Learning to Speak Like a Child: Reinforcing and Evaluating a Child-level Generative Language Model

Enoch Levandovsky
enochlevandovsky@u.boisestate.edu
Computer Science
Boise State University

Anna Manaseryan
annamanaseryan@u.boisestate.edu
Computer Science
Boise State University

Casey Kennington
caseykennington boisestate.edu
Computer Science
Boise State University

Abstract

A language model that can generate utterances that are appraised as being within a specific age range of a young child who is beginning their language learning journey can be useful in scenarios where child-level language is needed, for example in virtual avatars, interactions with individuals who have disabilities, or developmental robotics. In this paper, we focus on an age range that is not represented in prior work: emergent speakers. We use the CHILDES database to train and tune language models of different parameter sizes using a group relative policy optimization reinforcement learning regime. Our goal is to find the most coherent, yet child-like language model while keeping the number of parameters to as few as possible. We evaluate using metrics of coherency, "toddlerality," and an evaluation using human subjects who interact with two robot platforms. Our experiments show that even small language models (under 1 billion parameters) can be used effectively to generate child-like utterances.

1 Introduction

A rich setting of spoken dialogue is the interaction between children and their caregivers. The primary and fundamental site of language use is in this child-caregiver interaction, which is a co-located spoken dialogue setting (Smith and Gasser, 2005) beginning with first words, short phrases, and simple sentences that later complex syntax and abstract words build on. Early child speech is often produced in *installments*, requiring more dialogue turns to convey information. In contrast, most language models (LMs) work on full utterances—not installments—and have been trained on vast amounts of text, the majority of which likely comes from sources that are not written for children specifically, and certainly not in terms of direct transcriptions of child utterances.

Because LMs are trained primarily with text, the data reflects language written by and for people who are in literate age ranges, roughly 6 years old or older (the age when U.S. children begin learning how to read). This is an unfortunate oversight because language phenomena from child-target sources can be very insightful and useful. Romero and Razniewski (2022) showed that child texts help improve commonsense reasoning in LMs because child texts don't assume domain knowledge; instead, 'commonsense' explanations are clearly stated. An important resource for child-caregiver interactions is the CHILDES project (MacWhinney, 2000; Sanchez et al., 2019), a dataset of children-caregiver interactions with annotated transcriptions of speech as well as some gestures. CHILDES has been used as a resource for recent work including learning about word acquisition (Mahon et al., 2025), as well as early syntax and multlinguality (Fitzgerald, 2024). Furthermore, CHILDES has been used as a resource for pre-training LMs including BabyBERTa (Huebner et al., 2021) and for the BabyLM challenge, showing an increasing interest in compact, effective LMs that are trained on only 10 or 100 million tokens taken from child-produced spoken transcriptions.[1] One outcome of the BabyLM challenge was the Babystories model (Zhao et al., 2023) that showed improvements on downstream tasks using LMs that are reinforced with stories written for children. Taken together, this recent work focusing on child-level LMs shows that there are benefits to using child-level data.

An LM that can generate utterances finetuned to match the developmental language patterns of young children can be useful in scenarios requiring child-level communication. Applications include virtual avatars, interactions with individuals who have disabilities, or developmental robotics. Some LMs have been trained on transcriptions of child speech from young, emergent speakers (Zhao et al., 2023; Malik et al., 2024), but they have not been evaluated to generate utterances that are age appro-

[1] babylm.github.io

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 370–382
Aug 25–27, 2025.

priate for the youngest range of speakers, and furthermore have not been evaluated in interactive dialogue settings. In this paper we make two primary contributions: (1) we systematically explore how small LMs can effectively be reinforced on small amounts of child-level data and be able to generate utterances that are regarded as being at a specific child level, with focus on emergent, toddler-level speech (age 1-3), and (2) construct a LM with as few parameters and with as little data as necessary to produce child-level dialogue utterances. We systematize our evaluations using coherency (explained below), as well as a human evaluation on two robot platforms in a dialogue setting. Our results show that our method for reinforcement works effectively, enabling researchers to access a model that produces child utterances.

2 Related Work

Small Language Models LMs trained on huge datasets dominate, but many are turning to smaller LMs that are tuned to do specific tasks well. Small LMs that are more child developmentally plausible are the focus of the BabyLM challenge (Warstadt et al., 2023; Haga et al., 2024; Hu et al., 2024). Related work to ours include Zhao et al. (2023), a small model that was improved by using reinforcement learning on child texts; and Fields and Kennington (2023) and Bunzeck and Zarrieß (2023) both explore how transformer-based LMs can be effective, even with fewer than 10 million parameters on discriminative tasks. We follow this line of research, but instead of discriminative models, we work towards a harder generative model task that can produce child-level utterances which could be used for SDS dialogue.

Child-level Language Models LMs are increasingly used to identify early onset of certain linguistic abilities in children (e.g., determiner-noun onset (Alhama et al., 2024)). More general-purpose LMs have been trained on datasets derived from and designed for children including BabyBERTa (Huebner et al., 2021) and ChildGPT Romero and Razniewski (2022), though the latter was trained on child-targeted texts, not on transcriptions of child speech. As part of the BabyLM challenge, other recent work explored just how small transformer-based LMs could work on small amounts of data from CHILDES including GPT-wee (Bunzeck and Zarrieß, 2023) and Electra-Tiny (Fields and Kennington, 2023; Fields et al., 2023). However, these LMs are mostly discriminative, not generative.

