Modeling Overregularization in Children with Small Language Models

Akari Haga[1] **Saku Sugawara**[2] **Fukatsu Akiyo**[3] **Miyu Oba**[1]
Hiroki Ouchi[1] **Taro Watanabe**[1] **Yohei Oseki**[3]
[1]Nara Institute of Science and Technology
[2]National Institute of Informatics [3]The University of Tokyo
{haga.akari.ha0,oba.miyu.ol2,hiroki.ouchi,taro}@is.naist.jp
saku@nii.ac.jp
{akiyofukatsu,oseki}@g.ecc.u-tokyo.ac.jp

Abstract

The imitation of the children's language acquisition process has been explored to make language models (LMs) more efficient. In particular, errors caused by children's regularization (so-called overregularization, e.g., using *wroted* for the past tense of write) have been widely studied to reveal the mechanisms of language acquisition. Existing research has analyzed regularization in language acquisition only by modeling word inflection directly, which is unnatural in light of human language acquisition. In this paper, we hypothesize that language models that imitate the errors children make during language acquisition have a learning process more similar to humans. To verify this hypothesis, we analyzed the learning curve and error preferences of verb inflections in small-scale LMs using acceptability judgments. We analyze the differences in results by model architecture, data, and tokenization. Our model clearly shows child-like U-shaped learning curves for certain verbs, but the preferences for types of overgeneralization did not fully match the observations in children.

1 Introduction

Current LLMs require a huge amount of data for training, and this is an issue in terms of data collection cost and training time. In contrast, infants can acquire their first language with low resources. One of the LLMs, GPT-3, requires approximately 200 billion words for learning (Brown et al., 2020), whereas infants are required to learn only about 100 million words (Chomsky, 1959; Warstadt and Bowman, 2022). Recent studies have shown the benefits of mimicking human language acquisition. For instance, using child-oriented vocabulary and/or child-directed speech (CDS) as learning data improves learning efficiency (e.g., Huebner et al., 2021; Eldan and Li, 2023). In the language acquisition literature, there are many studies on children's errors and U-shaped learning curves (Bybee and Slobin, 1982; Pinker and Ullman, 2002; McClelland and Patterson, 2002). These errors and U-shaped learning curves may seem inefficient at first glance. However, it has been suggested that such characteristics are crucial for efficient language acquisition in children (Bowerman, 1982; Carlucci and Case, 2013). Based on this background, we believe that children make typical generalization errors due to their inductive biases. By modeling these errors, we can gain valuable insights into these biases. To analyze whether current LMs have a human-like learning process, we aim to test how well LMs reproduce the characteristics of children's errors. Our assumption here is that imitating human-like efficient learning processes necessitates imitating human-like errors as well. Therefore, to verify this hypothesis, we test the following two research questions:

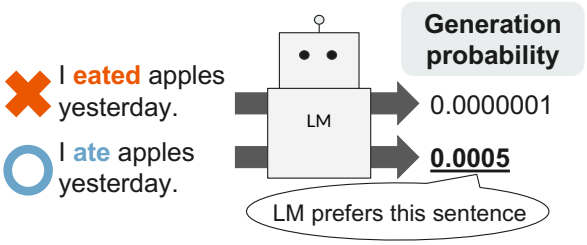


Figure 1: Our evaluation method for error preferences using minimal pair data. We can track the language model's learning of verb inflection by comparing the generation probabilities of two sentences, one with the correct past form and the other with the overregularized form.

- Do the current LMs exhibit a child-like U-shaped learning curve?
- Do current LMs make child-like mistakes?

Overregularization of acquired language knowledge is one of the typical phenomena observed in children's language acquisition. As a case study, we focus on the learning of past tense inflection

Findings of the Association for Computational Linguistics ACL 2024, pages 14532–14550

in English verbs, which is often analyzed to observe overregularization and investigate the learning curve inflection in neural LMs. In addition, previous studies showed that using CDS sorted by the child's age as learning data improves learning efficiency (Huebner et al., 2021). Therefore, we test whether the model shows human-like error preferences by training with CDS, using the model trained on Wikipedia as a comparison. Specifically, we created pairs of sentences with and without overregularization verbs and used them as shown in Figure 1 to evaluate what errors the model prefers during pre-training.

In the experiment, for certain verbs, our model shows U-shaped learning curves corresponding to the three stages observed in children. However, for other verbs, our model does not prefer the correct past tense until the end or does not show a U-shaped learning curve. Furthermore, even models that show a U-shaped curve have error type preferences different from children's. Thus, our results suggest that even recent LMs may not fully reproduce human characteristics in learning curves and error preferences in the learning process.

Our contributions are as follows:

1. We showed that using acceptability judgment allows us to analyze the overregularization of verbs in neural LMs.

2. We analyzed how well recent small language models (SLMs) capture children's learning curves and preferences for types of errors.

3. We analyzed the overregularization of the model by various data, tokenizers, architectures, and verb types and found that LMs reproduced a part of the characteristics of the child's errors in some verb types.

2 Related Work

Effects of Imitating Human Learning Processes Huebner et al. (2021) proposed BabyBERTa, a model for learning with age-ordered data of texts that humans encounter by 6 years old, and succeeded in improving learning efficiency. Eldan and Li (2023) created TinyStories, a dataset of short stories generated by GPT-3.5 and GPT-4 that contain only words normally understood by 3- to 4-year-olds. They improved the efficiency of learning by using TinyStories for training data.

Generalization in the Process of Language Acquisition in Children Overregularization is a common phenomenon of the language acquisition process in children (Marcus et al., 1992). For instance, children generalize the past tense of many English irregular verbs to the regular form at a certain stage of the learning process. This means children overregularize by adding *-d* and *-ed* to almost all verbs. There are three stages of language acquisition in children: (1) they only memorize inflection, (2) they overregularize inflection, and (3) they learn to use both irregular and regular verbs correctly. Note that in stage (2), children make mistakes not only with new words but also with words that they have been able to produce correctly. For this reason, the learning curve of children is said to be U-shaped (Rumelhart and McClelland, 1986). We analyze whether the LMs learning curve also follows a U-shape.

Modeling of First Language Acquisition Efforts to model the acquisition of English past tense, a prominent phenomenon characterized by overregularization, have been long-standing in machine learning (Corkery et al., 2019). Rumelhart and McClelland (1986) successfully trained a neural model to convert English irregular verbs into the past tense and observed a U-shaped learning curve that showed a child-like tendency to make errors. Then, Pinker and Prince (1988) showed many defects in this model. This criticism had a huge impact and popularized the idea that neural models cannot reproduce a child's language acquisition. Recently, Kirov and Cotterell (2018) have shown that the use of recent neural models eliminates most of the criticisms of Pinker and Prince.

In recent years, there has been a growing interest in the extent to which neural networks can capture the language acquisition process in children. Since 2018, several studies have attempted to imitate children's language acquisition using recent neural models (e.g., Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020). However, all of these studies have directly learned the transformation from verb lemma to past tense, which differs significantly from the setting of language acquisition in children learning from interactions with adults or conversations among adults (Yang, 2016). In this study, rather than directly learning inflectional morphology, we employ LMs trained in CDS to model the natural process of language acquisition in children.

3 Our Approach

In this study, we analyze the neural LMs using CDS as the training data and investigate what generalizations the model prefers at each step in the learning process. In previous studies, the model learned direct transformations of verbs from the base forms to the past forms, so it was sufficient to examine whether the generated past tense was correct. However, since this study does not directly learn the generation of past tense forms, an alternative evaluation method is necessary. Therefore, we adopt relative acceptability judgments for minimal pairs (Warstadt et al., 2020), as illustrated in Figure 1, to evaluate the model's preferences by comparing the generation probabilities. Such evaluation methods have become widely used in the context of LM probing (Linzen and Leonard, 2018). To assess children's error preferences, we create minimal pair data of children's errors, consisting of sentence pairs with and without overregularization. In this task, the model is forced to generate these pairs. If the probability of generating sentences with overregularization is higher, it indicates that overregularized forms are preferred over correct irregular forms; conversely, if it is lower, it indicates that overregularized forms are not preferred. We follow BLiMP (Warstadt et al., 2020) to generate sentences with over-regularization, which results in an evaluation dataset consisting of 1,000 minimal sentence pairs for the linguistic phenomenon we are investigating.

