Babysit A Language Model From Scratch: Interactive Language Learning by Trials and Demonstrations

Ziqiao Ma [*1] **Zekun Wang** [*1] **Joyce Chai** [1]

Abstract

Contrary to human language learning, recent advancements in large language models have primarily adopted a non-interactive training paradigm, and refined pre-trained models through feedback afterward. In this work, we aim to examine how corrective feedback from interactions influences neural language acquisition from the ground up through systematically controlled experiments, assessing whether it contributes to learning efficiency in language models. We introduce a trial-and-demonstration (TnD) learning framework that incorporates three distinct components: student trials, teacher demonstrations, and a reward conditioned on language competence at various developmental stages. Our experiments reveal that the TnD approach accelerates word acquisition for student models of equal and smaller numbers of parameters, and we highlight the significance of both trials and demonstrations. We further show that the teacher's choices of words influence students' word-specific learning efficiency, and a practice-makes-perfect effect is evident by a strong correlation between the frequency of words in trials and their respective learning curves.

1. Introduction and Background

Humans are social beings and we learn language from interactions (Vygotsky, 1934; Bruner, 1985; Palincsar, 1986; Kuhl, 2004; Tomasello, 2005). Long before children's linguistic skills are mature, they could engage in early forms of conversational exchange with others (Halliday, 1975; Clark, 2018). A critical component of social interactions in language acquisition is the *feedback* provided by the caregivers (Warlaumont et al., 2014). There have been extensive studies on a type of feedback that emphasizes the responses from caregivers, which offer corrections to possible errors in children's speech, encompassing variants like negative evidence, reformulations, or recasts, collectively referred to as *corrective feedback* (Farrar, 1992; Chouinard & Clark, 2003; Saxton et al., 2005; Hiller, 2016).

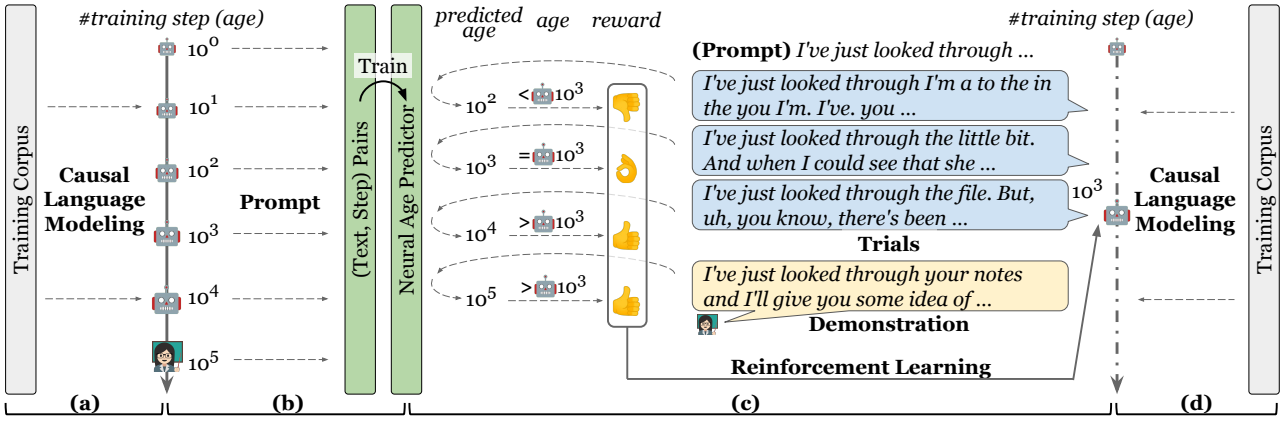
Unlike human learners who acquire language skills through feedback during interactions, most language models differ in terms of their inductive biases and data sources (Warstadt & Bowman, 2022). These models typically learn from massive text corpora using cross-entropy loss. Recently, several lines of cognitively motivated language modeling research have looked into the learnability and learning efficiency of language (Portelance et al., 2020; Chang & Bergen, 2022; Evanson et al., 2023). By incorporating non-linguistic inputs such as multimodal stimuli (Shi et al., 2019; Ma et al., 2023; Portelance et al., 2024) and/or communicative feedback (Nikolaus & Fourtassi, 2021; Zhu et al., 2022; Liu et al., 2022), recent studies have explored potential mechanisms that contribute to efficient language learning in (vision-)language models. Through controlled ablation studies (Warstadt & Bowman, 2022), these models can serve as proof of concept to verify cognitive mechanisms that are practically effective for machines, and generate hypotheses that are theoretically possible for human learners (Portelance, 2022; Portelance & Jasbi, 2023). We provide a detailed discussion of related work in Appendix A. In a similar spirit, we seek to investigate the role of explicit corrective feedback in neural language learning through controlled computational experiments. **Our goal is to determine if student trials and teacher demonstrations are effective in efficient word learning in training language models.**

2. Interactive Language Learning by Trials-and-Demonstrations (TnD)

We present a Trial-and-Demonstration (TnD) learning framework that streamlines the process (Figure 1). The TnD framework thus includes three components: a student model's *trials*, a teacher model's *demonstrations*, and *rewards*. This framework allows us to incorporate massive corpora to study modern transformer-based generative language models, offering a general and unrestricted approach to simulate interactive language learning with corrective feedback on massive corpora without human subjects to supervise the development of a language model.

[*]Equal contribution [1]Computer Science and Engineering Division, University of Michigan. Correspondence to: Ziqiao Ma <marstin@umich.edu>.

ICML 2024 Workshop on LLMs and Cognition, Vienna, Austria. 2024. Copyright 2024 by the author(s).



(a) Stage 1: Training a typical language model by causal language modeling.

(b) Stage 2: Training a neural age predictor from the trajectory of a typical language model.

(c) Stage 3: The student interactively learn from trials and demonstrations by a pre-trained teacher, score by an age-conditioned reward.

(d) Alternating between interactive learning and non-interactive learning.

Figure 1. The learning by trial-and-demonstration (TnD) framework. In stage 1, we start by training a language model with the causal language modeling objective. In stage 2, we prompt the models along the learning trajectory for (text, step) pairs and train a neural age predictor to predict the training step given a text. In stage 3, we use the final model in stage 1 as the teacher model. In an interactive step, the student model is prompted to complete a trial, and the teacher model is prompted to provide a demonstration. The trials and demonstrations are scored by an age-conditioned reward function (Eq. 6), and the student model updates the policy with reinforcement learning. The student alternates between interactive and non-interactive steps.

The student model and trials. We employ randomly initialized GPT-2 (Radford et al., 2019) as the student model for our investigation into language acquisition, leveraging its causal language modeling (CLM) objective and inherent generative capabilities for production-based learning. To encourage the student model to attempt text production, it is essential to provide an appropriate context. In each trial, we prompt the student with the first 5 tokens from a natural sentence, asking it to generate the continuation as a *trial*.

The teacher model and demonstrations. We utilize pre-trained language models as proxies for human language teachers. Employing language models as "caregivers" for language models offers two advantages. Firstly, it eliminates the need for humans to engage with a language model across thousands of iterations. Secondly, we can consistently control the behavior of the teacher model across experiments, and adjust the teacher's language behaviors by modifying its decoding strategies for language generation. The process of developing a teacher model is identical to the typical language model pre-training, as shown in Figure 1(a). We adopted the same GPT-2 architecture and pre-trained the model with the CLM objective for 100k steps, with all hyper-parameters following the default setup. To generate a natural language *demonstration* for the student's *trial*, we prompt the pre-trained teacher model with the same 5 tokens used for the student model, thereby obtaining the teacher's completion of the sentence.

The reward and reward model. Heuristic rewards like BLEU do not consider the developmental trajectory of language models, which is critical for simulating language acquisition. It's akin to human development where early words are celebrated as milestones, but prolonged reliance on initial language abilities can become a concern. We treat the number of training steps as the neural model's "age" (Chang & Bergen, 2022). A language model that generates fluent text at 500 steps, which typically emerges around 5,000 steps, should be rewarded for its accelerated learning. Conversely, if the language production quality in the student remains the same at 50,000 steps, it should be penalized, as shown in Figure 1(c). To train a neural age predictor, we prompt the teacher model at various training steps to generate text continuations, and fine-tune LLaMA-2-7B (Touvron et al., 2023) using these (text, step) pairs. More details are available in Appendix B.3. At step n, the student model π_θ parameterized by θ produces a sentence $\mathcal{S}$ of l tokens $t_1, \cdots, t_l$. We use the neural age predictor R_ϕ to estimate the logarithm of the expected training step $\hat{n}$ where this sentence typically emerges. Hence, the age-conditioned reward $r(\mathcal{S}, n)$ is given by $r := \log(\hat{n}/n) = R_\phi(\mathcal{S}) - \log n$.

Alternating interactive and non-interactive language learning As shown in Figure 1(c-d) and Algorithm 1, our TnD framework alternates between two forms of language learning: (1) interactive learning, in which corrective feedback is taken through reinforcement learning, utilizing rewards derived from both the student's trials and the teacher's demonstrations, and (2) non-interactive learning, which emulates the natural language exposure experienced by learners and is facilitated through causal language modeling as adopted by generative language models. During interactive learning, the goal is to view the language production in language models as actions within a vocabulary-defined action space. We use Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm with a clipped surrogate objective $\mathcal{L}_\theta^{\text{ppo}}$, and involve both the trial and demon-

stration into the batch for policy update. We adopt the causal language modeling objective for non-interactive learning and an alternating learning schedule over interactive and non-interactive language learning, i.e., perform $c = 3$ steps of causal language modeling, followed by $r = 1$ steps of reinforcement learning, in a continuous cycle. We refer to Appendix B and Appendix D.7 for details.

3. Experiment and Evaluations

Baselines. We designed our baselines for controlled ablation studies: (1) The `CLM` model, which adheres to the original GPT-2 pre-training with only CLM objective; (2) The `TnD` model, which implements the trial-and-demonstration framework described in Section 2; (3) The `Trial` model, which is the `TnD` framework with only student trials (no demonstrations); (4) The `Demo` model, which is the `TnD` framework with only teacher demonstrations (no trials). For each combination of corpus and baseline, the training process is conducted on 5 random seeds for 10k steps. We discuss more about other possible baseline setups and the hyper-parameters in Appendix C.

Training corpora and testing vocabulary. We repeat our study on two training corpora: the BookCorpus (Zhu et al., 2015) and the BabyLM Corpus. For evaluation, we specify two sets of vocabulary for evaluation. (1) The `CMN` set, consisting of common words that appear frequently in both corpora, covering a wide range of words and parts of speech. (2) The `CDI` set, consisting of words from the MacArthur-Bates Communicative Development Inventories (CDIs) (Fenson et al., 2006). We select these two vocabulary sets as `CMN` offers broader coverage while `CDI` is more frequency-neutral and is used to assess children's early vocabulary development.

