Dialogue Is Not Enough to Make a Communicative BabyLM (But Neither Is Developmentally Inspired Reinforcement Learning)

Francesca Padovani[1*] **Bastian Bunzeck**[2*] **Manar Ali**[2] **Omar Momen**[2]
Arianna Bisazza[1] **Hendrik Buschmeier**[2] **Sina Zarrieß**[2]
[1]Center for Language and Cognition (CLCG), University of Groningen
[2]CRC 1646 – Linguistic Creativity in Communication, Bielefeld University
`f.padovani@rug.nl` `bastian.bunzeck@uni-bielefeld.de`

Abstract

We investigate whether pre-training exclusively on dialogue data results in formally and functionally apt small language models. Based on this pre-trained `llamalogue` model, we employ a variety of fine-tuning strategies to enforce "more communicative" text generations by our models. Although our models underperform on most standard BabyLM benchmarks, they excel at dialogue continuation prediction in a minimal pair setting. While PPO fine-tuning has mixed to adversarial effects on our models, DPO fine-tuning further improves their performance on our custom dialogue benchmark.

1 Introduction

Large language models are capable of generating language with almost human-like fluency. To do so, however, they need unfathomable amounts of textual input as training data. In comparison, humans are highly sample-efficient learners and develop a full-fledged linguistic system from input that is orders of magnitude smaller. In the past, this sample efficiency has mostly been attributed to genetically pre-endowed priors (Chomsky, 1986; Berwick et al., 2011). More recently, the quantitative, usage-based turn in linguistics has focused on the importance of language use, interaction and grounding in the real world and more domain-general cognitive mechanisms for language learning (Tomasello, 2003, 2005; Behrens, 2021). Crucially, language is primarily a tool for communication (Fedorenko et al., 2024; Levinson, 2025), and therefore all acquisition processes must be conceptualized accordingly.

Lately, the BabyLM paradigm has emerged as a novel way of testing claims of learnability with little data, small language models and linguistically inspired evaluation tasks (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025). Although highly optimized models are indeed able to capture linguistic structure very accurately (e.g., Charpentier and Samuel, 2023; Tastet and Timiryasov, 2024), they are still trained on a wider variety of input registers than the main input modality of children, namely child-directed speech in dialogue. Observed in isolation, child-directed speech does differ tremendously from other input modalities, featuring many fragments, more questions and less canonical SV(X) sentences (Cameron-Faulkner et al., 2003; Bunzeck and Diessel, 2025). Despite Huebner et al. (2021) finding it to be conducive pretraining data for simplified benchmarks, more recent work has shown that its effects can be described as *mixed* at best (Padovani et al., 2025; Bunzeck et al., 2025).

One possible explanation for this discrepancy is that autoregressive language models, trained on a next-token prediction task, do not model the communicative aspects that are seen as crucial for language acquisition and underlie the fragmented nature of child-caregiver dialogue. Common data pre-processing protocols for BabyLMs split child-caregiver dialogues into isolated sentences, which effectively removes communicative context that is available and central for human learners. Therefore, we conceptualize the task of training a BabyLM differently: We train a small, autoregressive model[1] on dialogue triplets extracted from CHILDES (MacWhinney, 2000). As such, our model is not a model of the learner *per se*, but of the interaction and communication underlying the language learning process. Additionally, we apply different reinforcement learning paradigms to our model to make the 'child' component of the dialogue system more fluent and contextually appropriate when interacting with a 'caregiver' dialogue partner. In sum, we test the following ideas through this process: (i) How does a BabyLM trained only on child–caregiver dialogue perform?

*These authors contributed equally.

[1]Given its training on dialogue data only, throughout this paper we refer to the base model as `llamalogue`. All models and datasets can be found in this Huggingface collection.

Proceedings of the First BabyLM Workshop, pages 421–435
November 8, 2025

And (ii) Are there ways of teaching BabyLMs to be more communicative speakers via interaction and communication?

We find that (i) our base model pre-trained exclusively on child–caregiver dialogues maintains above-chance accuracy on formal linguistic competence, while achieving higher accuracy in predicting realistic communicative turns than a baseline autoregressive model. Moreover, (ii) directly aligning preferred child responses to caregiver utterances through DPO proves more effective than interactively fine-tuning the policy via PPO with a reward function, especially when evaluated on dialogue minimal pairs. However, none of these fine-tuning techniques improves performance on more formal benchmarks.

2 Related work

Learning exclusively from CDS While the standard English BabyLM corpus consists of approximately 30% child-directed speech, ample work exists on pretraining LMs from scratch on 100% child-directed speech (CDS). In a seminal paper, Huebner et al. (2021) showed that a small 5M-parameter BabyBERTa model, trained on 5M lexical tokens of child-directed speech, shows the same accuracy on Zorro (vocabulary-limited minimal pair tasks; Huebner et al., 2021) as the RoBERTa-base model with 125M parameters and trained on 30B words. Similar results are presented by Feng et al. (2024), who show that autoregressive models trained on CDS alone perform only slightly worse on Zorro than comparable architectures trained on Wikipedia data, synthetic data, or the BabyLM corpus. However, their CDS models underperform other models tremendously on semantic similarity benchmarks. Negative results are also reported by Yedetore et al. (2023), who show that autoregressive models trained on CHILDES data fail to reliably acquire hierarchical generalizations in question formation from declaratives, and rather prefer incorrect linear generalizations.

Expanding the CDS-only training paradigm to more languages than English, Salhan et al. (2024) find that developmentally-inspired curriculum learning strategies during pretraining improve scores on syntactic minimal pairs for models trained on English, French, German, Chinese, or Japanese CDS, outperforming models trained on Wikipedia data by over 10%. Conversely, Padovani et al. (2025) report less positive results. For many syntactic minimal pair benchmarks, their CDS models underperform in comparison to Wikipedia-trained models across different languages (English, German, French). Finally, Bunzeck et al. (2025) approximate German CDS on the level of utterance-level construction distributions. They also find that models trained on it are generally inferior to models trained on comparable Project Gutenberg data when evaluated on syntactic benchmarks, although the CDS models show moderate improvements on some word-level benchmarks.

