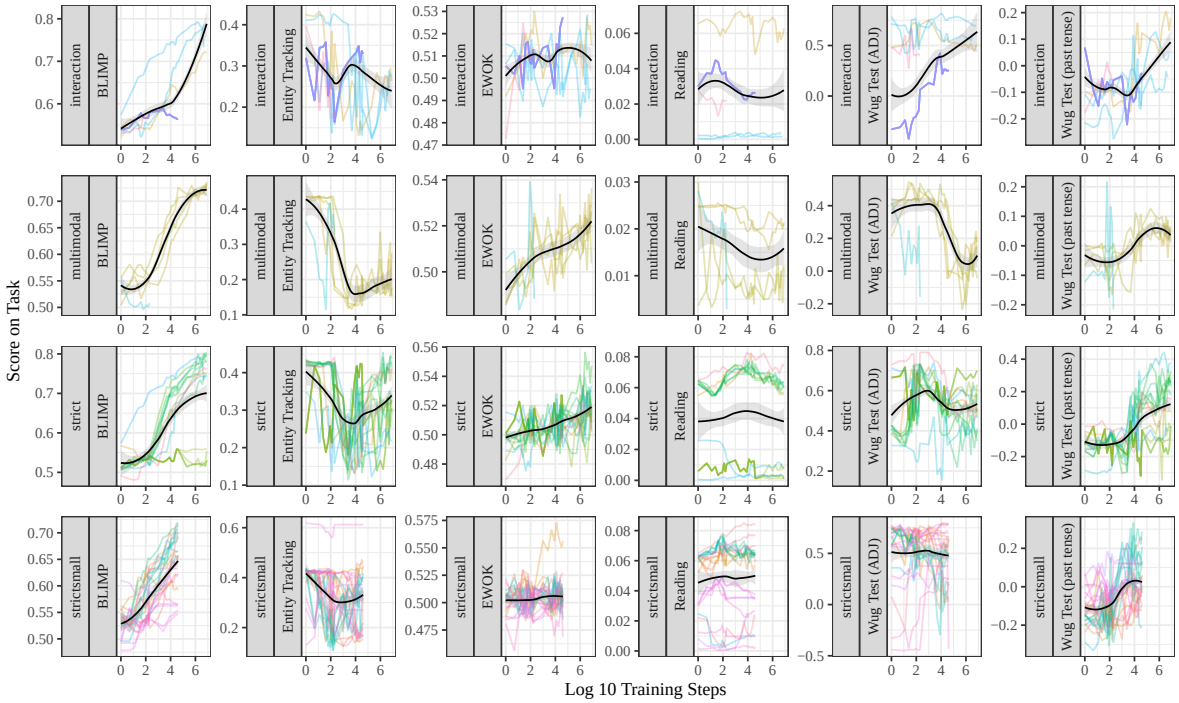

Figure 5: **Model Performance on Tasks over Training** Colors show scores for individual models; black lines show averages. Models generally show improvement for BLiMP and EWOK, while scores for reading-time predictions wug tests and entity tracking are more variable.

Track	# Models	# Participants
Strict-Small	15	9
Strict	7	4
Multimodal	1	1
Interaction	10	4
Total	*32*	*15*

Table 2: Total number of models and participants per track. This includes both participants in the challenge and workshop. Participants who submitted to multiple tracks are counted once in the total.

tion in the *Multimodal* track and received only one submission.

The breakdown of participants by affiliation and home country is as follows (submissions with multiple affiliations/countries are counted more than once): Germany (7), United States (6), England (5), Italy (3), Philippines (2), Switzerland (2), Denmark (2), Netherlands (2), Scotland (2), Japan (2), Sweden (2), Austria (2), Czechia (2), Turkey (1), India (1), Israel (1), Taiwan (1), Romania (1), Australia (1), Slovakia (1), South Korea (1), Ethiopia (1), Poland (1), Greece (1), Finland (1), Canada (1).

6.1 Winning Submissions

Human-likeness metrics were considered separate from accuracy metrics, such that a system could win either with respect to NLP task performance *or* human-likeness. We gave separate awards for both metrics.

***Strict* Track.** The winner of the human-likeness metric is CLASS-IT by Capone et al. (2025), which proposes to fine-tune small-scale LMs on a general instruction-following dataset. For the NLP tasks, the Simple Diffusion model by Kosmopoulou et al. (2025) is the winner; this is a diffusion *masked* language model.

***Strict-Small* Track.** The winner of the human-likeness metric is MoEP by Tapaninaho (2025), which employs modular mixtures of experts; this method achieves particularly high scores in the AoA task. For the NLP tasks, the AMLM-Hard-Decay model by Edman and Fraser (2025) is the winner; this method entails dynamically choosing which tokens in an input sequence to mask based on which are most difficult to predict according to the model.