Surprisal and learning curves. In line with previous work, we use the mean surprisal (log-perplexity) of a word to quantify the quality of the model's predictions for this word. For each occurrence of a word w, the surprisal is given by $-\log_2 P(w)$, and we average all occurrences. To visualize word acquisition from high surprisal to lower surprisal, we evaluate the student model throughout training and fit the surprisal values over logarithm training steps with a double-sigmoid function, leading to a learning curve for individual words and the overall vocabulary. We justify this choice and discuss more on the patterns in Appendix C.1.

Neural age of acquisition (nAoA). To evaluate the speed at which the student model acquires a word, we employ the neural age of acquisition (nAoA). Prior research (Chang & Bergen, 2022; Chang et al., 2023) has used a surprisal cutoff of 50% between the minimum and maximum surprisal levels, akin to the method used to determine children's age of acquisition (Braginsky et al., 2016). To further enhance the robustness of this metric, we average nAoA over different surprisal thresholds from 0.5 to 0.95 with a step of 0.05, denoted as `nAoA@[0.50:0.95]`.

Effective vocabulary size. Finally, we assess the effective vocabulary size relative to a test set of vocabulary. A word is deemed acquired at step n if `nAoA@0.50` $\leq n$. This approach yields a monotonically increasing curve that illustrates the growth of the effective vocabulary over time.

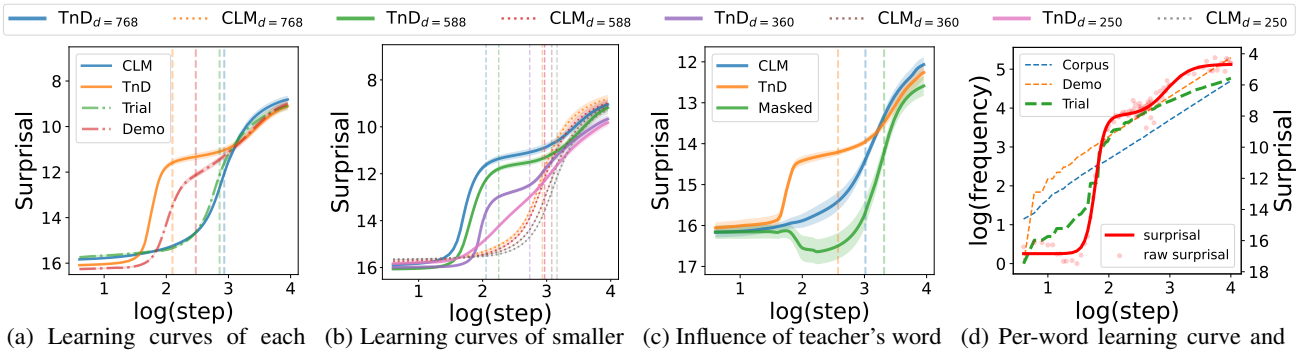
4. Main Results and Findings

In this section, we present our key findings using BabyLM corpus and `CDI` words for example. The complete tables, plots, and additional analysis are available in Appendix D.

Corrective feedback accelerates neural word acquisition. To evaluate `CMN` and `CDI` words on two corpora, we aggregate 5 random seeds and present the learning curves in Figure 6. Figure 2a reveals that the `TnD` learning framework significantly accelerates word acquisition in training, outperforming other baselines. This acceleration is attributed to the critical roles of both trials and demonstrations in the learning process. With only teacher demonstrations, the student model acquires words faster than with the plain `CLM` baseline alone, though not as rapidly as when active trials are incorporated in the `TnD` framework. Conversely, without the teacher's demonstrations, the student's trials in the wild do not yield a marked improvement, resulting in performance comparable to the `CLM` baseline. Figure 7, 8 and Table 2 present nAoA at different surprisal thresholds from 0.50 to 0.95 with a step of 0.05. We find that the `TnD` learning framework is particularly beneficial during the earlier stages of word acquisition, as it significantly outperforms the `CLM` baseline on `nAoA@0.50`, but is eventually on-par with the `CLM` baseline on `nAoA@0.95` towards the end of training. As a result, it can be seen from Figure 8 that students under the `TnD` framework quickly picked up a large volume of effective vocabulary, but eventually their vocabulary capacities have converged to `CLM`. Overall, our findings show that corrective feedback accelerates the student model's neural word acquisition process, yet the student eventually converges to the teacher model's performance.

Corrective feedback helps knowledge distillation for smaller student models The original student GPT-2 model has a dimension of $d = 768$ (12 attention heads each with a dimension of 64). We now keep all experimental setups untouched but smaller student models with dimensions of 588 (12×49), 360 (10×36), and 250 (10×25) respectively. Figure 2b shows that the efficiency in early language learning can still be observed. Each TnD model outperforms the CLM baseline of the same size, and even surpasses CLM baselines of large capacity in early steps.

Teacher's word preferences in demonstrations affect student To explore how the teacher model's word selection impacts students' language development, we repeat the experiments where a chosen set of 40 words for each test vocabulary is excluded from teacher demonstrations. During the language generation process by the teacher model,



(a) Learning curves of each baseline of CDI words on BabyLM Corpus.

(b) Learning curves of smaller models of CDI words on BabyLM Corpus.

(c) Influence of teacher's word preferences in CDI words on BabyLM Corpus.

(d) Per-word learning curve and cumulative word frequency for the word "have."

Figure 2. On 2 training corpora and 2 test vocabulary, we aggregate 5 random seeds for the plots. (a) The fitted learning curves of mean surprisal over $\log_{10}$ training steps, with nAoA@0.5 of each curve indicated by a vertical dashed line. (b) The fitted learning curves on smaller student models ($\texttt{TnD}_{d=588/360/250}$) of mean surprisal over $\log_{10}$ training steps, with nAoA@0.5 of each curve indicated by a vertical dashed line. (c) For the 40 words to be "masked out" from teacher demonstrations, we repeat the experiment with 5 random seeds and plot the learning curves of these words with those in CLM and TnD baselines. (d) Per-word learning curve and frequency. The dashed lines mark the log frequency of words (left y-axis) from each source. The solid line and dots mark the word surprisal (right y-axis).

POS	Freq.	BabyLM Corpus				BookCorpus			
		CMN		CDI		CMN		CDI	
		β	r	β	r	β	r	β	r
noun	trial	-0.36	-0.90	-0.25	-0.85	-0.38	-0.92	-0.31	-0.85
	demo	**-0.67**	-0.93	**-0.73**	-0.89	**-0.56**	-0.93	**-0.67**	-0.87
	corpus	-0.51	-0.93	-0.53	-0.88	**-0.56**	-0.93	-0.60	-0.87
pred	trial	**-0.70**	-0.90	**-0.72**	-0.86	**-0.49**	-0.92	**-0.54**	-0.88
	demo	-0.33	-0.93	-0.30	-0.90	**-0.49**	-0.92	-0.45	-0.90
	corpus	-0.22	-0.93	-0.19	-0.90	-0.45	-0.93	-0.40	-0.87
func	trial	**-0.67**	-0.93	**-0.72**	-0.92	**-0.67**	-0.94	**-0.59**	-0.93
	demo	-0.39	-0.92	-0.21	-0.90	-0.22	-0.91	-0.37	-0.87
	corpus	-0.17	-0.92	-0.25	-0.91	-0.35	-0.92	-0.35	-0.90

Table 1. For each POS category, we present the beta weights β and Pearson correlation r between their mean surprisal and cumulative word frequency over the course of training. These metrics are evaluated based on the student's trials, teacher's demonstrations, and the overall corpus frequency up to the current training step.

if a word from this set is to be decoded, we select the next best alternative, ensuring these words were never presented in teacher demonstrations. We depict the learning curves for these excluded words in Figure 2c and present the nAoA in Table 4. Our findings indicate that the absence of words from teacher demonstrations leads to slower learning speed for the student models, as evidenced by a higher nAoA.

Practice makes perfect in trials Finally, we conduct experiments to underscore the significance of the student's active trials in the process of word learning. A student model can learn a word from 3 sources: its own trials, teacher demonstrations, and exposure to the corpus. To determine which source contributes most significantly to learning, we begin by plotting the per-word learning curves alongside the cumulative frequency of word encounters in trials, demonstrations, and the corpus. We observe, interestingly, that the learning curves for certain words exhibit a pronounced correlation with the frequency of these words in trials, as exampled in Figure 2d and 11. We speculate that this pattern may be associated with the part of speech (POS) of the word, including nouns, predicates, and functional words, to explore the relationship further. To evaluate the impact of each source of word acquisition, we consider the cumulative word frequency as a predictor of the word surprisal, and carry out linear regressions complemented by likelihood ratio tests to determine the beta weights for each predictor, and summarize the results in Table 1. Negative beta weights signify a negative correlation, with a larger absolute value denoting a stronger association and contribution. Our analysis reveals that the cumulative frequency of words encountered in trials plays a significant role in the acquisition of functional words and predicates. However, this significant contribution does not extend to nouns, indicating a potential impact of active trials on different POS within the learning process. This finding is linguistically intuitive, as function words and predicates are words that require other dependent words to fully express their meaning (Gleitman, 1990) and thus require more practice. We posit that grounding language in the non-linguistic world is essential for acquiring the meanings of words, particularly for concrete noun (Ma et al., 2023).

5. Conclusion

We introduce a trial-and-demonstration (TnD) learning framework to examine the role of corrective feedback in neural word acquisition through systematically controlled experiments, assessing how the interplay between student trials and teacher demonstrations contributes to learning efficiency in neural language models. We find that (1) TnD learning accelerates neural word acquisition across student models of different sizes; (2) the teacher's choices of words influence students' word-specific learning efficiency; and (3) a practice-makes-perfect effect is evident by a strong correlation between the frequency of words in trials and their learning curves. Our findings confirm the crucial role of interaction in efficient neural word learning.

Acknowledgement.

This work was supported in part by NSF IIS-1949634 and NSF SES-2128623. The authors extend their special appreciation to Susan Gelman and Freda Shi for their valuable feedback (both communicatively and correctively). The authors would like to thank Run Peng, Yichi Zhang, Jacob Sansom, Zheyuan Zhang, Xuejun Zhang for proofreading and their helpful input during discussions, and all anonymous reviewers for their valuable feedback.

Limitations

Iterative setting. This experiment can be conducted iteratively by replacing the teacher model with the student model from previous iterations. While this iterative approach is intriguing, it introduces new complexities that require significant modifications to the current controlled ablation studies. We defer exploration of this approach to future work, as the current study focuses on examining the roles of trials and demonstrations.

