Language Models as Simulations of Early Language Acquisition: analysis of expressive vocabulary

Anonymous EMNLP submission

Abstract

Large language models (LLMs) have been shown to develop linguistic competence from mere exposure to language content, making them a promising avenue for investigating infants' language learning processes (Lavechin et al., 2023; Chang and Bergen, 2022). Nevertheless, LLMs typically require orders of magnitude more data than children, and language outcomes cannot be directly compared. Here, we introduce *machine-CDI*, a metric based on the learner's output to enable a direct comparison of machines and infants on their expressive vocabulary as a function of input quantity. This metric adapts the Communicative Development Inventories (Fenson et al., 2007; Frank et al., 2017), a normalized inventory of words to quantify child language development, to the evaluation set of language models. We illustrate machine-CDI by comparing the expressive vocabulary in infants and character language models (LSTMs and Transformers) trained on English audiobooks. The results show that language models approximately match the children's learning curves, although Transformers are delayed compared to LSTMs. A further analysis show that the models are more impacted by word frequency than children, with a large delay in acquiring low frequency words for models. This delay is found to be linked to the more general phenomenon of long tail truncation observed in language models, which makes them unable to learn words based on few shot observations. These results shed new light on the principles of language acquisition, and highlights important divergences in how humans and modern algorithms learn to process natural language.

1 Introduction

From a cognitive perspective, language models are of theoretical interest to test the distributional hypothesis of language acquisition according to which human children learn from the statistical patterns within language data (Boleda, 2020; Lenci, 2018; Saffran et al., 1996). Previous research has shown that language models can effectively simulate aspects of the language acquisition process, such as phoneme categorization (Lavechin et al., 2023), word acquisition prediction (Chang and Bergen, 2022), and grammatical development (Evanson et al., 2023; Lavechin et al., 2023; Pannitto and Herbelot, 2020). However, these studies have predominantly focused on qualitative analyses, often lacking detailed comparisons with real-world human data.

For a more quantitative approach to the simulation of language acquisition, we propose to match learning environment and outcome measures in the following two aspects. First, despite variations in socio-economic factors and cultural settings (Hart et al., 1997; Cristia et al., 2019), current estimates suggest that American English-learning children receive between 300 and 1,000 hours of speech input annually, totaling at most 30 million of words by the age of three. In contrast with modern language models trained on trillions of words(Hart et al., 1997; Cristia et al., 2019), we train models on developmentally plausible input, matched in quantity to the input that children are exposed to. Second, evaluation methods for language models should be consistent with those available form human dataset. Currently, human behavioral data are derived mainly from children's speech production (e.g., CHILDES) or parental reports (Communicative Development Inventories, hereafter *human CDI*) (MacWhinney and Snow, 1985; Fenson et al., 2007). In contrast, language model evaluations often involve zero-shot probing tasks, such as *spot-the-word* (Le Godais et al., 2017) or *grammatical acceptability judgments* (Warstadt et al., 2019), which, although inspired by psycholinguistic methods, are intrinsically different from production-based human data and typically rely on carefully designed probing sets.

To address these issues, we introduce *Machine-*

CDI, a lexical benchmark designed for direct comparison between infant language acquisition and language modeling. In this study, we (i) train vanilla character LSTMs and Transformers on developmentally plausible data; (ii) introduce a new metric to evaluate how the language models' generations fit into human acquisition patterns from the human CDI (Fenson et al., 2007; Frank et al., 2017), rather than relying on extrinsic evaluations or downstream tasks; and (iii) provide a comprehensive analysis on generations. Our findings reveal generally comparable curves in expressive vocabulary development between children and LSTMs, while Transformers show learning delays. A detailed examination on missing rates and out-of-vocabulary rate show that this is linked to the more general phenomenon of long tail truncation observed in language models, which makes them unable to learn words based on few shot observations. These findings provide new insights into the principles of language acquisition and highlight important differences in how humans and modern algorithms learn to process natural language.

2 Related Work

2.1 Language model as distributional learner

Recently, there has been substantial research applying language models to simulate language acquisition. The earliest study by (Rumelhart et al., 1986) investigated past tense verb form learning in phoneme-level neural networks, which was later replicated in character-level recurrent neural networks Kirov and Cotterell, 2018.

Inspired by "acceptability judgments" in psycholinguistic experiments, these models are often evaluated using zero-shot linguistic probes, i.e. comparing the estimated probability of legitimate sequences with that of matched implausible ones. Previous studies on infants' language acquisition have used this method to probe word-level acquisition ('spot-the-word') and sentence-level acceptability judgments (word: (Chang and Bergen, 2022; Vong et al., 2024); syntactic: (Evanson et al., 2023)). Notably, linguistic abilities in children and models are tested differently: models are explicitly tested on next-word prediction using a two-alternative forced-choice metric, while children are implicitly evaluated based on their spontaneous use of linguistic structures during natural speech. This critical difference addresses the necessity for a more comparable metric.