***Interaction* Track.** For the *Interaction* track, we have a single winner, BLM by Martins et al. (2025),

Model	Human-likeness	NLP score	Macro Average	Vision Average
STRICT				
Best Models				
CLASS-IT*	20.4	52.9	36.6	—
Simple-Diffusion†	12.6	58.4	35.5	—
Batchwise-convergent$_{\text{main}}$	12.3	48.6	30.4	—
BLaLM‡	23.6	54.4	39.0	—
Baselines				
GPT-BERT-causal$_{\text{MNTP}}$	22.5	62.0	42.3	—
GPT-BERT-causal$_{\text{AR}}$	**26.5**	60.2	43.4	—
GPT-BERT-mixed$_{\text{MNTP}}$	24.2	62.7	**43.5**	—
GPT-BERT-mixed$_{\text{AR}}$	23.2	60.1	41.6	—
GPT-BERT-masked$_{\text{MNTP}}$	18.5	**63.0**	40.8	—
GPT-BERT-masked$_{\text{AR}}$	23.8	57.8	40.8	—
GPT2	17.1	55.4	36.2	—
STRICT-SMALL				
Best Models				
MoEP*	**31.5**	53.2	**42.3**	—
AMLM-Hard-Decay†	8.4	**58.3**	33.3	—
GPT-BERT$_{\text{ACLM-6k-MNTP}}$	21.5	50.3	35.9	—
AMLM-Hard	26.7	58.0	**42.3**	—
Baselines				
GPT-BERT-causal$_{\text{MNTP}}$	17.4	56.7	37.1	—
GPT-BERT-causal$_{\text{AR}}$	18.7	56.1	37.4	—
GPT-BERT-mixed$_{\text{MNTP}}$	13.4	57.3	35.4	—
GPT-BERT-mixed$_{\text{AR}}$	16.6	54.3	35.4	—
GPT-BERT-masked$_{\text{MNTP}}$	9.5	57.3	33.4	—
GPT-BERT-masked$_{\text{AR}}$	18.9	54.0	36.5	—
GPT2	19.8	49.1	34.5	—
MULTIMODAL				
Best Models				
BitMar‡	-9.3	49.0	19.9	26.7
Baselines				
Flamingo	4.1	**55.8**	29.9	49.3
Git	**7.6**	54.7	**31.1**	**49.7**
INTERACTION				
Best Models				
BLM*†	**20.8**	**54.4**	**37.6**	—
CLASS-IT	20.4	52.9	36.6	—
llamalogue$_{\text{rfOLMo-score}}$	7.0	50.1	28.5	—
Baselines				
SimPO	20.4	54.1	37.3	—

Table 3: Human-likeness, NLP task, macro average, and vision scores for the best models and baselines per track for the challenge. Boldened results represent the best score per track. * are the track winners for the human-likeness score. † are the track winners for the NLP task score. ‡ are workshop papers, while other models are from the challenge.

who achieved the highest score in both the human-likeness and NLP task metrics. BLM employs Llama as an interactive teacher model; the student generates a completion to a story, and the teacher scores the generated completion based on coherence, readability, and creativity. These scores are propagated as training signals to the student via a reinforcement learning-based approach.

6.2 Outstanding Paper Awards

In addition to the BabyLM Challenge winners, we gave 3 outstanding paper awards to papers that were especially interesting and likely to have significant impact for those in the community. We considered papers from both the BabyLM Challenge and BabyLM Workshop.

Are BabyLMs Deaf to Gricean Maxims? A Pragmatic Evaluation of Data-Limited Language Models. (Askari et al., 2025) This paper introduces a benchmark for evaluating the sensitivity of cognitively plausible language models to Gricean maxims. Using maxim-adhering and maxim-violating examples, it is found that pragmatic abilities improve with scale,but also that models trained on 100M words fall well short of children's abilities. Reviewers appreciated that this work contributed to an underexplored evaluation dimension, the analyses, and the solid grounding in relevant literatures.

Looking to Learn: Token-wise Dynamic Gating for Low-Resource Vision-Language Modelling. (Ganescu et al., 2025) This paper analyzes how best to make use of multimodal (text-image) data when training on cognitively plausible text corpora. The authors explore token-wise gating, channel attention, and auxiliary contrastive training objectives; these yield multimodal models that outperform the baselines. This paper features thorough analyses and strong results, and also discusses ways in which the constraints of the BabyLM Challenge indirectly limit the performance of multimodal models.

Teacher Demonstrations in a BabyLM's Zone of Proximal Development for Contingent Multi-Turn Interaction. (Salhan et al., 2025a) This paper introduces ContingentChat, a framework for evaluating and improving the *contingency*, i.e., the relevance and meaningfulness of multi-turn dialogues between student and teacher models. The authors introduce a post-training pipeline based on the Switchboard corpus and a teacher model; the method improves the grammaticality and cohesiveness of small-scale language models' generations. Reviewers appreciated the strong grounding in the developmental psychology literature.

7 Discussion

High-level takeaways. While curriculum learning remains popular, the best-performing approaches were again based on modifications to the pretraining objective or the model architecture. Diffusion MLMs, reinforcement learning with a teacher model, and mixture-of-experts approaches were especially effective. Relatedly, we notice that model performance is not necessarily tied to the total amount of compute. The relationship between these two is plotted in Figure 6 and we only find a positive correlation for the *Interaction* track. As in previous challenges, surprisingly simple approaches like better data preprocessing or hyperparameter tuning also showed performance gains over simple baselines. Now that we have included human-likeness evaluations, we can more confidently state that these methods are effective not just for improving performance on NLP tasks, but also for building better cognitive models of language processing.

Training dynamics. Visualization of training dynamics is given in Figure 5. For all models, BLiMP performance increases with the number of pretraining words. WUG past-tense performance also scales with pre-training words, but far less monotonically: there is no change in performance for the first 10–50M words. Afterwards, a phase shift occurs, and WUG performance begins to increase more monotonically with the number of words in the training corpus. Perhaps this reflects a movement from overgeneralization or memorization toward true generalization; further analyses in this low-data setting would be interesting. Entity tracking shows what appears to be U-shaped scaling for *Strict* and to a lesser extent *Strict-Small* models (Wei et al., 2023), where performance starts high, drops, and then increases again. Other tasks like reading time prediction and WUG adjective performance do not demonstrate a strong relationship with number of pretraining words.

