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

Jonas Mayer Martins **Ali Hamza Bashir**
Muhammad Rehan Khalid **Lisa Beinborn**

University of Göttingen, Institute of Computer Science, Germany
firstname.lastname@uni-goettingen.de

Abstract

Children efficiently acquire language not just by listening, but by interacting with others in their social environment. Conversely, large language models are typically trained with next-word prediction on massive amounts of text. Motivated by this contrast, we investigate whether language models can be trained with less data by learning not only from next-word prediction but also from high-level, cognitively inspired feedback. We train a student model to generate stories, which a teacher model rates on readability, narrative coherence, and creativity. By varying the amount of pretraining before the feedback loop, we assess the impact of this interactive learning on formal and functional linguistic competence. We find that the high-level feedback is highly data efficient: With just 1 M words of input in interactive learning, storytelling skills can improve as much as with 410 M words of next-word prediction.

Models and data | Code repository

1 Introduction

HUMANS are storytelling animals (Gottschall, 2012; Campbell, 2008). From early myths to modern science, narratives have served not only as entertainment but also as cognitive tools to make sense of the world. Scientific models and historical accounts, personal and collective identities, and even abstract institutions such as currency, law, and national borders can all be understood as shared stories (Bruner, 1991). Through our capacity for language, we establish a communicative common ground to align intentions, construct shared realities, and thus cooperate at societal scales (Tomasello, 2008, 2014; Clark and Schaefer, 1989; Clark and Brennan, 1991).

In recent years, language models have achieved surprising proficiency in generating natural language. However, training these artificial neural networks with billions to trillions of parameters is inefficient (Wilcox et al., 2025). While modern supercomputers are trained on the order of 10^{13} words (DeepSeek-AI, 2025), a child is exposed to between 10^8 and 10^9 words by age 13, extrapolating from Gilkerson et al. (2017). How do children acquire language so efficiently? In this work, we explore one potential ingredient: enriching the learning signal for language models beyond classical next-word prediction (Stöpler et al., 2025).

Artificial and biological neural networks differ in structure and dynamics, yet both can acquire complex linguistic behavior (Evanson et al., 2023). The standard training objective for language models—next-word prediction—superficially resembles predictive processing (Clark, 2013; Ryskin and Nieuwland, 2023), but does not reflect the rich, interactive learning experienced by children. We hypothesize that incorporating high-level feedback can guide language models toward more efficient functional linguistic competence, i.e., coherent, pragmatic, and creative use of language (Mahowald et al., 2024).

While the human brain excels at finding patterns in sensory input—a capacity central to early language learning (Saffran, 2020)—children are more than just passive recipients of this input. Instead, they learn language in a social context, shaped by interaction and feedback from caregivers (Tomasello, 2008; Clark, 2018). This feedback includes both implicit cues, such as contingent responses and repetitions, and explicit forms, such as corrections and confirmations (Cheatham et al., 2015; Nikolaus and Fourtassi, 2023).

By contrast, traditional language modeling is fully self-supervised. External feedback is integrated only later, during fine-tuning for applied tasks, when the model receives feedback from labeled examples (Parthasarathy et al., 2024). More recently, reinforcement learning (RL) has been introduced to language modeling to better align

Proceedings of the First BabyLM Workshop, pages 454–468
November 8, 2025

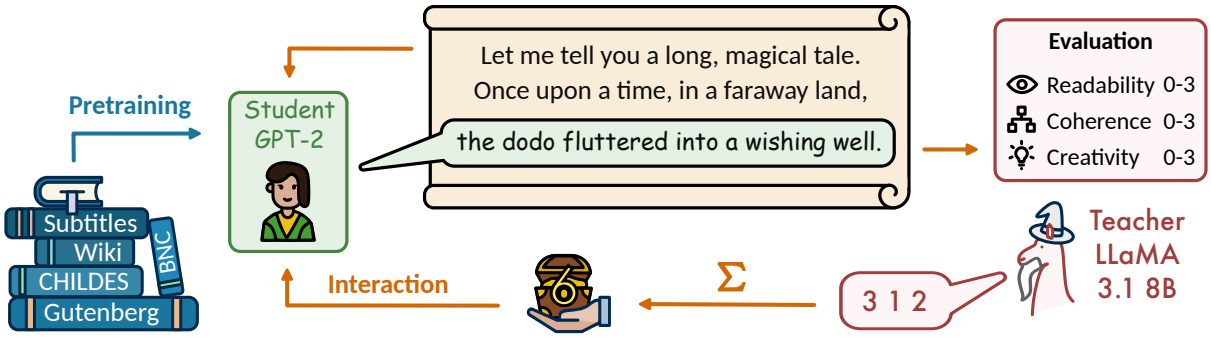

Figure 1: Schematic of the interactive learning setup with storytelling feedback. During pretraining, the student model optimizes next-word prediction on the BabyLM corpus. In the interaction stage, the student model completes a story prompt. A teacher model then evaluates the story on three criteria using a Likert scale from 0 to 3. The student receives the sum of these scores as a reward and updates its parameters to generate stories that maximize the expected reward.

model outputs with human preferences.

In this work, we replace part of the next-word prediction in pretraining by reinforcement learning in interaction with a teacher model, employing storytelling as a task that requires functional linguistic competence, see Fig. 1. After pretraining on the BabyLM corpus (Charpentier et al., 2025), the student model enters the interaction loop: First, the student generates a story from a generic snippet. Next, the teacher model judges the generated story with respect to readability, narrative coherence, and creativity. Finally, the student model receives the sum of the teacher scores as a reward and updates its parameters to maximize the expected reward.

We assess how high-level narrative and linguistic feedback impacts the student model's learning dynamics. Specifically, we demonstrate that partially replacing next-word prediction with interaction augments storytelling ability without compromising low-level linguistic generalization. Remarkably, with less than 1 M input words of interactive learning, storytelling skills improve as much as 410 M additional words of conventional pretraining. Finally, we examine how the amount of pretraining influences the effectiveness and dynamics of reinforcement learning for storytelling.

2 Interactive learning for small language models

Prior work on data efficiency in language modeling motivates alternative training objectives. Discussing storytelling as a lens for evaluating linguistic competence, we present interactive learning as a cognitively inspired approach to improving data efficiency and functional language skills in small models.