In sum, it can therefore be said that pre-training on CDS is only conducive to language model performance for highly specific benchmarks like Zorro (although results are inconsistent across studies) or in more specific training regimens like curriculum learning.

Cognitively/developmentally plausible RL Despite reinforcement learning, especially in the form of RLHF (Ouyang et al., 2022), being an integral part of modern language modeling practices, it has only very recently begun to get adopted in cognitively inspired modeling. Zhao et al. (2023) improve their small models trained on BabyLM data by constructing a RLHF dataset from human-annotated story continuations generated with regular GPT-2 and then reinforcing these storytelling capabilities of their models. While it does not improve performance on zeroshot benchmarks, it makes their models better base models for fine-tuning tasks.

In a more developmentally inspired fashion, Ma et al. (2025) generate text continuations from a student GPT-2 model and compare these to an already further trained teacher model. A reward signal is then generated from the model's estimated 'age' (*viz.* training steps), based on its continuations and the teacher continuation. This interactive learning is then interleaved with regular causal language modeling. Their interactive model outperforms regular autoregressive models on word acquisition, quantified as average surprisal for a set of test sentences.

Stöpler et al. (2025) introduce a training regime inspired by emergent communication research that again includes two language models: a speaker/child language model, and a listener/caregiver language model. In their setup, the speaker model has to summarize a passage, and the listener model has to answer a question solely based on the summary provided by the speaker. If the listener model (whose weights are frozen) answers correctly, a reward signal is used to update the speaker model. Although their reinforcement strategy changes speaker behavior, it does not improve performance on linguistic benchmarks.

Finally, Nikolaus and Fourtassi (2025) base their reward signal on an annotated dataset of CHILDES data with clarification requests in parental utterances, which often trigger children to use more "grammatical" language. For each utterance produced by their child language model, they predict if it would possibly beget such a clarification request, and reward productions that do not. They show that this process improves their models on the reinforcement goal of producing less "ungrammatical" utterances, but has mixed to no effects on grammar benchmarks like BLiMP (Warstadt et al., 2020) and Zorro.

3 Methodology

3.1 Pretraining

Data Our models are trained on dialogue data from the English CHILDES section. In a first preprocessing step, we clean the transcripts from CHILDES quite heavily by removing all extra- and paralinguistic information. Furthermore, we replace all unintelligible or otherwise incomplete utterances, for which annotations as to the intended word are available, with these intended words. Finally, we split all utterances that contain explicitly annotated pauses, as there is no clear distinction between such pauses and utterance boundaries marked by regular line breaks.

From these cleaned dialogues, we extract all utterance triplets (three consecutive turns) where at least two different speakers are involved. Furthermore, we enforce the triplets to contain at least five lexical words. This excludes triplets that only contain repetitions of single words or are otherwise light on lexical content. We leave the speaker tags in the data. A typical line from our data might therefore look as follows:

```
*CHI: all gone .
*MOT: where's the kitty ?
*CHI: all gone .
```

By using dialogue data only, we assume that the autoregressive pretraining process pushes our BabyLM to model contingent structure (responses depend on previous turns), learn turn-level coherence, and acquire some knowledge about implicit expectations in communication, e.g., that questions beget a response.

Base model We train a small 135M-parameter Llama model (Touvron et al., 2023a) on 10M lexical tokens from the aforementioned set of dialogue triplets. Our model features 16 layers, 16 attention heads and a hidden/intermediate size of 1024. We train the model for 10 epochs. As we found approximately 60k different lexical types in our data, we opt for a small vocabulary size to not store too many of these types holistically. We fit a BPE tokenizer on the training data to include 8k tokens. Crucially, we fit the tokenizer on the actual transcriptions only, not on the speaker tags. The speaker tags are added as additional tokens afterwards. In sum, with the inclusion of all speaker tags, this results in a vocabulary size of 8465.

3.2 DPO fine-tuning

As a first attempt to further align `llamalogue` with child-like, communicatively appropriate behavior, we employ Direct Preference Optimization (DPO; Rafailov et al., 2023), which is a preference-based training method that directly optimizes the model to prefer certain continuations over others. In our case, this procedure is supposed to guide our model to favor contextually appropriate utterances over random ones.

Naturalistic data As fine-tuning data, we construct a dataset of minimal dialogue pairs derived from another set of triplets not seen during pretraining and not used for validation. From these, we extract naturally occurring caregiver–child exchanges and derive contrastive, incorrect variants by replacing the child utterance with a randomly sampled one. To systematically control for confounds, we focus on minimal pairs that are matched in length (by number of words or subword tokens) and filter out pairs where the child utterance repeats words from the caregiver utterance, resulting in approximately 26 000 pairs. For DPO training, we select the word-matched minimal pairs, subsampling 18 000 examples for training, with the remaining 8000 examples held out for evaluation. Overall, the fine-tuning phase was conducted on a total of 245 480 tokens.

Synthetic data In addition to the real data, we generate a synthetic DPO dataset to probe the benefits of model-guided preference generation. Here, the caregiver's utterance is used as a prompt to Llama-3.2-3B (Touvron et al., 2023b), which generates a plausible child response. Incorrect alternatives are again randomly sampled from the original dataset.

You are a young child having a conversation with your mother. When your mother says something, you should answer as a typical and natural-sounding child. Do NOT repeat her words. Instead, give a new, relevant answer that shows understanding. Keep it short and child-like. *MOT: I think they just throw it on the side . *CHI:

Table 1: Zero-shot prompt to Llama-3.2-3B.

Here, we do not control for matched length, as the exact number of generated words is not easy to control. The child continuation is generated through an instructive prompt (cf. Table 1) designed to facilitate short and natural completions. In total, the synthetic training data is composed of 245 480 tokens.

The two fine-tuning datasets are available in our Huggingface collection. Representative examples from both datasets are provided in Appendix B.