Planned changes to future challenges. Ganescu et al. (2025) points out ways in which the provided vision embeddings for the multimodal track may constrain performance. Indeed, working with

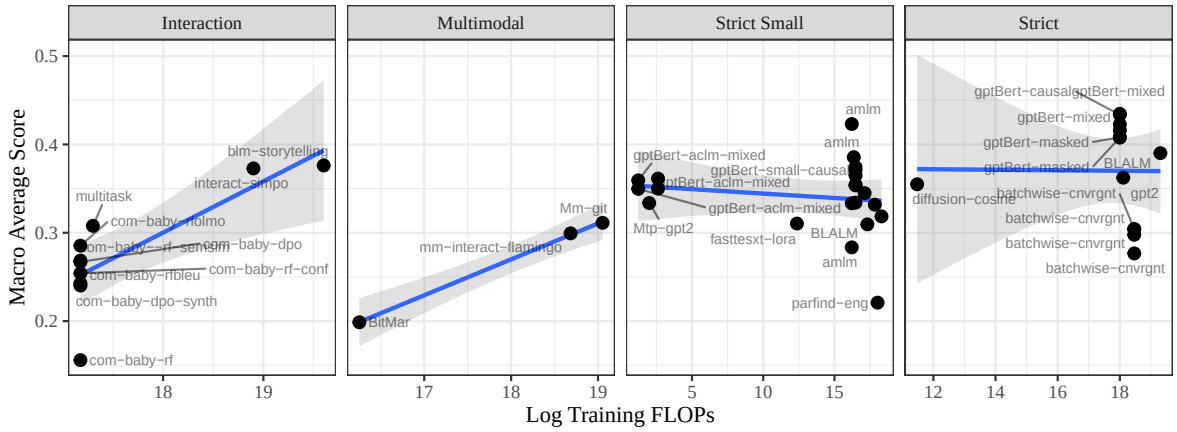

Figure 6: **Average score by flops used in training** We do not observe a strong relationship between the amount of compute used and the performance of the resulting model in the *Strict-Small* and *Strict* tracks.

these embeddings may be less straightforward than simply training the language and vision parts of a vision-language model from scratch. To encourage better performance and greater numbers of multimodal submissions in future challenges, we will consider moving to a dataset with a more open copyright license, such that participants will be able to train their own end-to-end models without needing to go through the current circumspect data download process.

8 Conclusion

The Third BabyLM Challenge shows that significant progress can be made in language modeling with academic-scale compute. With 32 models submitted from 26 countries, the challenge revealed several insights. Some agreed with the findings of previous years, for example, that training objective and architectural modifications were particularly effective. Some findings were novel this year, for example, that effective interactive approaches could be deployed using open-source teacher models. We also observed that the relationship between training FLOPs and performance was not nearly as strong this year as it was last year. Our controlled setup reveals that some approaches can outperform others based on methodological qualities distinct from how much compute they allow us to use.

Looking forward, we envision BabyLM continuing to evolve in its scope and focus. We hope to modify the multimodal evaluations to encourage more flexibility in future years. We also hope to continue exploring the value of tailored supervision and reinforcement learning-based approaches, as encouraged in the interaction track. While it is not currently as effective as simply pretraining on natural language corpora, we believe that this will continue to be a method of interest in both large- and small-scale language modeling research. By broadening our focus to include more language modeling methods, and by controlling for compute this year, we aim to inspire novel approaches that truly innovate beyond simply enabling greater compute to be spent. The strong participation and results this year suggest that the BabyLM community is well-positioned to pursue these ambitious goals, and ultimately to continue iterating towards the goal of human-like sample efficiency in language learning.

Acknowledgments

We are grateful to the BabyLM Challenge participants for making this challenge a consistent success. Our findings would not be nearly as interesting without their ambition and creativity, and their feedback on the logistics of the challenge itself, including the evaluation pipeline and training data. We are also grateful to the organizers of EMNLP for their efforts in hosting BabyLM this year.

Author Contributions

Primary Organizers Lu.Cha., Le.Cho., M.O.G., M.Y.H., J.L., J.J., A.M., C.R., E.G.W., Ad.Wi.

Pipeline implementation Lu.Cha., M.O.G., J.L., J.J.

Baseline model training Lu.Cha., M.O.G., J.L.

Communications with participants Le.Cho., E.G.W., M.Y.H

Training dataset compilation Al.Wa.

Reviewing submissions Lu.Cha., Le.Cho., M.Y.H., J.J., A.M., C.R., Al.Wa., E.G.W., Ad.Wi.

Initial draft on findings paper Lu.Cha., Le.Cho., M.O.G., M.Y.H., J.J., A.M., C.R., E.G.W., R.S.S.

Editing R.C., T.L., Ad.Wi., E.G.W., R.S.S.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *The 36th Conference on Neural Information Processing Systems*.

EUHID AMAN, Esteban Carlin, Hsing-Kuo Kenneth Pao, Giovanni Beltrame, Ghaluh Indah Permata Sari, and Yie-Tarng Chen. 2025. BitMar: Low-Bit Multimodal Fusion with Episodic Memory for Edge Devices. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Raha Askari, Sina Zarrieß, Özge Alacam, and Judith Sieker. 2025. Are BabyLMs Deaf to Gricean Maxims? A Pragmatic Evaluation of Sample-efficient Language Models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Ansar Aynetdinov and Alan Akbik. 2025. Babies Learn to Look Ahead: Multi-Token Prediction in Small LMs. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Necva Bölücü and Burcu Can. 2025. A Morpheme-Aware Child-Inspired Language Model. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Luca Capone, Alessandro Bondielli, and Alessandro Lenci. 2025. CLASS-IT: Conversational and Lecture-Aligned Small-Scale Instruction Tuning for BabyLMs. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

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.