Readability Readability is an area of research that attempts to automatically determine the appropriate age or grade level a child needs to be in order to comprehend a given text. For example, a short sentence *I am hungry* can be understood by very young children, while *The nomenclature of academic disciplines* requires more knowledge of abstract words. Approaches to the readability task include a number of formulae that use specific features such as number of sentences, average number of words in each sentence, and comparison to words on word lists maintained by educators, such as the Spache (Spache, 1953), Spache-Allen (Allen et al., 2022), and Flesch-Kincaid formula (Kincaid, 1975). Others have attempted to use LMs coupled with more linguistically-derived features with varying degrees of success (Lee et al., 2021). However, the literature on Readability focuses on children in their first years of education (Kindergarten through 5th grade in the U.S.), roughly beginning at age 6. What happens before that–children who are in their first years of speaking, ages 1-5—is what we focus on in this paper.

The most closely related to our work is Malik et al. (2024) controlling the language proficiency level of LM generation to a specific level, but focuses on Readability (they used Fleisch-Kincaid to guide the loss function to fine-tune a LM), whereas we are interested in children ages 1-3, younger than the target ages for Readability. Moreover, we are interested not in what kids are able to read, but what they are able to say that is age appropriate.

Group Relative Policy Optimization Group Relative Policy Optimization (GRPO) is a deep Reinforcement Learning (RL) training regime that enables fine-tuning of high-capacity models such as Large Language Models (LLMs) by leveraging reward models to guide learning (Shao et al., 2024; Ouyang et al., 2022). At each training step, GRPO uses feedback from the reward model to compute gradients and update the policy weights, allowing for stable, preference-aligned optimization in high-dimensional action spaces. Unlike other LM training regimes, GRPO only requires an input prompt; at each step the model generates N outputs, which are then scored by the reward models, and the model weights are updated accordingly.

Creating a robust reward model is critical to the

Age	0	1	2	3	4	5	6
# Utts	26K	118K	364K	220K	328K	81K	39K
word/utt	1.3	2.6	4.1	5.7	6.3	6.5	7.4

Table 1: Count of utterances for each child age 0-6

success of GRPO. A well-designed reward model must provide clear, incremental feedback to reliably guide the policy toward desirable behavior. However, poorly constructed reward models can suffer from local maxima, where the optimization process prematurely converges to suboptimal behaviors. This issue is particularly common in reward models based on frozen embeddings (e.g., BERT-based models), which may not generalize well or provide smooth reward signals. To address this, approaches such as Deepseek employ ensembles of reward models, as well as logic-driven or structural reward functions—such as regex-based heuristics—to improve signal diversity and reduce the risk of overfitting to a single flawed objective.

3 Method

Overview Our primary goal in this work is to understand how humans perceive and interact with robots that use a language model specifically tuned to produce utterances within the 1–3 year-old age range. A secondary focus is to explore and document the practical methods—specifically, reinforcement learning—that can be leveraged to achieve such child-like language production in LLMs. While model training details, including our use of GRPO, are important for reproducibility, our central interest lies in characterizing the human response to age-targeted robot speech.

To train an LM that can generate utterances within our targeted age range (1-3), we use a reinforcement learning (RL) regime of group relative policy optimization (GRPO), following Shao et al. (2024). As a preliminary step, we designed and fine-tuned a discriminator model which is to be used as our primary GRPO *reward model* (RM). This discriminator model is tasked with rewarding the GRPO fine-tuning process to encourage child-level (main target is 1-3 years of age). We use the appropriate caregiver questions extracted from the CHILDES dataset as the prompt inputs for the GRPO task and we expect the model to respond with the appropriate child language response. We explain each aspect of our method below.

Data The CHILDES dataset (MacWhinney, 2000; Sanchez et al., 2019) contains transcribed utterances of spoken interactions between children and their caregivers. Table 1 shows the data for each age, 1 through 6 years, for a total of 493,638 child utterances between those age ranges. We removed all additional non-spoken dialogue CHILDES annotations such as physical actions done by the child. We also added the preceding care-giver utterances prior to the child utterance to help with dialogue training.

The GRPO trainer task uses a set of user prompts as input, in our case a caregiver question. At each step, the trainer generates a batch of outputs and uses a set of reward models to compute the reward of each output. The GRPO finetunes (specifically, reinforces) the model to maximize the rewards using KL divergence. The technique is to create a well-defined discriminative model which can reward in the direction we choose, such as behave like a child.

RM-1 The primary RM will reinforce the generative model to behave as a toddler under the age of three, we will call it *Toddler-BERT*. To create *Toddler-BERT*, we fine-tuned a BERT (Kenton et al., 2017) (HuggingFace `bert-base-uncased` model and tokenizer) model using 90% of the data for fine-tuning and 10% for evaluation. For *Toddler-BERT* training, we discard the utterances given by the caregivers and focus only on the child utterances. We then split the labels into a binary classification task: everything below 3 years old as positive label and everything 3 years and older as the negative label (roughly a 44/56% split). This is done because we are ultimately interested in the age range below 3 years, and because 2 is the most represented age in the data, allowing a model to easier discriminate between 3 and above. Evaluation using all ages 1-8 resulted in unreliable accuracies for using the model in a RL regime (usually $<$76%) even for recent models (e.g., DeepSeek-R1-Distill-Qwen-1.5B), which illustrates that LMs are not trained on child data.