2.1 Scaling and parsimony

Large language models generally perform better with more parameters and more training data (Bahri et al., 2024). From a cognitive perspective, data parsimony is of particular interest. A child encounters orders of magnitude fewer words than large language models: Extrapolating from Gilkerson et al. (2017), we estimate that by age 13 a child has been exposed to around 100 million to 1 billion words—only a fraction of the input given to modern language models. Inspired by how children acquire language,[1] the BabyLM Challenge seeks to close this gap in data efficiency (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025). The findings from previous BabyLM challenges show that the most promising improvements in model performance come from changes in architecture and training objective (Warstadt et al., 2023; Hu et al., 2024).

We hypothesize that the next-word prediction objective—operating at the word or subword level—is too fine-grained to foster sufficient abstraction. In addition, next-word prediction requires multiple exposures to each word for effective learning and introduces frequency biases and anisotropy in model representations (Diehl Martinez et al., 2024; Godey et al., 2024). Achieving greater data efficiency may require a more comprehensive signal that incorporates high-level feedback.

2.2 Modeling storytelling

Humans communicate through stories and improve as storytellers by learning from interactive feedback. As a learning objective, storytelling is partic-

[1] We use language *learning* and *acquisition* interchangeably in this work.

ularly valuable because it requires *functional linguistic competence* (i.e., pragmatic use of language in real-world situations), as opposed to *formal linguistic competence* (i.e., knowledge of linguistic rules and patterns) (Mahowald et al., 2024). However, what defines a good story is difficult to formalize (Chhun et al., 2022) and existing metrics align poorly with human judgments (Guan et al., 2021).

Contemporary language models can produce fluent and grammatically correct stories but frequently struggle with coherence, creativity, and narrative structure (See et al., 2019; Xie et al., 2023). For example, models often fail at *entity tracking* (keeping track of facts about the world in a story), which is crucial for coherent stories (Kim and Schuster, 2023; Li et al., 2021). We propose that these functional skills can be improved by enriching the training objective with storytelling feedback.

2.3 Interactive learning

In multi-agent signaling games, the interactions of agents can lead to the emergence of novel communication protocols or even languages (Boldt and Mortensen, 2024; Bernard et al., 2024; Lazaridou et al., 2020). Also, cognitively inspired feedback can improve model performance (Nikolaus and Fourtassi, 2021; Saha et al., 2023; Stöpler et al., 2025). While previous work has explored various forms of feedback, our approach lets the student model generate stories freely in response to a writing prompt, while the teacher model provides high-level feedback on story quality.

Reinforcement learning, although a well-established method in machine learning, is relatively new to natural language processing (Parthasarathy et al., 2024; Havrilla et al., 2024). With regard to storytelling, reinforcement learning of sufficiently pretrained models appears surprisingly robust to sparse reward signals (Zhao et al., 2023; Wu et al., 2025). Unlike knowledge distillation, which approximates the function of a large language model through a model with fewer parameters (Dasgupta et al., 2023), our method uses textual feedback rather than probability distributions. This approach may be less computationally efficient, but it provides a more developmentally plausible reward signal, emulating student-teacher or child-caregiver interaction.

3 Methodology

As illustrated in Fig. 1, we model interaction as follows: A pretrained *student model* generates a story, which a *teacher model* then rates based on *evaluation instructions*. The teacher's scores serve as the reward signal for reinforcement learning via proximal policy optimization (PPO) (Parthasarathy et al., 2024).

Baselines We compare the student model against two baselines from the 2025 BabyLM challenge (Charpentier et al., 2025):

`1000M-pre` baseline: trained on 100 M unique words of the BabyLM corpus for 10 epochs with next-word prediction.

`SimPO` baseline: trained for 7 epochs with next-word prediction on the BabyLM corpus and 2 epochs interleaving next-word prediction with reinforcement learning. The reward is based on how similar the story completions of the student are to that of the teacher, providing corrective feedback.

Student model For our experiments, we use the same `GPT-2-small` architecture as the baseline for the student model and similar hyperparameters, see Appendix E.1. We divide the training into two stages:

`900M-pre` baseline: To stay within a word budget of 100 M words per epoch, we pretrain first on 90 % of the 100 M BabyLM corpus for 10 epochs.

`900M-RL` model: Subsequently, we do interactive learning with 1 M words of input. This yields fewer input words to the student model than the other baselines, namely, 901 M and 1,000 M words, respectively.

Teacher model Evaluating the quality of a story is a difficult task that requires both accurate judgments and computational efficiency. Based on pilot experiments, we select `Llama 3.1 8B Instruct` (Grattafiori, 2024).[2]

To mirror the student-teacher analogy, we keep the teacher model fixed throughout training.

Story generation To obtain a viable reward signal in reinforcement learning, we must elicit story-like outputs from the student model. We use the archetypal storytelling opening:

[2]Out of the three Llama Instruct models available for the Interaction Track of the BabyLM challenge (3.1 8B, 3.2 3B, and 3.2 1B), the largest one (Llama 3.1 8B Instruct) provides story scores with a reasonably high signal-to-noise ratio that aligned best with the developers' assessments of the story.

Student Model Input

Let me tell you a long, magical tale.
Once upon a time, in a faraway land,

Teacher feedback Defining the quality of a story is notoriously challenging (Chhun et al., 2022). Following Guan et al. (2021), we let the teacher model evaluate the student story on three criteria: readability, narrative coherence, and creativity.

Careful optimization of the teacher instructions was required for a strong and accurate learning signal, as language models are often highly sensitive to prompt phrasing (Chhun et al., 2022) and prone to label-induced biases (Saraf et al., 2025). During development, we refined the instructions to discourage shortcutting and ensure alignment with human judgment. We use rubrics to anchor the teacher's responses and provide examples of expected outputs. For each criterion, the teacher assigns a score from 0 (worst) to 3 (best), yielding robust and concise feedback. The full evaluation instructions are given in Appendix D.

Reward We use PPO to optimize the language model's policy for maximum expected reward. The reward R is calculated by combining the teacher scores $s_i \in \{0, 1, 2, 3\}$ for the three criteria i with a story length incentive based on the number of generated words L:

$$R = \frac{1}{1+\alpha}\left[\frac{1}{9}\sum_{i=1}^{3} s_i + \alpha\,\frac{L}{L_{\max}}\right] + r_{\mathrm{KL}}\,, \quad (1)$$

$L_{\max} = 100$ is the maximum allowed number of subword tokens (to normalize length), and $\alpha = 0.4$ controls the relative weight of the length bonus. The Kullback–Leibler (KL) divergence r_{KL} prevents the trained model from diverging too far from the pretrained baseline. See Appendix E.2 for full training parameters.