We perform one 10-epoch DPO fine-tuning run with `llamalogue` on each dataset with `trl` (von Werra et al., 2020). A learning rate of 5×10^{-6} is used, with a per-device batch size of 4, and 4-step gradient accumulation, resulting in an overall batch size of 16. Figure 1 shows the two loss and reward trends for the appropriate and random sentences of the fine-tuning datasets.

3.3 PPO fine-tuning

To steer the communicative behaviour of `llamalogue` more indirectly, we also fine-tune it using Proximal Policy Optimization (PPO; Schulman et al., 2017). To implement the notion of 'effective communication' for PPO, we needed to substantially simplify it. Developmental research has extensively characterized learning as involving a dynamic exploration–exploitation trade-off (Kim and Carlson, 2024; Gopnik, 2020; Nussenbaum and Hartley, 2019), in which children alternate between experimenting with novel behaviors (i.e., linguistic forms) and leveraging familiar patterns. However, operationalizing this *sweet spot* between exploration and exploitation as a computational reward function is inherently difficult. To formalize what constitutes a "successful" communicative turn, we explored a range of reward functions reflecting different aspects of communication: a BLEU-based reward, a semantic similarity reward, a quality score derived from an LLM, and an uncertainty-based reward measuring LLM confidence in processing child responses.

Our PPO pipeline requires real caregiver prompts as input for both `llamalogue` and a teacher LLM emulating a "good, communicative baby", therefore we extract caregiver utterances (minimum four tokens length) from unused segments of the preprocessed CHILDES dialogue triplets. We then prompt a teacher LLM, as before a Llama-3.2-3B (Touvron et al., 2023a), with these utterances, asking it to generate candidate responses simulating a short child-like answer that shows understanding of the caregiver utterance[2]. The prompt is the same as the one used for generating the DPO datasets (Table 1). The reward functions are then computed by comparing these teacher-generated responses to the output produced by `llamalogue` in response to the same utterance. Calculating the reward as an average over 10 generated responses proved to be noisy due to their variability, so we ultimately based the reward on the comparison between the model's output and the one single response generated by the teacher LLM.

1-gram BLEU Reward The BLEU-based metric (Papineni et al., 2002) captures surface-level lexical similarity. Specifically, we compute a smoothed unigram BLEU score (BLEU-1) between `llamalogue`'s response and the teacher LLM's reference answer with `nltk`. We apply smoothing to avoid zero scores. The resulting reward values range from 0 to 1.

Semantic Similarity Reward As a complementary approach to lexical overlap, we also implement a semantic similarity reward to promote contextually appropriate, meaningful responses. Specifically, we use the `all-MiniLM-L6-v2` model from `SentenceTransformers` (Reimers and Gurevych, 2019) to compute the cosine similarity between the BabyLM's response and the reference utterance generated by the teacher LLM. This similarity score, ranging from 0 to 1, encourages outputs that align semantically with high-quality examples.

LLM-generated Reward To further explore reward signals grounded in communicative quality, we prompt an LLM to directly assess `llamalogue`'s responses. Given a caregiver utterance and the generated child continuation of `llamalogue`, the LLM is instructed to assign a numerical quality score (from 0 to 5) based on contextual appropriateness and

[2]Examples can be found in Appendix C.

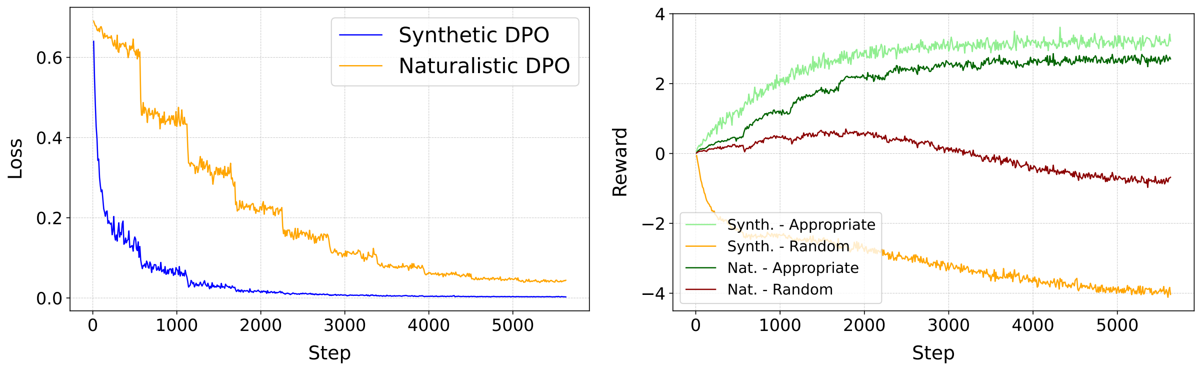

Figure 1: Training loss (left) and reward trends (appropriate vs. random) during training (right) for both DPO models.

<|system|> You are presented with a dialogue between a mother (MOT) and a child (CHI). Please rate how contextually appropriate and fluent the child's response is, on a scale from 0 (completely unfitting) to 5 (perfectly fine answer). If CHI answer is too short rate it low. <|end|> <|user|> MOT: It's like in the grocery store when go shopping . CHI: Mom, please let me choose the food for myself. <|end|> <|assistant|>

Table 2: Zero-shot prompt to OLMo.

fluency (see Table 2). After experimenting with various models, including Llama-3.2-3B and Nemotron-Research-Reasoning-Qwen-1.5B (Liu et al., 2025), we selected OLMo-2-1124-7B-Instruct (Olmo et al., 2025). This choice was motivated by the fact that OLMo consistently adhered to the requested output schema and avoided formatting anomalies that hindered automated reward extraction. The scalar score returned by OLMo is then used as the reward signal during each PPO step.