Additionally, previous studies applying language models to test cognitive hypotheses tend to make an analogy between training dynamics and language learning process. For instance, Chang and Bergen (2022) has shown the similarity of vocabulary learning curves between training steps and age of acquisition(AOA). One follow-up study showed that GPT-2 language models tend to acquire grammatical knowledge in a sequential order, which corresponds to what has observed from transcripts between children-parents (Evanson et al., 2023). However, most of the studies focus on qualitative analysis by making an analogy between children's developmental stages and training steps. This misalignment in time scales makes the trajectories less comparable. Subsequent research addressed this issue by training self-supervised models with varying input sizes to explicitly quantify human's developmental trajectories (Lavechin et al., 2023). Their study demonstrate analogous linear growth patterns in lexical test, initially suggesting the efficacy of language model for vocabulary development. However, there exists a discrepancy on the evaluation task, with the human reference data representing the proportion of children knowing the word, whereas the model is measured in probing task accuracy. What's more, the constructed test words do not directly reflect the distributional patterns of children's exposure to the words.

Therefore, the broad motivation of our work is to assess the distributional mechanisms in infants lexical acquisition (Saffran et al., 1996; Romberg and Saffran, 2010) using neural language models as distributional learners. If analogous distributional learning mechanisms were involved in children, then we would expect similar evaluation outcomes from the proposed lexical metrics.

2.2 Word representation in language models

Language models are typically trained in a way that take as inputs a series of token and output the predicted tokens. In text language models, the predicted tokens can be characters (Xue et al., 2022), entire words (Mikolov et al., 2013; Pennington et al., 2014), or word fragments, for example, byte pair encodings, (Sennrich et al., 2015)). In these two latter cases, the training of the LMs is done in two phases: first a tokenizer is learned, using spaces or punctuation to delimit the word boundaries, and the training corpus is tokenized; second the LM is trained with the token-prediction objec-

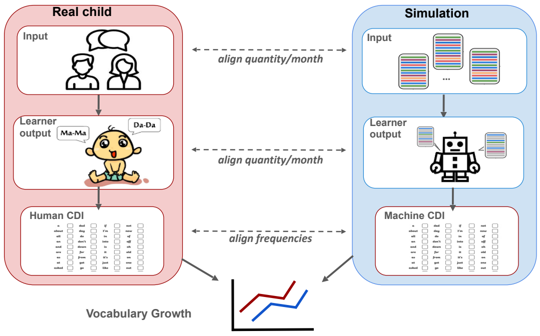

Figure 1: **Overview of Machine-CDI**. Models are fed with linguistic input in matched quantities compared to children and their output is also matched to their outputs. A list of test words is prepared for the machine, matched to the human-CDI in word frequency distibution, yielding comparable vocabulary growth curves.

tive. Tokenization assumes that beginning and ends of words can be identified prior learning, while infants typically acquire language from continuous speech without without knowing the correct linguistic labels like word boundaries a prior (Maye et al., 2002). Prior work (Sutskever et al., 2011; Graves et al., 2014; Hahn and Baroni, 2019; Nguyen et al., 2022; Boldsen et al., 2022; Yu et al., 2024), have found that character LMs can learn lexical, syntactic and semantic representations and do not need a prior segmentation in words. In such models, words are latent representations instead of being explicitly represented, making them a promising approach to simulate the process of lexical learning.

Therefore, we use characters as tokens and leave the language models to learn words in an unsupervised fashion in this work. Instead of applying models that are trained on speech or phonemes, we start with character language models to take the simplified invariant representations of word forms as input regardless of the acoustic variability. This might provide an upper bound of the overall model performance when compared with human data.

3 Method

We follow the reverse engineering approach in (Dupoux, 2018) where a simulation jointly models the input to the learner, the learner and the outcome measure in a quantitative fashion (Figure 1).

3.1 Metrics

3.1.1 Machine-CDI

We detail two key aspects of Machine-CDI that enable direct comparisons on the learning speed of human and machine: data quantity alignment and evaluation metric alignment.

Aligning input data quantity with infants' linguistic exposure Initially, we standardized the amount of training data to reflect the estimated speech exposure of each child, based on prior research of an average annual exposure of 1,000 hours of speech per child (Mendoza and Fausey, 2021). This number corresponds to an upper bound rather than an average when taking into consideration cross-linguistic variation , with the median being 500h/year, and the minimum around 60h/y or less(Cristia, 2023). We converted speech duration into corresponding word counts, assuming an average speech rate of 3 words per second, or 10,000 words per hour as is shown in Figure 2c. We then trained models with varying input sizes to represent infant cohorts at different developmental stages to ensure a realistic approximation of linguistic exposure.

Aligning evaluation metric We base our model evaluation task on human CDI (Fenson et al., 2007; Frank et al., 2017), a checklist of representative word samples used to measure word learning. Our evaluation tasks reported binary scores for a word set to align with parents' binary reports (Frank et al., 2017).

To construct the model's evaluation sets, we selected 520 words from the training set by matching both the number and frequency distribution of content words in the human CDI set. To estimate word frequency of human CDI set, we concatenated and cleaned all the transcripts from the English CHILD Language Data Exchange System (CHILDES) database, resulting in 14.5 million words of adult speech. For the machine CDI word set, frequency was derived from the largest language model training set, consisting of approximately 1 million word types from 30M words. We iteratively minimized the loss function between the frequency distributions of the human and machine CDI sets.