We fine-tune for 5 epochs (loss function=binary cross entropy, learning rate=2e-5, weight decay=0.01, batch size=1024, max length=32). This resulted in an accuracy of 79.85%, which is well above the most common baseline. We evaluated other models including those pre-trained on child data, such as Romero and Razniewski (2022) and larger models like GPT-2 (Radford et al., 2019), but

the former had worse accuracy, and while the latter did have slightly higher accuracy (0.02% relative improvement), we opt for BERT because it works nearly as well with far fewer parameters. Examples of *Toddler-BERT* score outputs are shown in Figure 6 of the appendix as RM-1.

After running GRPO with *Toddler-BERT* alone, we noticed that although the reward value of RM-1 reached >90% thresholds, the actual outputs were nonsensical and incoherent. As mentioned in prior work, we expected initial GRPO trained runs to reach undesired local maxima. To fix this, we will employ the following steps iteratively until desired results are achieved.

1. Add or correct RM training data.
2. Add an additional logic based RM eg. Regex clf
3. Add an additional BERT based RM eg. Coherence clf

RM-2 During our initial experiments, we observed that when using *Toddler-BERT*, our generative model lost coherency as RM-1 encouraged something akin to cognitive degradation (i.e., short utterances that were child-level, but responses to questions seemed random). This motivated us to finetune another BERT coherency model. We use it as a second reward model to stabilize coherency when finetuning our generative model with GRPO. We opted to create an additional coherence-based RM.

To collect the finetuning data for RM-2, we created a prompt engineering pipeline that takes batches of 16 parent-child conversation pairs from CHILDES. We then used Llama-3.3-70B to score the coherence of the child responses between 0.0 and 1.0.[2] We found that batched processing in one LLM query resulted in better consistency of scores. The exact prompts for this are in the Appendix.[3]

We applied similar fine-tuning parameters to RM-2 as we did to RM-1, except that we applied soft labels during the training portion. The model was trained for 5 epochs (loss function=BCEWithLogitsLoss, learning rate=2e-5, weight decay=0.01, batch size=150, max length=96). We introduce a larger max length as this model needs to support both child and caregivers utterances as input. Examples of RM-2 score outputs are shown in Figure 6 of the appendix.

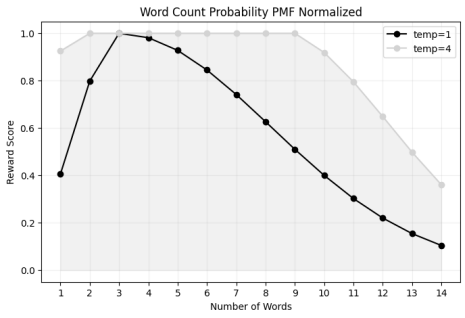

Figure 1: Reward scores given to sentences based of the distribution of words per sentences in CHILDES.

RM-3 After performing a GRPO training run with RM-1 and RM-2, we also found it necessary to encourage the model to respond with an appropriate number of words[4] To achieve this, we constructed a probability mass function (PMF) by computing Bayesian-based scores for various response lengths observed in our dataset. These scores were then min-max normalized to the range [0, 1] to ensure consistency. Additionally, we introduced a temperature parameter to control the sharpness of the distribution, enabling smoother length reward scores. An example RM-3 output is shown in Figure 1. To further encourage single-sentence responses, we applied the following function: $\frac{1}{\max(1,\text{number\_of\_punctuations})}$. Each completion is split into individual sentences in which then the number of words in each sentence is computed. Then a normalized PMF score based off the CHILDES dataset lengths is assigned; this is then scaled by the penalty function above, encouraging length appropriate, single-sentence responses.

After training the generative model with all the previous RMs, we noticed that the generative model would add in random baby words like *mommy, diaper, ball*, etc., even where it was not appropriate. We opted to adjust *Toddler-BERT* training data to add a few examples to discourage the random use of those words.

RM-4 Although at this point the generative model was mostly good 70% of the time, it was still not sufficient for human-robot interaction experiments. The model would produce relatively

[2] https://huggingface.co/RedHatAI/Llama-3.3-70B-Instruct

[3] A simpler approach would be to collect sentence pairs, randomizing half of the dataset pairs against each other, and labeling them for coherence with 1 and 0, respectively. However, our current strategy led to more stable GRPO training.

[4] RM-1 and RM-2 gave higher rewards to very long utterances, as might be expected as standard generative LLMs tend to produce long utterances.

non-contextual responses such as *Do you want to eat crackers -> Yes and play with train*. We take this as common for a child to assume context in parent-child interactions. But extreme cases are not desired during our human-robot interactions experiments. We believe this was the case due to poor quality care-giver questions given

Good quality prompts help generate relevant responses during GRPO training. To filter out helpful care-giver question utterances, we applied a similar approach to that of the RM-2 data collector, except that we prompted the LLM to score the **clarity** of the caregivers' utterance. This would effectively allow us to extract from the CHILDES dataset clear questions from the caregiver such as *What is your favorite color?* or *Do you want to play with blocks?* and filter out off-context questions or random text such as *and he starts jumping!* or *All three of them?*. Although RM-4 was not used as a reward model for GRPO, we use it to filter the top 10% helpful caregiver questions. We found that GRPO trained more stably as we increased the filter, but for diversity of the dataset, we kept it at 10%. Examples of RM-4 score outputs are shown in Figure 6 of the appendix.

Through iterations we found that these reward models and data filters were sufficient to reach our desired coherency and child response behavior.

3.1 Chosen LMs

At the time of the study, we found that the HuggingFace SmolLM2 as good candidates due to their small size and high benchmark performance compared to models of similar sizes. The models were released with parameters in increments of 135M, 360M, and 1.7B which allows us to apply the same fine-tuning techniques consistently on different model sizes, which we explain below.