We focus on a typical phenomenon of overregularization in children: the inflection of English past tense. For irregular verbs, we first create pairs of overregularization forms (e.g., write→writed) and correct past tense forms (wrote). Next, we create pairs of sentences containing the overregularization form and sentences containing the correct past tense. The overregularized forms are created by concatenating the bare form of the verb with *-d* or *-ed* (base+ed) and the past form of the verb plus *-d* or *-ed* (past+ed) in a rule-based manner. The list of overregularization forms created is shown in Table 3 in Appendix A.1.

4 Experiments

We train a small-scale version of GPT, an LM with incremental learning, on CDS using a character-level tokenizer and evaluate the model on our evaluation data of past-tense inflections in English verbs.

Inflection Forms	Examples
Correct	John **wrote** this article.
Regularize (base+ed)	John **writed** this article.
Regularize (past+ed)	John **wroted** this article.

Table 1: Examples of evaluation data when *write* is the target verb. "Correct" means the sentence contains the correct past tense, and "Regularize" means the sentence contains the overregularized form.

4.1 Training Data

We use CDS as training data to approximate a child's learning environment. To compare the training data, we added data from Wikipedia and used the following three patterns of small-scale data for training.

(i) AO-CHILDES (Huebner and Willits, 2021)

(ii) Wikipedia (Huebner et al., 2021)

(iii) AO-CHILDES+Wikipedia

We adopt the same settings as those used in Huebner et al. (2021) for training data (i) and (ii). (i) AO-CHILDES consists of about 5 million words of text in American English, collected from the CHILDES dataset (MacWhinney, 2014), which records CDS from conversations between children and adults chronologically.[1] (ii) is a dataset of 500,000 sentences randomly collected from the English Wikipedia corpus. (iii) is the combined and shuffled version of (i) and (ii). In each corpus, all sentences were lowercase, and sentences shorter than three words were excluded. We trained the model with five different seeds in all experiments and reported the averages of the results. Corkery et al. (2019) criticize the approach, while Kirov and Cotterell (2018) report the single best performance achieved by their model. In contrast, our results, reported with initialization using multiple seeds, are considered more reliable.

4.2 Minimal Pair Data

We created 1,000 sets of sentences containing the correct past tense, base+ed, and past+ed forms[2] using the method described in Section 3, and used these sets for evaluation. We adopted the BLiMP vocabulary because it contains labels that enable

[1]The AO-CHILDES dataset contains overregularized verb forms, but with only 0.49% of past tense verbs in this form, the results are not significantly impacted.

[2]Available at https://github.com/osekilab/SLM

Model	CLM/MLM	Layers	Heads	Embeddings	Intermediate size	Parameters
nanoGPT[3]	CLM	6	6	384	-	29.94M
nanoGPT 10.99M	CLM	3	3	192	-	10.99M
BabyBERTa (Huebner et al., 2021)	MLM	8	8	256	1024	8.52M

Table 2: Models used in our experiments. We use BabyBERTa as it can be trained on AO-CHILDES. To model children's errors, we also select nanoGPT in the GPT-2 family trained for incremental next-word prediction because humans process sentences incrementally. To match the number of parameters of BabyBERTa, we adjust nanoGPT to 10.99M parameters. Details can be found in Appendix A.3.

the production of various sentences. Additionally, we filtered out vocabulary that was not included in the AO-CHILDES dataset following Zorro[3]. Table 1 shows an example of creating evaluation data. We compute the model's probability of generating each sentence on the evaluation set. As an evaluation metric, the sentence to which the model assigns the highest probability is used as the model's preference. Children's U-shaped learning curves are observed for learned verbs, not for novel words. Thus, we would like to observe only a model's overregularization of the previously learned verbs. For this purpose, we only evaluated data in which the base and past forms of the target verb appeared in the training data at each step.

4.3 Models

When humans receive utterances as input, they listen to sentences from left to right, so it is natural to think of humans as processing sentences incrementally (Altmann and Mirković, 2009). In this experiment, to model children's errors in LMs, we train models in a setting that is more realistic than the human learning environment. Therefore, we select GPT-2 (Radford et al., 2019), one of the Causal LMs (CLMs) trained for incremental next-word prediction.

We also use small models because we train our models on small datasets to approximate the child's learning environment. nanoGPT[4] is widely used as a small-scale GPT implementation. The number of parameters for nanoGPT is approximately 30M[4], which is nearly 4 times larger than BabyBERTa (8M parameters) (Huebner et al., 2021) that was shown to efficiently learn from CDS. For this reason, we conducted experiments using an additional nanoGPT model with the number of parameters reduced to about 8M. As a comparison, we use BabyBERTa (Huebner et al., 2021), a masked LM (MLM) that is not incremental. The models used in this study and the number of parameters are listed in Table 2. We train all the models from scratch using the training data. Implementation details can be found in Appendix A.3.

We choose the character level tokenizer to approximate the modeling of past-tense inflection at the phonological level in the previous study (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020). As a comparison, we also experiment with subword-level tokenizers. The subword-level tokenizers are trained using Byte level BPE (Wang et al., 2019) for each training corpus.

Corkery et al. (2019) claimed the validity to interpret the results of each simulation with different seeds as the individual participants, rather than as the average behavior of all participants. We follow the practice and report the average of the results across multiple seeds for a fair comparison with human experiments that report averages across multiple participants.

5 Results

5.1 Correct Form vs. Overregularized Form

We show in Figure 2 the learning curves for the past tense inflection of the verb. The y-axis represents the proportion of evaluation data where the correct past tense was selected from sentences containing either the correct past tense or the overregularized form. The horizontal axis represents the steps of learning, while the vertical axis indicates a preference for the correct past tense forms when the value is above 0.5, and a preference for the overregularized forms when it is below 0.5.

Figure 2 shows that most of the learning curves when using the incremental character-level models appeared to be U-shaped around 0.2–4k steps, though, we believed that this was different from the U-shape of the children. This is because the