Teacher Confidence-based Reward To incorporate a measure of uncertainty into the reward signal, we implement a confidence-based metric: For each caregiver utterance x, we precompute the log-probabilities $\{\ell_i\}_{i=1}^{10}$ assigned by the frozen Llama-3.2-3B to the set of 10 reference child responses generated by that same model $\{y_i\}_{i=1}^{10}$, as explained in Section 3.3. During fine-tuning, the BabyLM's generated response $\tilde{y}$ is scored using the same teacher to obtain

$$\ell_{\text{baby}} = \log P_{\text{teacher}}(\tilde{y} \mid x)$$

Then we compute the normalized rank

$$\text{rank}(x, \tilde{y}) = \frac{1}{10} \sum_{i=1}^{10} \mathbf{1}\{\ell_i \le \ell_{\text{baby}}\} \in [0, 1],$$

and linearly map it to a PPO reward

$$r(x, \tilde{y}) = 2\,\text{rank}(x, \tilde{y}) - 1 \in [-1, 1].$$

This signal favors BabyLM outputs that the teacher assigned high likelihood and potentially bias the model towards more grammatical and distributionally expected utterances.

Training configuration In our experimental trials we rely on the default PPO training parameters provided by the `trl` library for all fine-tuned models, with the exception of the one trained using the Teacher Confidence-based reward. This reward caused higher variance in the reward values, making the KL control more sensitive. Therefore we set the KL penalty mode to `abs`, a lower learning rate of 5×10^{-6} and a small initial KL coefficient of 0.02 to weaken the penalty for policy updates in the early stage of training.

Moreover, the fine-tuning processes based on the first three PPO strategies employed a larger portion of the training data, as caregiver utterances inputs, from our original pre-processed set (220 000) compared to the model fine-tuned using the Teacher Confidence-based reward (150 000). In the latter case, we observed that a lower number of training steps was sufficient to achieve a consistent, significant improvement in the reward. We fine-tune for 3 epochs, with each epoch featuring 13 750 steps for the first three PPO strategies and 8645 steps for the Teacher Confidence-based reward. In terms of token usage, for the first three PPO strategies we estimate a total of 3 009 104 tokens, obtained by summing the

	Task	`llamalogue`	DPO Natural.	DPO Synth.	PPO Bleu	PPO SemSim	PPO LLM Score	PPO Conf.	Baseline
Zero-shot (Baby LM)	BLiMP	56.05	55.64	55.51	55.14	**56.36**	55.31	55.10	72.16
	BLiMP suppl.	51.06	49.97	**51.67**	51.33	51.48	50.58	49.45	61.22
	COMPS	51.62	51.51	**51.63**	50.66	51.58	51.25	51.59	—
	Entity tracking	30.66	32.66	31.29	16.20	34.64	**36.03**	34.05	28.06
	EWoK	50.19	50.12	**50.82**	49.65	49.62	50.12	50.81	51.92
	Read. (eye track.)	**3.88**	3.57	1.16	3.43	2.85	3.73	3.35	9.08
	Read. (self-paced)	1.43	1.35	0.44	**1.99**	1.04	1.30	1.14	3.5
	Wug adj.	0.45	0.52	0.16	0.13	0.01	**0.55**	0.41	38.5
	Wug past	-0.03	-0.01	-0.05	-0.15	-0.18	-0.01	-0.19	—
	AoA	-79.6	0	0	-80.1	0	-76.6	-78.7	—
FT	(Super)GLUE	51.82	51.72	51.77	51.12	52.10	51.69	**51.92**	67.91
Zero-shot (Add'l)	Lexical decision	40.3	40.5	**41.3**	40.7	39.7	40.2	40.8	57.2
	Zorro	**65.5**	64.8	62.7	62.5	64.7	65.2	63.7	77.7
	Dia. MP (Words)	64.3	**68.4**	64.9	62	61.1	60.6	63.7	58.1
	Dia. MP (Tokens)	63.8	**67.6**	64.3	61	63.6	62.5	62.4	57.9

Table 3: Full results for pre-trained and fine-tuned (FT) models. For each task, the best-performing model among those we pre-trained and fine-tuned (excluding the baseline) is shown in bold.

tokens from the prompts provided to `llamalogue` and the single ground-truth response generated by the teacher LLM. For the Teacher Confidence-based reward strategy, where ten teacher responses were used for confidence estimation, the total amounts to 9 903 146 tokens. In both cases, the overall token count remains well below the 100M-token limit specified by the BabyLM Challenge for the interaction track.[3]

3.4 Evaluation

Standard benchmarks For evaluation purposes, we rely on the BabyLM evaluation pipeline (Charpentier et al., 2025). As zero-shot evaluation, it includes minimal pairs tasks on the syntactic level (BLiMP, Warstadt et al., 2020) and on the semantic/world knowledge level (COMPS, Misra et al., 2023; EWoK, Ivanova et al., 2025; entity tracking, Kim and Schuster, 2023). Additionally, in further tasks, model probabilities/surprisal values are correlated with word-level age of acquisition (Chang and Bergen, 2022), cloze probabilities (De Varda et al., 2023), and preferences in morphological inflection for 'wug' words (Hofmann et al., 2025). Finally, the models are also evaluated through fine-tuning on a selection of tasks from GLUE and SuperGLUE (Wang et al., 2018, 2019).

Custom benchmarks To evaluate the models in a more holistic way, we include three additional minimal pair benchmarks. We (i) create a dialogue minimal pair set. As already described in Section 3.2, positive examples are created by simply matching parental utterances with children's answers, negative examples are sampled by matching the same parental utterances with unrelated child utterances. With this dataset, we aim to not only test the formal language skills of our models (as the BabyLM evaluations already do), but also their functional skills (Mahowald et al., 2024). Furthermore, we include (ii) Zorro (Huebner et al., 2021), a reduced version of BLiMP with a vocabulary restricted to words that occur in CHILDES, and (iii) the lexical decision dataset by Bunzeck and Zarrieß (2025), which contains word-level minimal pairs of words and non-words (e.g., *sending* and *monding*) as benchmarks that should be more tuned to the linguistic register found in our pretraining data.