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive.

Andrea de Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Lukas Edman and Alexander Fraser. 2025. Mask and You Shall Receive: Optimizing Masked Language Modeling For Pretraining BabyLMs. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Olivia La Fiandra, Nathalie Fernandez Echeverri, Patrick Shafto, and Naomi H. Feldman. 2025. Large Language Models and Children Have Different Learning Trajectories in Determiner Acquisition. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2016. Wordbank: an open repository for developmental vocabulary data*. *Journal of Child Language*, 44:677 – 694.

Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. Linguistic Units as Tokens: Intrinsic and Extrinsic Evaluation with BabyLM. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference*

on Empirical Methods in Natural Language Processing.

Eleni Fysikoudi, Sharid Loáiciga, and Asad B. Sayeed. 2025. Active Curriculum Language Modeling over a Hybrid Pre-training Method. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Bianca-Mihaela Ganescu, Suchir Salhan, Andrew Caines, and Paula Buttery. 2025. Looking to Learn: Token-wise Dynamic Gating for Low-Resource Vision-Language Modelling. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, and Weiwei Sun. 2025. BLiSS: Evaluating Bilingual Learner Competence in Second Language Small Language Models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Anita Gelboim and Elior Sulem. 2025. TafBERTa: Learning Grammatical Rules from Small-Scale Language Acquisition Data in Hebrew. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1).

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Patrick Haller, Jonas Golde, and Alan Akbik. 2025. Sample-Efficient Language Modeling with Linear Attention and Lightweight Enhancements. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 122:e2423232122.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *CoRR*, abs/2405.09605.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Despoina Kosmopoulou, Efthymios Georgiou, Vaggelis Dorovatas, Georgios Paraskevopoulos, and Alexandros Potamianos. 2025. Masked Diffusion Language Models with Frequency-Informed Training. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Nalin Kumar, Mateusz Lango, and Ondrej Dusek. 2025. Pretraining Language Models with LoRA and Artificial Languages. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Hyunji Lee, Wenhao Yu, Hongming Zhang, Kaixin Ma, Jiyeon Kim, Dong Yu, and Minjoon Seo. 2025. Understanding and Enhancing Mamba-Transformer Hybrids for Memory Recall and Language Modeling. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Sharid Loáiciga, Eleni Fysikoudi, and Asad B. Sayeed. 2025. Exploring smaller batch sizes for a high-performing BabyLM model architecture. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yilan Lu. 2025. Navigating the Design Space of MoE LLM Inference Optimization. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid, and Lisa Beinborn. 2025. Once Upon a Time: Interactive Learning for Storytelling with Small Language Models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Kate McCurdy, Katharina Christian, Amelie Seyfried, and Mikhail Sonkin. 2025. Two ways into the hall of mirrors: Language exposure and lossy memory drive cross-linguistic grammaticality illusions in language models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Sushant Mehta, Raj Dandekar, Rajat Dandekar, and Sreedath Panat. 2025. Unifying Mixture of Experts and Multi-Head Latent Attention for Efficient Language Models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Francesca Padovani, Bastian Bunzeck, Manar Ali, Omar Momen, Arianna Bisazza, Hendrik Buschmeier, and Sina Zarrieß. 2025. Dialogue Is Not Enough to Make a Communicative BabyLM (But Neither Is Developmentally Inspired Reinforcement Learning). In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Whitney Poh, Michael Tombolini, and Libby Barak. 2025. What did you say? Generating Child-Directed Speech Questions to Train LLMs. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *16th European Conference on Computer Vision*.

Rareș Păpușoi and Sergiu Nisioi. 2025. A Comparison of Elementary Baselines for BabyLM. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Matthew Theodore Roque and Dan John Velasco. 2025. Beyond Repetition: Text Simplification and Curriculum Learning for Data-Constrained Pretraining. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yamamoto Rui and Keiji Miura. 2025. FORGETTER with forgetful hyperparameters and recurring sleeps can continue to learn beyond normal overtfitting limits. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Suchir Salhan, Hongyi gu, Donya Rooein, Diana Galvan-Sosa, Gabrielle Gaudeau, Andrew Caines, Zheng Yuan, and Paula Buttery. 2025a. 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*.

Suchir Salhan, Richard Diehl Martinez, Zebulon Goriely, and Paula Buttery. 2025b. What is the Best Sequence Length for BabyLM? In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Loris Schoenegger, Lukas Thoma, Terra Blevins, and Benjamin Roth. 2025. Influence-driven Curriculum

Learning for Pre-training on Limited Data. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sam Shleifer, Jason Weston, and Myle Ott. 2021. Normformer: Improved transformer pretraining with extra normalization. *CoRR*, abs/2110.09456.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Ece Takmaz, Lisa Bylinina, and Jakub Dotlacil. 2025. Model Merging to Maintain Language-Only Performance in Developmentally Plausible Multimodal Models. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Alexander Tampier, Lukas Thoma, Loris Schoenegger, and Benjamin Roth. 2025. RecombiText: Compositional Data Augmentation for Enhancing LLM Pre-Training Datasets in Low-Resource Scenarios. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Alvin Wei Ming Tan, Chunhua Yu, Bria Lorelle Long, Wanjing Anya Ma, Tonya Murray, Rebecca D. Silverman, Jason D Yeatman, and Michael Frank. 2024. DevBench: A multimodal developmental benchmark for language learning. In *The 38th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Joonas Tapaninaho. 2025. MoEP: Modular Expert Paths for Sample-Efficient Language Modeling. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jannek Ulm, Kevin Du, and Vésteinn Snæbjarnarson. 2025. Contrastive Decoding for Synthetic Data Generation in Low-Resource Language Modeling. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Dan John Velasco and Matthew Theodore Roque. 2025. Rethinking the Role of Text Complexity in Language Model Pretraining. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *The 33rd Conference on Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

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*.