4 Experiment

Our experiment has two Steps: (1) use pre-trained LMs and determine the best, yet smallest architecture that we can use to effectively generate child utterances, and (2) using the best model size from (1) and pre-train a custom model using only child data. We explain each step in more detail.

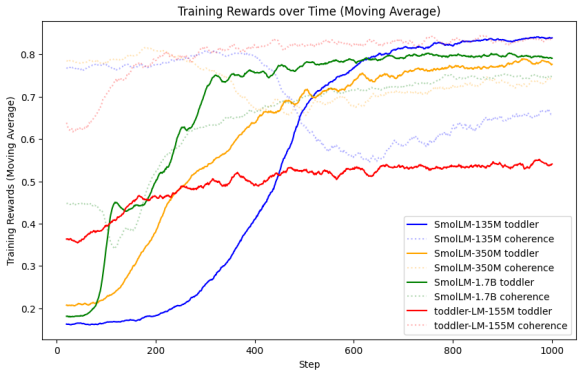

Figure 2: RM-1 and RM-2 counter balance each other. Giving too much weight to RM-1 (childish) can lead to non-coherent model whereas giving too much weight to RM-2 (coherence) can make the training process fail to allow the model to reduce cognitive behavior that of a child. We omit RM-3 results to remove clutter, however all models converged to 100% score with RM-3 due to its simplicity.

4.1 Step 1: Determining Best Model Using GRPO with Standard and Child-directed Utterances on Pre-trained LMs

In this step of our experiment, GRPO with Standard and Child-directed Utterances using Pre-trained LMs is designed to find the smallest model size needed for an LM to respond to a caregivers' question with coherent toddler-like utterance. At the time of the study there was no known language model designed to behave as a toddler. Thus for this experiment, we tune SmolLM2 models using GRPO with RM-1, RM-2, and RM-3 (explained above).

We use our GRPO RL regime introduced in Shao et al. (2024) to fine-tune SmolLM2 to produce child-level output with the following GRPO parameters: num_of_generations=8, batch_size=200, warmup_ratio=.1, max_completion_length=96, max_prompt_length=96, dtype=bfloat16, reward-weights=1.0, 0.2, 0.5 to RM-1, RM-2, RM-3 respectively. To allow for more efficient and smoother training, we used PEFT LoRa with the following parameters: rank=64, lora-alpha=64, target-module=[q,k,v,o,gate,up,down]. We ran this for 2000 steps, saving every 250th checkpoint.

Step 1 Results Figure 2 shows the training process across the 2000 steps for each of the pre-trained SmolLM models (135M, 350M, 1.7B parameters, respectively; for the moment, we ignore toddler-LM). Each model's toddler (i.e., child-level generated utterances) and coherence values change

over time; the objective is for both to be as high as possible, but since they tend to have counter objectives, we are looking for an optimal point where both are balanced. As expected, larger models become more child-like in their outputs sooner with higher coherence to begin with, but we also notice that the smallest model (135M parameters) reaches the highest child-level/toddler utterances, albeit with a lower coherence. The second smallest model (350M parameters) reaches a nice balance between the two, though the child-level utterance generation remains worse than the smaller model. These results inform us as to which architecture we should use for Step 2: For our custom, pre-trained model on child data, we opt for a model that is somewhere between the smallest and the second smallest models, which we focus on in the next section.

4.2 Step 2: Pre-training a Generative, child-level LM

In this Step, we develop a LM that has not been previously pre-trained; i.e., it begins from random parameters. The goal of this step is to examine how a LM trained exclusively on child-directed transcribed speech behaves compared to a model that has been pre-trained on large amounts of text containing adult-targeted examples, as we did in Step 1.

Task & Procedure We use data exclusively from CHILDES because that represents the target ages.[5] Given the limitations of our dataset size (approximately 14 million tokens from CHILDES after filtering comparing to 2 trillion token used by SmolLm-135M), we designed a more efficient LM architecture optimized for child-level conversation in only the English language rather than general knowledge. Key modifications include:

- reducing the vocabulary size to 8,196 tokens
- using a lower-case only tokenizer
- reducing max token length to a size of 256

Taking what we learned in Step 1, we initialized the model parameters to be sized between SmolLM-135M and SmolLM-360M with some key modifications that are itemized in Table 2.

[5]We considered using the BabyLM datasets for pre-training (Warstadt et al., 2023; Haga et al., 2024; Hu et al., 2024), which contain data from CHILDES, but some data that is out of the age range might also be included.

Parameter	Value
hidden_size	672
intermediate_size	1809
max_position_embeddings	256
num_attention_heads	12
num_hidden_layers	31
num_key_value_heads	4
rms_norm_eps	1.0e-05
rope_interleaved	false
rope_scaling	null
rope_theta	10000.0
tie_word_embeddings	true
vocab_size	8192

Table 2: Model Parameters for Experiment 2

We employed a three-stage training approach to develop our child-level language model pre-trained on CHILDES data:

Stage 1: Pre-training We used the Nanotron library with similar parameters to those used in SmolLM training. This produced an LM with basic language syntactic understanding. Key details of this stage include:

- Training continued until loss convergence just above 1.0
- Completed 25,000 training steps (approximately 64 epochs)
- Learning rate: 0.0025

Stage 2: Chat SFT After pre-training, our next goal was to train the model to learn to interact in a chat setting (which came for free in the pre-trained SmolLM models in Step 1). We modified the model and token configuration to adapt the same chat template and special tokens as SmolLM2 Instruct. We leveraged the unsloth library which had useful features that allowed us to train on responses only, eliminating the need to train the model to ask an adult question.