To bridge the gap between parental reports and observable language model performance, we aligned the vocabulary growth curve derived from children's speech in the CHILDES corpus with human CDI scores (see Figure 2b). We constructed the vocabulary growth curve by calculating the cumulative word frequency for each month and applying a constant count threshold to convert these results into binary scores. Due to data sparsity, the word frequency was recalibrated to approximate monthly speech production based on previous research on child vocalization duration. We selected the count threshold to best fit the human CDI growth curve, assuming that an optimal threshold would closely approximate the observed growth speed(see Figure 2c). Figure 2b shows that, on average, a child is expected to produce a correct word form approximately 60 times per month.

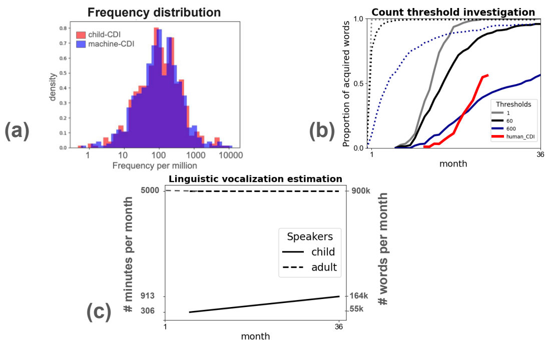

Figure 2: *Calibration method.* a. The distribution of frequency per million of the machine CDI (reference corpus: audiobooks) is matched to that of child-CDI (reference corpus: English-CHILDES). b. To decide whether a word is 'known' based on the speaker's output, we count instances of the word and apply a threshold. In plain blue, resulting vocabulary growth curves for different thresholds. In dotted blue, application of this criterion to the adult's own input. In red, parental reports of children vocabulary. c. Estimates of monthly parental output (input to the child) and child output.

3.1.2 Generation set analysis

To provide a detailed analysis of the generated dataset by language models, we conducted an examination of the *missing rate* and *out-of-vocabulary (OOV) rate*. Specifically, we generated an equal number of words using LSTM and Transformer language models, each trained on a 3.4 million-word corpus. The missing rate quantifies the proportion of words present in the training set but absent in the generated set, indicating the extent to which the generated data replicates the vocabulary of the training set. The OOV rate measures the proportion of words in the generated set that do not appear in the training set, thus assessing the model's ability to generalize beyond the training data. Additionally, we evaluate the true-word rate as the proportion of correct word forms in the generated set, to further assess the model's generalizability. For this evaluation, we use a combination of word lists from CELEX (Van Heuven et al., 2014), the Enchant Library [1], and Wiktionary [2] as a spelling checker.

To contextualize the characteristics of the model-generated text, we constructed two equal-sized reference sets respectively: the in-domain test set, which are selected from audiobook transcripts that are aligned in genres but not included in the training set, and an out-of-domain set consisting of child-directed speech from the English CHILDES corpus.

3.2 Developmentally plausible training set

Following STELA (Lavechin et al., 2022), we built a developmentally plausible training set from the orthographic transcripts of Librovox English audiobooks (Kearns, 2014), consisting subsets of transcripts of 400h, 800h, 1600h and 3200h respectively. Given our calibration of 1000h/year, this translates into 4.8m, 9.6m, 19.2m, and 38.4m, respectively.

3.3 Models

We applied two types of models on developmentally plausible datasets to simulate language acquisition: probabilistic language models (including Long Short-Term Memory models and decoder-only transformers 1), and a non-parametric Bayesian model: Chinese Restaurant Process(Gershman and Blei, 2012).

Chinese Restaurant Process (CRP) The non-parametric Bayesian CRP model clusters data by assigning probabilities. A new word joins an existing cluster with a probability proportional to the cluster size and starts a new cluster with a probability proportional to a parameter α (Gershman and Blei, 2012). We initialized the CRP model using a 3-gram language model derived from the developmentally plausible datasets.

Neural Network Architectures We employed two types of neural network architectures: decoder-only LSTMs (Hochreiter and Schmidhuber, 1997) and Transformers (Vaswani et al., 2017). Similar performance from both models would indicate that the learned patterns are robustly present in the data, not artifacts of a specific model architecture.

[1] https://pypi.org/project/pyenchant/
[2] https://www.wiktionary.org

For LSTMs, we used a three-layer architecture with an embedding layer of size 200, hidden layers of size 1024, and a feed-forward output layer of size 200, based on prior work (Lavechin et al., 2023). For the Transformer model, we experimented on different attention heads and decoding layers, and ended with 8 attention heads and 6 decoding layers that yielded the optimal perplexity.

3.4 Generations

The generation process for the Chinese Restaurant Process (CRP) model uses a concentration parameter, α, derived from the training data to simulate word occurrences. This approach ensures that each token is generated either as a new word or by incrementing the count of an existing word, thereby creating a corpus that reflects the token distribution dynamics as modeled by the CRP.

For LSTMs and Transformers, we employed both unprompted and prompted generation using temperature sampling. Temperature sampling adjusts the output logits by dividing them by a temperature parameter before sampling from the distribution. Higher temperature values make the distribution more uniform, increasing randomness. The number of prompts was matched to the number of sentences generated. Each prompt consisted of 3 words from the audiobook dataset that do not exist in the training set as a whole sequence. We also excluded sequences containing words in the machine CDI test set.

