CORPUS-CDI: comparing children and AI models learning rate of expressive vocabulary

Anonymous ACL submission

Abstract

Evaluating language models through the lens of human language acquisition requires developmentally grounded benchmarks that support direct model-child comparisons. We introduce CORPUS-CDI, a count-based lexical benchmark that compares language model generation with child word production across age groups. CORPUS-CDI adapts the MacArthur-Bates Communicative Development Inventories (CDI) by matching the size of training and generation sets and the word frequency distribution in the evaluation set. Using this benchmark, we evaluate character-level LSTMs and Transformers trained on child-equivalent amounts of data. While models acquire vocabulary with more data, they diverge from children: excessive lexical diversity, minimal improvement in word form accuracy, and slower vocabulary growth that lacks the characteristic acceleration of human development. Moreover, model generations overrepresent high-frequency function words and underrepresent low-frequency content words. These discrepancies cannot be addressed by temperature trade-offs during inference. Further frequency analysis of the evaluation set suggests that the mismatch might 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.

1 Introduction

Large language models (LLMs) have shown that a general-purpose architecture trained on large text corpora can acquire rich syntactic and semantic knowledge (Brown et al., 2020; Devlin et al., 2019; Touvron et al., 2023). Beyond transforming NLP applications (Bommasani et al., 2021), they also provide scalable tools to implement and test theories of early language acquisition (Dupoux, 2018; Evanson et al., 2023; Lavechin et al., 2022; Apidianaki et al., 2024; Hu et al., 2025), as they support the theory according to which children bootstrap into language learning by accumulating statistics from their language input (Saffran et al., 1996).

A central challenge when applying LLMs as models of the early language learner is establishing valid *quantitative* comparisons between the learning curves of real children and those of LLMs fed with comparable data. For instance, it has been claimed that LLMs require significantly more input data than children to reach the same level of performance (Frank, 2023). But measuring this 'data gap' depends critically on the development of sound metrics that can be applied to both humans and machines.

We focus on measures of vocabulary, as there is a wealth of developmental data which provides the human side of the comparison. We distinguish three methods to evaluate vocabulary applicable to both children and models: (i) **familiarity metrics**, which exploit LLMs' probabilistic outputs (e.g., surprisal or word–pseudoword preference) to probe lexical knowledge (Chang and Bergen, 2022; Ma et al., 2024; Lavechin et al., 2025); (ii) **cross-modal matching**, which assesses word–referent grounding using multimodal training and behavioral benchmarks (Vong et al., 2024; Tan et al., 2024; Nikolaus and Fourtassi, 2021b), the equivalent of looking while listening in children (Swingley, 2011); and (iii) **production metrics**, which analyze model-generated text with the same tools applied to child production corpora. Familiarity and cross-modal methods are informative but either data-sparse (infant behavioral paradigms (Hallé and de Boysson-Bardies, 1994; Ngon et al., 2013) or inapplicable to unimodal LMs; whereas production metrics allow broad, quantitative comparisons provided model input exposure is appropriately matched to child input.

To support such comparisons, we introduce CORPUS-CDI, a benchmark aligning child expressive vocabulary with model generation. CORPUS-

CDI calibrates three key dimensions: (1) training exposure, scaled from 6 to 36 months based on daylong recordings (Cristia et al., 2019); (2) cumulated generated word count, calibrated to children's monthly word production rates; and (3) calibrated test vocabulary, matched in part-of-speech and frequency distribution to the CDI word list (Frank et al., 2017).

Using CORPUS-CDI, we evaluate character-level LSTMs and Transformers trained on child-directed speech and audiobook transcripts. We assess vocabulary acquisition across three complementary dimensions: lexical diversity (type-token ratio), word form accuracy (non-word rate), and vocabulary size (CORPUS-CDI score). Our analyses reveal divergences from children: models show excessive diversity, minimal improvement in word form accuracy, and slower vocabulary growth. Frequency analyses suggest these gaps stem from long-tail truncation, with low-frequency words underrepresented and high-frequency function words overrepresented compared to children.