To arrive at the data we used at this stage, we selected the top N% most coherent child and parent sentences (using high scores from both the RM-2 and RM-4 reward models from Step 1). To ensure gradual adaptation to the chat format, we implemented a curriculum with progressively higher quality examples and decreasing learning rates, depicted in Table 3.

These parameters were carefully tuned until we consistently achieved a loss of 0.45, when the model started yielding coherent chat responses.

Subset Quality	Learning Rate	Epochs
Top 10%	9e-4	2
Top 5%	8e-4	14
Top 2.5%	7e-4	7
Top 1.25%	6e-4	3

Table 3: Training Stage 2 Settings by Subset Quality

Model	Accuracy
Gold	0.641
SmolLM-1.7B	0.643
SmolLM-350M	0.733
SmolLM-135M	0.729
Ours	0.740

Table 4: 1000 Questions model accuracies.

Stage 3: GRPO Optimization In the third and final stage, we applied the GRPO RL approach from Step 1 with minor modifications. We reduced the learning rate to 1e-5 and doubled the LoRa rank. We trained for 1000 steps and through manual analysis, we selected the checkpoint which yielded the best results for coherency and child-level 'toddler' utterance generation. Interestingly, while GRPO did not dramatically improve the model's performance as in Step 1, it was able to improve the model beyond Stage 2.

Metrics & Baselines The challenge with our very young child-level model (2-3 years of age) is meant to produce outputs that have limited vocabulary and syntactic structure. This means that using standard benchmarks for evaluating our LM is not applicable here. For our custom metric, we use RM-4 to extract the 1000 most useful caregiver questions from CHILDES. We then use the models we created in Step 1 and Step 2 from section 4 of this paper to generate simulated responses for every question. We then constructed an LLM prompt to score the coherency of the responses between 1-5 for every model and including the score of the original child response. We use ChatGPT 4.1[6] to score the models using the a prompt shown in Figure 5 of the appendix. These metrics are sufficient for this experiment, but we further evaluate with human participants, described in Section 5.

Results Figure 2 shows how our LM (termed *toddler-LM in the figure*, pre-trained on CHILDES data and reinforced through Stages 1-3 compares with the pre-trained SmolLM models. While not as effective in terms of child-level utterance generation (due likely to the vastly smaller amount of training data), it nonetheless results in high coherence, which is a reasonable tradeoff.

Table 4 shows the accuracy results of our evaluation on 1000 caregiver-child question-answer pairs. Though we used ChatGPT to score if responses were coherent answers, this result gives us an insight into the overall capability of the models to answer questions coherently. We see that the best pre-trained model that we reinforced through our GRPO regime is SmolLM-350M, but the best overall performing model is our 155M parameter model that was pre-trained on CHILDES data. This is a very positive result, as our smaller model works better according to this metric, but more evaluation is needed to answer if the model can be used in an interactive dialogue task. In the following section, we describe a human evaluating that compares our model with the SmolLM-350M model, the two highest-performing models.

5 Human Evaluation

Our offline metrics show that our child-level generative models are coherent and speak at a toddler-level age. To further evaluate both models, we performed a human evaluation.

Robots & System Because humans assign adult-level expectations to dialogue systems that can speak (Plane et al., 2018), it is important that we evaluate models on multiple robot platforms, where one platform is perceived as child-age, and another that appears older. We opt for the Cozmo robot which has been shown to have child-level qualities including morphology and voice (Plane et al., 2018). For the 'older' robot, we opt for Misty II since it is much larger than Cozmo, but still smaller than a human toddler. We use the default voice for both Misty II and Cozmo.

We used the rrSDS system for building dialogue systems on robotic platforms, which is built on Retico (Michael and Möller, 2019; Kennington et al., 2020). Our system consisted of a USB microphone, Whisper automatic speech recognition (ASR) (Radford et al., 2022), an rrSDS/Retico LM module that takes ASR transcriptions as text input and returns a response that is then uttered on the robot (each robot has a rrSDS control module). We test two models that resulted from the above Ex-

[6] https://platform.openai.com/docs/models/gpt-4.1

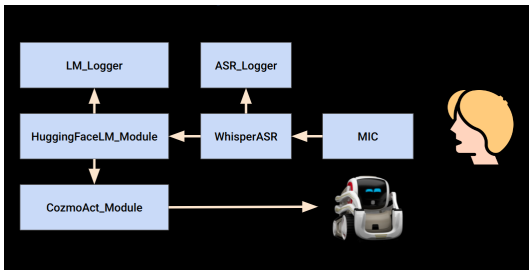

Figure 3: Depiction of our dialogue system that enables human participants to interact with a robot that uses a child-level LM to generate utterances.

periment, SmolLM-350M and our toddler model (155M parameters, pre-trained on CHILDES). The system is depicted in Figure 3. All modules that produce output are logged into a logging module for later analysis.