Jason Wei, Najoung Kim, Yi Tay, and Quoc Le. 2023. Inverse scaling can become U-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15580–15591, Singapore. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schütze, Kemal Oflazer, and David R. Mortensen. 2023. Counting the bugs in chatgpt's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6508–6524. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive

power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

L'uboš Kriš and Marek Suppa. 2025. SlovakBabyLM: Replication of the BabyLM and Sample-efficient Pretraining for a Low-Resource Language. In *Proceedings of the 3rd BabyLM Challenge at the 2025 Conference on Empirical Methods in Natural Language Processing*.

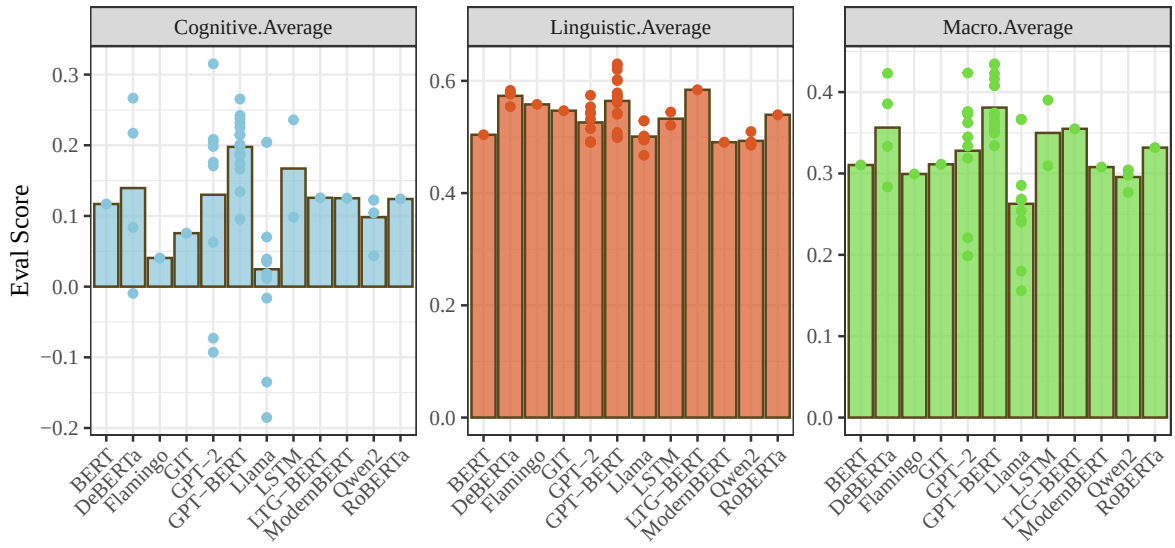

Figure 7: **Scores by backbone architecture** As with last year, we find that the GPT-BERT model consistently leads to stronger performance. Also consistent with previous years, we find that DeBERTa and LTG-BERT lead to strong performance as well.

A Additional Results

B HuggingFace Repository Names for Baseline Models

Model	HuggingFace Repository
STRICT	
GPT-BERT-causal	BabyLM-community/babylm-baseline-100m-gpt-bert-causal-focus
GPT-BERT-mixed	BabyLM-community/babylm-baseline-100m-gpt-bert-mixed
GPT-BERT-masked	BabyLM-community/babylm-baseline-100m-gpt-bert-masked-focus
GPT2	BabyLM-community/babylm-baseline-100m-gpt2
STRICT-SMALL	
GPT-BERT-causal	BabyLM-community/babylm-baseline-10m-gpt-bert-causal-focus
GPT-BERT-mixed	BabyLM-community/babylm-baseline-10m-gpt-bert-mixed
GPT-BERT-masked	BabyLM-community/babylm-baseline-10m-gpt-bert-masked-focus
GPT2	BabyLM-community/babylm-baseline-10m-gpt2
MULTIMODAL	
Flamingo	BabyLM-community/babylm-multimodal-baseline-flamingo
Git	BabyLM-community/babylm-multimodal-baseline-git
INTERACTION	
SimPO	BabyLM-community/babylm-interaction-baseline-simpo

Table 4: HuggingFace Repositories for the baseline models separated by tracks.

C Interaction External Model Correction Prompt

The prompt used for the external model to correct student model generations is shown in Figure 9.