4 Results

4.1 Human-model comparison

Protracted development of language models

The results from our lexical benchmark, illustrated in Figure 3, reveal significant dependencies of vocabulary growth curves on model architectures. The CRP models consistently demonstrate higher vocabulary sizes across all months compared to probabilistic language models. In contrast, probabilistic language models, including LSTMs and Transformers, exhibit slower vocabulary growth, a trend that persists regardless of decoding temperatures and prompt types. This difference highlights a fundamental limitation in their ability to mimic human-like vocabulary expansion.

A closer examination reveals that LSTM models align more closely with human vocabulary growth curves than Transformers, particularly in unprompted generations. Manual analysis of generated utterances indicates that unprompted generations from Transformers frequently suffer from repeated characters. The architectual influence on model fitness is less pronounced but still present in prompted generations, but the effect is reverse in two architectures: with the LSTM fitness with human vocabulary growth interferred with the prompts; the Transformer's fitness largely increased by the prompts.

Variations in temperature during the generation process yield similar trends across different experiment settings, with closest fitness observed for temperature settings around 1. Lower temperatures produce more deterministic outputs, while higher temperatures result in more random and noisy generations. These results highlight the importance of model architecture and generation settings in simulating human language acquisition, suggesting that incorporating mechanisms to handle memory and context appropriately could enhance the vocabulary learning capabilities of probabilistic language models.

Frequency effect So far, our findings indicate that language models acquire lexical knowledge less effectively than humans, regardless of decoding methods and prompts. One potential reason for this discrepancy is the models' difficulty with infrequent words. To investigate this, we decomposed the CDI words into six frequency bands, each containing equal number of words, and fitted sigmoid functions for each frequency band(Chang and Bergen, 2022). We then calculated the estimated month for each frequency band to 80% of known words.

As shown in Figure 4, child speech is less influenced by input word frequency than language models, whereas the expressive vocabulary growth of language models is significantly affected by lower-frequency words across all experimental settings. This frequency effect varies by model architectures, with Transformer models require considerably more training data to reproduce the same proportion of infrequent words compared to LSTM models.

Additionally, lower temperatures exacerbate this effect, likely due to the altered probability distribution generated by the output layers of the language models. These observations suggest that while humans learn words more uniformly across frequencies, probabilistic language models struggle with lower-frequency words, and their performance

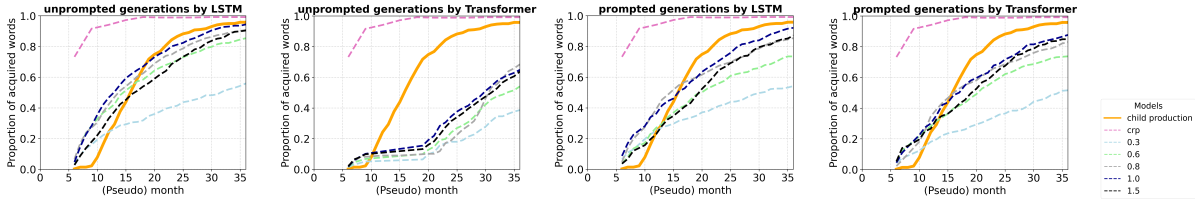

Figure 3: **Vocabulary growth curves for children and models.** Words are considered known if produced more than 60 times. For LSTM and Transformer models, unprompted and prompted generation were sampled at different temperatures. We also plot the curves for an accumulator model and a CRP model.

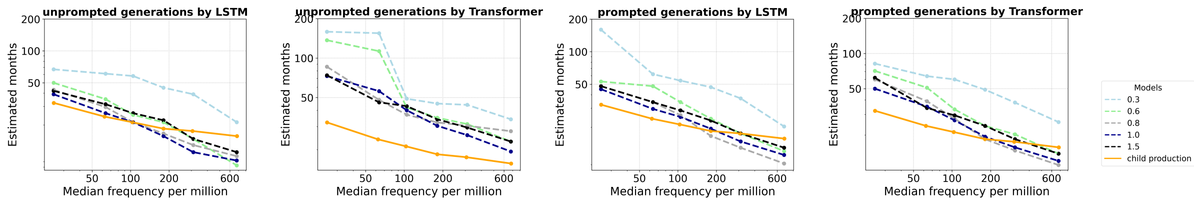

Figure 4: **Estimated age of acquisition as a function of word frequency.** Growth curves for 6 frequency bins of child- and machine-CDI word lists were fitted with a sigmoid and number of months to reach 80% of known words was computed.

is further influenced by architectural choices and decoding methods.

4.2 Generation set analysis

Notably, the frequency effects observed in the Machine CDI are evaluated on a selected subset though matched carefully with human-CDI set. It is unclear whether this effect is influenced by the randomness of sampling. To address this, we expand our analysis on the generations using the same size of training set, focusing on models trained on 3.4M words as a case study. Specifically, we investigated the potential reasons for the human-model discrepency on vocabulary growth curve. Our analysis examined whether the discrepancy arises from the omission of infrequent words in the training data (missing rate) and whether it can be mitigated by the models generating novel sequences (OOV rate).