Task & Procedure We recruited 19 participants to come into our lab and interact with our robots. Each participant interacted with two of four possible robot+model pairs assigned at random:

- Cozmo + SmolLM-child (10 participants)
- Cozmo + Ours-child (6 participants)[7]
- Misty II + SmolLM-child (11 participants)
- Misty II + Ours-child (8 participants)

The robots were set on a table and participants were seated near the table so they could be face-to-face with the robot. We programmed the robots to perform small random movements to ensure to the participants that the robots were still functioning even when not speaking. Moreover, constant, small random movements are closer to what humans expect from each other, children in particular. After signing an informed consent,[8] participants were tasked with speaking with the robot. We made sure that participants understood that they needed to help drive the conversation by asking questions, but we did not prime them to assume that the robot would act a certain age. After interacting with the robot for 15-20 minutes, they filled out a survey.

Metrics We used the standard human-robot interaction Godspeed Questionnaire (Bartneck et al., 2009) which consists of 21 pairs of contrasting characteristics, each with a 5-point Likert scale between them. Some examples below:

- anthropomorphism (e.g., *artificial* vs. *lifelike*)
- likability (e.g., *unfriendly* vs. *friendly*)
- intelligence (e.g., *incompetent* vs. *competent*)
- interpretabilitiy (e.g., *confusing* vs. *clear*)

[7]The discrepancy in participants was partially a result of infrastructure failure with the Cozmo robot

[8]This study was approved by IRB.

Robot+model	human vocab	robot vocab	human AoA	robot AoA
Cozmo+SmolLM	3767	441	4.49	3.64
Cozmo+Ours	624	324	4.34	4.03
Misty+SmolLM	4219	450	4.38	3.8
Misty+Ours	758	368	4.34	4.02

Table 5: Results of objective measures.

Following Plane et al. (2018), in addition to these questions, we also ask participants about their perceptions of the robots' age and grade level. We also asked that the participants rate coherence (specifically, if the responses 'made sense') on a scale of 1-5 (5=very coherent). We also include average *age-of-acquisition* (AoA) values for words spoken by participants and generated by the models. The AoA dataset is a list of words and the estimated age when those words are learned (for example, *puppy* is learned at 3 years four months on average, whereas *quadratic* is learned at 14 years old).

In addition to questionnaires, we analyze the logs to determine several objective measures: the vocabulary of the participants and each robot+model, and the average AoA value for each robot+model pair. We hypothesize that the vocabulary of the human participants will be larger than that of the robot+model pairs, that the AoA values will be higher for the human-participant words compared to the robot+model words, and we anticipate that our model will be regarded as younger and more coherent than SmolLM-350.

Results Table 5 shows the objective results from the log files. We see overall that the vocabulary of the participants is much higher than that of the robots for all models, but particularly for our model, which is what we expect: children do not have as big of a vocabulary as adults. However, we can see from the average AoA scores that the participants did talk at a more child-friendly level to the robots, but overall the participants had higher average AoA scores than the robots. However, the SmolLM model had an overall average lower AoA score. This is somewhat unexpected, as our model was trained only on child data, but it does illustrate the effectiveness of the fine-tuning and RL regime for both models.

Table 6 shows a selection of results for survey questions (results for all questions can be

Question	mean	std
Robot/Model:	**Cozmo**	**SmolLM**
age	3.4	1.57
grade	Pre/Kind	N/A
coherence	3.0	0.82
artific/lifelike	3.3	0.82
dislike/like	3.6	0.97
unintel/intelligent	3.11	0.78
Robot/Model:	**Misty**	**SmolLM**
age	4.27	3.22
grade	Pre/Kind	N/A
coherence	2.45	0.52
artific/lifelike	3.0	0.77
dislike/like	3.0	0.77
unintel/intelligent	2.82	0.4
Robot/Model:	**Cozmo**	**Ours**
age	3.2	1.1
grade	Pre/Kind	N/A
coherence	3.33	0.82
artific/lifelike	3.5	0.55
dislike/like	3.8	0.84
unintel/intelligent	3.17	0.47
Robot/Model:	**Misty**	**Ours**
age	3.5	2.38
grade	Pre/Kind	N/A
coherence	2.12	0.64
artific/lifelike	1.88	0.64
dislike/like	2.75	1.04
unintel/intelligent	3.12	0.64

Table 6: Selection of Godspeed and additional question responses mean/standard deviations of Likert 1-5 scales. Higher Godspeed responses denote positive attributes (e.g., intelligence vs. unintelligence).

seen in the Appendix). We see that participants rated our model overall as closer to the target age (3.2 years for Cozmo and 3.5 years for Misty). All robot+model pairs were regarded as Pre-Kindergarten/Kindergarten aged (though Misty was sometimes rated higher; 2nd grade for example). There was no real difference in coherence between models, though participants rated Cozmo on average as higher than Misty. Cozmo is overall regarded by participants as more life-like, likable, and intelligent than Misty, despite housing the same models. This is an interesting result: humans likely expected Misty to act more mature/older, but the child-like speaking of Cozmo matched expectations of the participants.

Analysis & Discussion Even though the SmolLM model showed lower overall AoA levels for its vocabulary compared to our model, some examples illustrate why a LM pre-trained on child-level data might be preferable. In more than one case, the SmolLM model generated utterances that the robots spoke which were well above what a 2-3 year old child would be expected to be able to talk about (e.g., the participant mentioned cars and it responded with an utterance about Formula 1 racing). The model also spoke Spanish when asked about languages other than English; which is not a negative thing, but the model wasn't tuned to speak at a child-level in Spanish. We leave multilingual evaluation for future work.