[3] https://github.com/phueb/Zorro
[4] https://github.com/karpathy/nanoGPT

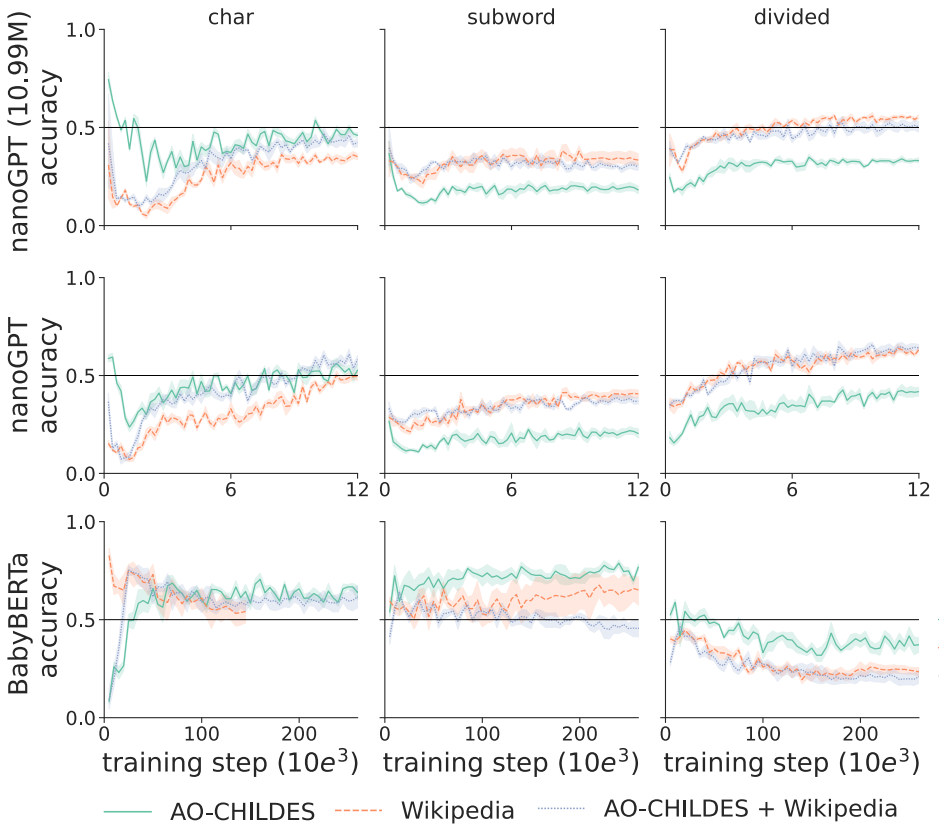


Figure 2: Learning curves of past tense inflection for nanoGPT and BabyBERTa. The columns indicate the tokenizer used. "divided" refers to a subword-level tokenizer that splits inflected forms into multiple tokens to facilitate inflection learning. Accuracy shows the proportion of correct past tense preferences. Points above 0.5 indicate a preference for the correct past tense, while points below 0.5 indicate a preference for the overregularized form. The learning curve of character-level nanoGPT forms a U-shape but lacks the Stage 3 of children's learning. Overall accuracy remains low even at the end of training. Due to variations in max steps depending on the dataset and type of tokenizer, the endpoints of the lines differ.

correct response rate in the early steps of model learning is around 50% or lower, except in the case of the 10.99M model using AO-CHILDES as training data, which is different from the Stage 1 for children who can correctly use the learned verbs. Moreover, at the end of the learning phase of the model, the preference for the correct past tense does not significantly exceed for the overregularized form, which is different from the latter part of the Stage 3, when the child can use the correct past tense almost perfectly.

In the character-based CLM models, the ratio of correct past tense and overregularized form preferences is nearly random at the end of the learning. In training BabyBERTa, as an MLM, the overregularized form was preferred more than 80% except in the early stages of learning. Additionally, the learning curves in this experiment were unstable for most of the training period. From these results, we assume that the model and dataset used in this study may not have been capable of acquiring the correct past tense inflection. The subword-level tokenizer we used does not split past tense inflected forms, and the model may not learn the rules of the word form transition. To confirm this hypothesis, we checked the subword-level tokenizer and found that the past tense of many verbs is not split and is marked as a single token. To resolve the problem, we modified the subword-level tokenizer to split past tense inflected forms and added "divided" as a result of experimentation. As a result, the percentage of correct past tense preferred increased at the end of the learning. However, the percentage was still around 60%, which does not correspond to the children's Stage 3. We have also not previously resolved the low correct response rate of the character-level models.

All of the above results indicate that the neural model does not reproduce the learning curve of children. However, since these results are aver-

aged over all verbs, a child-like U-shaped learning curve may be observed for some verbs. In the next section, we show a more detailed analysis of the results.

5.2 Results in each Verb Type

Bybee and Slobin (1982) investigated the rate at which overregularization is applied to irregular forms in the natural speech of children aged 1.5–5 years and found that children's overregularization trends differed by verb type. They define verb types by the way the verb changes phonetically during inflection. Table 4 in Appendix A.2 shows the verb types they defined. In addition, Rumelhart and McClelland (1986) claimed that the generalization tendency of the neural model is not affected by verb frequency but by verb type for most of the learning period. For a detailed analysis of the impact of the verb types on generalization tendencies, we also report the percentage of correct responses for each verb type for a detailed analysis of the impact of these verb types on generalization tendencies and how well the model reproduces human overregularization tendencies. Table 4 also shows the verbs used in the evaluation data of this experiment.

Table 3 shows the rates of overregularized form produced by children as shown by Bybee and Slobin (1982) and the rates of overregularized forms preferred by character-level models trained on AO-CHILDES for each verb type. Table 3 also shows that, for some verb types, the model can be trained to prefer the correct past tense 80% of the time by the end of learning. On the other hand, even at the end of the learning, five verb types (II, III, V, VI, and VII) showed a greater preference, and the Verb type II even showed a 90% preference for the overregularized forms.

We found that the learning curve of the Verb type IV partially corresponds to the three stages of language acquisition in children. Figure 4a shows the learning curve when evaluated only with verbs of the Verb type IV. Each line in the figure shows the results for five different seeds. The learning curve in IV around 1–2k steps corresponds to the Stage 1 where the children can select the correct form of the past tense with learned verbs in the early stages of learning. The 2–4k steps correspond to the Stage 2, where the children prefer overregularized forms even with learned verbs. After 4k steps correspond to the Stage 3, where the correct past tense is gradually preferred.[5] However, the phase in the 0–1k steps where the LM prefers a more overregularized form does not correspond to the three stages of the child. Since the Verb type IV has a high percentage (90%) of correct past tenses selected in observations of children, it is considered to be easy to learn the correct past tense. Thus, our results suggest that while the model failed to learn the correct form for many verb types, it may be able to learn the correct past tense for the Verb type IV. However, the learning curve for IV verbs showed oscillations even at the end of learning, suggesting that learning is unstable.

For the type I, which showed a high preference for the correct past tense, the learning curve did not form a U shape but it did correspond to the other observations for children. Kuczaj (1977) showed that children produce the correct past tense for unchanged verbs (Type I) more than for irregular verbs with vowel change (Type III–VIII). Our results show that LMs prefer the correct past tense for type I, and this is consistent with the children. However, for the other verb types, there was no correspondence between the model and the children. We found that for some types the learning process is very unstable. We also found that for some types past tense inflection cannot be learned. We show the results of the learning curves for each verb type in Appendix A.4. This result indicates that there are some verb types for which it is difficult for the model to correctly learn past tense inflection.

5.3 Error Types

Overregularization of children's verbs includes the addition of -ed at the end of the original form (base+ed) and the addition of -ed at the end of the past tense (past+ed) (Kuczaj, 1977). Kuczaj (1977) observed that children generally produce base+ed more often than past+ed. Additionally, when children grow up and rarely overregularize, they produce past+ed more than base+ed. Rumelhart and McClelland (1986) claim that their model reproduced this trend. However, it is considered that children produce past+ed when they believe the past tense of the verb to be the original form (Pinker and Prince, 1988). Therefore, Pinker and Prince (1988) criticizes Rumelhart and McClelland's model as not appropriate for producing past+ed because the original form of the verb is explicitly given as input.

[5] The performance did not fully improve at the end of learning because the results varied by verb, possibly due to edit distance. See Section 6 for details.