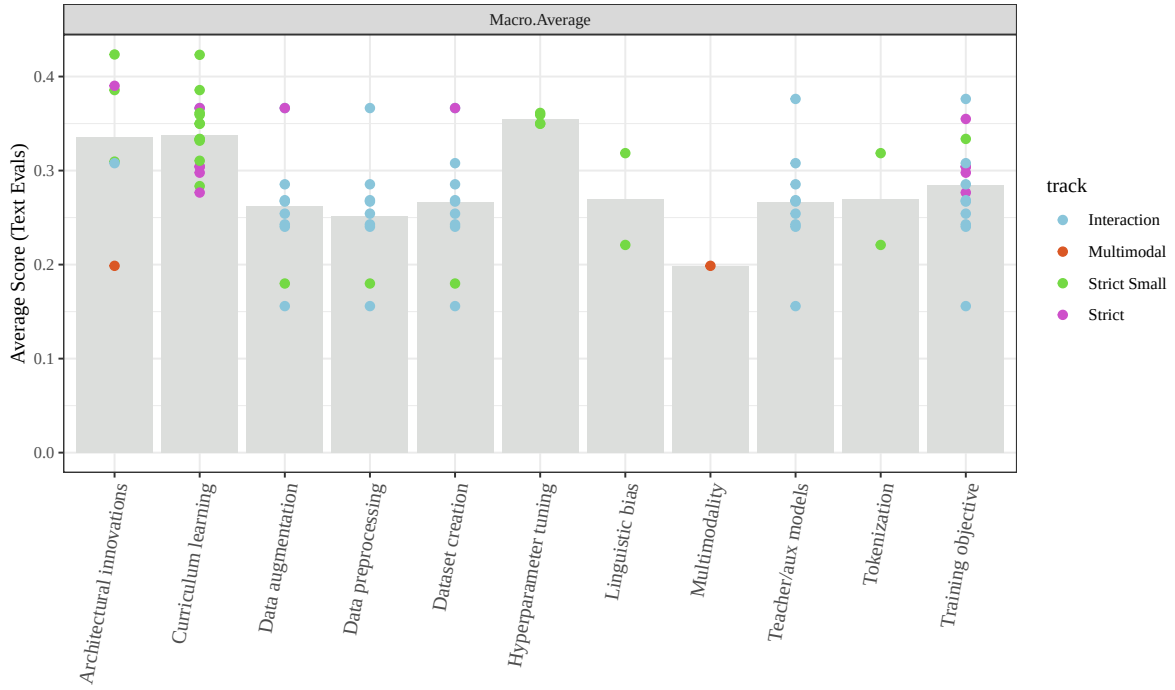

Figure 8: **Scores by approach taken** As with previous years, we find higher scores for models that employ architectural innovations. We also find higher scores for models that employ hyperparameter tuning.

Correction Prompt:
[User] You will be given a partial text (labeled "Partial Text") and a completion of said text produced by a student of English (labeled "Student Completion"). Your goal is to produce a corrected version of the student's completion. This corrected version should be grammatically correct, coherent and relevant to the initial partial text. If the student's response is incomprehensible, output your own independent completion. You should only provide your own completion without any added commentary or feedback.

Partial Text: <student input>
Student Completion: <student completion>

Now produce your own completion of the Partial Text. Do not include any external commentary.

[Assistant] Partial Text: <student input>
Corrected Completion:

Figure 9: The prompt given to the teacher model to sample corrected versions of the student's completions.

D Detailed Findings across the BabyLM Workshop and Challenge Submission

We synthesized results across the Workshop and Challenge tracks, examining each paper in terms of its data operations, training and optimization strategies, architectural choices, evaluation dimensions, release artifacts, developmental plausibility, multilingual scope, and use of MoE/sparsity mechanisms, along with additional factors including interaction and feedback methods, tokenizer family, objective variants, data provenance, selection policies, and competence-related effects. At BabyLM scales (≈10M-100M tokens), small design choices, such as tokenization, sequence length, or optimizer cadence, often yielded gains comparable to or greater than architectural modifications (Salhan et al., 2025b). Below we summarize the dominant empirical patterns that emerged across submissions.

D.1 Data: Selection Over Ordering, and the Shape of "Helpful" Synthetic Data

Selection Outperforms Human-Curated Ordering. Across multiple studies, model-driven selection criteria proved substantially more effective than human-designed curricula or naïve ordering strategies. Influence-based curricula (Schoenegger et al., 2025) improved performance by prioritizing examples that have the greatest effect on model predictions, while active surprisal-based selection (Fysikoudi et al.,

2025) dynamically focuses training on inputs the model finds most uncertain. Notably, gains arise not from a particular ordering direction (e.g., easy-to-hard), but from grouping data by similar influence or uncertainty levels, which stabilizes learning dynamics and improves generalization under strict data constraints.

Simplification of corpus text helps when balanced with diversity and aligned with model capacity. In several submissions, LLM-assisted simplification of existing corpus text improved sample efficiency and accelerated convergence under constrained token budgets. However, this was only true when simplification was applied as *augmentation* rather than replacement (Velasco and Roque, 2025; Roque and Velasco, 2025). Smaller models (under 200M parameters) benefited from simple-to-complex curricula, consistent with classical "starting small" effects; in contrast, relatively larger BabyLM-scale models (300M-1B) achieve higher downstream accuracy when simplified and original text were *interleaved*, indicating that simplification operated as a form of regularization.

Conversely, simplification improved linguistic knowledge transfer and zero-shot generalization only when the diversity and semantic coverage of the original corpus were preserved. Narrow simplification strategies that target a single linguistic feature like inserting only pedagogical questions, led to overfitting towards stylistic cues and reduced robustness on evaluation tasks (Poh et al., 2025). This shows that simplification is not inherently beneficial: its value lies in enhancing *coverage density* rather than constraining style.

Synthetic data is most effective when it complements rather than replaces natural text. Across both the tracks, two forms of synthetic augmentation consistently improved performance under fixed token budgets. Contrastive synthetic data, which is generated using paired Good/ Bad completions improved reasoning and robustness more reliably than vanilla synthetic sampling by providing explicit discriminative signals (Ulm et al., 2025). The effectiveness of this approach depended critically on maintaining diversity and balancing synthetic and natural data. Compositional, corpus-internal augmentation strategies, such as recombining semantically compatible sentence fragments, improved entity tracking, morphology, and several NLU metrics when synthetic data made up approximately half of the pretraining corpus (Tampier et al., 2025). Performance declined when synthetic data dominated the corpus, underscoring the need to ground augmentation in authentic linguistic distributions.