4 Results

4.1 Base model evaluation

We evaluate our base model after being trained for 10 epochs. We compare `llamalogue` to the baseline model `babylm-interaction-baseline-simpo`[4] provided by the BabyLM organizers for the *interaction* track. Our model performs worse than this baseline model in almost every BabyLM evaluation task, except entity tracking (cf. Table 3). In comparison to other models submitted to the *strict-small* track, our model performs particularly worse on

[3]The code for DPO and PPO experiments can be found at these two Github repositories: https://github.com/fpadovani/communicative_baby_ppo and https://github.com/fpadovani/communicative_baby_dpo

[4]https://huggingface.co/BabyLM-community/babylm-interaction-baseline-simpo

BLiMP and AoA prediction, whereas scores for EWoK, COMPS, (Super)GLUE or the different wug tests are undercut by several other submissions. Therefore, `llamalogue` is not a generally bad language model, but its pretraining peculiarities have a non-straightforward effect on performance.

With respect to the custom benchmarks the results are more nuanced. For example, on the *dialogue minimal pairs* task, which aligns closely with the pre-training goal of `llamalogue`, it exhibits a clear advantage over the baseline comparison (63–64% vs. 57–58%). Our model also achieves a reasonable accuracy of 65.5% on *Zorro*. Nevertheless, it is clearly outperformed by the interactive baseline (77.7%) which was trained on the full BabyLM data. Our base model also falls behind the interactive baseline on the *lexical decision* task, performing quite far (40.3%) below chance level.

4.2 Fine-tuned models

4.2.1 BabyLM evaluations

Like the `llamalogue` base model, our fine-tuned models show overall lower performance on almost all of the zero-shot BabyLM Challenge tasks than the baseline model and the other models submitted to the interaction track. For BLiMP, all model variants score substantially below the baseline's 72.16%, with results clustering around chance level. The highest score is achieved by the Semantic Similarity model at 56.36%. Similar trends hold for BLiMP supplementary, where the gap to the baseline remains notable. Surprisingly, for entity tracking our models improve over the baseline of 28.06, with the best score (36.03) achieved by a model fine-tuned with OLMo Score. For EWoK, scores are near chance level, in accordance with the baseline model. Reading-based tasks (eye-tracking and self-paced reading) show much lower alignment with human patterns than the baselines. The Wug adjective and Wug past morphological generalization tasks yield near-zero or negative correlations across all models, far from the baseline model score of 38.5 for Wug adjective, underscoring persistent difficulty in capturing human-like morphological generalization. For AoA, after a closer look at metric computation, we find that only very few data points (1–5 words) are considered. This is due to an unpassed condition on the parameters of the fitted sigmoid function within AoA computation in the evaluation pipeline. Limited data points lead to either a score of zero or a strong negative correlation; hence, these results can be misleading. Overall, while entity tracking shows a modest improvement over the baseline, most linguistic and psycholinguistic tasks still reveal substantial gaps. The usefulness of our models for fine-tuning is not affected by reinforcement learning, indicated by (Super)GLUE scores that do not change drastically and also remain lower than for the baseline model (cf. also Appendix D).

4.2.2 DPO reward and custom evaluations

As shown in the right plot of Figure 1, the reward assigned to acceptable and unacceptable utterances begins to diverge early in the fine-tuning process. This separation is particularly pronounced in the case where the acceptable sentence is artificially generated by the LLM, suggesting a stronger initial reward signal and a more stark contrast between both continuations. Interestingly, this tendency is not confirmed by performance on *dialogue minimal pairs*. Although both DPO models improve upon the base model with regard to this measure, the effect of the synthetic data is rather low (increase of approximately 0.5%).

In contrast, the model fine-tuned on real caregiver–child interaction data scores approximately 4% higher than the base model and the model fine-tuned on artificially generated child utterances. This suggests that, although LLM-generated utterances may be more grammatical and exhibit greater syntactic and lexical variety than real data found in CHILDES, the model fine-tuned on synthetic data is less apt at predicting real minimal pairs derived from genuine interactions. The natural data is clearly superior to synthetic data when trying to optimize for this task. For Zorro, the naturalistic model maintains performance comparable to `llamalogue`, whereas the synthetic model shows slightly lower accuracy.

4.2.3 PPO reward and custom evaluations

During PPO fine-tuning, we observe occasional instability[5] in the training process. To ensure consistency in evaluation, we assess all models at the end of the first epoch, after a single full pass over the novel data. For the OLMo-based score, the training process shows a sharp reward decline before completing the first full epoch. Therefore, we select an earlier checkpoint (5000 steps) for evaluation, under the assumption that these 5000 steps still

[5] Including abrupt drops in reward and unexpected script crashes before completion.

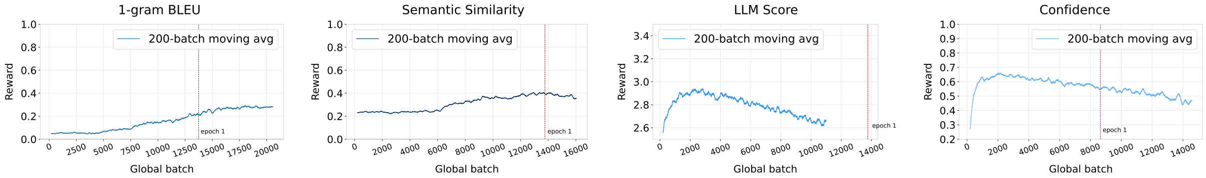

Figure 2: Reward trends over training steps for four reward metrics: 1-gram BLEU, Semantic similarity, LLM score, and Teacher Confidence-based. Vertical line marks the end of epoch 1. For LLM score and confidence, the y-axis range has been restricted to enhance visibility of trends, and does not represent the full possible reward scale.

provide a meaningful degree of fine-tuning before instability occurs.

1-gram BLEU Reward The reward starts off very low and remains low for a substantial number of steps before beginning to increase steadily. Given that this is a unigram-based metric focused on token overlap between the generated utterance and a reference, a slow and gradual increase is actually desirable, a sharp rise could lead the model to simply replicate the caregiver's utterances. For *Zorro*, this model achieves the lowest score among all those evaluated, and it also ranks among the least accurate models on the *dialogue minimal pairs*. Although it is a word-based metric, no further improvements on the *lexical decision* data can be reported.