Figure 5 illustrates that the overall missing proportion of word types generated by language models is higher than that of the reference in-domain test set across various settings. This suggests that language models tend to omit a significant proportion of word types in their generated sets.Additionally, lower temperatures result in higher missing rates across different experimental settings.

The comparison with OOV rates reveals a substantial gap between the proportion of missing

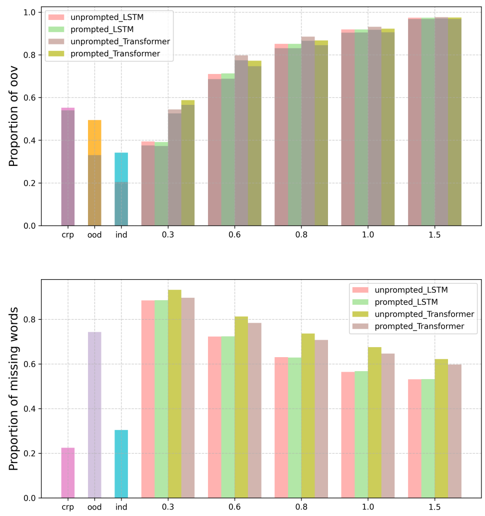

Figure 5: **OOV and word missing rates** Top: the proportion of out-of-vocabulary token types among the generation types; the shaded part shows the proportion of non-words. Below: the proportion of missing token types out of train token types

word types and the amount of novel sequence types generated. Moreover, Figure 5 highlights a very high non-word rate across different generation sets, indicating that current models struggle to generalize through compositional rules.

We further examined whether the missing words are influenced by their frequency in training set. Figure 6 shows that most missing words are in lower-frequency bands, which indicates LMs' de-

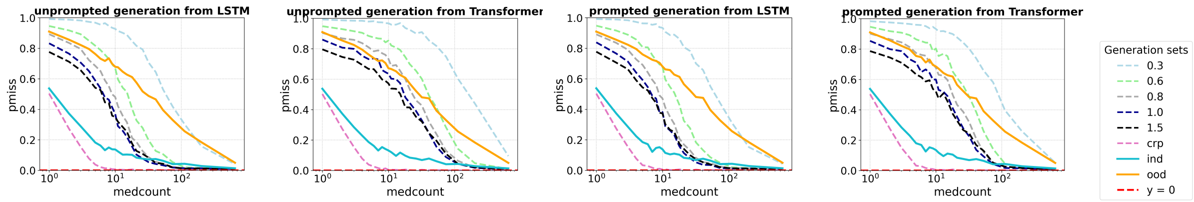

Figure 6: **Missing rates as a function of word count in input corpus.** The subfigures show the missing rates for unprompted and prompted generations from LSTMs and Transformers, respectively. Perfect memorization corresponds to no missing word types from the generation sets, with y=0 as the baseline.

ficiency in reproducing words in tail distributions. Further comparison across different temperatures reveals a negative relationship between temperatures and the proportion of missing words. Similarly, as shown in Figure 7, OOV words predominantly appear in low-frequency bands across different experimental settings. This suggests that while language models exhibit some degree of generalization, this effect is minimal and limited to a few instances.

5 Discussion

In this study, we assess the distributional mechanisms in infants language acquisition (Saffran et al., 1996; Romberg and Saffran, 2010) using neural language models as distributional learners. Our results demonstrate that a purely distributional learner trained on text only approximately reproduce human's expressive vocabulary growth.

We found two main differences. First, the models are much more influenced by word frequency compared to children. This yield a delay in word acquisition for low frequency words. Further analyses show that very low frequency items (seen less than 10 times in the training corpus) tend to be overwhelmingly missed by the language models. Our findings suggest that while current language models approximate the statistical properties of their training data, this does not necessarily imply generating the desired underlying data distribution across various model architectures and decoding methods. This echoes prior research on language model's memorization, in which a log-linear trend between the number of duplicates in the training data and the extent of verbatim memorization (Carlini et al., 2022; Razeghi et al., 2022; Kandpal et al., 2022). In contrast, evidence show that children can learn new words in a few shot fashion, suggesting that they may use different learning mechanisms (e.g., episodic memory), not available in LMs. Prior study using non-parametric knowledge to capture long-tail information has shown a promising avenue to simulate the episodic memory mechanism(Kandpal et al., 2023). Further investigation needs to be done on cognitive plausibility. Second, the models tend to produce a large quantity of novel word forms (more than 80% of the word forms) , the vast majority of which are nonwords. This corresponds to the well known tendency of LLMs to 'hallucinate' (Ji et al., 2023). In only a small fraction of the cases, these hallucinations are actual words, obtained through the combinatory recomposition of known words or morphemes. In contrast, infants do not produce many nonwords, and these nonwords tend to be due to be mispronunciations of real words.

