Is Child-Directed Speech Effective Training Data for Language Models?

Steven Y. Feng, Noah D. Goodman, Michael C. Frank

Stanford University

Overview

- Human children learn on magnitudes less language data than LM.
- Is this a function of the child's data or learning algorithm?
- One possibility is that the data is curricularized to support learning.

Research Questions:

- Can child-directed speech serve as effective training data for LM?
- Does global developmental or local discourse ordering affect LM?

Datasets

We train GPT-2 and RoBERTa on five datasets (≈29m words each):

- **CHILDES**: Natural conversations with children.
- **BabyLM**: Mixture of texts from different domains.
- **Wikipedia**: Expository text from Wikipedia.
- **OpenSubtitles**: Movie & TV transcriptions.
- **TinyDialogues (TD)**: To provide a fully grammatical and curricularized conversation dataset with restricted vocab, we used GPT-4 to synthesize ≈130k child-directed conversations that differ by child age, type, participants, length, and content.

Experiments

We also test two ordering effects on CHILDES & TinyDialogues:

- **Global order (child age)**: Conversations ordered by age (ascending), reversed, or randomly shuffled.
- **Local order**: Utterances within conversations ordered or shuffled.

Repeated buckets for global order: divide dataset into b buckets, e.g. A, B, C. Train on each n times before advancing: i.e. $An\ Bn\ Cn$.

Evaluation Metrics

- **Zorro**: Syntactic & grammar knowledge. Distinguish between minimal pairs of child-vocab sentences with grammatical contrasts.
- **Word Similarity (WS)**: Semantic knowledge. Spearman corr. between human and model judgments on word pair similarities.

TinyDialogues Dataset Examples

Age	Example
2	**Babysitter**: Look, the pool is all empty. All gone! | **Toddler**: Water all gone? | **Babysitter**: Yes, all gone. We let the water out. It went down, down, down. | **Toddler**: Why? | **Babysitter**: So the pool can dry. We don't want it to get yucky. (...)
5	**Teacher**: Alright, everyone, it's time to clean up! Child, can you please help me by putting the crayons back in the box? | **Child**: Yes! I can do that. The box is empty so I'll fill it up! | **Teacher**: Thank you, that's very helpful. Make sure the lids are on tight so they don't dry out. | **Child:** I did it! Look, they're all inside now. | **Teacher:** Great job! (...)
10	**Dad**: Once upon a time, in a faraway kingdom, there lived an earless rabbit who loved to make pancakes. | **Child**: An earless rabbit? How could he hear if he wanted to flip the batter? | **Dad**: Well, you see, this rabbit had a special talent. He could feel the vibrations of the batter sizzling on the pan. When it was time to flip, he'd give it a perfect toss. | **Child:** That's so cool! Did the rabbit have any friends? | **Dad:** Yes! His best friend was a turtle (...)
15	**Girlfriend**: Hey, so what's the plan for this history project video? | **Teenager**: We need to make a mini-documentary about the industrial revolution. I was thinking we could start by showing how machines changed production, like how they used to churn butter by hand before. | (...) | **Younger Sibling**: Can I help too? I want to be in the video! | **Teenager**: Sure, you can help us set up the scenes. But no forcible taking over, okay? We need to work together as a team. | **Younger Sibling**: I promise I'll be good! Can I churn the butter for the scene? (...)

Results (GPT-2)

Dataset Results

Dataset	Zorro (%)	WS
CHILDES	78.3 ± 0.5	0.24 ± 0.01
TD	78.5 ± 0.8	0.42 ± 0.01
Wikipedia	78.2 ± 0.6	0.32 ± 0.02
OpenSubtitles	81.0 ± 1.0	0.38 ± 0.00
BabyLM	82.9 ± 1.0	0.42 ± 0.01

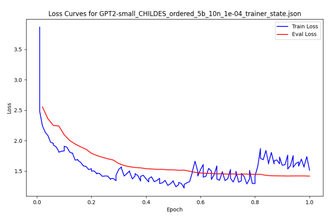
Global Ordering Results

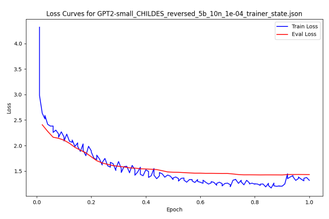
Dataset	Order	Zorro (%)	WS
CHILDES	Age	75.6 ± 1.2	0.20 ± 0.01
CHILDES	Reverse	77.6 ± 1.3	0.20 ± 0.01
CHILDES	Random	76.9 ± 1.1	0.19 ± 0.01
TD	Age	78.2 ± 0.1	0.32 ± 0.01
TD	Reverse	77.7 ± 0.2	0.32 ± 0.01
TD	Random	79.5 ± 2.1	0.34 ± 0.01

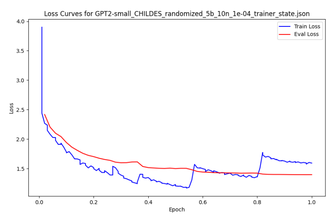
Local Ordering Results

Dataset	Order	Zorro (%)	WS
CHILDES	Normal	78.3 ± 0.5	0.24 ± 0.01
CHILDES	Random	77.3 ± 1.0	0.19 ± 0.01
TD	Normal	78.5 ± 0.8	0.42 ± 0.01
TD	Random	78.4 ± 0.8	0.42 ± 0.00

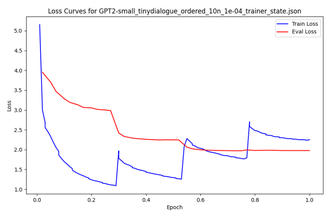
Convergence Behavior (GPT-2)

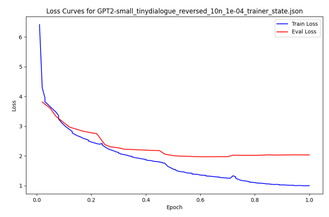


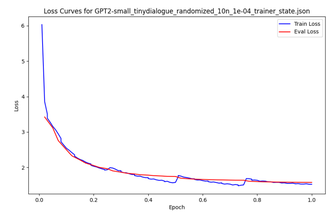




Train (blue) & val (red) loss of CHILDES repeated buckets (b=5,n=10). Order: age, reverse, random.







Train (blue) & val (red) loss of TD repeated buckets (b=4,n=10). Order: age, reverse, random.

Findings & Conclusions

- Diverse data sources (BabyLM) provide better learning for language models than purely child-directed speech data.
- Synthetic child-directed speech data (e.g. TinyDialogues) is more effective than natural (e.g. CHILDES).
- Global developmental ordering has minimal impact, while local discourse coherence affects model learning on natural data.
- Other aspects of children's learning, not simply the data, are responsible for their efficient language learning.
- The child's learning algorithm is substantially more data-efficient than current language modeling techniques.

Email: syfeng@stanford.edu

Code & Data: github.com/styfeng/TinyDialogues