6 Conclusion

In this paper, we modeled, trained, and evaluated child-level LMs. We evaluated multiple architectures using existing, pre-trained SmolLM models, and from that we determined the smallest architecture to pre-train a custom model on child-level data from CHILDES. Our training reinforced models to speak at a young child level (around 3 years of age), younger than prior models in the literature. Our human evaluations using the models on two robot platforms further illustrate their usefulness in human-robot interaction settings, but more importantly provided additional evaluation about the perceived age, coherency, and competence of the models. The results suggest that the small model trained only on child-level data is comparable to a larger model trained on much larger amounts of adult-level text, though using an existing pre-trained model and tuning it using our method to speak at child-level can be just as effective, albeit using a larger model.

For future work, we will incorporate the child-level LM into a human-robot interaction robot task where the robot and the human collaborate and the robot must learn from the human as it interacts.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 2140642. Finally, we would like to thank Beth Grenz for her help with human participants.

References

Raquel G Alhama, Ruthe Foushee, Dan Byrne, Allyson Ettinger, Afra Alishahi, and Susan Goldin-Meadow. 2024. Using computational modeling to validate the onset of productive determiner-noun combinations in english-learning children. *Proc. Natl. Acad. Sci. U. S. A.*, 121(50):e2316527121.

Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. 2022. Supercalifragilisticexpialidocious: Why using the "right" readability formula in children's web search matters. In *Advances in Information Retrieval*, pages 3–18. Springer International Publishing.

Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 7–18, Singapore. Association for Computational Linguistics.

Clayton Fields and Casey Kennington. 2023. Exploring transformers as compact, data-efficient language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Clayton Fields, Osama Natouf, Andrew McMains, Catherine Henry, and Casey Kennington. 2023. Tiny language models enriched with multimodal knowledge from multiplex networks. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 19–29, Singapore. Association for Computational Linguistics.

Kira Fitzgerald. 2024. From child speech to computational insights: Exploring the syntax of bilingual language acquisition.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. *arXiv [cs.CL]*.

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. *arXiv [cs.CL]*.

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.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. 2020. rrSDS: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21st Annual SIGdial Meeting on Discourse and Dialogue*, Virtual. Association for Computational Linguistics.

Ming-Wei Chang Kenton, Lee Kristina, and Jacob Devlin. 2017. BERT: Pre-training of deep bidirectional transformers. *arXiv:1810.04805*.

J P Kincaid. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*.

Bruce W Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian MacWhinney. 2000. The CHILDES project. *Computational Linguistics*, 26(4):657–657.

Louis Mahon, Omri Abend, Uri Berger, Katherine Demuth, Mark Johnson, and Mark Steedman. 2025. A language-agnostic model of child language acquisition. *Comput. Speech Lang.*, 90(101714):101714.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From tarzan to tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15670–15693, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thilo Michael and Sebastian Möller. 2019. ReTiCo: An open-source framework for modeling real-time conversations in spoken dialogue systems. In *Tagungsband der 30. Konferenz Elektronische Sprachsignalverarbeitung 2019*, ESSV, pages 134–140, Dresden. TUDpress, Dresden.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv [cs.CL]*.

Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. 2018. Predicting perceived age: Both language ability and appearance are important. In *Proceedings of SigDial*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Julien Romero and Simon Razniewski. 2022. Do children texts hold the key to commonsense knowledge? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10954–10959, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behav. Res. Methods*, 51(4):1928–1941.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv [cs.CL]*.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artif. Life*, (11):13–29.

George Spache. 1953. A new readability formula for primary-grade reading materials. *Elem. Sch. J.*, 53(7):410–413.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

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

A Appendix

A.1 Godspeed Question mean/std responses for Cozmo+SmolLM2

Godspeed Question	mean	std
fake/natural	3.00	0.82
machinelike/humanlike	2.50	0.53
artificial/lifelike	3.30	0.82
dead/alive	2.50	0.97
stagnant/lively	3.62	1.19
mechanical/organic	3.40	0.97
inert/interactive	2.33	0.71
apathetic/responsive	3.80	0.42
dislike/like	3.60	0.97
unfriendly/friendly	3.89	0.78
unkind/kind	4.20	0.63
unpleasant/pleasant	4.00	0.82
awful/nice	4.00	0.82
incompetetent/competent	4.00	0.67
ignorant/knowledgeable	2.50	1.18
irresponsible/responsible	2.70	1.06
unintelliget/intelligent	3.11	0.78
foolish/sensible	2.90	0.88
anxious/relaxed	2.89	1.05

A.2 Godspeed Question mean/std responses for Cozmo+Ours

Godspeed Question	mean	std
fake/natural	3.33	0.82
machinelike/humanlike	2.83	0.75
artificial/lifelike	3.50	0.55
dead/alive	2.50	0.84
stagnant/lively	3.50	0.55
mechanical/organic	3.67	1.03
inert/interactive	2.67	1.37
apathetic/responsive	4.00	0.63
dislike/like	3.80	0.84
unfriendly/friendly	4.00	0.63
unkind/kind	3.83	0.75
unpleasant/pleasant	3.67	0.82
awful/nice	3.50	1.05
incompetetent/competent	3.67	0.82
ignorant/knowledgeable	3.33	1.03
irresponsible/responsible	2.83	0.75
unintelliget/intelligent	3.17	0.41
foolish/sensible	3.17	0.75
anxious/relaxed	3.17	0.41