2 Related Work

2.1 Developmental Trajectories Across Linguistic Levels

Prior work has explored the alignment between language model learning and child language development across multiple linguistic levels. A central challenge is mapping model training to developmental time. Early studies used training checkpoints as proxies for age, tracking when models acquire words (Chang and Bergen, 2022) or syntactic structures (Evanson et al., 2023). These approaches, however, repeatedly expose models to identical input batches, unlike children who experience increasingly varied linguistic input over time. To address this, Lavechin et al. (2023) varied input size to model developmental trajectories explicitly, showing correlations between zero-shot lexical accuracy and children's vocabulary growth. Beyond vocabulary, research has examined phonological development via word segmentation (Goriely et al., 2024; Goriely and Buttery, 2025), morphological acquisition through overregularization patterns (Haga et al., 2024), and syntactic learning trajectories (Evanson et al., 2023). Yet, most prior studies calibrate only input quantity, leaving output production and evaluation frequency uncontrolled. Our work aligns training exposure, production quantity, and evaluation frequency to enable direct comparison between model and child productive vocabulary.

2.2 Vocabulary Evaluation

Evaluating vocabulary acquisition in models has followed two main strategies: developing developmentally inspired metrics or applying existing metrics to child-relevant tasks (Kosoy et al., 2023; Nikolaus and Fourtassi, 2021a; Jiang et al., 2023). Notably, Chang and Bergen (2022) first introduced the MacArthur-Bates Communicative Development Inventory (CDI) to evaluate when models acquire individual words during training. Their findings highlight that models rely disproportionately on input frequency, with more variance in acquisition timing than child age-of-acquisition norms. This frequency dependence aligns with extensive developmental evidence showing frequency effects across all linguistic levels (Ambridge et al., 2015), though children also leverage cross-situational learning and social-pragmatic cues unavailable to text-only models.

The learnability gap between models and children may partly stem from the Zipfian distribution of child-directed speech, which confers advantages for human learners: children acquire novel words more effectively under Zipfian than under uniform input distributions (Wolters et al., 2024; Lavi-Rotbain and Arnon, 2023, 2022). Cross-linguistic studies further show that contextual surprisal predicts age of acquisition for function words and predicates but not nouns (Portelance et al., 2023), suggesting that distributional learning mechanisms affect lexical categories differentially. Building on this work, we construct frequency-matched evaluation sets to control for the potential influential factors from frequency effects.

2.3 Calibrated Evaluation Frameworks

Direct comparisons between models and children require alignment across training data, learning objectives, and evaluation tasks(Warstadt and Bowman, 2022). The BabyLM Challenge (Warstadt et al., 2023; Hu et al., 2024) established shared tasks by restricting model training to 10M or 100M words to approximate child exposure at ages 2 and 13. Despite these constraints, results remain inconsistent: some studies find child-directed speech helps (Feng et al., 2024), while others observe Wikipedia outperforming CHILDES (Bunzeck et al., 2025).

Input representation is another consideration:

most models use BPE tokens (Huebner et al., 2021; Evanson et al., 2023), whereas infants perceive continuous acoustic speech (Dupoux, 2018). We use character-level inputs, forcing models to learn word boundaries unsupervised from orthographic sequences (Bunzeck et al., 2024; Goriely et al., 2024). This provides an upper bound on performance compared to phoneme-based approaches.

Regarding production metrics, lexical diversity is typically measured with type-token ratio (TTR), which is highly sample-size dependent (Richards, 1987) and requires fixed-size windows for comparability (McKee et al., 2022). Temperature sampling further introduces tradeoffs between diversity and quality, with no single setting optimizing both (Verine et al., 2025; Sclar et al., 2024).

Our framework addresses these challenges via three calibrations: (1) age-appropriate training exposure from daylong recordings (Cristia et al., 2019), (2) production quantity matched to children's monthly output, and (3) frequency-matched evaluation sets reflecting child-directed speech patterns. Together, these calibrations enable a developmentally grounded evaluation of expressive vocabulary.