Experimental setup We first pretrain the `GPT-2-small` student model on 90 % of the BabyLM corpus for 10 epochs. To track the learning dynamics, we save checkpoints at logarithmically spaced intervals (1 M, 2 M, . . . , 10 M, 20 M, . . . , 100 M, 200 M, . . . , and 900 M words seen by the model). The final checkpoint constitutes our `900M-pre` baseline.

To assess the amount of pretraining necessary for efficient RL, we start the reinforcement learning from selected checkpoints (20 M, 50 M, 90 M, 200 M, 500 M, 900 M)[3] and train for 1 M words in 331.2k interactions (that is, 331.2k stories told), with evaluation checkpoints every 100k words.

During reinforcement learning, we log the stories, story length, teacher scores, as well as the KL divergence. The figures in Section 5.2 report Gaussian-smoothed batch averages ($\sigma = 30$ with batch size 360), unless otherwise noted.

4 Evaluation setup

We use the evaluation pipeline of the 2025 BabyLM Challenge (Charpentier et al., 2025). It comprises nine zero-shot diagnostic benchmarks and seven task-specific datasets that require model fine-tuning (see Appendix A).

Zero-shot diagnostics This suite evaluates the linguistic and conceptual capabilities of the language model by comparing its language modeling probabilities to human judgments. Minimal pair tasks are used to assess whether the model assigns higher probability to the more acceptable sentence. Each pair consists of two minimally contrastive sentences that isolate a certain phenomenon relating to syntactic and semantic grammaticality (`BLiMP`), dialogue and question processing (`BLimP supplement`), world knowledge about physical and social concepts (`EWoK`), and property inheritance (`COMPS`). In addition, the probabilities are correlated with human ratings for morphological properties of pseudo-words (`WUGs`), and to age-of-acquisition labels (`AoA`). Context integration capabilities of the model are tested by evaluating the proportion of the variance in eye-tracking (`Eye-T`) and self-paced reading (`SPR`) signals that is predictable from the surprisal of the model and by the accuracy of predicting the final state of an entity (entity tracking, `ET`) after a series of operations described as natural language discourse.

Task-specific fine-tuning The applicability of the model for downstream tasks is evaluated by its task-specific accuracy after supervised fine-tuning for question answering (`BoolQ` and `MultiRC`), natural language inference (`MNLI` and `RTE`), paraphrase recognition (`MRPC` and `QQP`), and coreference resolution (`WSC`). In the results, the fine-tuning tasks are summarized as `GLUE`. See Appendix E.3 for fine-tuning parameters.

[3]The tags (e.g., 900 M) refer to the number of pretrained words, not model size.

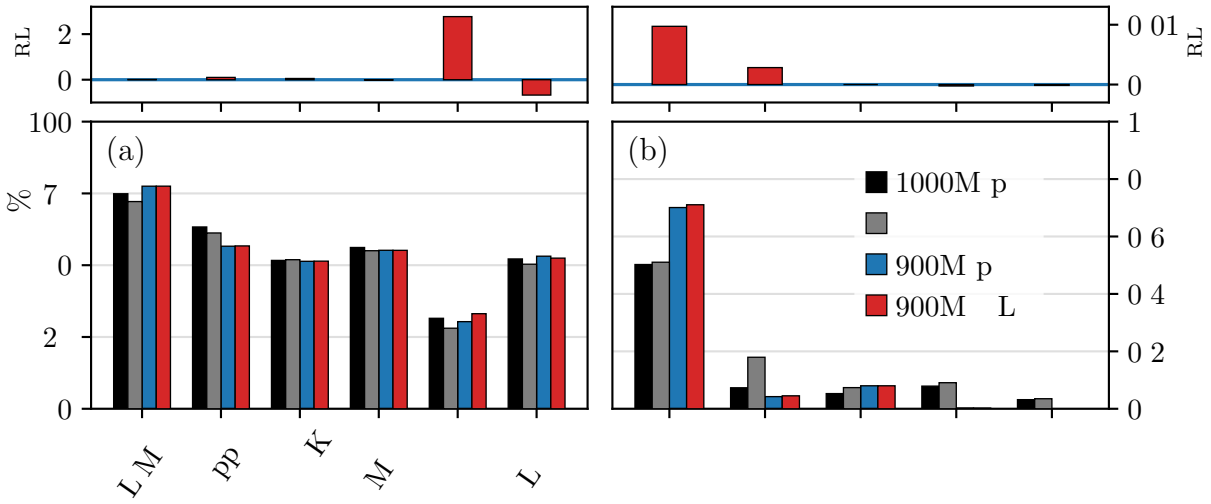

Figure 2: Evaluation on the BabyLM tasks, cf. Table 1. The bottom panels show the (a) accuracy and (b) correlation or partial correlation on the respective tasks for the next-word prediction baseline `1000M-pre`, the interaction baseline `SimPO`, and the model before and after interactive reinforcement learning (RL). The top panels indicate the difference of the 900 M model before and after interaction learning (in percentage points on the left, correlation differences on the right). `GLUE` encompasses all fine-tuning tasks.

5 Results and discussion

We first examine the effect of our interaction model on formal linguistic competence as assessed by the BabyLM evaluation pipeline. We then analyze how storytelling skills improve through reinforcement learning, and explore the training dynamics.

5.1 Formal linguistic competence

We evaluate formal linguistic competence using the BabyLM tasks, comparing our model pretrained on 900 M words before (`900M-pre`) and after (`900M-RL`) interactive reinforcement learning. We also compare with a baseline pretrained on 1,000 M words (`1000M-pre`), and an interaction baseline with a different training objective (`SimPO`). The results are summarized in Fig. 2; for detailed values, see Table 8.

We observe that the two baselines, `1000M-pre` and `900M-pre`, achieve similar performance on most tasks. This suggests that the missing 10% of the pretraining corpus and thus 100 M additional words in pretraining have little effect on formal linguistic competence.

Strikingly, as shown in the top panels in Fig. 2, the accuracy on entity tracking (`ET`) increases the most, from 30.3 % to 33.1 %, and correlations on the two `WUG` tasks improve marginally. Although the teacher reward was not tailored to any of these tasks, improved entity tracking likely reflects the importance of maintaining narrative coherence—specifically, keeping track of characters and objects—in storytelling. Accuracy on the `GLUE` benchmark drops slightly by 0.7 percentage points after interaction. Notably, interactive reinforcement learning does not affect most other BabyLM tasks.