The reward. Our study is limited by the use of a single reward model focused on corrective feedback. More realistic scenarios should also encompass communicative feedback, with the success of communication serving as a reward. Additionally, Thorndike (1911) proposed the idea that a child might instinctively feel satisfaction from producing a sound that echoes a meaningful memory. The design of such an intrinsic reward model is cognitively intriguing, and could aid in scaling student models without an external reward model, offering potential benefits for engineering applications.

The reward model. We employ a robust language model (LLaMA-2-7B) as a reward model to concentrate on the roles of trials and demonstrations without concerns about reward quality. Future research should explore the impact of using less accurate reward models.

The tokenizer. One limitation of our approach is the reliance on the Byte Pair Encoding (BPE) tokenizer (Sennrich et al., 2016) inherited from GPT-2. Ideally, the method should be tokenizer-free to facilitate the learning of early language elements such as sound effects and animal sounds, e.g., "baa-baa" in CDI (Fenson et al., 2006), which can be crucial for a more natural and foundational language acquisition process.

Other languages. The present study focuses on English as the subject of investigation due to the available corpus resources. Future research should consider exploring other languages.

Impact Statement

This work does not depend on human annotators or human subjects for interactive experiments. We leverage open datasets and model-generated content for training that could contain biases and sensitive contents inherited, which may cause fairness issues in the final model when applied to practical applications. Future research should be done to look into these issues, potentially by designing fairness-aware reward models.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Baroni, M. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language*, pp. 1–16, 2022.

Bates, E., Camaioni, L., and Volterra, V. The acquisition of performatives prior to speech. *Merrill-Palmer quarterly of behavior and development*, 21(3):205–226, 1975.

Bianchi, F., Greco, C., and Tagliabue, J. Language in a (search) box: Grounding language learning in real-world human-machine interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4409–4415, 2021.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N., and Turian, J. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8718–8735. Association for Computational Linguistics, November 2020.

Bitchener, J. and Ferris, D. R. Written corrective feedback in second language acquisition and writing. *Routledge, Taylor & Francis Group*, 2011.

Blevins, T., Gonen, H., and Zettlemoyer, L. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3575–3590, 2022.

Bohn, M. and Frank, M. C. The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, 1:223–249, 2019.

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *Cognitive Science*, 2016.

Brown, R. Derivational complexity and order of acquisition. *Cognition and Development of Language*, 1970.

Bruner, J. Child's talk: Learning to use language. *Child Language Teaching and Therapy*, 1(1):111–114, 1985.

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

Chang, T. A., Tu, Z., and Bergen, B. K. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *arXiv preprint arXiv:2308.15419*, 2023.

Chouinard, M. M. and Clark, E. V. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669, 2003.

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

Clark, H. H. *Using language*. Cambridge university press, 1996.

El Tatawy, M. Corrective feedback in second language acquisition. *Studies in Applied Linguistics and TESOL*, 2 (2), 2002.

Ellis, R., Loewen, S., and Erlam, R. Implicit and explicit corrective feedback and the acquisition of l2 grammar. *Studies in second language acquisition*, 28(2):339–368, 2006.

Evanson, L., Lakretz, Y., and King, J.-R. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12205–12218, 2023.

Farrar, M. Negative evidence and grammatical morphemes. *Developmental Psychology*, 28:91–99, 1992.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., and Bates, E. Macarthur-bates communicative development inventories. *PsycTESTS Dataset*, 2006.

Gleitman, L. The structural sources of verb meanings. *Language acquisition*, 1(1):3–55, 1990.

Halliday, M. A. K. Learning how to mean. In *Foundations of language development*, pp. 239–265. Elsevier, 1975.

He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., and Ostendorf, M. Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1621–1630, 2016.

Hiller, S. Corrective feedback in first language acquisition. Master's thesis, Universiteit van Amsterdam, 2016.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Kuhl, P. K. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.

Lazaridou, A., Pham, N. T., and Baroni, M. Towards multi-agent communication-based language learning. *arXiv preprint arXiv:1605.07133*, 2016.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Liu, A., Zhu, H., Liu, E., Bisk, Y., and Neubig, G. Computational language acquisition with theory of mind. In *The Eleventh International Conference on Learning Representations*, 2022.

Ma, Z., Pan, J., and Chai, J. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 524–544, 2023.

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

Madotto, A., Namazifar, M., Huizinga, J., Molino, P., Ecoffet, A., Zheng, H., Yu, D., Papangelis, A., Khatri, C., and Tur, G. Exploration based language learning for text-based games. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1488–1494, 2021.

Marcus, G. F. Negative evidence in language acquisition. *Cognition*, 46(1):53–85, 1993.

Narasimhan, K., Kulkarni, T. D., and Barzilay, R. Language understanding for text-based games using deep reinforcement learning. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pp. 1–11. Association for Computational Linguistics (ACL), 2015.

Nikolaus, M. and Fourtassi, A. 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, pp. 391–407, 2021.

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

OpenAI. Chatgpt: Optimizing language models for dialogue, Nov 2022. URL https://openai.com/blog/chatgpt/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Palincsar, A. S. The role of dialogue in providing scaffolded instruction. *Educational psychologist*, 21(1-2):73–98, 1986.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Portelance, E. *Neural Network Approaches to the Study of Word Learning*. PhD thesis, Stanford University, 2022.

Portelance, E. and Jasbi, M. The roles of neural networks in language acquisition. *PsyArXiv*, 2023.

Portelance, E., Degen, J., and Frank, M. C. Predicting age of acquisition in early word learning using recurrent neural networks. In *CogSci*, 2020.

Portelance, E., Frank, M. C., and Jurafsky, D. Learning the meanings of function words from grounded language using a visual question answering model. *Cognitive Science*, 48(5):e13448, 2024.

Qu, S. and Chai, J. Y. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:247–277, 2010.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Saxton, M. Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language*, 20(60):221–252, 2000.

Saxton, M., Backley, P., and Gallaway, C. Negative input for grammatical errors: Effects after a lag of 12 weeks. *Journal of child language*, 32(3):643–672, 2005.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I. R., Eisenstein, J., et al. The multiberts: Bert reproductions for robustness analysis. In *International Conference on Learning Representations*, 2021.

Senju, A. and Csibra, G. Gaze following in human infants depends on communicative signals. *Current biology*, 18 (9):668–671, 2008.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725. Association for Computational Linguistics (ACL), 2016.

Shi, H., Mao, J., Gimpel, K., and Livescu, K. Visually grounded neural syntax acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1842–1861, 2019.

Smith, L. and Yu, C. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106 (3):1558–1568, 2008.

Snow, C. E., Pan, B. A., Imbens-Bailey, A., and Herman, J. Learning how to say what one means: A longitudinal study of children's speech act use. *Social Development*, 5(1):56–84, 1996.

Stein, K., Harter, L., and Geiger, L. Shapelurn: An interactive language learning game with logical inference. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pp. 16–24, 2021.

ter Hoeve, M., Kharitonov, E., Hupkes, D., and Dupoux, E. Towards interactive language modeling. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, 2022.

Thorndike, E. *Animal intelligence: Experimental studies*. New York: Macmillan, 1911.

Tomasello, M. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press, 2005.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsuji, S., Cristia, A., and Dupoux, E. Scala: A blueprint for computational models of language acquisition in social context. *Cognition*, 213:104779, 2021.

Vygotsky, L. S. *Thought and language*. MIT press, 1934.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. 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*, pp. 353–355, 2018.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., and Oller, D. K. A social feedback loop for speech development and its reduction in autism. *Psychological science*, 25(7): 1314–1324, 2014.

Warstadt, A. and Bowman, S. R. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pp. 17–60, 2022.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., et al. 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*, 2023.

Weston, J. E. Dialog-based language learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., and Stoyanov, V. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13738, 2023.

Yuan, X., Côté, M.-A., Fu, J., Lin, Z., Pal, C., Bengio, Y., and Trischler, A. Interactive language learning by question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2796–2813, 2019.

Yurovsky, D. and Frank, M. C. Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, 20(2):e12349, 2017.

Zhang, H., Yu, H., and Xu, W. Interactive language acquisition with one-shot visual concept learning through a conversational game. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2619, 2018.

Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., et al. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023.

Zhu, H., Bisk, Y., and Neubig, G. Language learning from communicative goals and linguistic input. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Related Work and Discussions

A.1. Interaction in neural language learning

Researchers have emphasized the role of interaction in computational models of language (Bisk et al., 2020; Tsuji et al., 2021). Preliminary efforts have been conducted under specific constraints, such as in domain-specific scenarios (Qu & Chai, 2010; Weston, 2016; Bianchi et al., 2021; Stein et al., 2021; Madotto et al., 2021) or considering particular types of dialogue acts (Zhang et al., 2018; Yuan et al., 2019). More recently, a series of studies have approached language acquisition through the lens of multimodal referential games (Lazaridou et al., 2016; Zhu et al., 2022; Liu et al., 2022), emphasizing the importance of pragmatic inference and communicative feedback in speaker-listener interactions. Early work investigated scenarios where the teacher models actively select training data to optimally assist a passive student learner (ter Hoeve et al., 2022). Nikolaus & Fourtassi (2021) adopted setups where student models learn by producing language and receive feedback of communicative success using the BLEU score (Papineni et al., 2002). However, these models do not receive teachers' demonstrations or corrections. Our work diverges from these studies as we focus on symmetric teacher-student interactions that require language production from the student and explicit corrective feedback from the teacher, which goes beyond the simple speaker-listener roles. We also adopt massive corpora to study modern transformer-based generative language models, a general approach without domain restrictions. Recent advancements in reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) mark a breakthrough in interactive language learning. This work leverages the implementation infrastructure of RLHF, but diverges in significant ways. Whereas RLHF aims to align a pre-trained language model with human preferences, our objective is to "babysit" a language model from scratch. We refer to existing surveys (Kaufmann et al., 2023; Zheng et al., 2023).

A.2. Interaction in human language learning

To unveil the role of social interactions in language acquisition, a prominent body of research has focused on *pragmatic inference* – that children exhibit the ability to refine their linguistic knowledge by inferring the communicative intents of others (Senju & Csibra, 2008; Yurovsky & Frank, 2017; Bohn & Frank, 2019). Besides children's pragmatic inference about the speakers, many researchers have argued for the importance of caregivers' *feedback* to the language produced by children (Warlaumont et al., 2014), in the form of descriptions, explanations, corrections, etc, to human language development. For example, extensive efforts have been made to examine the effects of *communicative feedback* on language acquisition, both in developmental psychology (Bates et al., 1975; Snow et al., 1996; Nikolaus & Fourtassi, 2023) and computational modeling (Nikolaus & Fourtassi, 2021; Liu et al., 2022). This type of feedback emphasizes the explicit negotiation of mutual understanding with the conversational partner to achieve and maintain common ground (Clark, 1996). While communicative feedback emphasizes the success and failure of communication, the feedback we study in this study is more akin to *corrective feedback*. This type of feedback involves responses from caregivers, which offer corrections to possible errors in children's speech, including variants such as negative evidence, reformulations, or recasts (Farrar, 1992; Saxton, 2000; Chouinard & Clark, 2003; Saxton et al., 2005; Hiller, 2016). Although corrective feedback is shown to be helpful in second language acquisition (El Tatawy, 2002; Ellis et al., 2006; Bitchener & Ferris, 2011), researchers largely dispute its availability and effectiveness in human first language acquisition (Brown, 1970; Marcus, 1993). While our findings suggest that corrective feedback through demonstrations can enhance efficient neural language learning, these results should not be generalized to human language learning, where demonstrations are much less frequent in the noisy feedback that children typically receive.