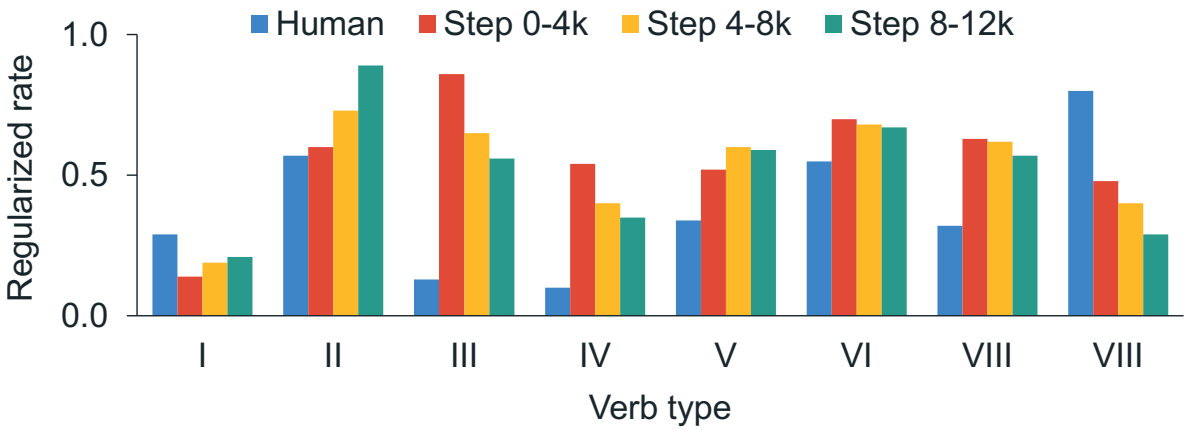
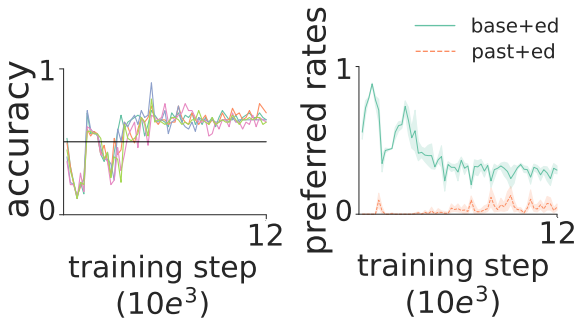


Figure 3: Rate of preference for overregularized form. From left to right: human preferences, model preferences at learning steps 0-4k, 4k-8k, and 8k-12k. For the Verb type I, both humans and models consistently prefer the correct past tense at all stages.



(a) Learning curve (b) Error preferences

Figure 4: Learning curves of verb inflection and error preferences for the Verb type IV on the character-level nanoGPT model trained on AO-CHILDES. The model's learning curves corresponded to the three stages of children's learning. However, the model's error preferences differed from those of children.

However, since our approach does not explicitly provide the original form of the verb in the model, a comparison between base+ed and past+ed preferences would be useful. Hence, we analyze whether the models learned in the CDS show preferences for past+ed and base+ed similar to children. Figure 5 shows the percentage of pairs in which the character-based model trained with AO-CHILDES preferred base+ed and past+ed among the correct and overregularized forms. We excluded verbs that have the same base form and past form, such as shut and upset. As shown in Figures 4b and 5, the incremental model nanoGPT preferred base+ed over past+ed at all learning points, including the Verb type IV with its U-shaped learning curve. This general preference for base+ed matches children's behavior, but continuing to prefer base+ed until the end does not. Additionally, BabyBERTa generally preferred past+ed. The only exception was BabyBERTa trained on Wikipedia, which initially preferred base+ed and later shifted to past+ed, matching the observations in children.

6 Discussion

How Well Do LMs Reproduce Children's U-Shaped Learning Curve? As we saw in Section 5, the learning curves of the models trained with CDS show that, for certain verb types, there are U-shaped curves corresponding to the three stages observed in children. For other certain verbs, even if the models do not show U-shaped learning curves, the accuracy still matches children's observations. However, in many of the verb types, the model did not reproduce the errors preferred by the children. Even the Verb type IV, which shows the most child-like trend in the model, did not perfectly reproduce the child's learning curve.

However, as shown in Figure 6, the performance varied depending on the verb. Specifically, verbs with a small character-level edit distance to their past tense form exhibited high performance and clear U-shaped learning curves. Based on these findings, our model generalizes in a human-like manner in the learning of some verbs. This observation of a U-shaped curve in sub-regular verbs is consistent with previous studies (Rumelhart and McClelland, 1986). Additionally, Kirov and Cotterell (2018) reported that the trend of the U-shaped curve was not observed in all irregular verbs. Therefore, the results of our study are consistent with their findings. Detailed results by verb can be found

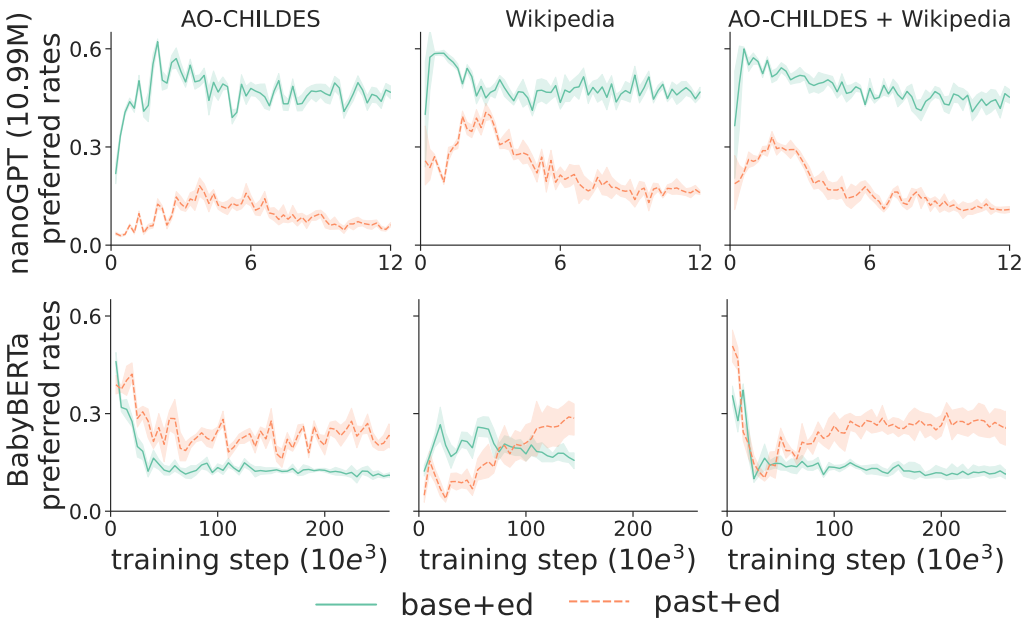


Figure 5: Preferences for overregularized forms base+ed and past+ed in the character-level models. The columns indicate the training data. BabyBERTa trained on Wikipedia shows overgeneralization preferences similar to children. However, in the setting that showed a U-shaped learning curve, no models demonstrated these errors. Due to variations in max steps depending on the dataset, the endpoints of the lines differ.

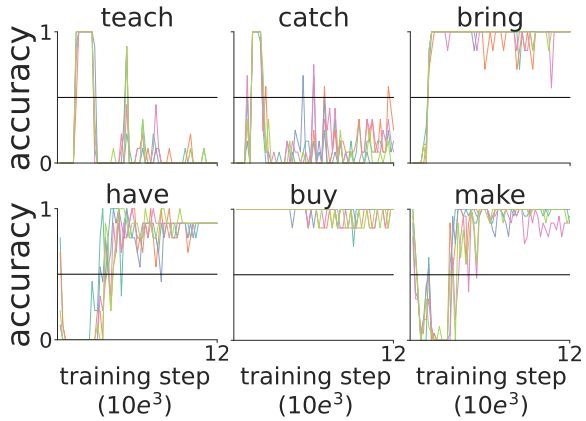


Figure 6: Learning curve for each verb in the Verb type IV on the character level nanoGPT (10.99M) trained with AO-CHILDES. The model shows clear U-shaped curves and high performance at the end of training for *have* and *make*. Each line in the figure shows the results for five different seeds.

in Appendix A.5. These results indicate that the model's performance is not entirely influenced by the frequency of the verbs. When we measured the correlation between the performance for each verb and the verb frequency, it did not show a clear correlation. Detailed correlations can be found in Appendix A.5.1.