The `SimPO` baseline, despite being exposed to more words during interaction, does not differ much from the baselines and performs slightly worse than `1000M-pre` on `BLiMP`, `BLiMP Supplement`, `ET`, and `GLUE`.

As shown in panel (b) the metrics measuring alignment with psycholinguistic data (`AoA`, `Eye-T`, and `SPR`) have less consistent trends than the accuracy-based scores in panel (a) and the correlations of all models are below 0.1. The `WUG-A` task has a high correlation between 0.5 and 0.7 for all models.

In summary, two observations stand out: First, omitting 10 % of training data (`900M-pre` vs. `1000M-pre`) does not significantly affect the performance on the formal linguistic competence captured by the BabyLM tasks. Second, adding only 1 M additional words of interactive reinforcement learning after pretraining maintains those competences and even improves entity tracking.

5.2 Storytelling

How does interactive learning affect the learning dynamics of a small language model? We first

explore the storytelling performance itself and the data efficiency of the learning setup. Next, we dive deeper into the learning dynamics of the individual storytelling criteria, the influence of the number of pretraining words and the interaction progress.

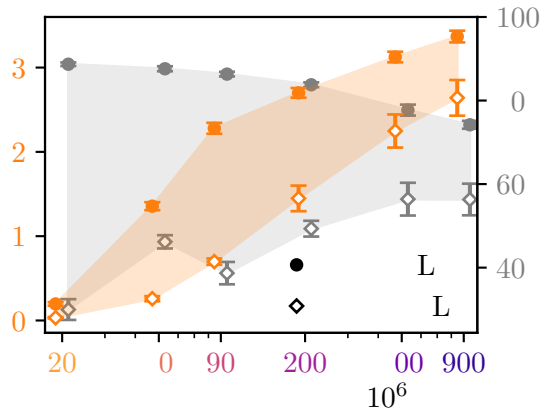

Figure 3: The effect of interactive reinforcement learning (RL) for models with increasing number of pretraining words on two variables: average teacher score (orange, left axis) and story length (gray, right axis). Error bars indicate the standard deviation of the first and last 20 batch averages, respectively. Orange and gray data points are slightly offset horizontally to avoid overlap.

Storytelling skills As shown in Figure 3, after the reinforcement learning (RL) interaction phase, the student models produce stories that are both longer and rated higher by the teacher. At first glance, this indicates that the models successfully learn to optimize the reward, which combines the teacher score and a bonus for story length. However, the extent of the improvement depends strongly on the amount of pretraining.

Specifically, models with more pretraining produce higher-scoring stories, both before and after RL: The 20 M model initially produces short and after RL long stories that the teacher scores almost zero throughout. In contrast, the 90 M and 200 M models show the greatest increase in teacher score, while the most pretrained model, 900 M, gains less from RL, although it ultimately achieves the highest absolute scores. Interestingly, the 900 M model also produces the shortest stories after RL, despite earning the highest ratings, which suggests that it relies least on story length as a shortcut.

In Appendix C, we provide a random sample of stories from the first, middle, and last third of interactions, as well as the best story, for the 90 M and 900 M models. The anthology of all stories produced by the models is available as a Hugging Face dataset.

Data efficiency We find that interactive learning is remarkably data efficient: After RL, the 90 M model receives an average teacher score of 2.3 that outperforms that of the 500 M model before storytelling interaction. Thus, 1 M words of interactive learning achieve the same improvement as 410 M extra words in pretraining. This result aligns with the findings of Wu et al. (2025) and Zhao et al. (2023), who demonstrate that LLMs learn with surprising efficiency in reinforcement learning. This robustness to sparse reward signals—such as the fixed student input in our setup—can be attributed to knowledge of the target domain acquired through sufficient pretraining. In our case, this finding agrees with our observation that a certain amount of pretraining is required before reinforcement learning can meaningfully enhance storytelling skills.

Story quality We analyze the distribution of teacher scores across the criteria used for evaluating the student model's stories. Figure 4 (a) shows the evolution of each criterion's score with the number of interactions for the six models with different amounts of pretraining. To emphasize underlying trends and filter out high-frequency fluctuations of the data, we apply a Gaussian filter.

Overall, the scores for all three criteria increase over time. As illustrated in Fig. 4 (b), models with more pretraining perform better on all criteria. Notably, readability emerges as the hardest criterion, for which even the 900 M model rarely attains two points, while performance in creativity and narrative coherence is substantially better across all models. The limited improvements on readability, which reflects superficial fluency, fit the observation from Section 5.1 that, for example, grammatical knowledge (as measured by BLiMP) is not much affected by the interactive RL, but creativity and coherence improve instead.

Fig. 4 also shows that the 20 M model fails to achieve higher teacher scores except for a minor gain in creativity. Models pretrained for 90 M and 200 M words gain the most on all criteria, whereas more pretraining leads to diminishing returns in teacher scores.[4]

[4]Considering the entropy per word, see Appendix B, we find that it is dominated by the amount of pretraining, with little change during interactive RL. This indicates that improvements in storytelling cannot simply be attributed to changes in output diversity.