3 Methods

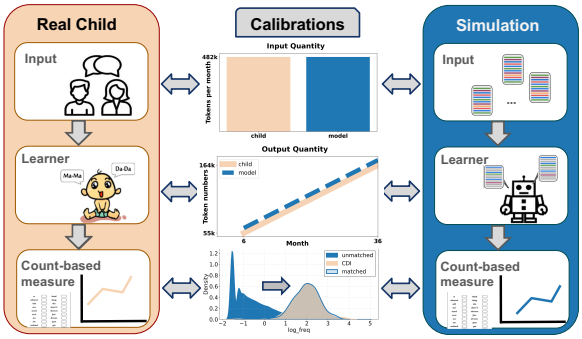

Figure 1: **Overview of the CORPUS-CDI benchmark framework.** (a) Training and generation sizes are aligned with age-appropriate estimates of linguistic exposure and production from 6–36 months (Cristia et al., 2019). (b) Evaluation sets match the frequency distribution of MacArthur–Bates CDI words in child-directed speech (English-CHILDES).

3.1 CORPUS-CDI Benchmark Overview

The CORPUS-CDI benchmark enables direct comparison between language model generation and child word production through three methodological calibrations to address common mismatches in prior work (Figure 1):

(1) Age-appropriate training exposure: Training sets are scaled to match cumulative linguistic input from 6–36 months of age estimated from daylong recordings (Cristia et al., 2019).

(2) Production quantity matching: Generation sizes are calibrated to children's monthly word production rates estimated from longitudinal corpora (Cristia et al., 2019) to ensure comparable sample sizes for vocabulary assessment.

(3) Frequency-matched evaluation: Evaluation vocabularies are drawn from the MacArthur–Bates Communicative Development Inventory (CDI; Fenson et al. 2007) and frequency-matched to the same words in English CHILDES corpora.

We evaluate models along three dimensions: *lexical diversity* (type–token ratio), *word form accuracy* (non-word rate), and *productive vocabulary size* (CORPUS-CDI score).

3.2 Training Data

We construct two training corpora representing different linguistic environments:

Audiobook corpus: English audiobook transcripts from LibriVox (Kearns, 2014), representing narrative-rich, syntactically complex adult-directed language. **Child-directed speech (CDS) corpus:** English CHILDES transcripts (MacWhinney, 2000), representing the naturalistic register of early child input.

Age-appropriate scaling Following Cristia et al. (2019), we estimate exposure from 250 hours at 6 months to 1,500 hours at 36 months of age. This corresponds to approximately 500 hours of speech input per year, representing a mid-range estimate of children's linguistic environments. Converting speech hours to words using a standard rate of 3 words per second yields training sets ranging from 2.89M words (6 months) to 17.35M words (36 months). Developmental datasets are constructed every 6 months by randomly sampling the target quantity from each corpus (see Table 1). Given English CHILDES contains a total of 15.6M words of caregiver inputs, the available data is equivalent to those at the 30-month level, which thus serves as the maximum training input.

Preprocessing All text is lowercased, with spaces and basic punctuation (periods, commas,

apostrophes) preserved as word boundary markers. No tokenization or subword segmentation is applied, forcing models to infer word boundaries directly from raw character sequences.

3.3 Model Architectures and Training

We train character-level autoregressive models under two architectures:

LSTM Three layers with 200-dimensional character embeddings, 1024-dimensional hidden states, and 200-dimensional output projections, following Lavechin et al. (2023).

Transformer GPT-2-style decoder-only architecture with 12 transformer layers, 512-dimensional hidden states, 12 attention heads per layer, and 0.05 dropout rate, following Goriely et al. (2024).

Training procedure Models are trained to minimize cross-entropy loss using the Adam optimizer (learning rate 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.98$) with inverse square root learning rate decay and 10,000 warmup steps. We use a batch size of 40,960 characters and train until validation perplexity converges. For each age-corpus pair, two models are trained with different random seeds and non-overlapping data subsets. Training details are provided in Table 2.

Generation procedure At inference time, we generate text using temperature sampling with $T \in \{0.3, 0.6, 1.0, 1.5\}$. Generation begins with a beginning-of-sequence (BOS) token and continues until reaching the target word count corresponding to children's estimated monthly production at that developmental stage (see Table 3 in Appendix). Character sequences are segmented into words using space and punctuation boundaries. For months 32–36, outputs are generated from the same model checkpoints.

3.4 Evaluation Set Construction

Base vocabulary and POS filtering We start with the MacArthur-Bates Communicative Development Inventory (CDI) (Fenson et al., 2007), a standardized checklist of 680 English words with normative age-of-production data. To focus on content words, we apply POS tagging using a RoBERTa-based SpaCy toolkit. For each word, POS labels are computed across multiple contexts, and majority voting is applied separately in each corpus (CDI and audiobook) to determine the dominant POS. Function words (articles, pronouns, prepositions, auxiliaries) are excluded, retaining nouns, verbs, adjectives, and adverbs.

