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]
CLCG, University of Groningen | CRC 1646, Bielefeld University f.padovani@rug.nl, bastian.bunzeck@uni-bielefeld.de

General Idea: simulating more fluent and contextually plausible spoken interactions between child and caregivers

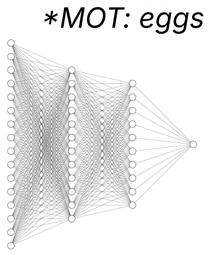
LLAMALOGUE

(model trained on dialogue data, triplets of three consecutive turns)

**MOT: oh yes .*
**CHI: make it flat .*
**MOT: okay .*

**MOT: what're you making ?*
**CHI: a house .*
**MOT: oh very nice .*

**MOT: what do you think that is ?*
**CHI: eggs .*
**MOT: eggs*



AUTOREGRESSIVE MODELING
Llama model trained on 10M of CHILDES tokens for 10 epochs.

DPO (Direct Preference Optimization)

1. Naturalistic data

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 .

2. Synthetic data 🤖

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 .

PPO (Proximal Policy Optimization)

What constitutes a "successful" communicative turn?

Experimenting with four reward functions:

- 1-gram BLEU Reward
- Semantic Similarity Reward
- A quality score derived from an LLM (Llama-3.2-3B)
- An uncertainty-based measuring LLM confidence in processing child responses

🤖 A **teacher LLM** simulates short child-like answer that shows understanding of the caregiver utterance.

🤖 Prompt to Llama-3.2-3B

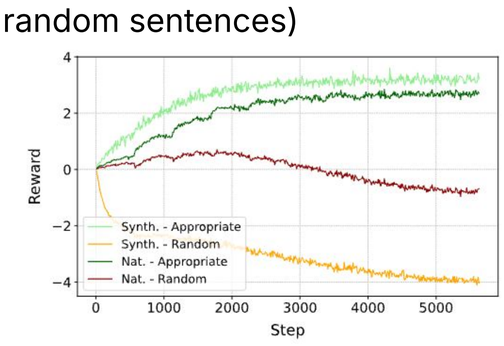
▶ Reward functions are computed by comparing the teacher-generated responses to the output produced by *llamalogue* when prompted with the same utterance.

▶ Also the synthetic sentences used in the DPO dataset are generated with the same prompt.

> 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:

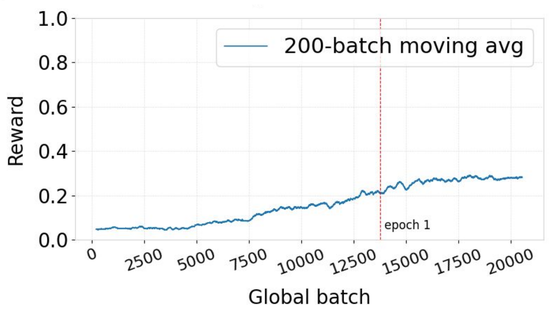
DPO Finetuning

Reward trends (for the appropriate and random sentences)

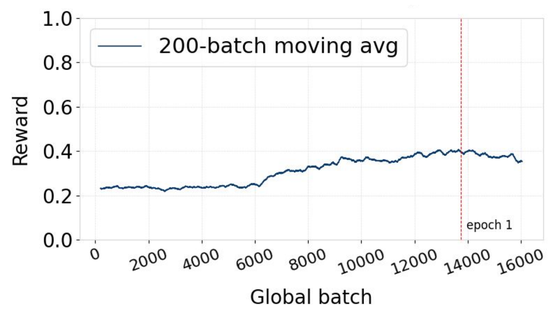


PPO Finetuning

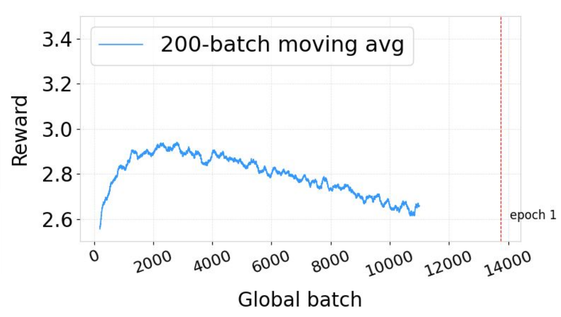
▶ **nltk** to compute a smoothed unigram **BLEU score** (BLEU-1) between llamalogue's response and teacher reference answer



▶ **all-MiniLM-L6-v2** model from SentenceTransformers to compute the cosine similarity



▶ **Olmo LLM** generate numerical quality **score** (from 0 to 5) based on contextual appropriateness of child response

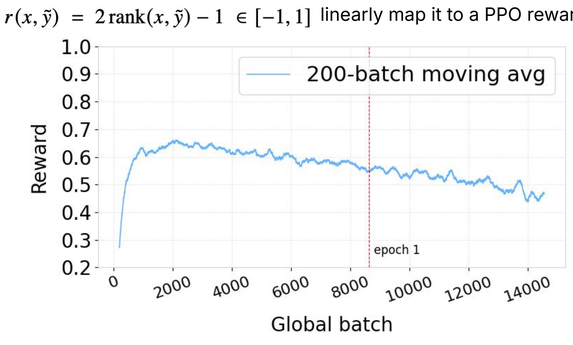


▶ **Confidence based reward**

$\ell_{\text{baby}} = \log P_{\text{teacher}}(\tilde{y} \mid x)$ log-likelihood for 10 LLM-answers

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

$r(x, \tilde{y}) = 2\,\text{rank}(x, \tilde{y}) - 1 \in [-1, 1]$ linearly map it to a PPO reward



Evaluation

BabyLM Challenge evaluation suite: both zero-shot and fine-tuning
Custom Benchmarks: dialogue minimal pairs, Zorro and lexical decision task.

	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

Results

BabyLM Challenge evaluation suite

1. **Llamalogue** underperforms the BabyLM interaction baseline on most formal benchmarks, **except for entity tracking**.
2. It scores lower on **BLiMP** and **AoA prediction**, with modest results on **EWoK, COMPS, (Super)GLUE**, and **wug tests** (compared to other small-track submissions).

Custom Benchmarks

1. On dialogue minimal pairs task - that aligns with the pre-training goal of llamalogue - it exhibits a clear advantage over the baseline.
2. Our model achieves accuracy of 65.5% on Zorro, but is outperformed by the baseline.
3. Llamalogue also falls behind the interactive baseline on the lexical decision task.

"We find that pre-training small LMs on child–caregiver dialogues somehow preserves formal linguistic competence and that Direct Preference alignment (DPO) better captures conversational intent than PPO (Proximal Policy Optimization), but neither improves formal language skills."

*These authors contributed equally