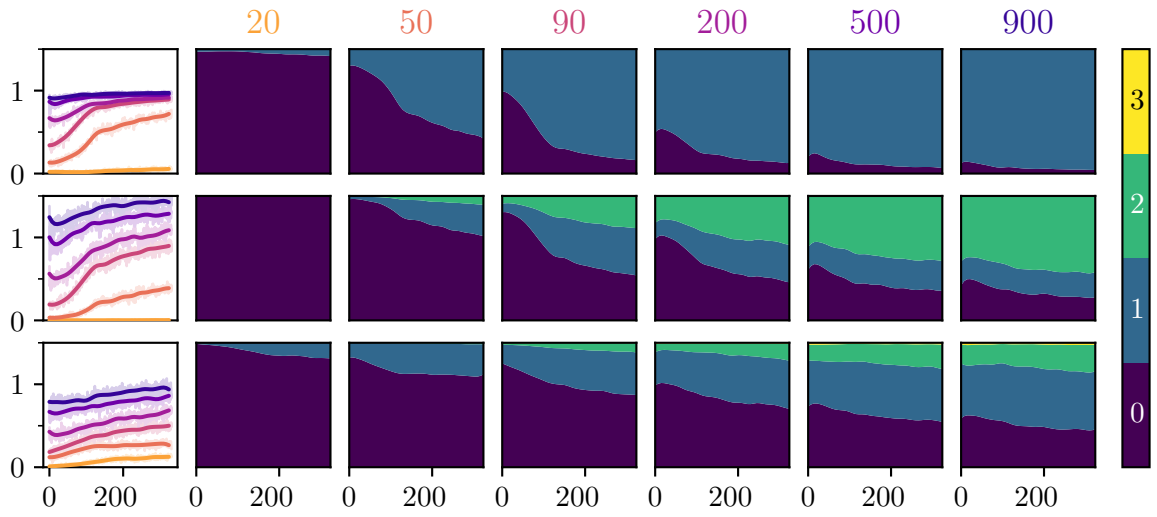

Figure 4: Teacher score over the course of RL interaction for models with increasing pretraining: 20 M, 50 M, . . . , 900 M words. (a) Teacher scores by criterion. Shaded regions show averages per batch, solid lines are Gaussian-smoothed batch averages. (b) Distribution of teacher scores over time.

Learning dynamics Figure 5 (a) illustrates the evolution of teacher score, story length, and KL divergence over the number of interactions. Across all models, both teacher score and story length increase most rapidly until 100k interactions, after which improvements continue but at a slower pace. This deceleration is also reflected in Fig. 4. KL divergence, which quantifies the similarity of the RL-trained model to its pretrained baseline, increases during early training and then stays constant around $\mathrm{KL} = 6$, a convergence determined by the adaptive KL scheduling of PPO. Deviating from this plateau would compromise the total reward signal, thus constraining policy updates. Notably, models with more pretraining, like 500 M and 900 M words, exhibit a decrease in story length after KL convergence before increasing again, potentially signaling a delayed adaptation of the model's reward prediction as these models adjust to changes in the slope of KL divergence.

Fig. 5 (b) combines the trajectories of the different models along three dimensions: story length, teacher score, and number of interactions. These trajectories define a surface, which we approximate with a one-dimensional linear interpolation (surface with blue to yellow gradient). The upper two diagrams in panel (a) correspond to projections of the trajectories, connecting the nonlinear effect of pretraining on the evolution of these variables.

Fig. 5 (c) completes the picture with a projection onto the plane of teacher score and story length, collapsing the dimension of interactions. This view reveals how models with different pretraining navigate the trade-off between story length and teacher score. The 20 M model shows a limited slope, improving primarily in story length. This indicates a threshold: models pretrained on fewer than 50 M words cannot leverage interactive feedback, which implies that some amount of pretraining is necessary for a viable reward signal. In contrast, the 90 M and 200 M models exhibit pronounced improvement in both dimensions. Models with even more pretraining like 500 M and 900 M display diminishing returns, consistent with Fig. 3. Overall, the 90 M model benefits most from interactive learning.

6 Conclusion

Our experiments demonstrate that interactive feedback is highly data efficient for storytelling: With just 1 M words of additional input, storytelling skills reach the equivalent of an additional 410 M words of next-word prediction in pretraining. This result highlights the data inefficiency of next-word prediction and might explain why children acquire language with far less input than today's large language models.

We find that interactive reinforcement learning primarily enhances narrative coherence and creativity, while leaving surface-level fluency—measured by the BabyLM tasks—largely unchanged. An improvement in entity tracking aligns with the training objective focused on storytelling.

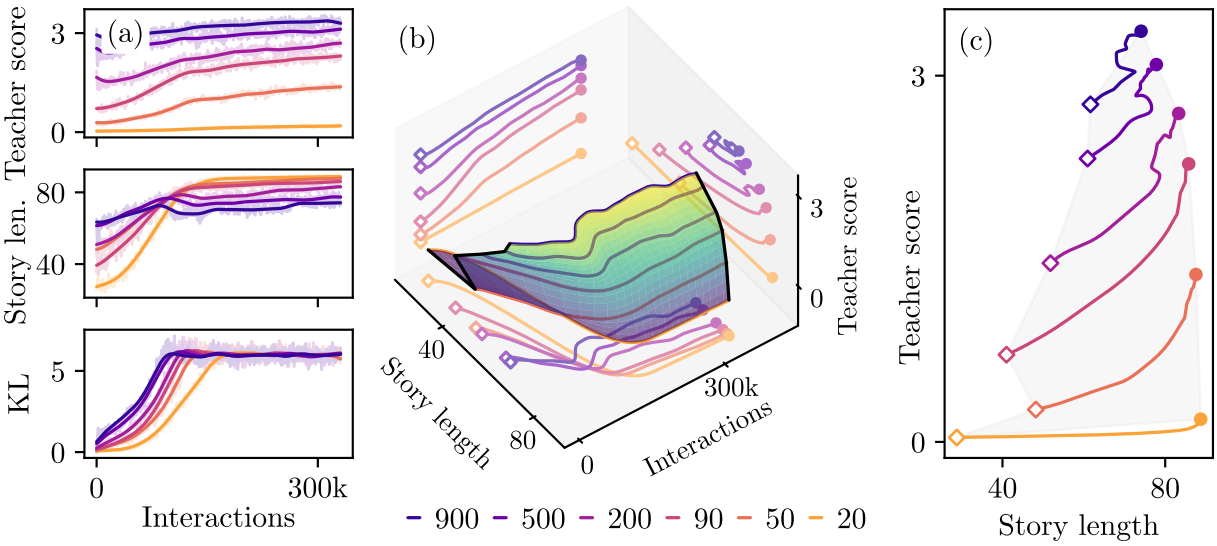

Figure 5: Learning dynamics of the reinforcement learning (RL). (a) Teacher score, story length, and KL divergence by interaction number. The shading shows the average per batch, the solid lines are Gaussian-smoothed batch averages. (b) Training trajectories visualized as a manifold with projections in the dimensions of story length (0 to 90 words) and teacher score (0 to 9 points) and number of interactions. (c) Trajectories in the phase space of teacher score and story length.

Our analysis reveals that models with less pretraining tend to exploit story length as a shortcut, whereas those with 90 M and 200 M words of pretraining benefit the most from interactive learning. Models with more pretraining suffer from diminishing returns from interaction. Notably, we identify a threshold: between 20 M and 50 M words of pretraining are necessary for the model to benefit from interactive reinforcement learning. Examining the nature of this threshold and its parallels to language acquisition in children presents an intriguing avenue for future research.