Semantic Similarity Reward The reward increases gradually during training, similarly to what is observed for BLEU. However, the overall improvement across training steps is modest, and the reward values remain relatively low. On *Zorro*, the model's accuracy stays roughly at the level of `llamalogue`. Additionally, performance on the *dialogue minimal pairs* shows a slight decline of a few percentage points compared to the pre-trained model. The score on the *lexical decision* task is the lowest observed among all the fine-tuned models.

LLM-generated Reward Here, the reward increases during the very early phase of fine-tuning, although only by approximately 0.5 on a scale ranging from 0 to 5. This limited growth indicates that the OLMo model used to assign the reward rarely utilized the full range of available values. In particular, scores of 0 or 5 were almost never assigned to generated utterances. Starting from around step 3000, the reward begins to decline steadily. The model evaluated at checkpoint-5000 maintains a relatively strong performance on *Zorro*. However, similar to the previous two PPO models, there is a decrease in accuracy on the *dialogue MP* task compared to `llamalogue`. Zorro and *lexical decision* scores stay roughly equivalent to the base model.

Teacher Confidence-based Reward The initial reward being around 0.2 means the `llamalogue` reply was already above the median among the teacher's ten candidates. At the start of fine-tuning the reward increased quickly, probably due to the initially small KL coefficient value. During fine-tuning, the reward rose to around 0.6, meaning the fine-tuned model beats roughly eight of the teacher's alternatives. After epoch one, the reward curve had a slight dip to around 0.5. On the *lexical decision* task, the model is roughly on par with `llamalogue`, but lower on *Zorro* and (slightly) *dialogue MP*.

5 Discussion and Conclusion

How can these slightly underwhelming results be explained? First, we need to emphasize that our dialogue-only models, trained on child-directed and child speech, are exposed to a smaller vocabulary (Snow and Ferguson, 1977) and simpler structures (Genovese et al., 2020) than found in adult speech (although complex structures are occasionally found in CDS, they are rare, cf. Cameron-Faulkner et al., 2003). As the benchmarks included in BabyLM target broader lexical and syntactic variation in the input, there is a slight mismatch between our data and the evaluation data. The accuracy on lexically restricted Zorro, for example, is much higher than the one reported for BLiMP. More generally speaking, these results also align with previous findings on other models trained on CDS only (cf. Padovani et al., 2025; Bunzeck et al., 2025). Where our models excel is the domain of dialogue minimal pairs. There, they outperform the base model by a margin of 10%. While it is not overly surprising that our model masters a task that aligns 100% with its pre-training goal and the shape of its data, learning dialogue coherence is still far from easy. Judging contingency and coherence without lexical overlap requires a different kind of linguistic knowledge than syntactic phenomena like island effects – exactly the kind of knowledge our model picks up.

With respect to the performance of our fine-tuned models, it is important to note that our results align with previous studies (Liu and Fourtassi, 2025; Stöpler et al., 2025), which all found no significant improvements on grammatical or similar benchmarks after interaction-driven fine-tuning. Such fine-tuning with a specific, pragmatics- or communication-based goal in mind has so far only shown to improve performance on benchmarks that also test for this goal. Our DPO fine-tuning, which directly optimizes preference for correct answers, does have a positive effect on the model preferring such answers from a held-out test set. In contrast, more generalized optimization for communicatively appropriate generations with PPO does not have this effect. It remains open to further inquiry whether our scoring methods might be too abstract. After all, they are only indirectly aligned with all the different evaluation measures we want to optimize for (correct grammar, world knowledge, approximation of human reading behaviour, AoA estimation, etc.). Also, if the one, singular answer that we compare with our generation in all PPO training regimens is too distant to the generated answer (semantically, pragmatically, lexically, etc.), then the provided training signal might steer the model's weights into incorrect directions or leads to it getting stuck in local optima (exemplified by the non-monotonic reward trends).

Finally, as the differences between DPO with naturally occurring and synthetically generated answers are quite large for the dialogue MP performance, this hints towards a shortcoming of current LLMs: despite generating language that superficially resembles CDS being easy, generating authentic interactions is actually hard. For example, Feng et al. (2024) generate synthetic dialogues which differ tremendously from real caretaker-child interactions – the utterances are not fragmentary, highly verbose and complex. Räsänen and Kocharov (2024) train a CDS model from scratch, which approximates many statistical tendencies of CDS, but often generates nonsensical or ungrammatical utterances. While our model did not perform well on the general BabyLM benchmarks, a first qualitative inspection of its generative capabilities showed that it can actually continue dialogue in a plausible-looking way. Here, further experimentation with dialogue-based models is clearly needed.

Limitations

This study has several limitations that should be acknowledged. First – as previously discussed – the training data is narrowly focused on child-directed and child speech, which, while intentional for our research goals, constrains the model's lexical diversity and syntactic variety. This domain-specific bias limits its generalization to broader linguistic contexts, as evidenced by weaker performance on benchmarks that target a wider range of grammatical phenomena such as BLiMP. The incorporation of adult–adult dialogue into our training regimen might be a promising direction for future research. However, our primary objective in this study was to optimize the child component's conversational turns in dialogic interactions with caregivers, while testing if this also enhances secondary objectives like semantic relevance, common-sense reasoning, and linguistic competence. In child language development, these abilities emerge through interleaved phases/periods characterized by imitation and strong reliance (exploitation) on parental input, and others dominated by exploration of self-generated abilities and emergent capacities. Transposed to the context of a reward function guiding model competencies over time, this developmental dynamic could, for example, suggest the use of a curriculum-based reward schedule across fine-tuning steps. Such a schedule could involve intensifying the reward signal during certain stages and attenuating it during others, or alternatively optimizing different aspects of verbal production at distinct developmental phases of the model. Notably, our study did not incorporate such a curriculum in the reward design, which may have limited the effectiveness of the PPO fine-tuning. It would be interesting for future work to explore this direction and assess whether exploration/exploitation reward patterns inspired by human developmental trends could yield greater benefits for model fine-tuning.