These differences could be interpreted in terms both of learning mechanisms and input. Children's linguistic experience is grounded in multi-modal experience. Research shows that children as young as ten months old learn word-object pairings, mapping novel words onto perceptually salient objects (Pruden et al., 2006). By the age of two, they integrate social cues such as eye gaze, pointing, and joint attention (Çetinçelik et al., 2021). Our findings suggest that these grounded and interactive experiences could impact child word acquisition in ways that cannot be fully explained by linguistic signals alone. Additionally, the communicative nature of the language environment provides a more dynamic context where infants receive feedback from caregivers. Studies on reinforcement learning in multi-agent communication tasks highlight the importance of these non-distributional properties for achieving more human-like natural language understanding. For example, research by Chevalier-Boisvert et al. (2018), Lazaridou et al. (2016), and Zhu et al. (2020) emphasizes the role of interaction and feedback in language learning.

In this paper, we have described how lexical eval-

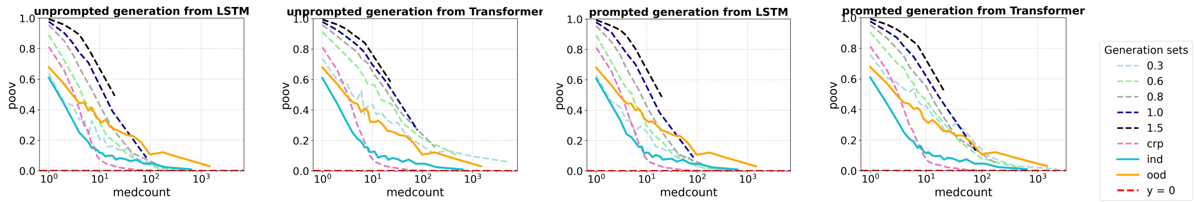

Figure 7: **OOV rates as a function of word count in generated corpus.** The subfigures show the OOV rates from unprompted and prompted generations from LSTMs and Transformers, respectively. This score reflects the novelty of the generation sets.

uation metrics have been carefully designed to evaluate language models trained on developmentally plausible text corpus. Notably, we only focused on the word form inspection, which might inflate the model performance. Even with the upper bound of the model performance, the models are delayed for expressive vocabulary. And we found a stable frequency effect across different language model architectures and decoding settings. We show that this is linked to the more general phenomenon of long tail truncation observed in language models, which makes them unable to learn words based on few shot observations. These results shed new light on the principles of language acquisition, and highlights important divergences in how humans and modern algorithms learn to process natural language.

6 Limitations

One limitation lies in discrepancies between infant behavioral measures derived from parental reports and those evaluated on our model. This difference potentially accounts for differences observed between estimates from CHILDES transcripts and parental reports. On the one hand, the model's evaluation test, focusing on word form segmentation from input data. In contrast, parental criteria may involve the proper usage of the test words, which typically involve a broader scope of linguistic knowledge on semantic and syntactic levels. On the other one hand, the CHILDES transcripts, though pre-processed carefully to remove all the human annotations, it is possible that incomplete word forms are completed and normalized by human annotators, which may cause the inflated language performance in the transcript.Also, some subsets interview procedure's richness may boost children's expressivity beyond everyday speech, potentially leading parents to underestimate vocabulary in CDI inventories. What's more, CHILDES transcripts aggregate data from multiple children, while parental reports are averaged on single child. Also, we calibrated the word counts based on the estimated vocalization length. This might result in duplicated counts on children's production. Nevertheless, all these differences might inflate the lexical scores obtained from transcripts. Notably, we apply the exactly the same post-process on the model's generation and also compare model's generations with CHILDES references. Therefore, this might exert a trivial influence on CHILDES-model difference.

Another limitation lies in the usage of character-level input rather than speech input. Characters preserve the invariant form of words, and space or punctuation indicate word boundaries. Hence the models we tested correspond to an upper bound of what could be found with some realistic models based on speech inputs (Lavechin et al., 2023, 2024), where word forms are variable and not delimited with clear boundaries. Further studies are needed to evaluate speech-LMs (Lakhotia et al., 2021; Nguyen et al., 2024) and address the technical difficulty of transcribing the speech output of such models in a format that can be applied to our machine-CDI benchmark.

Ethics Statement

Use of human data: While we did not collect any new human data ourselves, many of our analyses involved the use of prior datasets within the CHILDES database. All of these datasets were collected in accordance with IRB policies at the institutions of the data collectors, and all followed standard practices in obtaining informed consent and deidentifying data.

References

Maria Boldsen, Per Anker Jensen, Lars Kai Hansen, and Ole Winther Andersen. 2022. Perceptual representations vs character embeddings in cross-lingual analysis. *arXiv preprint arXiv:2203.12345*.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Melis Çetinçelik, Caroline F Rowland, and Tineke M Snijders. 2021. Do the eyes have it? a systematic review on the role of eye gaze in infant language development. *Frontiers in psychology*, 11:589096.

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

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2018. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*.

Alejandrina Cristia. 2023. A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, 26(1):e13265.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? *arXiv preprint arXiv:2306.03586*.

Larry Fenson et al. 2007. Macarthur-bates communicative development inventories.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2017. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694.

Samuel J Gershman and David M Blei. 2012. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.12345*.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.

Michael Hahn and Marco Baroni. 2019. Neural models learn morphological, syntactic and semantic aspects from unsegmented text. *arXiv preprint arXiv:1902.12345*.

Betty Hart, Todd R Risley, and John R Kirby. 1997. Meaningful differences in the everyday experience of young american children. *Canadian Journal of Education*, 22(3):323.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR.