Limitations

While storytelling RL is highly data efficient, it is by no means computationally efficient: RL on 1 M input words took 20 GPU hours per model, because it involves generating 20 M words of student output for the stories. For comparison, 900 M words of pretraining amounted to less than 10 GPU hours.

Moreover, our analysis focuses on the learning dynamics. We leave a detailed study of the student stories—how content, register, vocabulary, and syntax evolve through interaction—for future work. Mechanistic interpretability methods could also provide insights into how training affects internal model representations.

Furthermore, we weight the three evaluation criteria of the teacher equally, but these weights can be adapted during RL to implement a form of curriculum learning.

Our teacher rewards serve as a heuristic for story quality. Further validation using benchmarks like `OpenMEVA` (Guan et al., 2021) or human annotations would strengthen this approach.

We used a fixed input for story generation, but more diverse corpora (e.g., BabyLM (Charpentier et al., 2025), TinyStories (Eldan and Li, 2023), or WritingPrompts (Fan et al., 2018)) could affect learning outcomes; each with its own tradeoffs regarding narrative content and diversity.

Ethics statement

Importantly, computational language models are not faithful representations of human cognition and should not be anthropomorphized. Rather, they are tools for informing hypotheses about language learning, which should ultimately be tested on human studies.

While the BabyLM challenge targets more sustainable training regimes, model development still requires considerable computing resources. Model development and final training took about 140 kcore-hours in total. Pretraining took 2 hours on 4 A100 GPUs. RL learning took 20 hours on 1 A100 GPU for each of the six RL models (5 - 10 kcore-hours per model).

Acknowledgements

We thank the reviewers for their input. We thank Eva Beck for helpful discussions. Lisa Beinborn's research is partially supported by an *Impulsprofessur* grant from the *zukunft.niedersachsen* program and by a VENI grant (Vl.Veni.211C.039) from the Dutch National Science Organisation (NWO). The authors gratefully acknowledge computing time provided to them at the GWDG HPC cluster.

References

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2024. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121.

Luisa Bentivogli, Ido Kalman Dagan, Hoa Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *TAC 2009 Workshop*.

Timothée Bernard, Timothee Mickus, and Hiroya Takamura. 2024. The emergence of high-level semantics in a signaling game. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics (ACL).

BNC Consortium. 2007. British National Corpus, XML edition.

Brendon Boldt and David Mortensen. 2024. A review of the applications of deep learning-based emergent communication. arXiv:2407.03302. *Transactions on Machine Learning Research*, arXiv:2407.03302.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Joseph Campbell. 2008. *The Hero with a Thousand Faces*, 3rd edition. Bollingen series XVII. New World Library, Novato, Calif.

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 Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. BabyLM turns 3: Call for papers for the 2025 BabyLM workshop. arXiv:2502.10645. *arXiv preprint*.

Gregory Cheatham, Margarita Jimenez-Silva, and Hyejin Park. 2015. Teacher feedback to support oral language learning for young dual language learners. *Early Child Development and Care*, 185:1452–1463.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Andy Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics (ACL).

Eve V. Clark. 2018. Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, 14(3):170–185.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, Levine John M., and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. Cost-effective distillation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354, Toronto, Canada. Association for Computational Linguistics (ACL).

Andrea Gregor De Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

DeepSeek-AI. 2025. DeepSeek-V3 technical report. arXiv:2412.19437. *arXiv preprint*.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5999–6011, Miami, Florida, USA. Association for Computational Linguistics (ACL).

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases.

In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent English? arXiv:2305.07759. *arXiv preprint*.

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: Do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics (ACL).

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics (ACL).

Martin Gerlach and Francesc Font-Clos. 2020. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta. Association for Computational Linguistics.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol. 1, San Francisco, CA, USA. IEEE.

Jonathan Gottschall. 2012. *The Storytelling Animal: How Stories Make Us Human*. Houghton Mifflin Harcourt, Boston.

Aaron et. al. Grattafiori. 2024. The Llama 3 herd of models. arXiv:2407.21783. *arXiv preprint*.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics (ACL).

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. arXiv:2403.04642. *arXiv preprint*.

Valentin Hofmann, Leonie Weissweiler, David Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19):e2423232122.

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 (ACL).

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 Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. arXiv:2405.09605. *arXiv preprint*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics (ACL).

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 (ACL).

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics (ACL).

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics (ACL).

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 (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Brian MacWhinney. 2014. *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*, 3rd edition. Psychology Press, New York.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

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 (ACL).

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407, Online. Association for Computational Linguistics (ACL).

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. arXiv:2408.13296. *arXiv preprint*.

Rachel Ryskin and Mante S. Nieuwland. 2023. Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, 27(11):1032–1052.

Jenny R. Saffran. 2020. Statistical language learning in infancy. *Child Development Perspectives*, 14(1):49–54.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? Teacher explanations improve students via personalization. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 62869–62891, Red Hook, NY, USA. Curran Associates Inc.

Muskan Saraf, Sajjad Rezvani Boroujeni, Justin Beaudry, Hossein Abedi, and Tom Bush. 2025. Quantifying label-induced bias in large language model self- and cross-evaluations. arXiv.2508.21164. *arXiv preprint*.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics (ACL).

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.

Lennart Stöpler, Rufat Asadli, Mitja Nikolaus, Ryan Cotterell, and Alex Warstadt. 2025. Towards developmentally plausible rewards: Communicative success as a learning signal for interactive language models. arXiv:2505.05970. *arXiv preprint*.

Michael Tomasello. 2008. *Origins of Human Communication*. The MIT Press.

Michael Tomasello. 2014. The ultra-social animal. *European Journal of Social Psychology*, 44(3):187–194.

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 (ACL).

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*, 8:377–392.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David 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, pages 6508–6524, Singapore. Association for Computational Linguistics (ACL).

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics (ACL).

Haoze Wu, Cheng Wang, Wenshuo Zhao, and Junxian He. 2025. Model-task alignment drives distinct RL outcomes. arXiv.2508.21188. *arXiv preprint*.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics (ACL).

Xingmeng Zhao, Tongnian Wang, Sheri Osborn, and Anthony Rios. 2023. BabyStories: Can reinforcement learning teach baby language models to write better stories? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 186–197, Singapore. Association for Computational Linguistics (ACL).