We then discuss the oscillations of the learning curve. Plunkett and Marchman (1991) argued that when interpreting U-shaped developmental patterns, it is important to distinguish between macro and micro U-shaped curves to quantify the acquisition of inflectional systems such as the English past tense. They noted that many irregular forms oscillate between correct and overregularized forms with micro U-shaped curves. Kirov and Cotterell (2018) reported the results of learning verb past tense inflection with a recent neural model and claimed that the learning curve does not show a macro U-shape, but a micro U-shape. Our results in Section 5 also show oscillations. However, learning in a transformer is more unstable, especially with small models or training data, and our results also show that the learning curve always oscillates at any point in the learning. We therefore conclude that the oscillations indicated by our results do not necessarily reproduce the characteristics of children's errors. Our results better replicate children's language acquisition compared to Kirov and Cotterell (2018), not by showing oscillations in the learning curve, but by demonstrating a macro-level U-shaped learning curve. Instability in the learning process makes it difficult to conduct reliable analysis. In the future, we plan to try to eliminate the instability to make a more reliable analysis.

Why Cannot LMs Prefer the Correct Past Tense? As we saw in Section 5, LMs did not per-

fectly select the correct past tense, even at the end of the training phase; some verbs showed unstable learning curves, and some selected overgeneralized forms until the end. The training data included the correct past tense, and the overregularized form, i.e., the negative example, was a non-existent word that should not have been included in the training data. Therefore, it should be easy to select the correct past tense at the end of the training, but LMs could not.

One possibility for the cause of this problem is that the model does not capture the meaning of the task. Because not all the sentences in the evaluation data have words related to the past tense (e.g., yesterday), the model may not understand the task of assigning a higher score when the correct past tense is used. If this is the case, it is natural for the model to assign higher generation probabilities to sequences that are closer to the original verb forms that occur frequently in the training data.

Another possible cause is that the setting of model training may not be sufficient for the model to learn the correct past tense. Our results suggest that the learning of small-scale models is generally unstable and that CDS is not enough for models to learn the generalization of correct past tense. The CDS data we used for training, AO-CHILDES, was taken from dialogue data between children and their parents. In our results, Wikipedia articles probably substituted for children's input other than CDS, making learning slightly easier. The results of the subword-level model with divided inflectional forms in Figure 2 support this possibility. However, while Wikipedia articles provide additional information, it is not the same as the content of conversations between adults that would be input for children. More training data or more parameters may be necessary for the model to learn generalization. We would like to confirm this possibility in future experiments.

Note that the model of Kirov and Cotterell (2018) achieves nearly 100% accuracy on the training data, outperforming our model. However, it is important to note that their model learns to transduce verbs to past tense forms, which is a relatively simpler task. In contrast, our task involves training a language model on CDS and evaluating it by requiring the model to assign higher probabilities to sentences with the correct past tense as acceptability judgments. This distinction highlights that their model addresses an easier task compared to ours. Therefore, we cannot conclude that our model is inferior to theirs based on this comparison alone.

How Well do LMs Reproduce Children's Error Types? LMs generally preferred base+ed to past+ed when making errors, and past+ed was rare, both consistent with the observations of the children. However, in all settings, including the Verb type IV (Figure 4b), which was observed to correspond with the three stages of children's language acquisition, LMs showed no correspondence with the trend in the production of error types in children as shown by Kuczaj (1977). Our results suggest that even recent LMs, which have shown efficiency improvements with CDS, cannot reproduce the error type of the children's learning process. Although we only considered two error types in our experiments, the previous studies (e.g., Kirov and Cotterell, 2018) analyzed other error types such as copying. Covering these other error types in the analysis may provide further insights.

7 Conclusion

We carried out the analysis of verb overregularization in neural LMs by acceptability judgment and reported the trend of overregularization in current neural LMs. Our model shows a macro U-shaped curve corresponding to the three stages of children's language acquisition, not just oscillations. Therefore, our model better replicates children's language acquisition than Kirov and Cotterell (2018)'s model. Furthermore, the error types our model generally prefers match Kuczaj (1977)'s observations of children. However, our results differ from children's observations, which show a shift in error type preferences as learning progresses. Additionally, our results do not match those of Rumelhart and McClelland (1986) and Kirov and Cotterell (2018), which show that models use nearly 100% correct past tense for many verbs by the end of learning.

To solve these problems and conduct a more detailed analysis, we will use a larger model or training data more similar to a child's input. In addition, we plan to analyze the conditions required to replicate children's error type preferences. We also plan to conduct phonological-based experiments, a wug test (Berko, 1958) to test the regularization performance of new words, and evaluations across multiple error phenomena.

Limitations

We adopted character-level tokenization to approximate phonological-level modeling for ease of implementation. However, this is a different setting from previous studies on modeling the learning of word inflection (e.g., Rumelhart and McClelland, 1986; Kirov and Cotterell, 2018; Corkery et al., 2019; McCurdy et al., 2020). To provide more appropriate modeling of lexical inflection and to compare the results with those of children, we plan to experiment at the phonological level.

Ethics Statement

There might be a possibility that the texts we used (Wikipedia) are socially biased, despite their popular use in the NLP community. At the least, the AO-CHILDES data we used for training are preprocessed for anonymization.

Acknowledgements

Special thanks also go to the coauthors for the interesting comments and energetic discussions. This work was supported by JSPS KAKENHI Grant Numbers JP21H05054, 22K17954, and 24KJ1700, and JST PRESTO Grant Numbers JPMJPR21C2 and JPMJPR20C4.

References

Gerry TM Altmann and Jelena Mirković. 2009. Incrementality and prediction in human sentence processing. *Cognitive science*, 33(4):583–609.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Melissa Bowerman. 1982. Starting to talk worse: Clues to language acquisition from children's late speech errors. In *U shaped behavioral growth*, pages 101–145. Academic Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Joan L Bybee and Dan I Slobin. 1982. Rules and schemas in the development and use of the english past tense. *Language*, 58(2):265–289.

Lorenzo Carlucci and John Case. 2013. On the necessity of u-shaped learning. *Topics in cognitive Science*, 5(1):56–88.

Noam Chomsky. 1959. A review of B. F. Skinner's verbal behavior. *Language*, pages 26–58.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

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

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Philip A Huebner and Jon A Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, volume 75, pages 279–331. Elsevier.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Stan A Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of verbal learning and verbal behavior*, 16(5):589–600.

Tal Linzen and Brian Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. *In Proceedings of the 40th Annual Conference of the Cognitive Science Society*, page 692–697.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Gary F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*, pages i–178.

James L McClelland and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11):465–472.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Steven Pinker and Michael T Ullman. 2002. The past and future of the past tense. *Trends in cognitive sciences*, 6(11):456–463.

K Plunkett and V Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, 38(1).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of english verbs. *Parallel Distributed Processing: Explorations in the microstructure of cognition*, page 216–271.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

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.

Charles Yang. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

A Appendix

A.1 Overregularized Verb Forms

Table 3 shows the automatically created base+ed and past+ed forms used in our experiments.

A.2 Verb Types Definition

Table 4 shows verb types based on phonological changes in past tense inflection, as defined by Bybee and Slobin (1982).

A.3 Implementation Details

Table 5 details the experimental settings. Most parameters are adopted from the original nanoGPT and BabyBERTa. The vocabulary size in Table 5 refers to the subword-level model, while the char-level model, is the number of symbol types appearing in the training data.

A.4 Learning Curve for each Verb Type

Figures 7, 8, and 9 show the learning curves for each verb type when the character level models are trained with AO-CHILDES. Each line in the graph shows the results of the trials with different seeds.

A.5 Learning Curve for each Verb

Figures 10, 11, and 12 show the learning curves for each verb when the character level nanoGPT (10.99M) is trained with AO-CHILDES. Each line in the graph shows the results of the trials with different seeds.

For some verbs (*lose*, *have*, *make*, *read*, *break*, *wake*, and *come*), we observed a U-shaped learning curve similar to that of children. Additionally, we observed high performance for these verbs towards the end of the learning process. These verbs share the characteristic of having a small character-level edit distance when converting to their past tense forms. These results are likely since our experiment used a character-level model rather than a phonological one.