Frequency stratification Filtered CDI words are divided into six frequency bands based on their occurrence in child-directed speech (adult input in English CHILDES), with an equal number of words per band.

Audiobook subset selection For a comparable evaluation, we construct a corresponding audiobook subset using the same POS filtering and frequency banding. Words are selected such that their frequency distribution mirrors the filtered CDI evaluation set.

This process yields two aligned datasets: i.) Filtered CDI evaluation set with 420 words in six frequency bands; and ii.) frequency-matched audiobook subset with the same size per frequency band. This enables a developmentally grounded comparison between model generations and children's expressive vocabulary.

3.5 Evaluation Metrics

We assess model performance using three complementary metrics capturing lexical diversity, form accuracy, and expressive vocabulary.

Type-Token Ratio (TTR) Measures lexical diversity as the ratio of unique word types to total word tokens. Following Räsänen and Kocharov (2024), we compute TTR over fixed-size chunks of 3,500 words to control for sample size effects, then average across all chunks.

Non-word Rate Quantifies word form accuracy as the proportion of generated word types not found in comprehensive English dictionaries: CELEX (Van Heuven et al., 2014), the Enchant library,[1] and Wiktionary.[2] A word is classified as valid if it appears in any of these three resources. This metric captures models' ability to produce orthographically valid English words.

CORPUS-CDI Score Adapts the human CDI vocabulary assessment to model generations through a count-based criterion. We mark a word in the evaluation set as "known" if it appears at least τ times in the generated sample. To determine τ, we analyze CHILDES child productions and fit the resulting vocabulary growth curve to the empirical CDI trajectory. We select $\tau = 60$, which yields

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

slopes most consistent with observed developmental patterns. The CORPUS-CDI score is computed as:

$$\text{CDI} = \frac{|\{w \in \text{eval set} : \text{count}(w) \geq \tau\}|}{|\text{eval set}|}$$

where the evaluation set includes 420 words. This value represents the proportion of age-appropriate vocabulary that the model generates with sufficient frequency to be considered "known." Sensitivity analyses for alternative thresholds ($\tau \in \{1, 500\}$) are shown in Figure 6.

4 Results

4.1 Developmental Trajectories

Figure 2 presents model performance across three complementary metrics as a function of cumulative training exposure (6-36 months), compared against children's developmental trajectories from CHILDES production data. All results use temperature $T = 1.0$ as the default setting.

Lexical diversity Both LSTM and Transformer models exhibit higher type-token ratios (TTR) than children across developmental stages. Models trained on audiobooks show slightly higher TTR than those trained on child-directed speech, reflecting broader lexical variety in narrative text. Unlike children, whose TTR gradually increases with age, model TTR remains relatively stable. This suggests models generate more lexically diverse outputs from early on, but does not increase substantially with additional training data.

Word form accuracy Models and children show different developmental patterns in non-word production rates. Children begin with high non-word rates that decrease sharply by 24 months, reflecting the gradual development of articulatory control and phonological mastery. In contrast, both model architectures maintain relatively stable non-word rates across all training quantities, showing minimal improvement with increased data exposure. Transformers produce slightly fewer non-words than LSTMs, but neither captures the steep improvement observed in children. This divergence likely reflects a critical limitation of text-based models: they cannot capture the development of articulatory-motor control that drives children's decreasing error rates.

Vocabulary growth CORPUS-CDI scores show that children's vocabulary grows sigmoidally, with rapid acceleration during the vocabulary spurt. Models, by contrast, display roughly linear growth, with Transformers slightly outperforming LSTMs and audiobook-trained models outperforming CDS-trained ones. Critically, neither architecture reproduces the accelerating trajectory characteristic of human vocabulary development even as training data accumulates as the estimated quantity.