👶[CHILD]	*what you doing round here ?*	
👩[MOTHER]	*what **are** you doing round here ?*	(Corrective)

Figure 3. A natural example of corrective feedback in child-parent conversation from the CHILDES corpus (MacWhinney, 2000).

A.3. Human cognition and language models

Although language models and human language learners differ in their inductive biases and data sources (Baroni, 2022; Warstadt & Bowman, 2022), several works have looked into the learnability, proficiency, and efficiency of neural language learning, for example, the relationships between word surprisal in language models to various psycholinguistic variables (Portelance et al., 2020; Ma et al., 2023). Recent efforts have shifted towards exploring the developmental trajectories of language models, rather than their end performance (Sellam et al., 2021; Blevins et al., 2022; Biderman et al., 2023; Xia et al., 2023), sparkling further investigations into the developmental aspects of psycholinguistics using computational approaches (Chang & Bergen, 2022; Chang et al., 2023; Evanson et al., 2023). The scientific rationale behind this is that these models can serve as hypothesis generators or proofs of concept, verifying mechanisms that are practically effective for machines and potentially feasible for human learners (Portelance, 2022; Portelance & Jasbi, 2023). We echo that the benchmark outcomes of language models themselves are insufficient (Baroni, 2022; Portelance, 2022), and researchers need to control the factors that may have contributed to the model's learning process, particularly through conducting ablation studies (Warstadt & Bowman, 2022). In this study, we follow this spirit and investigate corrective feedback through systematic computational experiments, assessing its role in neural word acquisition.

B. Trials-and-Demonstrations (TnD) Details

B.1. Relationship with RLHF

Reinforcement learning enables language systems to learn from feedback in the form of rewards from games (Narasimhan et al., 2015; He et al., 2016) or heuristic scores (Ranzato et al., 2016; Nikolaus & Fourtassi, 2021). Recent advancements in reinforcement learning from human feedback (RLHF) have generated considerable excitement, especially in its application to large language models such as ChatGPT (OpenAI, 2022). Reinforcement learning is employed to align the model's policy with human preferences, utilizing human-annotated preference data (Ziegler et al., 2019; Ouyang et al., 2022) or AI models acting as proxies for human judgment (Bai et al., 2022; Lee et al., 2023). We refer to Kaufmann et al. (2023) and Zheng et al. (2023) for more details. Our work leverages the implementation infrastructure of RLHF but diverges in significant ways: whereas RLHF aims to align an existing language model with human preferences, our objective is to "babysit" a language model from scratch using reinforcement learning, specifically to model the process of receiving and integrating corrective feedback.

B.2. Preliminaries

Inspired by recent work in RLHF, we use Proximal Policy Optimization (PPO) (Schulman et al., 2017) to train the student model.[1] Consider a model π_θ whose the current state s_i is a sequence of tokens $s_i = t_1, \cdots, t_i$, and receives a reward r_i. We outline the key components and refer to Zheng et al. (2023) for more details.

Clipped surrogate objective. The clipped surrogate objective is defined as:

$$\mathcal{L}_\theta^{\text{pg}} = \mathbb{E}_i\big[\min\big(\Pi_i(\theta), \text{CLIP}\big(\Pi_i(\theta), \epsilon\big)\big)A_i\big] \tag{1}$$

where

$$\Pi_i(\theta) := \frac{\pi_\theta(t_{i+1} \mid s_i)}{\pi_{\theta_{\text{old}}}(t_{i+1} \mid s_i)} \tag{2}$$

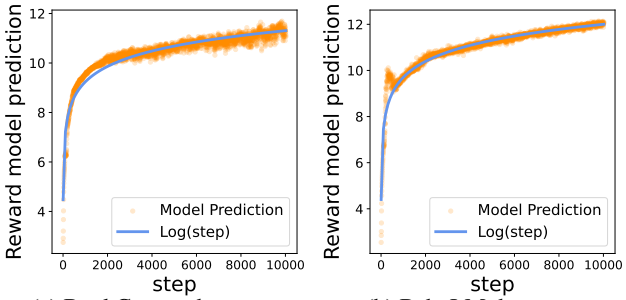
The CLIP function clips the value within $(1-\epsilon, 1+\epsilon)$, which regularizes the policy from drastic changes to ensure robustness.

Generalized advantage estimation. The advantage function A_i at step i is estimated by the Generalized Advantage Estimation (GAE) algorithm (Schulman et al., 2016) for a balanced bias-variance trade-off:

$$A_i = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta_{i+k} \tag{3}$$

where γ is the discount factor, λ is a hyperparameter controlling the trade-off between bias and variance, and $\delta_i = r_i + \gamma V(s_{i+1}) - V(s_i)$ is the temporal difference (TD) error. The reward r is defined in **??**.

[1]We use the TRL library to train language models with reinforcement learning.



(a) BookCorpus dataset. (b) BabyLM dataset.

Figure 4. We sample reward model predictions at different steps and compare them to ground truth logarithm. The reward models are satisfactory as the model predicted age/step highly overlaps with the true age/step.

Proximal policy optimization. During the same time, PPO also estimates and optimizes its value function $V_{\theta_{vh}}$ with MSE loss:

$$\mathcal{L}_{\theta_{vh}}^{\text{value}} = \mathbb{E}_i\left[\left(V_{\theta_{vh}}(s_i) - \hat{V}_i\right)^2\right] \tag{4}$$

where $V_{\theta_{vh}}(s_i)$ is the estimated value, and $\hat{V}_i$ is the target value from GAE. Differing from conventional implementation, our approach didn't employ a reference model for the KL penalty. This deviation emphasizes the language model's evolutionary nature during its training, enabling more significant updates and eliminating biases from over-adherence to a reference model. The final reinforcement learning loss is a linear combination:

$$\mathcal{L}_\theta^{\text{ppo}} = \mathcal{L}^{\text{pg}}(\theta) + c \cdot \mathcal{L}_{\theta_{vh}}^{\text{value}} \tag{5}$$

where $c \in [0, 1]$.

B.3. The reward and reward model

To train a neural age predictor, we utilize the developmental trajectory of the teacher model by saving over 100 checkpoints at various training steps. For each checkpoint, we select 25,000 contexts (each consisting of 5 tokens) from the test set and prompt the model to generate continuations for each context. This process generates a dataset comprising over 2.5 million (text, step) pairs. We then use this dataset to fine-tune a LLaMA-2-7B language model (Touvron et al., 2023), incorporating a 1-layer linear head for regression. At step n, the student model π_θ parameterized by θ produces a sentence $\mathcal{S}$ of l tokens $t_1, \cdots, t_l$. We use the neural age predictor R_ϕ to estimate the logarithm of the expected training step $\hat{n}$ where this sentence typically emerges. Hence, the age-conditioned reward $r(\mathcal{S}, n)$ is given by:

$$r := \log(\hat{n}/n) = R_\phi(\mathcal{S}) - \log n \tag{6}$$

B.4. Alternating interactive and non-interactive language learning

As shown in Figure 1(c-d), our TnD framework alternates between two forms of language learning: (1) interactive learning, in which corrective feedback is taken through reinforcement learning, utilizing rewards derived from both the student's trials and the teacher's demonstrations, and (2) non-interactive learning, which emulates the natural language exposure experienced by learners and is facilitated through causal language modeling as adopted by generative language models.

Algorithm 1 TRIAL-AND-DEMONSTRATION

Input: student model π_θ, teacher model $\hat{\pi}_\varphi$, reward model R_ϕ, alternating schedule (c, r), training corpus $\mathcal{C}$.
for $n, \mathcal{S}$ **in enumerate** $(\mathcal{C})$ **do**
  $t_1, \cdots, t_l = \text{TOKENIZE}(\mathcal{S})$
  if $n\%(c+r) \leq c$ **then**
    // Non-interactive learning
    Gradient descent $\nabla_\theta \mathcal{L}_\theta^{\text{clm}}([t_1, \cdots, t_l])$ (Eq. 7)
  else
    // Interactive learning
    $\mathcal{S}_{\text{trial}} \leftarrow \text{PROMPT}\big(\pi_\theta, [t_1, t_2, \cdots, t_5]\big)$
    $r_{\text{trial}} = R_\phi(\mathcal{S}_{\text{trial}}) - \log n$
    $\mathcal{S}_{\text{demo}} \leftarrow \text{PROMPT}\big(\hat{\pi}_\varphi, [t_1, t_2, \cdots, t_5]\big)$
    $r_{\text{demo}} = R_\phi(\mathcal{S}_{\text{demo}}) - \log n$
    Train batch $B = (S_{\text{trial}}, r_{\text{trial}}) \cup (S_{\text{demo}}, r_{\text{demo}})$
    Policy update $\nabla_\theta \mathcal{L}_\theta^{\text{ppo}}(B)$ (See B.2)
  end if
end for
Output: π_θ

B.4.1. INTERACTIVE LANGUAGE LEARNING.

Reinforcement learning has been applied to language models, such as the success of games (Narasimhan et al., 2015; He et al., 2016), heuristic scores (Ranzato et al., 2016; Nikolaus & Fourtassi, 2021), and models of human preferences (Ziegler et al., 2019; Ouyang et al., 2022). The key is to view the language production in language models as actions within a vocabulary-defined action space. Formally, the model π_θ is given a context of 5 tokens (i.e., the initial state $s_0 = \{t_1, \cdots, t_5\}$) to produce the next token (i.e., the next action $a_1 = t_6$). It lands in the next state $s_1 = \{t_1 \cdots t_6\}$, and this process repeats until the sentence concludes. Following this, rewards can be computed according to Eq. 6 for the student's trials and the teacher's demonstrations. The goal is to maximize the expected return (i.e., the expected cumulative future reward) following π_θ along the interaction, with the trials and demonstrations both in the training batch. Inspired and taking the computational infrastructure in recent advances in reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), we use Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm with a clipped surrogate objective $\mathcal{L}_\theta^{\text{ppo}}$. We applied two modifications in the implementation of the RLHF algorithm, which include the involvement of demonstration in policy update, and the removal of KL-divergence.