Furthermore, our fine-tuning phases with DPO and PPO were conducted without a previous extensive hyperparameter search. As a result, the (in-)effectiveness of our proposed reward functions and their learning dynamics remain open for further exploration. Importantly, our project is intended as a pilot study. While we put more emphasis on the comparison of and experimentation with a broad variety of studies, future work should place greater emphasis on systematically identifying the optimal hyperparameters for each reward function prior to training, thereby ensuring that observed effects can be more confidently attributed to the reward design itself rather than possibly suboptimal fine-tuning setups.

Supplementary Materials

In addition to being conveniently available on Huggingface and GitHub, the long-term accessibility of the datasets, models, and code for the DPO and PPO experiments is ensured via a data publication on Zenodo: https://doi.org/10.5281/zenodo.17253651

Acknowledgments

BB, MA, OM, HB and SZ acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): CRC 1646/1 2024 – 512393437 projects A02, A05, and B02. FP and AB were supported by the Talent Programme of the Dutch Research Council (grant VI.Vidi.221C.009).

References

Heike Behrens. 2021. Constructivist approaches to first language acquisition. *Journal of Child Language*, 48(5):959–983.

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242.

Bastian Bunzeck and Holger Diessel. 2025. The richness of the stimulus: Constructional variation and development in child-directed speech. *First Language*, 45(2):152–176.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 169–186, Vienna, Austria. ACL.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shapes of BLiMP: Syntactic learning curves in language models are not uniform, but sometimes unruly. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 39–55, Gothenburg, Sweden. ACL.

Bastian Bunzeck and Sina Zarrieß. 2025. Subword models struggle with word learning, but surprisal hides it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 286–300, Vienna, Austria. ACL.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive Science*, 27(6):843–873.

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

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

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts bert. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 210–224, Singapore. ACL.

Noam Chomsky. 1986. *Knowledge of Language. Its Nature, Origin and Use.* Praeger, New York, NY, USA.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8281–8297, Dublin, Ireland. ACL.

Andrea Gregor De Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. ACL.

Giuliana Genovese, Maria Spinelli, Leonor J. Romero Lauro, Tiziana Aureli, Giulia Castelletti, and Mirco Fasolo. 2020. Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features. *Journal of Child Language*, 47(1):22–44.

Alison Gopnik. 2020. Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, 375(1803):20190502.

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

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. ACL.

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. ACL.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. Elements of World Knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. *Preprint*, arXiv:2405.09605.

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

Seokyung Kim and Stephanie M. Carlson. 2024. Understanding explore–exploit dynamics in child development: Current insights and future directions. *Frontiers in Developmental Psychology*, 2:1467880.

Stephen C. Levinson. 2025. *The Interaction Engine: Language in Social Life and Human Evolution*. Cambridge University Press, Cambridge, UK.

Jing Liu and Abdellah Fourtassi. 2025. Benchmarking LLMs for mimicking child-caregiver language in interaction. *Preprint*, arXiv:2412.09318.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *Preprint*, arXiv:2505.24864.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. ACL.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2025. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 991–1010, Albuquerque, NM, USA. ACL.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum, Mahwah, NJ, USA.

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

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

Mitja Nikolaus and Abdellah Fourtassi. 2025. Modeling children's grammar learning via caregiver feedback in natural conversations. *Preprint*, OSF:e6kv3_v1.

Kate Nussenbaum and Catherine A. Hartley. 2019. Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, 40:100733.

Jeffrey Olmo, Jared Wilson, Max Forsey, Bryce Hepner, Thomas Vincent Howe, and David Wingate. 2025. Features that make a difference: Leveraging gradients for improved dictionary learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7609–7619, Albuquerque, NM, USA. ACL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. *Preprint*, arXiv:2505.23689.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. ACL.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran.

Okko Räsänen and Daniil Kocharov. 2024. Age-dependent analysis and stochastic generation of child-directed speech. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, pages 5102–5108, Rotterdam, The Netherlands.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. ACL.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Catherine E. Snow and Charles A. Ferguson, editors. 1977. *Talking to Children: Language Input and Acquisition*. Cambridge University Press, Cambridge, MA, USA.

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

Jean-Loup Tastet and Inar Timiryasov. 2024. BabyLlama-2: Ensemble-distilled models consistently outperform teachers with limited data. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 292–301, Miami, FL, USA. ACL.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA, USA.

Michael Tomasello. 2005. Beyond formalities: The case of language acquisition. *The Linguistic Review*, 22(2-4):183–197.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran, Vancouver, Canada.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. ACL.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–6, Singapore. ACL.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9370–9393, Toronto, Canada. ACL.

Xingmeng Zhao, Tongnian Wang, Sheri Osborn, and Anthony Rios. 2023. BabyStories: Can reinforcement learning teach baby language models to write better

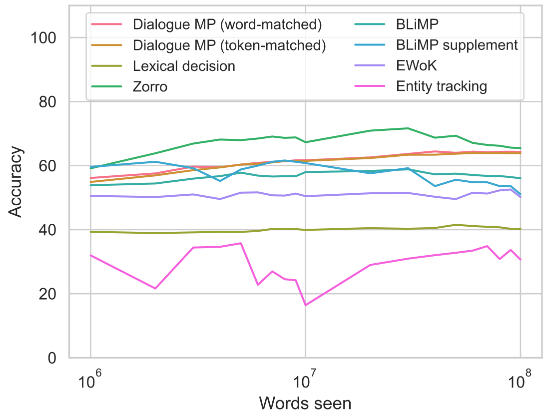

Figure 3: Learning trajectories for our base model across pre-training for 10 epochs. Note that the *x*-axis is log-scaled to make the very early training dynamics more visible.

stories? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 158–169, Singapore. ACL.

A Learning trajectories across pretraining

To trace the learning process of `llamalogue`, we continually evaluate it during pretraining. We benchmark ten checkpoints across the first epoch (so after each 1M token set has been seen by the model once, until 10M tokens are reached) and then nine further checkpoints over the remaining nine epochs. We visualize the development of performance on eight different minimal pair sets in Figure 3.