Jodi Kearns. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

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

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Marvin Lavechin, Maureen de Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2024. Modeling early phonetic acquisition from child-centered audio data. *Cognition*, 245:105734.

Marvin Lavechin, Maureen De Seyssel, Hadrien Titeux, Hervé Bredin, Guillaume Wisniewski, Alejandrina

Cristia, and Emmanuel Dupoux. 2022. Can statistical learning bootstrap early language acquisition? a modeling investigation.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. Babyslm: language-acquisition-friendly benchmark of self-supervised spoken language models. *arXiv preprint arXiv:2306.01506*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Gaël Le Godais, Tal Linzen, and Emmanuel Dupoux. 2017. Comparing character-level neural language models using a lexical decision task. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 125–130.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Brian MacWhinney and Catherine Snow. 1985. The child language data exchange system. *Journal of child language*, 12(2):271–295.

Jessica Maye, Janet F Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111.

Jennifer K Mendoza and Caitlin M Fausey. 2021. Quantifying everyday ecologies: Principles for manual annotation of many hours of infants' lives. *Frontiers in psychology*, 12:710636.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tu Anh Nguyen, Maureen de Seyssel, Robin Algayres, Patricia Roze, Ewan Dunbar, and Emmanuel Dupoux. 2022. Are word boundaries useful for unsupervised language learning? *arXiv preprint arXiv:2210.02956*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: evaluating the acquisition of grammar from limited input data. *arXiv preprint arXiv:2010.04637*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shannon M Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Elizabeth A Hennon. 2006. The birth of words: Ten-month-olds learn words through perceptual salience. *Child development*, 77(2):266–280.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Alexa R Romberg and Jenny R Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914.

David E Rumelhart, Geoffrey E Hinton, James L McClelland, et al. 1986. A general framework for parallel distributed processing. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(45-76):26.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. 2024. Megabyte: Predicting million-byte sequences with multiscale transformers. *Advances in Neural Information Processing Systems*, 36.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020. Babywalk: Going farther in vision-and-language navigation by taking baby steps. *arXiv preprint arXiv:2005.04625*.

months	# words	# char	# utt
4-8	3.4M	17.7M	0.3M
9-18	7.0M	36.6M	0.7M
19-28	14.7M	76.7M	1.4M
29-36	27.7M	144.9M	2.6M

Table 1: Statistics of train data

Hyperparameter	**Value**
Max sequence length	40960
Batch size	40960
Learning rate	0.0001
Learning rate scheduler	inverse sqrt
Warmup steps	10000
Optimizer	Adam
Adam-beta1	0.9
Adam-beta2	0.98
Dropout	0.1
LSTM hyperparameter	**Value**
decoder layers	3
hidden size	1024
embedding dimension	200
Transformer hyperparameter	**Value**
Transformer layers	3
Intermediate hidden size	2048
Attention heads	8
Attention dropout	0.1

Table 2: Language model training hyperparameters.

A Appendix

A.1 Training data Details

Table 1 shows details of training dataset. All the digits and punctuation are removed and all the characters are lower-cased, with special tokens inserted as word boundaries. Language model training hyperparameters are listed in Table 2. Each model was trained on four A40 GPUs.

A.2 Lexical diversity across different sets

We investigated linguistic diversity of different test sets. Figure 8 shows the type-token ratios of different sets. The CHILD-directed speech is less lexically diverse compared with other sets, which corresponds to previous language acquisition study that caregivers tend to repeat same words to scaffold lexical learning.

The overall inspection of the generated data patterns correspond to prior observation of declining lexical diversity of generated data (Guo et al., 2023). And the decreased lexical diversity might

train data	3.4M	7.0M	14.7M	27.7M
child production	aha	bye bye mommy	he eaten toes	it is there mom
LSTM(un	wounded	for woman says	smile some that	professor at once busy
Trans(un)	to	o he had been	him to bring	aaia with all his
Prompt	it tries to	one side only	smile when you	but while giving
LSTM(prompted)	make	of his wife	king of him	thing they must never
Trans(prompted)	me	horse the wilderness	will have me	the greater i

Table 3: Examples of generated sequences. The boundary marker is replaced with blank space for the ease of reading. We show the generations with the temperature = 1.0 as examples

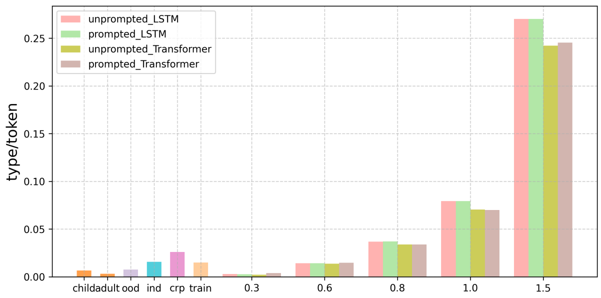

Figure 8: *Type/token ratios in different datasets*

stem from the large proportion of missing words, which might not necessarily be compensated by the amount of OOV words.