A.3 Godspeed Question mean/std responses for Misty+SmolLM2

Godspeed Question	mean	std
fake/natural	2.45	0.52
machinelike/humanlike	2.00	0.63
artificial/lifelike	3.00	0.77
dead/alive	2.36	0.67
stagnant/lively	3.18	0.87
mechanical/organic	2.91	1.04
inert/interactive	1.82	0.60
apathetic/responsive	3.45	0.82
dislike/like	3.00	0.77
unfriendly/friendly	3.27	0.79
unkind/kind	3.55	0.69
unpleasant/pleasant	3.55	0.69
awful/nice	3.64	0.67
incompetetent/competent	3.73	0.65
ignorant/knowledgeable	2.64	1.03
irresponsible/responsible	2.73	0.65
unintelliget/intelligent	2.82	0.40
foolish/sensible	2.82	0.87
anxious/relaxed	3.00	0.77

A.4 Godspeed Question mean/std responses for Misty+Ours

Godspeed Question	mean	std
fake/natural	2.12	0.64
machinelike/humanlike	1.75	0.71
artificial/lifelike	1.88	0.64
dead/alive	1.75	0.46
stagnant/lively	2.88	0.83
mechanical/organic	2.88	0.64
inert/interactive	2.00	0.93
apathetic/responsive	3.38	0.74
dislike/like	2.75	1.04
unfriendly/friendly	2.88	0.64
unkind/kind	3.38	1.06
unpleasant/pleasant	3.00	0.93
awful/nice	2.75	0.89
incompetetent/competent	3.25	0.71
ignorant/knowledgeable	2.38	1.06
irresponsible/responsible	2.12	0.99
unintelliget/intelligent	3.12	0.64
foolish/sensible	2.12	0.83
anxious/relaxed	2.62	0.74

A.5 Prompt examples

```
You are an expert linguist and cognitive scientist.
Critically evaluate each short conversational exchange
between an adult and a child on three dimensions:

1. Adult utterance coherence (0.0–1.0):
   Measures if the adult's line makes sense by itself.
   Example – High (0.9): "What's your favorite toy?"

2. Child response coherence (0.0–1.0):
   Measures if the child's reply relates reasonably to the
adult.
   Example – High (0.8): Adult: "Do you want a snack?" Child:
"Yes, please!"

3. Child strict context coherence (0.0–1.0):
   Measures if the child's reply accurately addresses the
specific context set by the adult.
   Example – High (0.9): Adult: "How many cookies are here?"
Child: "Three."

Return exactly one JSON object mapping each id to its three
scores, for example:

{
  "0": {
    "adult_utterance_coherence": 0.85,
    "child_response_coherence": 0.80,
    "child_response_strict_context_coherence": 0.75
  },
  "1": {
    "adult_utterance_coherence": 0.30,
    "child_response_coherence": 0.25,
    "child_response_strict_context_coherence": 0.10
  }
}

Only output the JSON. No explanations or extra text.

{% endraw %}

{% for convo in conversations %}
id: {{ convo.id }}
adult_utterance: "{{ convo.adult_utterance }}"
child_utterance: "{{ convo.child_utterance }}"

{% endfor %}
```

Figure 4: Prompt used to collect soft labels for Reward Model 3 and 4.

```
Here is the csv

{{csv_text}}

Process this and create a CSV and output 7 columns.
1. Index
2. text (original text, make sure to put it in quotes to
make it parsable)
3. CHI_response_score
4. SmolLM-1.7B_score
5. SmolLM-350M_score
6. toddler-LM-155M_score
7. SmolLM-135M_score


Look at the text column and determine whether the
responce is a coherent correct responce. Tate the answer
accuracy. Provide me with a value between 1 and 5.

Give it to me in csv format with all 1000 rows.
```

Figure 5: Prompts used to collect score coherency of the Experiment 1 models.

```
Example 1:
  care_giver: what did you have for dessert for lunch?
    RM-4 question clarity score: 0.33
  original (CHILDES): a ice cream sandwich.
    RM-1 Toddler Score: 0.34
    RM-2 Coherence Score: 0.79
  SmolLM-350M-GRPO: want ice cream
    RM-1 Toddler Score:   0.69
    RM-2 Coherence Score:   0.51

Example 2:
  care_giver: what did we do with the toy?
    RM-4 question clarity score: 0.28
  original (CHILDES): we put it in his cage.
    RM-1 Toddler Score: 0.23
    RM-2 Coherence Score: 0.80
  SmolLM-350M-GRPO: put toy in toy box
    RM-1 Toddler Score:   0.84
    RM-2 Coherence Score:   0.76

Example 3:
  care_giver: where is your bag?
    RM-4 question clarity score: 0.39
  original (CHILDES): I didn't bring it. it's in the car.
    RM-1 Toddler Score: 0.53
    RM-2 Coherence Score: 0.79
  SmolLM-350M-GRPO: me carry bag
    RM-1 Toddler Score:   0.80
    RM-2 Coherence Score:   0.64

Example 4:
  care_giver: what did you find?
    RM-4 question clarity score: 0.53
  original (CHILDES): look. Elmo.
    RM-1 Toddler Score: 0.75
    RM-2 Coherence Score: 0.24
  SmolLM-350M-GRPO: me found a toy ball
    RM-1 Toddler Score:   0.77
    RM-2 Coherence Score:   0.78

Example 5:
  care_giver: what did you put on the pizza?
    RM-4 question clarity score: 0.35
  original (CHILDES): cheese and cheese
    RM-1 Toddler Score: 0.46
    RM-2 Coherence Score: 0.62
  SmolLM-350M-GRPO: put pizza on pizza
    RM-1 Toddler Score:   0.80
    RM-2 Coherence Score:   0.67
```

Figure 6: Output examples from RM-1, RM-2, and SmolLM-350M-GRPO