A.5.1 Correlation between Verb Frequency and Accuracy

We considered the possibility that model performance could vary based on verb frequency in the training data. Therefore, we calculated the correlation between each verb's frequency in the training data and its accuracy during evaluation. As shown in Figure 13, we observed correlations of 0.42 for nanoGPT and -0.29 for BabyBERTa, both trained with AO-CHILDES using a character-level tokenizer. This indicates that the character-level nanoGPT model trained with AO-CHILDES is more influenced by frequency information than other settings. However, none of the models showed a strong correlation, indicating that performance does not entirely depend on verb frequency.

	Overregularized form			Overregularized form	
Correct form	base+ed	past+ed	Correct form	base+ed	past+ed
ate	eated	ated	lost	losed	losted
bent	bended	bented	made	maked	maded
bit	bited	bitted	met	meeted	metted
bought	buyed	boughted	read	readed	readed
bred	breeded	bredded	rode	rided	roded
broke	breaked	broked	sang	singed	sanged
brought	bringed	broughted	sank	sinked	sanked
built	builded	builted	sat	sitted	satted
came	comed	camed	shook	shaked	shooked
caught	catched	caughted	shot	shooted	shotted
chose	choosed	chosed	shrank	shrinked	shranked
drank	drinked	dranked	shut	shutted	shutted
drew	drawed	drewed	sold	selled	solded
drove	drived	droved	spent	spended	spented
fell	falled	felled	spoke	speaked	spoked
fled	fleed	fledded	spun	spinned	spunned
forgave	forgived	forgaved	stole	stealed	stoled
forgot	forgetted	forgotted	stood	standed	stooded
fought	fighted	foughted	struck	striked	strucked
found	finded	founded	swept	sweeped	swepted
froze	freezed	frozed	taught	teached	taughted
got	getted	gotted	threw	throwed	threwed
grew	growed	grewed	took	taked	tooked
had	haved	hadded	tore	teared	tored
heard	heared	hearded	understood	understanded	understooded
held	holded	helded	upset	upsetted	upsetted
hid	hided	hidded	went	goed	wented
hurt	hurted	hurted	wept	weeped	wepted
kept	keeped	kepted	woke	waked	woked
knew	knowed	knewed	won	winned	wonned
led	leaded	ledded	wore	weared	wored
left	leaved	lefted	wrote	writed	wroted
lit	lighted	litted			

Table 3: Overregularized forms of verbs used in the acceptability judgments with our minimal pair data.

Type	Description	Verbs used in our experiments
I	Verbs that do not change at all to form the past tense	shut, upset, hurt
II	Verbs that change a final /d/ to /t/ to form the past tense	build, bend, spend
III	Verbs that undergo an internal vowel change, and also add a final /t/ or /d/	flee, leave, lose, hear, sell, weep keep, sweep
IV	Verbs that undergo vowel change, delete a final consonant and add a final /t/	bring, have, buy, make, catch, teach
V	Verbs that undergo an internal vowel change and whose stems end in a dental	meet, light, stand, hide, ride, write shoot, read, sit, get, fight, hold understand, breed, lead, find, bite eat, forget
VI	Verbs that undergo a vowel change of /ɪ/ to /æ/ or to /ʌ/	sing, drink, shrink, sink, win, spin
VII	All other verbs that undergo an internal vowel change	speak, freeze, drive, tear, shake come, wear, choose, fall, strike take, wake, forgive, break, steal
VIII	All verbs that undergo a vowel change and that end in a diphthongal sequence	throw, go, grow, know, draw

Table 4: The verb types by Bybee and Slobin (1982) and the verbs used in our evaluation. The verb types are defined based on the phonological changes during verb inflection.

	nanoGPT (10.99M)	nanoGPT	BabyBERTa
parameters	10.99M	29.94M	8.52M
layers	3	6	8
heads	3	6	8
embeddings	192	384	256
intermediate size	-	-	1,024
block size	256	256	-
batch size	64	64	16
epochs	10	10	10
max step	12K	12K	260K
vocabulary size	8,192	8,192	8,192
maximum sequence length	128	128	128
dropout	0.2	0.2	0.1
peak learning rate	1e-3	1e-3	1e-4
warm-up steps	6K	6K	24K

Table 5: Implementation details. Most parameters are adopted from the original nanoGPT and BabyBERTa.

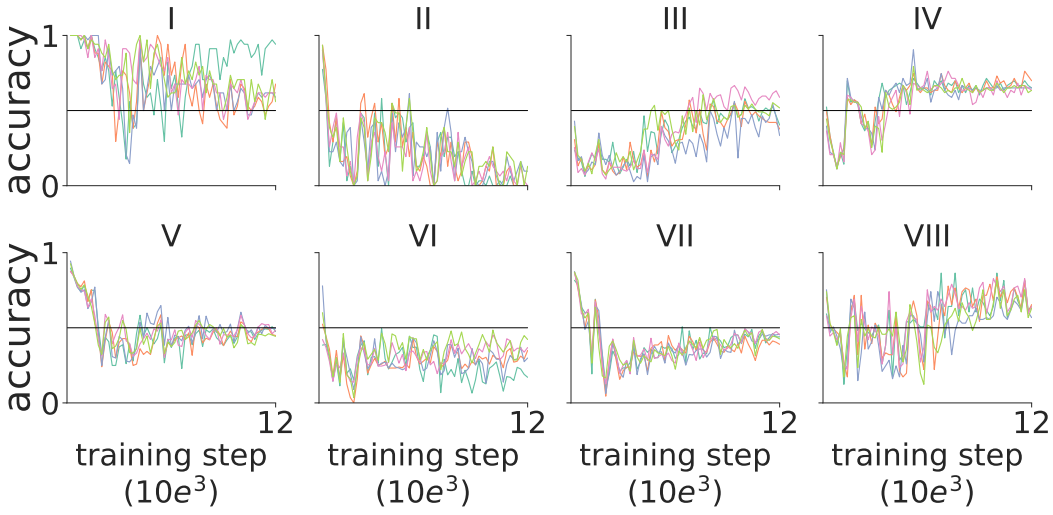


Figure 7: Learning curves for each verb type on the character-level nanoGPT (10.99M) trained on AO-CHILDES. For the Verb type IV, our result showed a U-shaped learning curve corresponding to the three stages of children's learning.

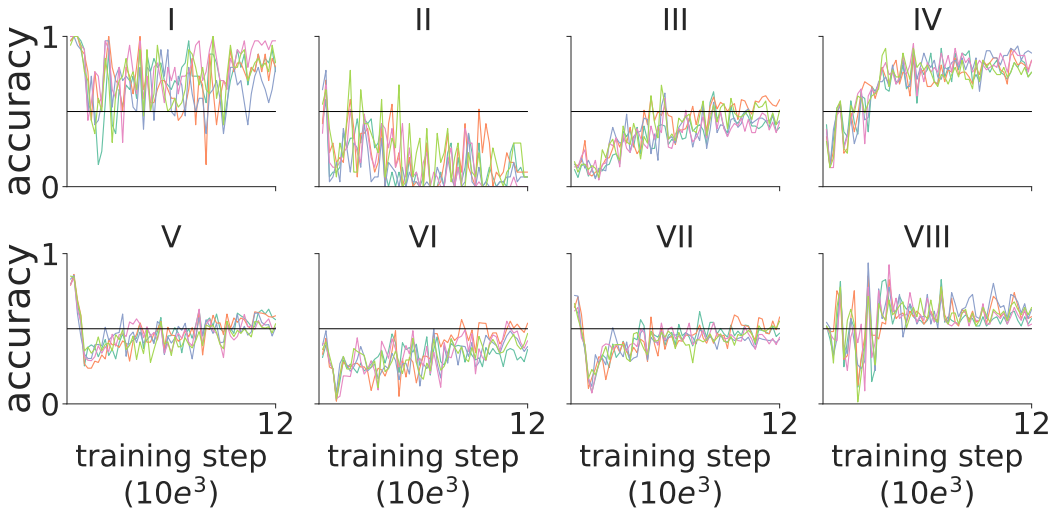


Figure 8: Learning curves for each verb type on the character-level nanoGPT trained on AO-CHILDES.