Demonstration in policy update. The original PPO algorithm only learns from the agent's own trials, i.e., the sentences it generated before. We expand the training batch to add teacher's demonstrations: At each step, we sample the sentences generated by the latest student model and also collect those from the teacher model with the same prompts. Subsequently, we regard both of them as the training batch, and apply the policy update to improve the student model. Intuitively, our goal is to encourage the student model to imitate and repeat the teacher's demonstration during training. Through the reward disparity between their own trial and the demonstration, we motivate the student model to learn a better language generation policy.

Removal of KL-divergence objective. The loss function in the conventional RLHF algorithm involves a KL-divergence term between the updated student model and a reference model, which is usually a fine-tuned language model or the initial student model before gradient updates. The goal is to penalize the learned policy that largely deviates from the referenced policy. Different from the conventional approach, this penalty is not preferred in the TnD setting. This change encourages the student model to explore unfamiliar words during the training, which enables relatively significant updates, as well as eliminates biases from over-adherence to a reference model.

B.4.2. NON-INTERACTIVE LANGUAGE LEARNING.

While our interactive language learning replicates active engagement with language through corrective feedback, it's also crucial to simulate human's passive exposure to language, as emphasized in prior psychological (Smith & Yu, 2008) and computational (Nikolaus & Fourtassi, 2021) studies. We implement this with the causal language modeling objective, which is adopted by most generative language models. Consider a sentence $t_1, \cdots, t_l$ in the corpus with l tokens. The causal language modeling objective is to predict the next token t_{i+1} given the previous context $t_{\leq i}$ by minimizing:

$$\mathcal{L}_\theta^{\text{clm}} = -\sum_{i=1}^{l-1} \log P(t_{i+1} | t_{\leq i}; \theta) \tag{7}$$

B.4.3. ALTERNATING INTERACTIVE AND NON-INTERACTIVE LEARNING.

Following the setup of Nikolaus & Fourtassi (2021), we adopt an alternating learning schedule over interactive and non-interactive language learning, i.e., perform c steps of causal language modeling, followed by r steps of reinforcement learning, in a continuous cycle (Algorithm 1). We select $c = 3$ and $r = 1$, respectively, leading to 1 PPO update following every 3 CLM updates. We justify this design choice and present experiments on other hyper-parameters in Appendix D.7.

C. Experiments, Results, and Discussions

C.1. Evaluation

Testing vocabulary. We specify two sets of vocabulary for evaluation.

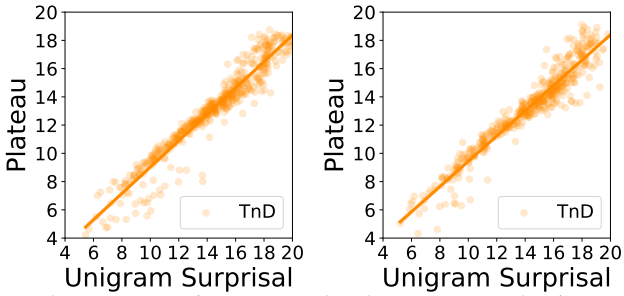
- The `CMN` set, consisting of common words that appear frequently in both corpora, covering a wide range of words and parts of speech.
- The `CDI` set, consisting of words from the MacArthur-Bates Communicative Development Inventories (CDIs) (Fenson et al., 2006). Following Portelance et al. (2020), we excluded any items comprising multiple words (such as "choo-choo") from our dataset, as the tokenizer would treat these as distinct items.

We select these two vocabulary sets as `CMN` offers broader coverage while `CDI` is more frequency-neutral and is used to assess children's early vocabulary development. Following Chang & Bergen (2022), we remove words with less than 100 occurrences in the evaluation set of each corpus to ensure reliable results, and keep at most 512 samples for each word. This leads to 309 common words in `CMN`, as well as 345 words in `CDI` for BookCorpus and 243 words in `CDI` for BabyLM Corpus. We include the list of words and the evaluation datasets within the code.[2]

Surprisal and learning curves. In line with previous work, we use the mean surprisal (log-perplexity) of a word to quantify the quality of the model's predictions for this word. For each occurrence of a word w, the surprisal is given by $-\log_2 P(w)$, and we average all occurrences. To visualize word acquisition from high surprisal to lower surprisal, we evaluate the student model throughout training and plot the surprisal values over logarithm training steps, leading to a learning curve for individual words and the overall vocabulary. We observe a similar pattern reported by Chang & Bergen (2022) that learning curves tend to level off at a local plateau, which aligns with the unigram surprisal. This phenomenon renders the single-sigmoid model unreliable for capturing the complexity, and we adopt a double-sigmoid function to fit the learning curve. We run regression between the plateau of double-sigmoid curves and the unigram surprisals calculated from all sources of word occurrences (Figure 5). We find a strong correspondence between the plateau and unigram frequency, suggesting that the double-sigmoid function is a better option than the single-sigmoid function to fit learning curves. To analyze longer learning curves of up to 1M steps, more complex functions such as linear GAMs have been adopted (Chang et al., 2023). For our purposes, a double-sigmoid function suffices.

Neural age of acquisition (nAoA). To evaluate the speed at which the student model acquires a word, we employ the neural age of acquisition (nAoA). Prior research (Chang & Bergen, 2022; Chang et al., 2023) has used a surprisal cut-off of 50% between the minimum and maximum surprisal levels. This is akin to the method used to determine children's age of acquisition, where the cutoff is set at the point when 50% of children are observed to produce a word (Braginsky et al., 2016). To further enhance the robustness of this metric, we average nAoA over different surprisal thresholds from 0.5 to 0.95 with a step of 0.05, denoted as `nAoA@[0.50:0.95]`. It's important to note that nAoA serves as a metric to assess the speed at which a model acquires a word, rather than the quality of word learning. A model might learn a word quickly, indicated by a low nAoA, yet not master it effectively, indicated by a high surprisal. Combining both metrics offers a thorough evaluation of word learning.

[2]Our code and data will be made publicly available upon the acceptance of this work.



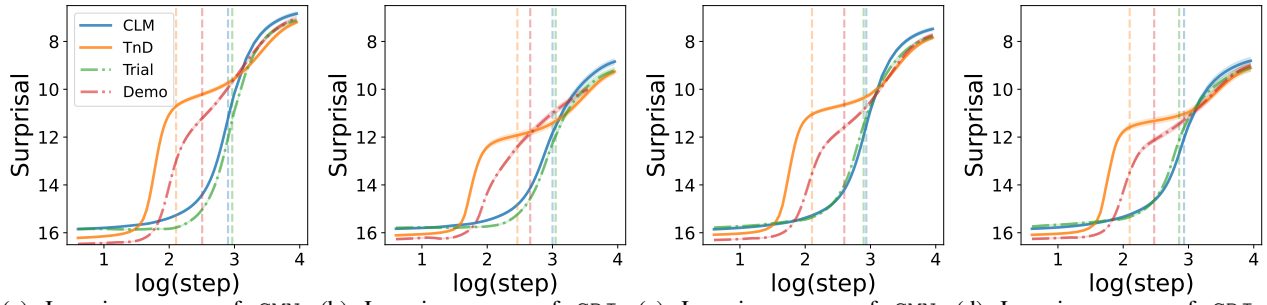
(a) Plateau v.s. unigram surprisal in BookCorpus. (b) Plateau v.s. word unigram surprisal in BabyLM.

Figure 5. Sigmoid plateau v.s. unigram surprisal.

C.2. Detailed Setups and Reproducibility

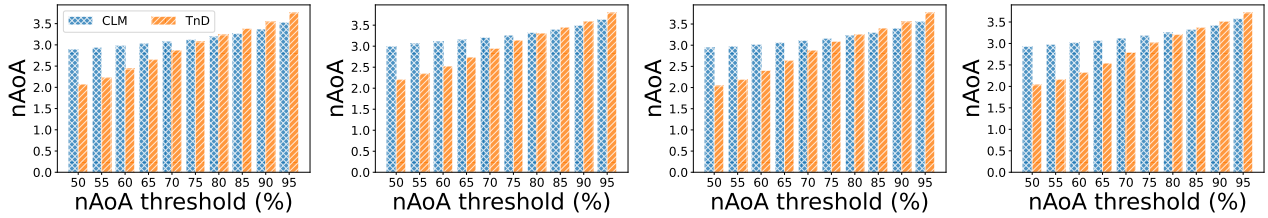
Training details. We randomly initialize the GPT-2 student models. For each combination of corpus and baseline, the training process is conducted 5 times for 10k steps, each with a different random seed. We utilize the top-k decoding strategy for language generation, setting $k = 20$. The learning rate for reinforcement learning is set to $2e^{-5}$. We follow the default setting for other PPO hyper-parameters, such as a clip range ($\epsilon = 0.2$) from the TRL library. All other hyper-parameters for causal language modeling, including the learning rate set at $1e^{-4}$ and a batch size of 128, remain consistent with the training setup of GPT-2 by Chang & Bergen (2022).

Checkpointing. We save the intermediate steps: [2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410, 420, 430, 440, 450, 460, 470, 480, 490, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000].



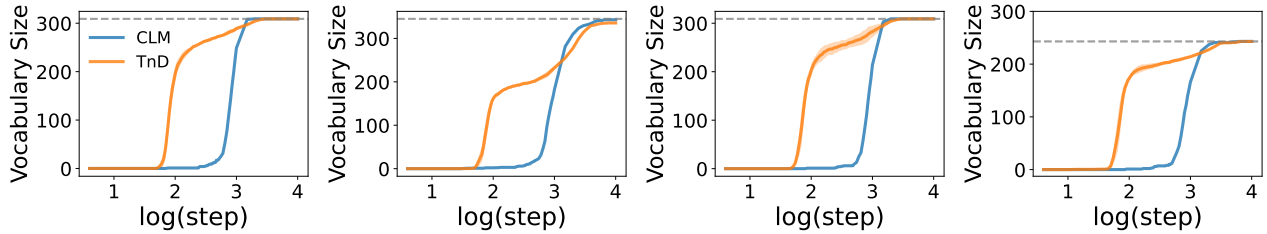
(a) Learning curve of CMN words on BookCorpus. (b) Learning curve of CDI words on BookCorpus. (c) Learning curve of CMN words on BabyLM. (d) Learning curve of CDI words on BabyLM.

Figure 6. On 2 training corpora and 2 test vocabulary, we aggregate 5 random seeds and present the fitted learning curves of mean surprisal over $\log_{10}$ training steps, with nAoA@0.5 of each curve indicated by a vertical dashed line.



(a) nAoA of CMN words on BookCorpus. (b) nAoA of CDI words on BookCorpus. (c) nAoA of CMN words on BabyLM. (d) nAoA of CDI words on BabyLM.

Figure 7. On 2 training corpora and 2 test vocabulary, we aggregate 5 random seeds and present the neural age of acquisition (nAoA) at different surprisal thresholds from 0.5 to 0.95 with a step of 0.05.