Across all three metrics, models exhibit qualitatively different developmental patterns from children. They show excessive lexical diversity, minimal improvement in word form accuracy, and slower vocabulary growth that lacks the characteristic acceleration of human development. These differences persist across both architectures and training corpora, suggesting they reflect fundamental limitations in how current language models learn from sequential exposure rather than artifacts of specific model choices. In the following sections, we investigate two potential explanations for these gaps: whether alternative temperature settings during decoding can better align model behavior with children (Section 4.2), and whether the observed differences relate to models' frequency-dependent learning patterns (Section 4.3).

4.2 Temperature Tradeoff

The developmental gaps observed in Section 4.1 raise an important question: could these differences arise from suboptimal generation parameters rather than inherent learning limitations? Temperature sampling controls generation stochasticity, with lower temperatures favoring high-probability outputs and higher temperatures increasing diversity. We evaluate whether any temperature setting can align models with children across all three metrics.

Figure 3 shows the relationships between metrics as temperature varies from 0.3 to 1.5. Each point represents the mean performance across all developmental stages for a given architecture-corpus-temperature combination.

TTR-CDI trade-off Models exhibit a clear trade-off: increasing temperature simultaneously raises type-token ratio (TTR) and CORPUS-CDI scores. Low temperatures produce TTR values closer to children's range but severely underrepresent vocabulary size, whereas high temperatures improve vocabulary coverage at the cost of excessive lexical diversity. No temperature setting positions

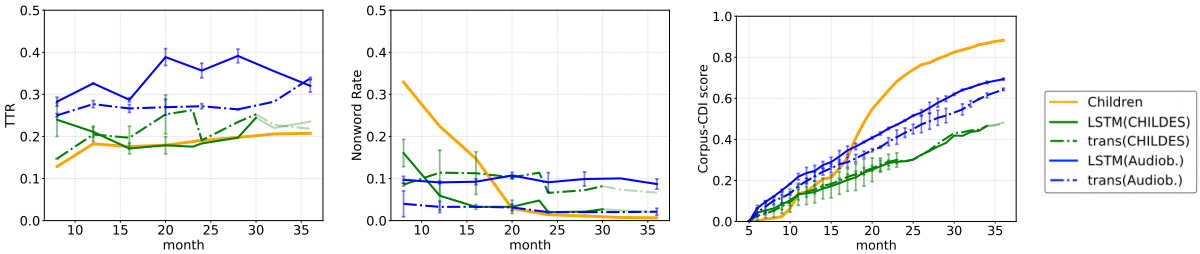

Figure 2: **Core metrics comparing child and model vocabulary development across age groups.** Left: Type–token ratio (TTR) measuring lexical diversity. Middle: Non-word type rate indicating the proportion of invalid word forms. Right: CORPUS-CDI scores tracking vocabulary acquisition (words counted as known if produced >60 times). Results are shown for LSTM and Transformer models trained on audiobooks and child-directed speech (CDS), compared with CHILDES child production data. Error bars denote 95% confidence intervals across disjoint training subsets. For the final months (32–36), outputs are generated from the same model checkpoint and shown with lighter transparency.

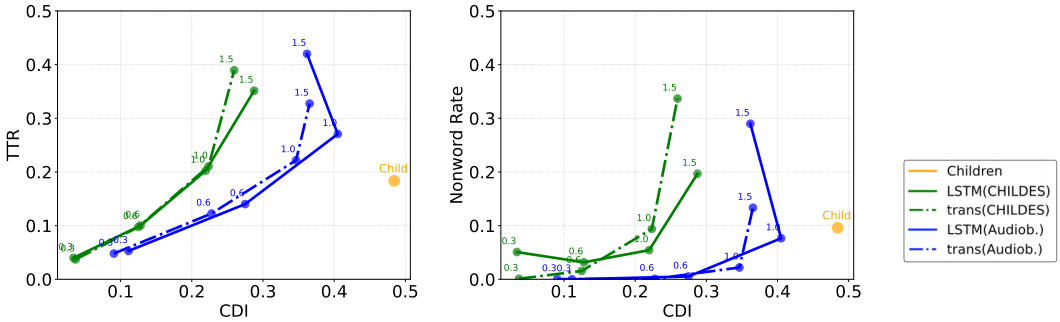

Figure 3: **Temperature trade-off across different metrics; temperature settings (0.3-1.5)** Left: Trade-off between type-token ratio and CDI score. Right: Relationship between non-word rate and CDI score. Points show mean scores across models, with numbers illustrating temperature values.

models within the child-like range for both metrics. Audiobook-trained models consistently overshoot children's TTR, while CDS-trained models come closer but still fall short in vocabulary size. This indicates that the repetitive, high-frequency vocabulary characteristic of early child speech cannot be recovered through decoding adjustments alone.