Submissions to both tracks converged on the same principle: hybrid data regimes, where synthetic and natural text were interleaved or mixed in controlled ratios, consistently outperformed purely synthetic or purely natural corpora at BabyLM scale.

D.2 Objectives and Training: Small Knobs, Big Effects

A consistent theme across submissions was that modifying the learning objective itself often yielded gains comparable to scaling data or model size. Several papers demonstrated that the choice and scheduling of the pretraining loss function directly shaped sample efficiency and downstream generalization. For example, *Mask and You Shall Receive* (Edman and Fraser, 2025) introduced an adaptive masked language modeling objective in which harder-to-predict tokens were masked more frequently, leading to improved performance on morphology-sensitive evaluations such as the WUG test. Similarly, *Babies Learn to Look Ahead* (Aynetdinov and Akbik, 2025) showed that incorporating multi-token prediction improved entity tracking and discourse modeling in models as small as 130M parameters trained on 10M tokens, with curriculum-based scheduling outperforming static variants. Alternative objectives were also explored, including diffusion-based language modeling (Kosmopoulou et al., 2025), where frequency-aware noise schedules yielded performance competitive with GPT-BERT hybrids, and parameter-efficient strategies such as pretraining on artificial formal languages followed by LoRA adaptation (Kumar et al., 2025), which outperformed full-parameter training on morpho-syntactic benchmarks.

Another set of findings showed that training cadence and procedural hyperparameters can rival architectural changes in their effect on model quality. *Exploring Smaller Batch Sizes* (Loáiciga et al., 2025) reported that reducing effective batch size while using gradient accumulation improved generalization on BLiMP and MSGS benchmarks, suggesting that increased optimization noise benefits small-data learning.

Complementary work (Rui and Miura, 2025) demonstrated that periodically resetting optimizer states allowed models to continue improving beyond conventional convergence points, yielding lower validation loss on both Baby10M and Baby100M settings. Additionally, *What's the Best Sequence Length?* (Salhan et al., 2025b) found that the optimal sequence length was highly dependent on both architecture and task: longer contexts benefited analogy and entity tracking tasks in state-space models, whereas shorter sequences were sufficient for syntactic generalization in transformer-based architectures.

D.3 Architectures: Efficiency, Linear-Time Models, and Sparse Routing Mechanisms

One key finding across submissions is that architectural modifications which reduce attention complexity or introduce structured sparsity often yield measurable gains under our strict data and compute constraints, particularly when paired with appropriate optimization. Linear and state space model (SSM)-based token mixers acted as alternatives to full self-attention. Haller et al. (2025) replaced self-attention with an mLSTM token mixer, combining them with lightweight modifications such as sliding-window attention and short convolutions, improved zero-shot performance and training stability, especially when using the Muon optimizer instead of AdamW. Lee et al. (2025) further demonstrated that hybrid architectures combining state space models with attention yield complementary strengths: sequential hybrid architectures performed better on short-context tasks, while parallel architectures with cross-attention achieved better long-context recall.

Sparse and routed architectures were also investigated from both deployment and learning perspectives. On the systems side, *Navigating the Design Space of MoE Inference Optimization* (Lu, 2025) evaluated expert offloading, quantization, and distillation strategies for serving mixture-of-experts (MoE) models under memory and latency constraints, finding that dynamic expert offloading can maintain model quality while reducing hardware requirements. On the learning side, other work (Tapaninaho, 2025) introduced token-routed sparse paths across modular transformer blocks and reported faster early learning and improved strict-small benchmark scores relative to a dense GPT-2 baseline, though with later-phase stability trade-offs. Mehta et al. (2025) presented a combined MoE and latent attention architecture that reduced KV-cache memory while maintaining competitive perplexity, suggesting that MoE-style routing and compression mechanisms can be jointly leveraged to improve efficiency.

D.4 Tokenization and Morphology

A consistent pattern across both Workshop and Challenge submissions is that tokenization choices exert disproportionately large effects in our constrained setups (*Strict*, *Strict-Small*), often rivaling objective or architectural modifications. Models using morphology-aware tokenizers demonstrated substantial gains in entity tracking and world knowledge tasks. One submission that compared BPE with rule-based and unsupervised morphological tokenization reported improvements of approximately 20% on EWoK and 40% on entity tracking when morpheme segmentation was applied, indicating that linguistically grounded token boundaries directly support better generalization in small models (Bölücü and Can, 2025). Curriculum-based introduction of morphology yielded mixed results: it added modest improvements for GPT-BERT architectures but degraded BLiMP performance in GPT-2 variants.

Multiple papers evaluated the redistribution of linguistic competence induced by tokenizer choice. Systems trained with BPE typically achieved the strongest syntactic acceptability judgments, while morphology or syllable-aware tokenizers improved semantic generalization and discourse tracking (Fusco et al., 2025; Păpușoi and Nisioi, 2025). These findings show that tokenization implicitly moves models toward different linguistic capacities, even when architecture and training data are held constant.

Evidence from multilingual model training with morphologically rich languages further supported the role of tokenization. In Hebrew, a compact RoBERTa-style model trained with morphology-aware representations achieved competitive grammatical judgments despite a reduced data budget (Gelboim and Sulem, 2025). In Slovak, a replication study found that token inflation caused by applying English-trained BPE tokenizers, increased the number of tokens per sentence and effectively reduced the usable data budget. In this setting, tokenization appeared as the single highest-leverage intervention, surpassing curriculum and architecture in impact (Ľuboš Kriš and Suppa, 2025).