(a) Effective vocabulary size in CMN words on BookCorpus. (b) Effective vocabulary size in CDI words on BookCorpus. (c) Effective vocabulary size in CMN words on BabyLM. (d) Effective vocabulary size in CDI words on BabyLM.

Figure 8. On 2 training corpora and 2 test vocabulary, we aggregate 5 random seeds and evaluate the effective vocabulary size over $\log_{10}$ training steps. The dashed lines mark the tested vocabulary size.

Method	nAoA↓	BabyLM Corpus		BookCorpus	
		CMN	CDI	CMN	CDI
CLM	@.5	$2.94_{\pm 0.01}$	$2.93_{\pm 0.01}$	$2.90_{\pm 0.01}$	$3.00_{\pm 0.01}$
	@[.5:.95]	$3.18_{\pm 0.01}$	$3.19_{\pm 0.01}$	$3.14_{\pm 0.01}$	$3.26_{\pm 0.01}$
Trial	@.5	$2.90_{\pm 0.01}$	$2.86_{\pm 0.01}$	$2.97_{\pm 0.01}$	$3.05_{\pm 0.01}$
	@[.5:.95]	$3.20_{\pm 0.01}$	$3.17_{\pm 0.01}$	$3.21_{\pm 0.01}$	$3.30_{\pm 0.01}$
Demo	@.5	$2.60_{\pm 0.02}$	$2.47_{\pm 0.03}$	$2.51_{\pm 0.02}$	$2.66_{\pm 0.02}$
	@[.5:.95]	$3.15_{\pm 0.01}$	$3.08_{\pm 0.02}$	$3.05_{\pm 0.01}$	$\mathbf{2.97}_{\pm 0.02}$
TnD	@.5	$\mathbf{2.10}_{\pm 0.02}$	$\mathbf{2.10}_{\pm 0.03}$	$\mathbf{2.11}_{\pm 0.02}$	$\mathbf{2.46}_{\pm 0.03}$
	@[.5:.95]	$\mathbf{2.95}_{\pm 0.02}$	$\mathbf{2.87}_{\pm 0.03}$	$\mathbf{2.90}_{\pm 0.02}$	$3.11_{\pm 0.02}$

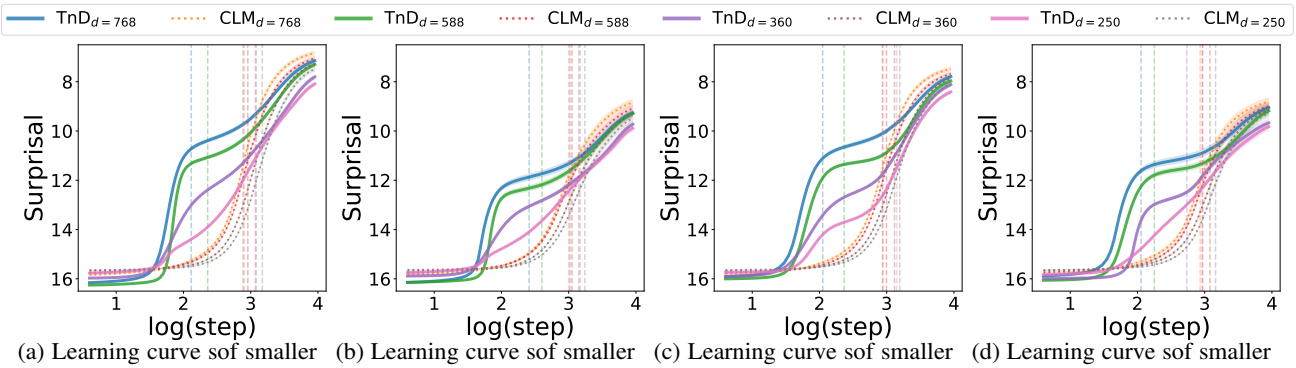
Table 2. For each baseline and setup, we report the neural age of acquisition (nAoA) with standard errors at 0.50 cutoff and averaged over surprisal thresholds from 0.50 to 0.95 with a step of 0.05.

Computational resources. In TnD training, each experiment to train a student model using TnD is conducted on 2 A40 GPUs for 36 hours. In teacher model pre-training, we distribute the computation over 4 A40 GPUs with batch size 32 per device for 20 hours. To fine-tune the neural age predictor, we use a LLaMA-2-7B model (Touvron et al., 2023) with regression head on our developmental trajectory (text and step pairs) dataset. To ensure the prediction quality and save computation resources, we fine-tune our model using mixed precision on all parameters. The training is distributed over 8 A40 GPUs with batch size 8 per device with fully-shared data parallel. The learning rate is $5e^{-5}$.

D. Complete Results and Findings

D.1. Corrective feedback accelerates neural word acquisition

To evaluate CMN and CDI words on two corpora, we aggregate 5 random seeds and present the learning curves in Figure 6, the neural age of acquisition in Figure 7, and the effective vocabulary size in Figure 8. Figure 6 reveals that the TnD learning framework significantly accelerates word acquisition in training, outperforming other baselines. This acceleration is attributed to the critical roles of both



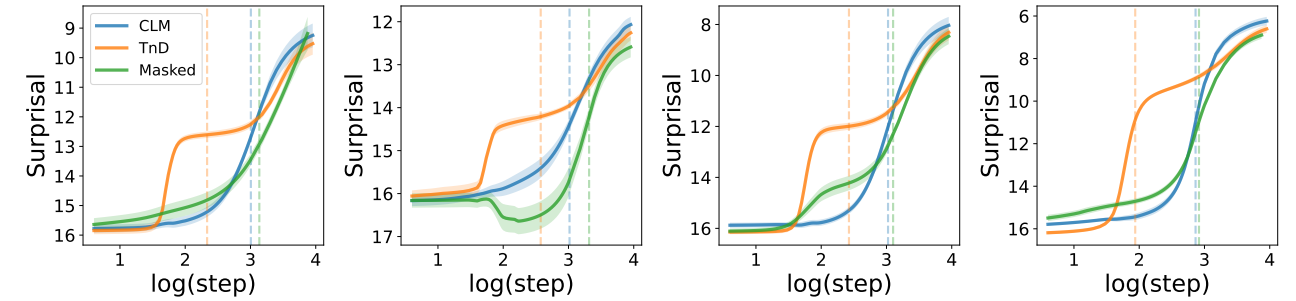
(a) Learning curve sof smaller models of CMN words on BookCorpus.

(b) Learning curve sof smaller models of CDI words on BookCorpus.

(c) Learning curve sof smaller models of CMN words on BabyLM.

(d) Learning curve sof smaller models of CDI words on BabyLM.

Figure 9. On 2 training corpora and 2 test vocabulary, we further present the fitted learning curves on smaller student models ($\texttt{TnD}_{d=588/360/250}$) of mean surprisal over $\log_{10}$ training steps, with `nAoA@0.5` of each curve indicated by a vertical dashed line.



(a) Influence of teacher's word preferences in CMN words on BabyLM.

(b) Influence of teacher's word preferences in CDI words on BabyLM.

(c) Influence of teacher's word preferences in CMN words on BookCorpus.

(d) Influence of teacher's word preferences in CDI words on BookCorpus.

Figure 10. For the 40 words to be "masked out" from teacher demonstrations, we repeat the experiment with 5 random seeds and plot the learning curves of these words with those in CLM and TnD baselines.

trials and demonstrations in the learning process. With only teacher demonstrations, the student model acquires words faster than with the plain CLM baseline alone, though not as rapidly as when active trials are incorporated in the TnD framework. Conversely, without the teacher's demonstrations, the student's trials in the wild do not yield a marked improvement, resulting in performance comparable to the CLM baseline. We refer to Figure 14 and 15 in the Appendix for the ridgeline and scatter plot of words and their nAoA.

Figure 7 and Table 2 present nAoA at different surprisal thresholds from 0.50 to 0.95 with a step of 0.05. We find that the TnD learning framework is particularly beneficial during the earlier stages of word acquisition, as it significantly outperforms the CLM baseline on `nAoA@0.50`, but is eventually on-par with the CLM baseline on `nAoA@0.95` towards the end of training. As a result, it can be seen from Figure 8 that students under the TnD framework quickly picked up a large volume of effective vocabulary, but eventually their vocabulary capacities have converged to the CLM baseline as expected. We present the ridgeline and scatter plot of words and their neural age of acquisition (nAoA) in BabyLM Corpus (Figure 14) and BookCorpus (Figure 15). The TnD model performs on par with the CLM model (See Appendix D.2). Overall, our findings show that corrective feedback accelerates the student model's neural word acquisition process, yet the student eventually converges to the teacher model's performance. This contradicts Nikolaus & Fourtassi (2021)'s previous observation that combining production and perception-based language learning from the start will deteriorate performance. We hypothesize that this discrepancy may result from using the BLEU score as a proxy for rewards from communicative feedback in their work, while our setup incorporates a more explicit form of corrective feedback.

D.2. Downstream evaluation on NLU tasks

We evaluate the final model on downstream natural language understanding (NLU) tasks. Specifically, we fine-tune the final CLM and TnD models (from the BabyLM corpus) on the BabyLM round 1 NLU evaluation set, which is based on (Super)GLUE (Wang et al., 2018; 2019). Table 3 shows that the TnD model performs on par with the CLM model, slightly better overall. We observe that the TnD model did significantly better on the Recognizing Textual Entailment (RTE, which requires determining inferential relationships between hypothesis and premise) task, but underperforms on the Question-Answering NLI (QNLI, which requires comprehending longer paragraphs) task.

Model	Average	CoLA	SST-2	MRPC	QQP	MNLI	MNLI$_{mm}$	QNLI	RTE	BoolQ	MultiRC	WSC
CLM + BabyLM	67.2	**69.8**	**84.4**	**76.2**	**79.1**	67.2	68.0	**68.5**	48.5	63.6	52.0	61.4
TnD + BabyLM	**67.5**	66.6	82.5	75.4	79.0	**67.4**	**69.0**	61.0	**60.6**	**64.9**	**54.5**	**61.4**
CLM + BookCorpus	65.5	**67.6**	**85.4**	73.2	**78.3**	66.0	67.1	60.1	45.5	65.3	51.0	**61.4**
TnD + BookCorpus	**65.8**	67.5	83.3	**77.3**	77.5	**66.7**	**67.4**	**66.4**	**45.5**	**67.5**	**51.2**	53.0

Table 3. We present the NLU evaluation results based on the BabyLM Challenge, which consists of tasks selected from (Super)GLUE. We fine-tune the final models developed on both BookCorpus and BabyLM datasets.