The worst performance can be observed for the entity tracking evaluation – performance does not stabilize at all and oscillates between 20–40%, which means that our model actively disprefers correct continuations. The same goes for the lexical decision data, where our model consistently scores around 40%. Performance on EWoK stays around the chance baseline as well. Interestingly, our model surpasses 60% on the BLiMP supplement data around approximately 7M tokens, after which performance deteriorates again. Similarly, BLiMP performance increases slightly early on, but then also stabilizes at a low level. Accuracy scores on Zorro, the scaled-down derivative of BLiMP that only contains words also occurring in CHILDES, are generally higher and improve until the third epoch of training, after which they deteriorate again. The only stable, monotonically improving learning trajectory can be observed for our dialogue minimal pairs. This, however, is not overly surprising, as this testing paradigm aligns closely with the pretraining goal of `llamalogue`. Viewed in conjunction with our general results, these learning trajectories further corroborate the fact that the general BabyLM evaluation measures are not very suitable for our models, as the decreasing learning trajectories hint towards our models not being undertrained and because comparable studies of learning dynamics overwhelmingly report power-law like curves (cf. Huebner et al., 2021; Liu et al., 2021; Choshen et al., 2022; Bunzeck and Zarrieß, 2024; Padovani et al., 2025).

B DPO Datasets

Table 4 shows a sample of sentences from the dataset we used to fine-tune the model with DPO. The appropriate and random sentences are matched in terms of token length, and both come from the distribution of sentences actually observed in CHILDES.

Table 5 was also used to fine-tune `llamalogue` with DPO. In contrast to the previous case, the appropriate sentences here are synthetic, artificially generated by Llama-3.2-3B, and their length is not matched to that of the random counterparts.

C PPO Reference Child Responses

In Table 6, we show a sample of 3 prompts used during fine-tuning and the 10 ground truth answers generated by Llama-3.2-3B when it is asked to simulate a child responding to a caregiver's sentence, using the prompt shown in detail in Table 1.

D (Super)GLUE results

We report the results for the SuperGLUE tasks in Table 7. Here, we can generally report that fine-tuning with DPO and PPO has only very little effect on our models' advantages for further fine-tuning. In comparison to the baseline model trained on the whole BabyLM corpus, they are generally worse base models for fine-tuning on (Super)GLUE.

Prompt (MOT)	Appropriate (CHI)	Random (CHI)
what is that ?	it looks like a gun .	you do it like that .
pull the string .	and where do they hook it ?	do you know what it was ?
I think they just throw it on the side .	you know what Mom ?	I get this hole .
what are you playing with huh toys ?	there's a dog .	there's the sports .
the bottom ones come off .	want to know what ?	we stole the brush .
can you say that ?	okay the hungry hungry caterpillar .	yeah I want that too .
what is it ?	a baby caterpillar !	I'm just pretending .
what is it ?	I don't .. know !	put my dress down .

Table 4: Examples of naturalistic DPO dialogue pairs. Each row shows a caregiver's utterance (MOT) as prompt, with corresponding appropriate and random child responses (CHI) used for preference training.

Prompt (MOT)	Appropriate (CHI)	Random (CHI)
what is that ?	it is a ball.	you do it like that .
pull the string .	i can do it!	do you know what it was ?
I think they just throw it on the side .	the garbage truck.	I get this hole .
what are you playing with huh toys ?	i'm playing with this toy, it's a nice toy.	there's the sports .
the bottom ones come off .	the bottom ones come off .	we stole the brush .
can you say that ?	yes, mother can say that.	yeah I want that too .
what is it ?	i don't know.	I'm just pretending .
what is it ?	a big tree.	put my dress down .

Table 5: Examples of synthetic DPO dialogue pairs. Each row shows a caregiver's utterance (MOT) as prompt, with corresponding appropriate child responses (CHI) generated using the Llama-3.2-3B teacher model, and random responses from naturalistic mismatches.

Prompt (MOT)	Generated child-like responses (LLaMA-3B)
I was wondering who told you your neck needed washing.	1. It was a frog. 2. The neck was washed, so I asked the dog to lick it. 3. No one told me. 4. I don't know, mom. 5. I don't know, but I think I'll wash it now. 6. I don't know, I just felt it needed washing. 7. I think it was you. 8. It was the person in the mirror. 9. Mommy told me to wash my neck. 10. Mother, I was wondering who told you my neck needed washing.
I think they're all friends together, aren't they?	1. Yes, mother, I think so. 2. I think they're all friends together, aren't they? 3. I think they're all friends together, aren't they? 4. Yes, they are. 5. Yes, mother, they are. 6. Yes, I think so. 7. Yes, they are, Mom. 8. I think they're all friends together, aren't they? 9. Yes, I think so, aren't they? 10. Yes, they are friends.
You're a nice girl.	1. you're a nice girl too. 2. I'm sorry, I'm not a nice girl, I'm a bad girl. 3. Yes, I am. 4. I am a nice girl. 5. thank you, mom. 6. yes, I am a nice girl. 7. thank you mother. 8. I'm nice too. 9. thank you, mom. 10. yes I am.

Table 6: Caregiver prompts and ten possible child-like answers generated by the Llama-3.2-3B model.

		DPO		PPO				
Task	`llamalogue`	**Natural.**	**Synth.**	**Bleu**	**SemSim**	**LM Score**	**Conf.**	**Baseline**
BoolQ (acc)	64.04	64.04	64.04	64.04	64.04	64.04	64.04	68.38
MNLI (acc)	35.17	35.17	34.92	34.60	34.82	35.23	34.92	61.04
MRPC (F1)	80.95	80.95	80.95	81.31	80.95	81.31	80.95	83.61
QQP (F1)	10.28	10.28	10.17	5.55	11.13	10.37	11.17	71.82
RTE (acc)	53.24	52.52	53.24	53.24	54.68	51.80	53.24	61.15
MultiRC (acc)	57.55	57.55	57.55	57.55	57.55	57.55	57.55	65.92
WSC (acc)	61.54	61.54	61.54	61.54	61.54	61.54	61.54	63.46

Table 7: SuperGLUE results.