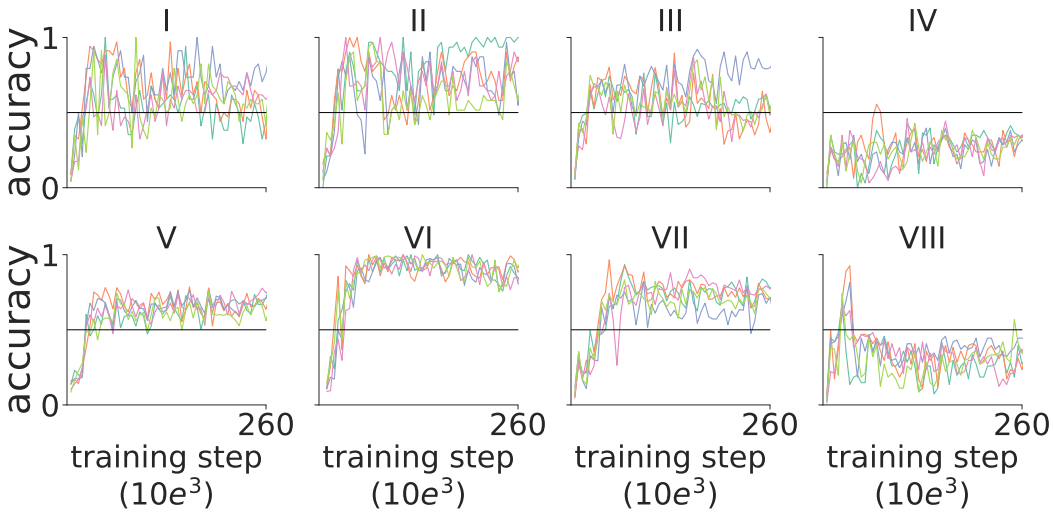


Figure 9: Learning curves for each verb type on the character-level BabyBERTa trained on AO-CHILDES.

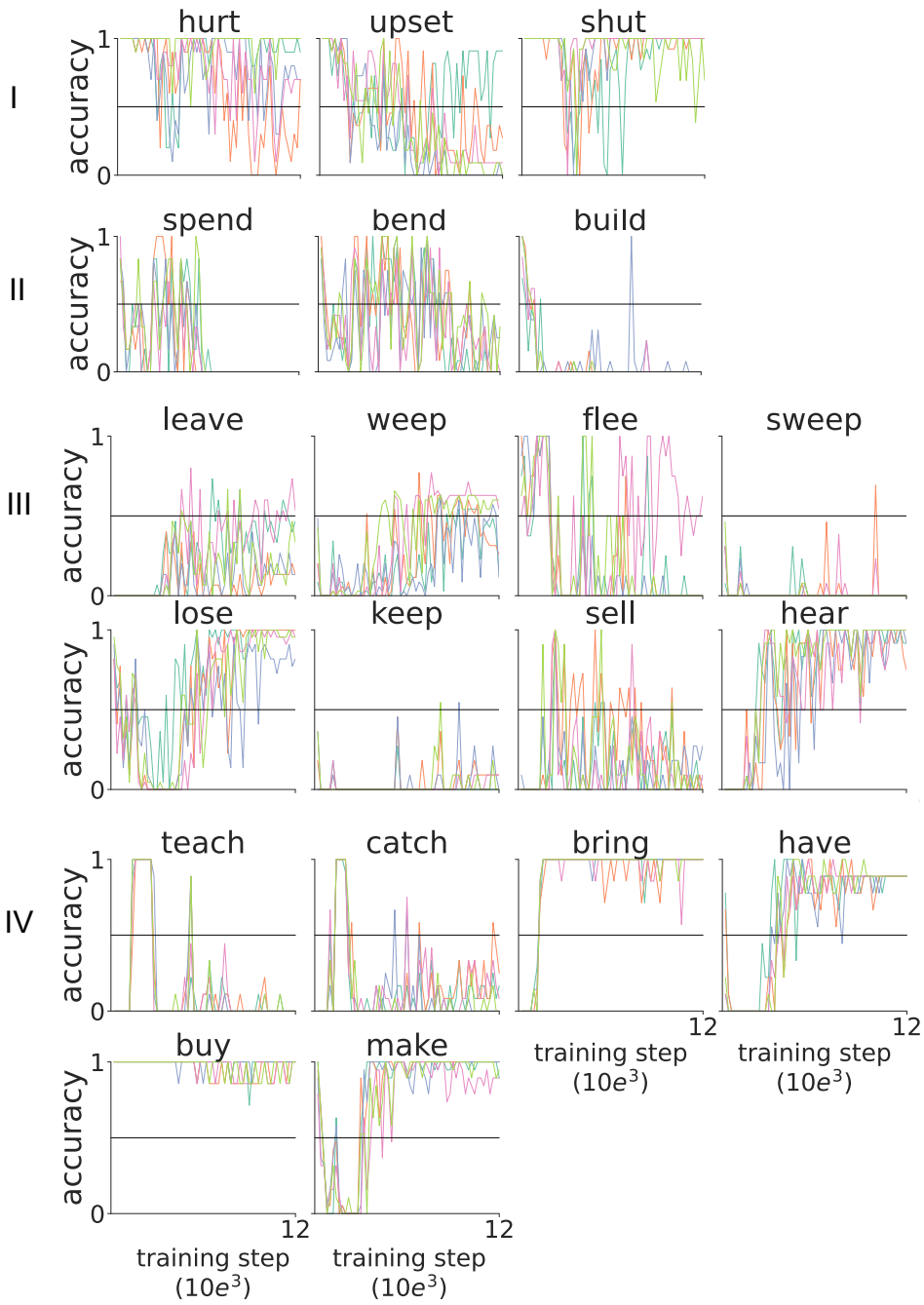


Figure 10: Learning curve for each verb of the Verb type I–IV on the character level nanoGPT (10.99M) trained on AO-CHILDES. Our results show U-shaped learning curves, similar to children, on the verbs *lose*, *have*, and *make*.