D.3. Corrective feedback helps knowledge distillation for smaller student models

The current age-based reward design assumes the student and teacher models are of the same size. This section investigates if such a reward function could distill linguistic knowledge to smaller student models. The original student GPT-2 model has a dimension of $d = 768$ (12 attention heads each with a dimension of 64). We now keep all experimental setups untouched but smaller student models with dimensions of 588 (12×49), 360 (10×36), and 250 (10×25) respectively. Figure 9 shows that the efficiency in early language learning can still be observed. Each TnD model outperforms the CLM baseline of the same size, and even surpasses CLM baselines of large capacity in early steps.

D.4. Teacher's word preferences in demonstrations affect student

To explore how the teacher model's word selection impacts students' language development, we repeat the experiments where a chosen set of 40 words for each test vocabulary is excluded from teacher demonstrations. To ensure fluent generation, we maintain the presence of essential functional words so these words don't appear in the 40 chosen words. During the language generation process by the teacher model, if a word from this set is to be decoded, we select the next best alternative, ensuring these words were never presented in teacher demonstrations. We depict the learning curves for these excluded words in Figure 10 and present the nAoA in Table 4 (in Appendix). Our findings indicate that the teacher model's word choices significantly influence the efficiency of word acquisition by the student model. The absence of words from teacher demonstrations leads to slower learning speed for the student models, as evidenced by a higher nAoA, although the student models are ultimately able to learn these words from the corpus and their trials.

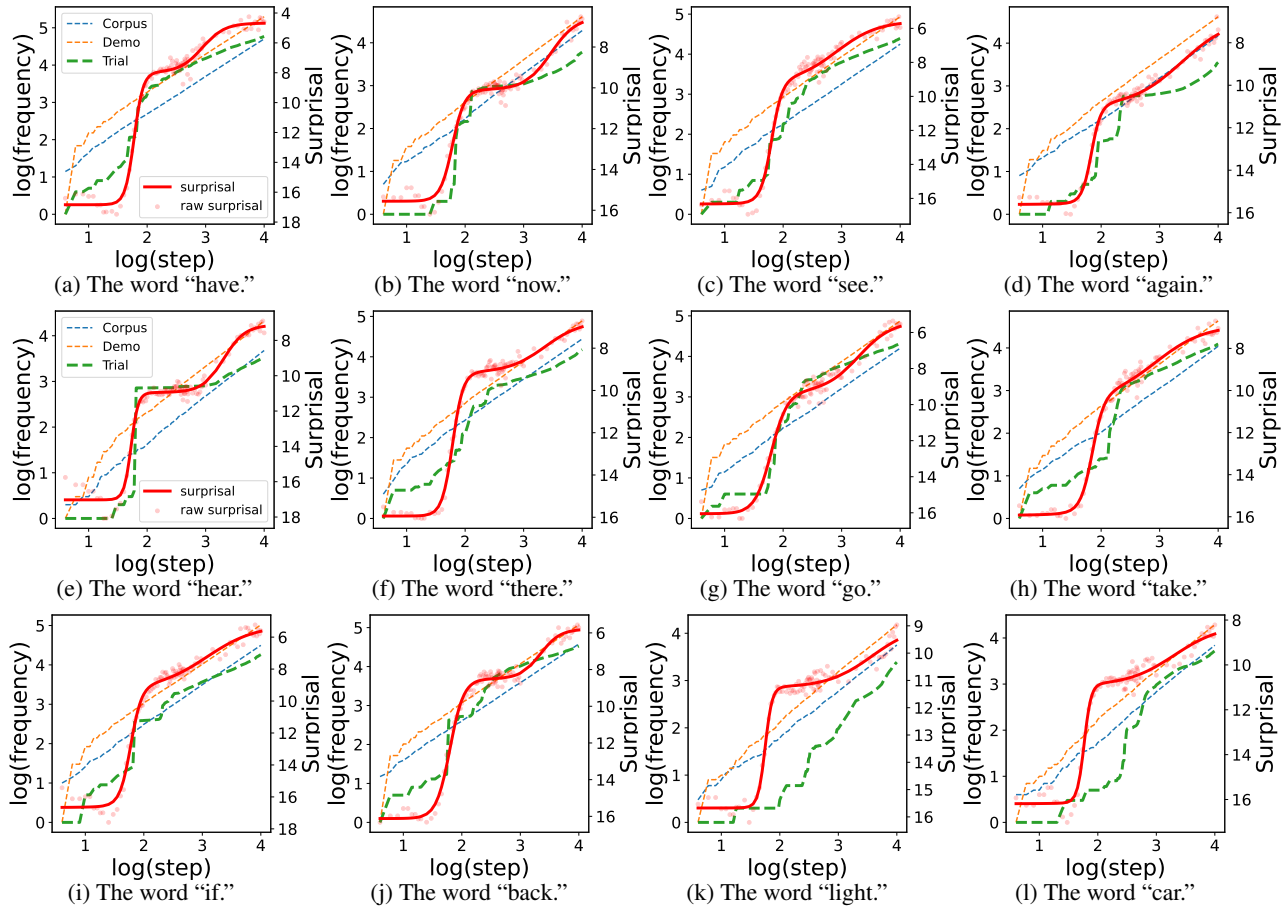
D.5. Practice makes perfect in trials

Finally, we conduct experiments to underscore the significance of the student's active trials in the process of word learning. A student model can learn a word from 3 sources: its own trials, teacher demonstrations, and exposure to the corpus. To determine which source contributes most significantly to learning, we begin by plotting the per-word learning curves alongside the cumulative frequency of word encounters in trials, demonstrations, and the corpus. We observe, interestingly, that the learning curves for certain words exhibit a pronounced correlation with the frequency of these words in trials, as exampled in Figure 11. Predicates and function words, such as "back" and "go", have a high correlation between surprisal and cumulative word frequency on the student model's trial, shown in Figure 11(a-j). Nouns such as "light" and "car", however, depict a less correlation between the word's learning curve and trial's frequency, shown in Figure 11(k-l).

Method	nAoA↓	BabyLM Corpus		BookCorpus	
		CMN	CDI	CMN	CDI
CLM	@.5	$3.01_{\pm 0.03}$	$3.02_{\pm 0.07}$	$3.02_{\pm 0.02}$	$2.86_{\pm 0.01}$
	@[.5:.95]	$3.29_{\pm 0.02}$	$3.37_{\pm 0.03}$	$3.28_{\pm 0.02}$	$3.09_{\pm 0.01}$
TnD	@.5	$\mathbf{2.34}_{\pm 0.15}$	$\mathbf{2.57}_{\pm 0.14}$	$\mathbf{2.43}_{\pm 0.12}$	$\mathbf{1.94}_{\pm 0.02}$
	@[.5:.95]	$\mathbf{3.18}_{\pm 0.07}$	$\mathbf{3.22}_{\pm 0.08}$	$\mathbf{3.22}_{\pm 0.05}$	$\mathbf{2.76}_{\pm 0.04}$
TnD (Masked)	@.5	$3.14_{\pm 0.09}$	$3.32_{\pm 0.04}$	$3.10_{\pm 0.04}$	$2.92_{\pm 0.02}$
	@[.5:.95]	$3.49_{\pm 0.03}$	$3.50_{\pm 0.03}$	$3.42_{\pm 0.02}$	$3.21_{\pm 0.02}$

Table 4. For the 40 words to be "masked out" from teacher demonstrations, we repeat the experiment with 5 random seeds and compared their nAoA with standard errors to those observed in CLM and TnD baselines.

We speculate that this pattern may be associated with the part of speech (POS) of the word. Following Portelance et al. (2020), we delve into this phenomenon by focusing on specific subsets of words, including nouns, predicates, and functional words, to explore the relationship further. To evaluate the impact of each source of word acquisition, we consider the cumulative word frequency as a predictor of the word surprisal, and carry out linear regressions complemented by likelihood ratio tests to determine the beta weights for each predictor. Upon analysis, we identify significant collinearity between word frequency in teacher demonstrations and the corpus, as indicated by a high variance inflation factor (VIF), while the word frequency in student trials is less intertwined, exhibiting a moderate VIF below 10. We thus calculate the beta weights separately for pairs (trial, demo) and (trial, corpus), then compute the average beta weights for trials. Together with the Pearson correlation, we summarize the results in Table 1. Negative beta weights signify a negative correlation, with a larger absolute value denoting a stronger association and contribution. Our analysis reveals that the cumulative frequency



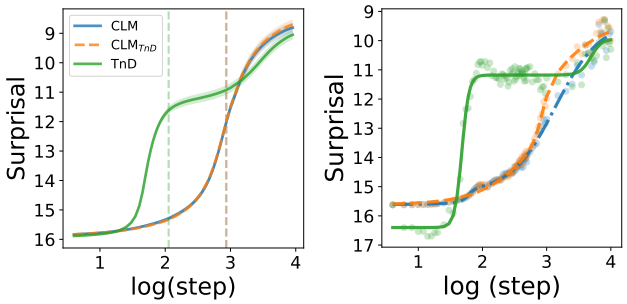
(a) The word "have." (b) The word "now." (c) The word "see." (d) The word "again."

(e) The word "hear." (f) The word "there." (g) The word "go." (h) The word "take."

(i) The word "if." (j) The word "back." (k) The word "light." (l) The word "car."

Figure 11. Examples of per-word learning curves and cumulative word frequency in BookCorpus. The dashed lines mark the log frequency of words (left y-axis) from each source. The solid line and dots mark the word surprisal (right y-axis).

of words encountered in trials plays a significant role in the acquisition of functional words and predicates. However, this significant contribution does not extend to nouns, indicating a potential impact of active trials on different POS within the learning process. This finding is linguistically intuitive, as function words and predicates are words that require other dependent words to fully express their meaning (Gleitman, 1990) and thus require more practice. We posit that grounding language in the non-linguistic world is essential for acquiring the meanings of words, particularly for concrete noun (Ma et al., 2023).



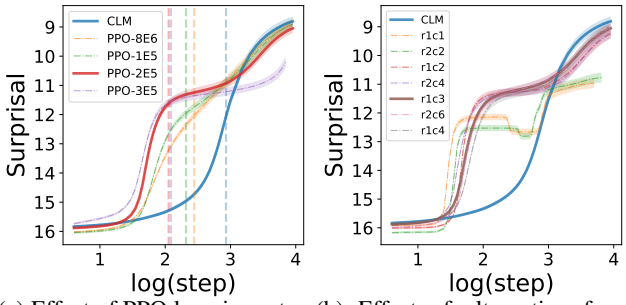
(a) Overall surprisal v.s. training step on BabyLM. (b) Surprisal v.s. training step on word "fine".

Figure 12. Comparison between CLM on model predicted text, CLM on train corpus, and TnD. Evaluated on CDI words on BabyLM.