D.5 Interaction, Feedback, and Alignment: Learning Beyond Pretraining Tokens

A small proportion of submissions that targeted interaction or feedback showed that alignment signals can act as efficient substitutes for large-scale pretraining. One line of work focused on dialogue alignment using minimal preference pairs. Padovani et al. (2025) show that fine-tuning with Direct Preference Optimization (DPO) on child-caregiver dialogue minimal pairs improved pragmatic choice behavior, leading to higher accuracy on communicative benchmarks, even though zero-shot language modeling metrics such as BLiMP and EWoK remained unchanged. In contrast, Proximal Policy Optimization (PPO), had mixed effects and occasionally destabilized model behavior. Another submission framed narrative generation as an interactive learning problem. In the storytelling setup, a teacher model assigned feedback on readability, coherence, and creativity to student-generated stories (Martins et al., 2025). With fewer than one million interactive tokens, the student model achieved gains comparable to a model trained on 100M tokens, particularly in narrative cohesion and entity tracking, while retaining performance on formal linguistic benchmarks.

A related submission introduced teacher demonstrations as aligned continuations in multi-turn dialogue. Salhan et al. (2025a) showed that models trained on "edited" responses provided by a teacher language model produced more contextually contingent and cohesive dialogue turns than baseline autoregressive models. This work further demonstrated that post-training on preference pairs improved multi-turn interaction quality without requiring large additional corpora. Together, these studies show that structured interaction primarily implemented through preference alignment (reinforcement-style scoring, or teacher demonstrations) helps induce qualitative gains in communicative and functional language use at a fraction of the token cost of additional pretraining.

D.6 Evaluation Beyond Grammar: Pragmatics, Learner Profiles, Developmental Trajectories, and Multimodal Trade-offs

Some submissions expanded the BabyLM evaluation landscape beyond traditional grammatical benchmarks, introducing new metrics for assessing pragmatic competence, second-language developmental profiles, learning trajectories, and multimodal efficiency.

Several works evaluated models on pragmatic reasoning. In one of the outstanding papers, BabyLM-scale models were assessed on a benchmark grounded in Gricean maxims (Askari et al., 2025). The authors found that while models trained on 100M tokens outperformed those trained on 10M, all models lagged behind child-level performance, with the largest deficits observed in the maxim of Quantity, indicating continued difficulty in evaluating informativeness. This pattern held even when other maxims, such as Quality and Relation, and showed moderate improvement with scale.

Another direction evaluated models through the lens of second-language acquisition. The BLiSS benchmark (Gao et al., 2025) introduced a large-scale evaluation of learner-like grammatical competence, using minimal pairs derived from annotated L2 corpora and organized by CEFR proficiency level and learner L1. Models were assessed on their ability to distinguish learner errors from corrections, revealing systematic differences across training regimes. Tokenization choice and transfer learning strategies significantly influenced alignment with bilingual learner profiles.

Developmental comparisons were explicitly explored in work on determiner acquisition trajectories. One submission (Fiandra et al., 2025) compared intermediate training checkpoints of BabyLM models with speech samples from children, showing that children consistently produced indefinite determiners first, while models acquired definite determiners earlier. This divergence suggests that models optimize for frequency and predictability rather than cognitive developmental salience. McCurdy et al. (2025) demonstrated that both language exposure statistics and memory constraints contribute to model behavior, but neither factor alone accounted for human-like processing across languages.

Finally, multimodal submissions examined the interaction between vision-language grounding and linguistic competence. One study showed that multimodal pretraining reduced performance on text-only grammatical benchmarks, but that merging the parameters of a multimodal model with those of a text-only model through weighted interpolation partially restored language-focused performance while maintaining multimodal capabilities (Takmaz et al., 2025). Another submission introduced a low-bit multimodal

fusion model with episodic memory (AMAN et al., 2025), demonstrating that aggressive quantization and memory augmentation allowed on-device deployment while preserving basic multimodal reasoning, albeit with reduced fine-grained linguistic fidelity.

Collectively, these evaluations extend the BabyLM paradigm beyond formal grammatical competence, revealing distinct dimensions of pragmatic inference, bilingual developmental alignment, cognitive trajectory modeling, and multimodal trade-offs.

D.7 Challenge versus Workshop Contributions

The Challenge track primarily focused on mechanisms for increasing sample efficiency under fixed token constraints. Submissions explored adaptive objectives such as difficulty-aware masked language modeling and multi-token prediction, as well as diffusion-based language modeling, demonstrating that modifying the learning signal can recover performance otherwise dependent on larger datasets. Several entries adopted parameter-efficient pretraining strategies, including the use of artificial structural priors followed by LoRA adaptation, and others demonstrated that small-scale interactive feedback could enhance communicative behavior with fewer than one million additional tokens. Corpus-internal augmentation methods, such as compositional recombination, were also introduced as an alternative to external synthetic generation, enabling performance gains while adhering to the Challenge's data budget constraints.

By contrast, the Workshop track broadened the evaluation and systems landscape. Submissions introduced new axes of measurement beyond grammatical competence, including pragmatic informativeness, bilingual learner profiles, and developmental trajectories derived from longitudinal human acquisition data. Other work focused on architectural and deployment efficiency, exploring sparse mixture-of-experts models, latent attention mechanisms, multimodal alignment, and model merging techniques designed to restore linguistic competence after multimodal pretraining.

Contributions to both tracks are in spirit of the BabyLM goals: rethinking data use rather than increasing data volume.