A Evaluation

We use the evaluation pipeline of the 2025 BabyLM Challenge (Charpentier et al., 2025). In Table 1, we provide an overview of the evaluation data.

B Entropy

Figure Fig. 6 shows that the average entropy per word increases slightly at the beginning of training, staying mostly constant until the end of reinforcement learning, but the entropy is otherwise not substantially correlated with story length or teacher score. Pretraining, on the other hand, has a strong influence on the entropy per word.

C Sample stories

Best story by reward and example stories—randomly sampled from the first, second, and last third of RL training—are listed in Table 2 for 90 M pretrained words and Table 3 for 900 M pretrained words. Interestingly, the best story for 900 M pretrained words is a meta-story—a story about a story—that directly appeals to the teacher evaluation by describing a "great story". Each model produced about 20 M words during RL training, which amounts to about 50,000 pages. The full anthology is available as a Hugging Face dataset.

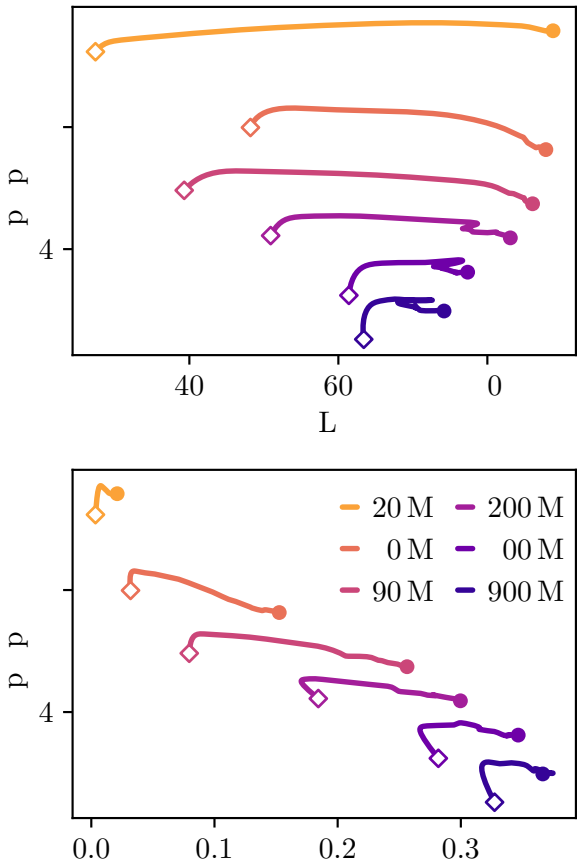

Figure 6: Entropy per word by story length and teacher score during interactive RL for different amounts of pretrained words. An empty diamond marks the start of a trajectory, a filled circle the end.

Setting	Dataset	Prediction Task	Evaluation Metric	Reference
Zero-shot	BLiMP	Grammatical acceptability	Accuracy	Warstadt et al. (2020)
	Suppl.	Discourse acceptability	Accuracy	Warstadt et al. (2023)
	EWOK	Conceptual knowledge	Accuracy	Ivanova et al. (2025)
	COMPS	Property knowledge	Accuracy	Misra et al. (2023)
	WUG-A	Morphol. generalization (adj.)	Spearman's ρ	Weissweiler et al. (2023)
	WUG-P	Morphol. generalization (verbs)	Spearman's ρ	Hofmann et al. (2025)
	AoA	Age of acquisition	Pearson's ρ	Chang and Bergen (2022)
	Eye-T	Eye-tracking fixations	Squ. partial corr. $\mathrm{p}R^2$	De Varda et al. (2024)
	SPR	Reading times	Squ. partial corr. $\mathrm{p}R^2$	De Varda et al. (2024)
	ET	Entity Tracking	Accuracy	Kim and Schuster (2023)
Fine-tuning	BoolQ	Question answering	Accuracy	Clark et al. (2019)
	MultiRC	Question answering	Accuracy	Khashabi et al. (2018)
	MNLI	Natural language inference	Accuracy	Williams et al. (2018)
	RTE	Entailment	Accuracy	Bentivogli et al. (2009)
	MRPC	Paragraph identification	Accuracy	Dolan and Brockett (2005)
	QQP	Question similarity	Accuracy	Iyer et al., (2017)
	WSC	Coreference resolution	Accuracy	Levesque et al. (2012)

Table 1: Overview of evaluation datasets in the BabyLM pipeline.

(1, 2, 3) **Length:** 82 **Reward:** 0.71 **#** 318,619

when it was given to the sun. Did he tell you that I wrote a letter, then that the moon was not frightened of telling her story? She introduced her red coat into the eyes of his safety. Perhaps she loved that, though she had seen sad old laughter. She now told her primary harmon they in spite of the name as she had never seen a star light thunder escort. She found herself there a dragon doing her best at it

(0, 0, 0) **Length:** 25 **Reward:** 0.07 **#** 43,434

she will possess me a descend-weller repwined feast. Of this of the Saxon divinity and for fifty months will occur; " Mr. March reppainted .</s>

(1, 1, 1) **Length:** 67 **Reward:** 0.43 **#** 217,857

hitherto hathine before twenty-three of people had lost their Senator Ragnzan leave in the decision, formally on the fact that he would make-up payment. After Commonwealth negotiated against a white elector offenders, Captain Spndez Junpped with the alliance and denied the suggestion which was only a vote from him: he seems to men such as England and Wales, but he said not to have been a party,

(1, 2, 1) **Length:** 71 **Reward:** 0.52 **#** 321,361

which whenever it may not be trusted. 'Cause she really would feel about that she might have recognized her lyrics. I encourage her. The love is not yet permitted to pronouncemy to the contrary." She was in a state that it was a small, sweet gift. She kept pieces of the discussion, and she heard her dear face. Her newence was based upon her key to the audit that she had

Table 2: Examples of generated stories from the 90 M model: overall best story (top) ranked by teacher and length reward, plus random samples (seed 42) from the first, second, and last third. Prompt: "*Let me tell you a long, magical tale. Once upon a time, in a faraway land*".