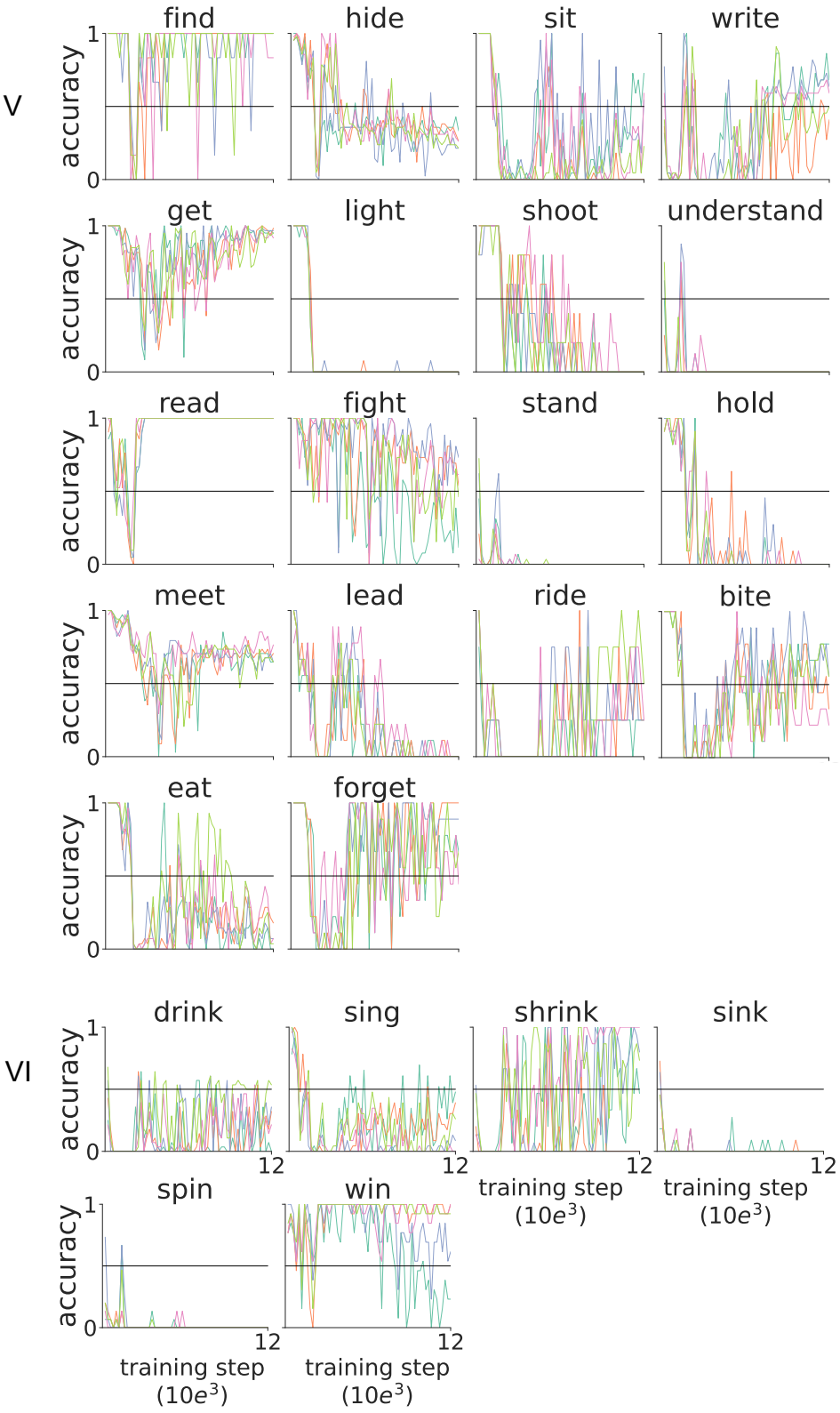


Figure 11: Learning curve for each verb of the Verb type V–VI on the character level nanoGPT (10.99M) trained on AO-CHILDES. Our results show a U-shaped learning curve, similar to children, on the verb *read*.

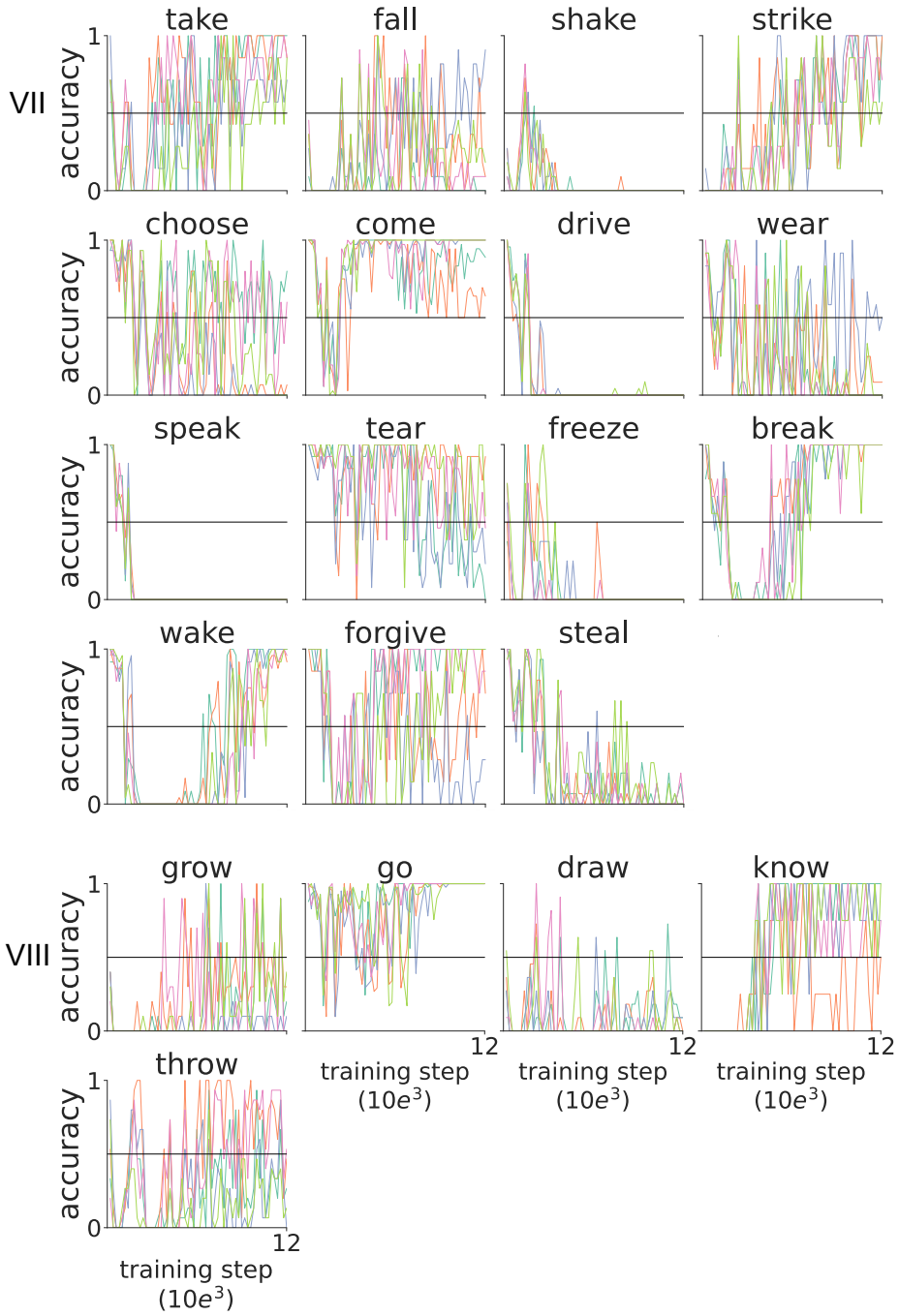


Figure 12: Learning curve for each verb of the Verb type VII-VIII on the character level nanoGPT (10.99M) trained on AO-CHILDES. Our results show U-shaped learning curves, similar to children, on the verbs *break*, *wake*, and *come*.

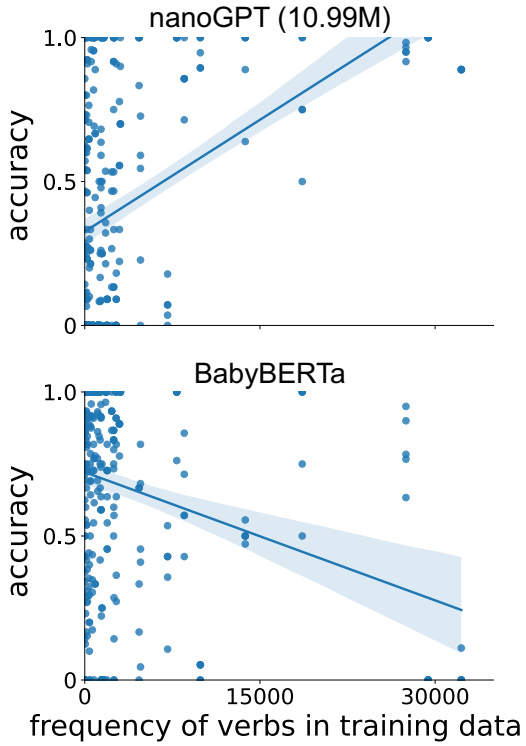


Figure 13: Correlation between the frequency of each verb in the training data and its accuracy. The results from training with AO-CHILDES are reported. We observed correlations of 0.42 for nanoGPT and -0.29 for BabyBERTa. None of the models showed a strong correlation, indicating that performance does not entirely depend on verb frequency.