D.6. Discussion and experiment on other possible baselines

Causal language modeling using model-generated texts. One possible baseline is to re-run the CLM baseline using teacher/student-generated texts, rather than the corpus text, at a non-interactive step. This experiment is controlled over the language input to the CLM and TnD baselines, but is not a fair setting for CLM as the student's trials are usually of poor quality in initial steps. We replace the PPO updates in TnD with causal language modeling on both texts from the student model's trial and the teacher model's demonstration. We find that the resulting CLM_{TnD} baseline achieves an almost identical overall learning curve as CLM (Figure 12b).

Using the corpus sentences as demonstrations. While it is possible to directly use the ground truth sentences from the training corpus as demonstrations (ter Hoeve et al., 2022), it can be very difficult to adapt the teacher model's behaviors for our controlled studies. As a result, we use a



(a) Effect of PPO learning rate on `CDI` words on BabyLM.

(b) Effect of alternating frequency on `CDI` words on BabyLM.

Figure 13. The robustness of the learning curve results over other hyper-parameters, e.g., learning rates and alternating schedules.

Dataset	URL	License
BookCorpus	Link	MIT License
BabyLM	Link	MIT License
Models	**URL**	**License**
GPT-2	Link	MIT License
LLaMA-2	Link	Llama License

Table 5. License information for our experiment.

pre-trained language model as the proxy for demonstration generation, rather than using the original text. In our preliminary experiments, using ground truth sentences from the training corpus as demonstrations do not lead to a noticeable difference from model-generated demonstrations.

D.7. The robustness of results and findings over hyper-parameters

Learning rate To evaluate the robustness of results over different PPO learning rates, we run our experiment with learning rate = $8e^{-6}, 1e^{-5}, 2e^{-5}, 3e^{-5}$ (Figure 13a). We find that a smaller learning rate results in a later age of acquisition compared to a higher learning rate. Nevertheless, our findings remain robust across different learning rates, although a higher learning rate can lead to unstable training and poorer end performance.

Alternating frequency We run our experiment on different alternating frequencies to study its impact on the `TnD` framework. To perform a systematic study, we group the alternating schedule $[c, r]$ by their PPO/CLM steps ratio. We experiment with the settings when the ratio equals to 1 ($[1, 1]$, $[2, 2]$), 2 ($[2, 1]$, $[4, 2]$), 3 ($[3, 1]$, $[6, 2]$), and 4 ($[4, 1]$), as presented in Figure 13b. We find that the alternating frequencies under the same ratio lead to similar performance. Our findings are robust over different alternating frequencies, except the ratio $= 1$ leading to unstable training and poor end performance.

E. Limitations and Licenses

E.1. Artifacts and licenses

Our work largely relies on publicly available datasets such as BookCorpus (Zhu et al., 2015) and BabyLM (Warstadt et al., 2023), and pre-trained models such as GPT-2 (Radford et al., 2019) and LLaMA-2 (Touvron et al., 2023). We strictly follow the LLaMA license and limit the scope of the LLaMA model to academic research only. We report a list of licenses for all datasets and models used in our experiment in Table 5.

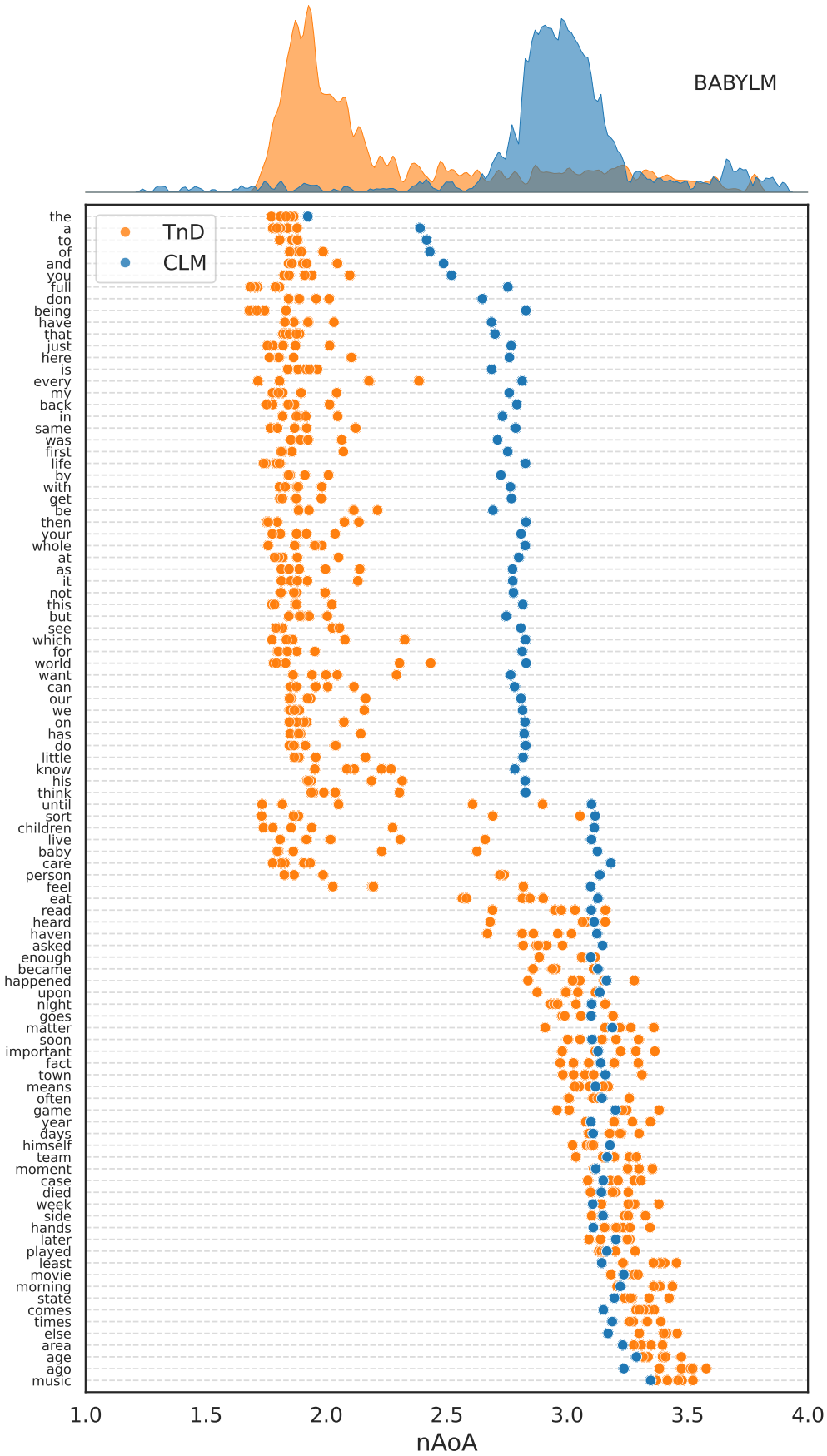


Figure 14. The ridgeline and scatter plot of words and their neural age of acquisition (nAoA) in BabyLM Corpus.

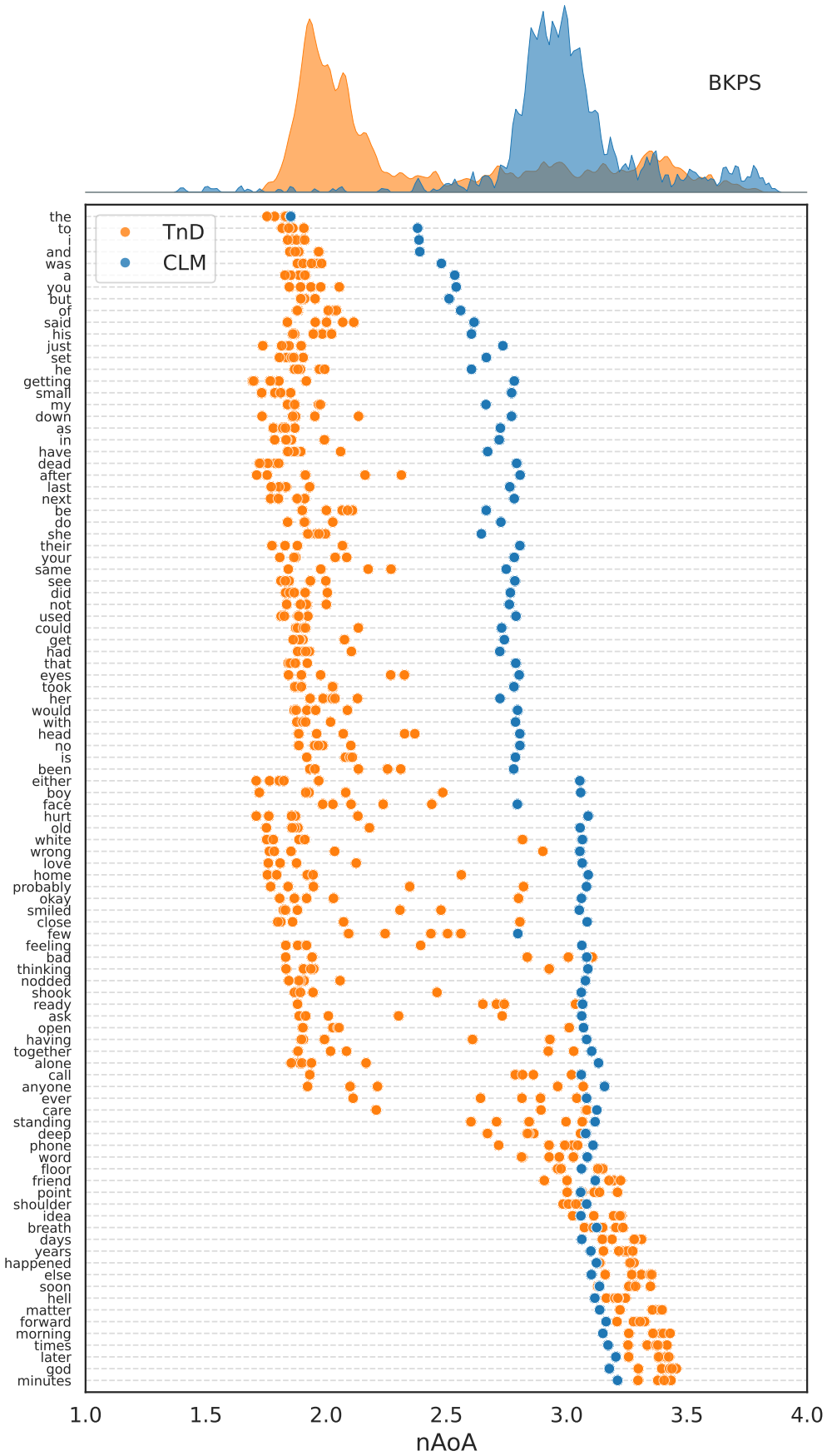


Figure 15. The ridgeline and scatter plot of words and their neural age of acquisition (nAoA) in BookCorpus.