(2, 3, 3) **Length:** 76 **Reward:** 0.85 **#** 8,821

everything seemed as if it were a fairy tale—so much as you have told us a great story that we cannot bear to believe. There was something, however, very different from the description of the whole story, most surprising. If the story reached out, then, and it was always the story at the end, it was always something inexpressibly a story that made a great deal of sense. But it was all a mistake and wonder

(1, 2, 2) **Length:** 76 **Reward:** 0.61 **#** 43,434

there lived old men like me, the flowers of Jed serene rose from every mountain and valley, and they grew wildly from one to another. The inhabitants of those under one tree and the branches in the other, had been bewitched in their own world by their native dwelling; but having enabled them to approach them they had for good to take an active place among mankind. They resided many weary people, but had always a

(1, 0, 0) **Length:** 63 **Reward:** 0.26 **#** 217,857

there lived an unquenchable king. But cannot the words wear off, and for ten days he was forced to stake his head on one native bird-cage, while the hunters came in for us the fish-skin preservation. He was a dread of poor little war-birds, and a more likeable wickedness so lonesome in proportion to his cruel fangs as a young bird devouring li

(1, 0, 0) **Length:** 67 **Reward:** 0.27 **#** 321,361

immature and upland boy, he was dazzled by the tremendous overlooks of his race had lighted. He met a spirit who had been there all the day to bespeak in the midst of many years, and answered: "Hear him, Don Carlos, from there he lent it to reality; He is a different kind of drunken-looking man; I consider him much creamered after his teeth." In the same

Table 3: Examples of generated stories from the 900 M model: overall best story (top) ranked by teacher and length reward, plus random samples (seed 42) from the first, second, and last third.

D Evaluation instructions

```
You are a helpful teacher grading a student story. Be nice!
Only evaluate the student story itself, not the story prompt.
Given the student's word limit of about 80 words,
score the story on each of these three categories separately
on a scale from 0 to 3,
where 0 is the worst and 3 is the best.

Readability:
0 - Frequent and severe grammar errors; difficult to understand.
1 - Noticeable grammar errors; mostly understandable.
2 - Few minor grammar errors; well-formed overall.
3 - Correct grammar; well written.

Narrative Coherence:
0 - No story: completely incoherent or too short.
1 - No logical flow, confusing narrative.
2 - Mostly coherent story and not cut off.
3 - Coherent and logically structured story.

Creativity:
0 - Dull or incomprehensible.
1 - Somewhat creative; mostly predictable.
2 - Fairly creative and engaging.
3 - Highly original, imaginative, and engaging.

If the student story is empty ("") or less than a full sentence,
you must give the score 0 0 0!

Provide your scores, separated by single spaces, in the format:
Readability, Narrative, Creativity = _ _ _

Respond ONLY with this sequence of three numbers
without any extra text or explanation.

Story Prompt:
`{{story_prompt}}`

Student Story:
"`{{student_completion}}`"

Readability, Narrative, Creativity =
```

Source	Ratio	Domain	Reference
BNC	8%	Dialogue	BNC Consortium (2007)
CHILDES	29%	Dialogue, child-directed	MacWhinney (2014)
Proj. Gutenberg	26%	Fiction, nonfiction	Gerlach and Font-Clos (2020)
OpenSubtitles	20%	Dialogue, scripted	Lison and Tiedemann (2016)
Simple Eng. Wiki.	15%	Nonfiction	–
Switchboard	1%	Dialogue	Godfrey et al. (1992), Stolcke et al. (2000)

Table 4: Composition of the BabyLM corpus.

E Model parameters

BabyLM corpus The composition of the BabyLM corpus is listed in Table 4. It comprises 100 M words, of which we use 90% for pretraining and tokenization.

E.1 Pretraining

Model and training The model parameters are listed in Table 5. The vocab size of the tokenizer is 16,000 to match the baseline `1000M-pre` and the interaction baseline `SimPO`, which have vocab size 16,384. We use different values for seed, batch size, gradient accumulation, and learning rate compared with the baselines.

Hyperparameter	Value
Number of epochs	10
Context length	512
Batch size	16
Gradient accum. steps	4
Learning rate	0.0005
Number of steps	211,650
Warmup steps	2,116
Gradient clipping	1
Seed	42
Optimizer	AdamW
Optimizer β_1	0.9
Optimizer β_2	0.999
Optimizer ε	10^{-8}
Tokenizer	ByteLevelBPE
Tokenizer vocab size	16,000
Tokenizer min. frequency	2

Table 5: Hyperparameters used for pretraining.

E.2 Reinforcement learning

See Table 6.

Parameter	Value
Student context length	512
Seed	42
Batch size	360
Student sampling temp.	1
Top k	0
Top p	1
Max. new tokens (student)	90
Teacher model	Llama 3.1 8B Instr.
Teacher context length	1,024
Student sampling temp.	0.2
Max. new tokens (teacher)	6
Gradient acc. steps	1
Adapt. KL control	True
Init. KL coef.	0.2
Learning rate	1×10^{-6}
Student input limit	1 M words

Table 6: PPO Training Hyperparameters. Other parameters defaults of TRL 0.9.4.

E.3 Fine-tuning

See Table 7.

Hyperparameter	Value
Number of Epochs	10
Batch Size	16
Learning Rate	3×10^{-5}
Warmup percentage	6 %
Optimizer	AdamW
Weight decay	0.01
Scheduler	cosine
Dropout	0.1

Table 7: Hyperparameters used for fine-tuning.

F BabyLM evaluation results

See Table 8.

Task	1000M-pre	SimPO	900M-pre	900M-RL
`BLiMP`	74.88	72.16	77.52	**77.53**
`Suppl.`	**63.32**	61.22	56.62	56.72
`EWOK`	51.67	**51.92**	51.36	51.41
`COMPS`	**56.17**	55.05	55.20	55.18
`ET`	31.51	28.06	30.34	**33.11**
`GLUE`	52.18	50.35	**53.14**	52.46

Task	1000M-pre	SimPO	900M-pre	900M-RL
`WUG-A`	0.502	0.510	0.701	**0.711**
`WUG-P`	0.073	**0.179**	0.042	0.045
`AoA`	0.053	0.074	**0.080**	**0.080**
`Eye-T`	0.079	**0.091**	0.003	0.002
`SPR`	0.032	**0.035**	0.000	0.000

Table 8: BabyLM task scores for the four models from Fig. 2. Accuracy metrics are reported as percentages, WUG-A/P as Spearman's ρ, AoA as Pearson's ρ, Eye-T and SPR as partial correlations $\mathrm{p}R^2$. Bold indicates the best model for each task.