A.3 Fitted vocabulary growth curves

The figures below show the fitted sigmoid curves across different consitions

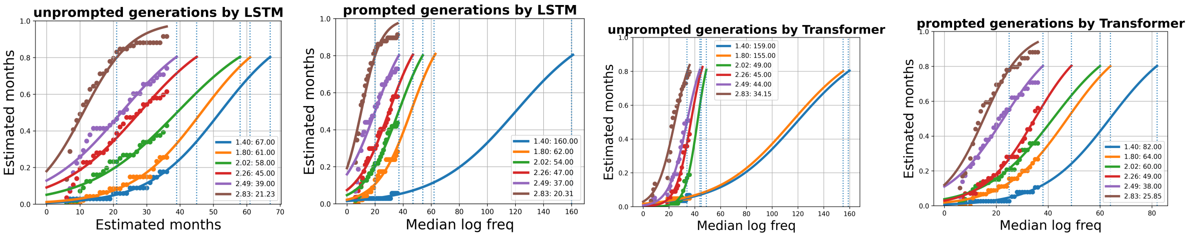

Figure 9: **Fitted sigmoid curves of models in temperature of 0.3.**

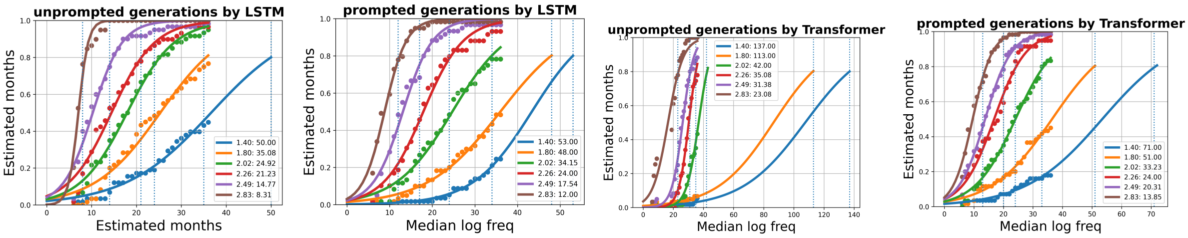

Figure 10: **Fitted sigmoid curves of models in temperature of 0.6.**

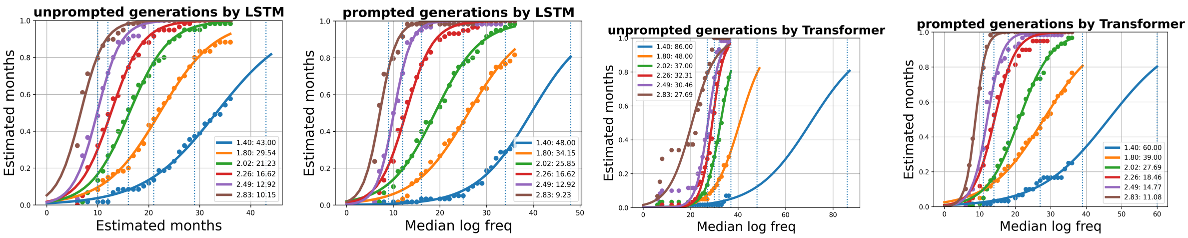

Figure 11: **Fitted sigmoid curves of models in temperature of 0.8.**

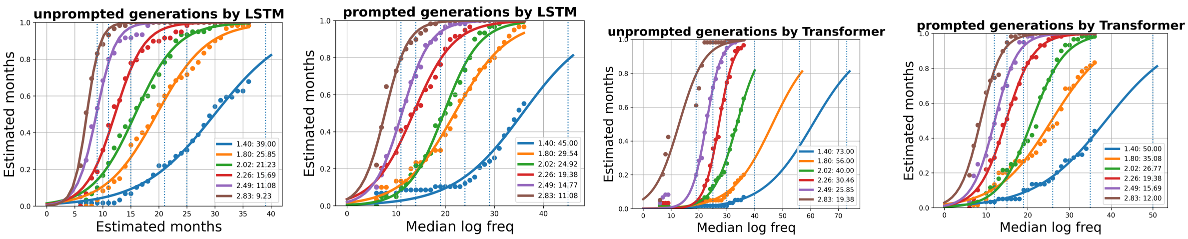

Figure 12: **Fitted sigmoid curves of models in temperature of 1.0.**

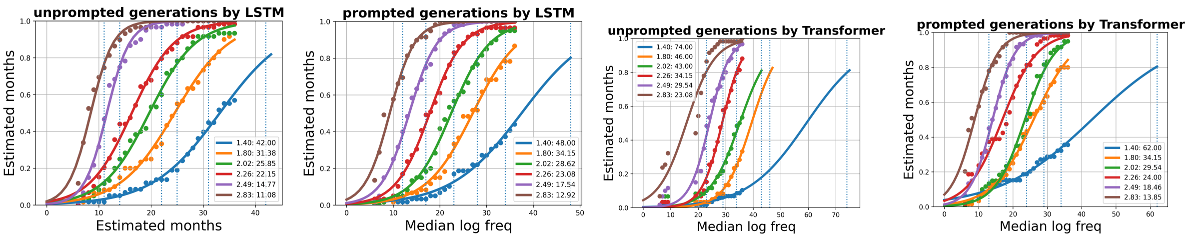

Figure 13: **Fitted sigmoid curves of models in temperature of 1.5.**