Non-word-CDI trade-off Higher temperatures increase non-word rates for both architectures. While some CDS-trained models at intermediate temperatures approximate children's average non-word rates, this alignment is misleading: children's non-word rates decrease sharply over development, whereas models remain largely static. Moreover, achieving this partial alignment comes at the expense of TTR matching, suggesting that it is still implausible to simultaneously satisfy multiple developmental constraints.

These results demonstrate that temperature tuning cannot resolve the developmental misalignment between models and children. The gaps reflect fundamental differences in how models and children balance vocabulary size, lexical diversity, and production accuracy. Post-hoc decoding adjustments cannot substitute for differences in underlying learning and production mechanisms. In the next section, we examine whether these differences relate to frequency-dependent learning patterns.

4.3 Frequency effect

The analyses so far show that models lag far behind children in vocabulary acquisition, and this gap cannot be closed by tuning generation parameters. We now ask whether this delay is uniform across input frequency. To test this, we divided the vocabulary into six frequency bands based on child-directed speech and tracked the proportion of words acquired in each band (the same as above analysis). For each band, we fit sigmoid growth curves and extracted the exposure point at which models reached 80% mastery, matching the maximum rate observed in CDI norms. This "months-to-mastery" measure quantifies how quickly different frequency bands are learned as a function of cumulative training exposure.

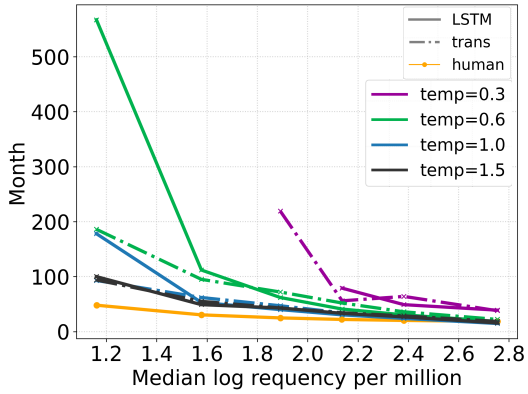

Figure 4: **Extrapolated months across different datasets for models trained on audiobooks**. x-axis: log frequency per million in training data; y-axis: estimated months. Missing points of of the lower frequency bands of temperature =0.3 comes from 0

Divergent frequency sensitivity Figure 4 shows that children acquire words at relatively uniform rates across frequency bands, with only modest variation between high- and low-frequency vocabulary. By contrast, both models display considerable frequency dependence. High-frequency words are acquired relatively quickly, but acquisition times for low-frequency words increase by orders of magnitude. This indicates that models rely heavily on repeated exposure, whereas children generalize more effectively from sparse input.

Temperature modulation Temperature settings systematically modulate frequency effects. At low temperatures ($T = 0.3$), models virtually fail to acquire rare words, as deterministic decoding prevents them from being generated sufficient number of times considered as "known". Higher temperatures ($T = 1.0$–1.5) alleviate this issue by surfacing rare words more often, but the improvement remains limited: even under the most favorable conditions, models require many times more exposure than children to acquire low-frequency vocabulary. Thus, temperature affects severity but not the underlying bias.

Long-tail truncation These results point to long-tail truncation as the primary mechanism underlying models' vocabulary acquisition deficits. Language models are trained to minimize cross-entropy loss, which optimizes for predicting the most probable next character given context. This objective inherently prioritizes frequent patterns: prediction errors on common words contribute more to loss than errors on rare words, leading models to allocate representational capacity proportionally to frequency.

Children, by contrast, do not optimize for predictive accuracy. Their learning mechanisms such as explicit memory for individual word-learning episodes (Markson and Bloom, 1997), cross-situational inference(Smith and Yu, 2008), and social-pragmatic cues (Tomasello, 2000) allow them to acquire words from sparse observations without requiring proportional exposure. This qualitative difference in learning objective, rather than mere data inefficiency, explains why children succeed in the long tail where models consistently fail.

4.4 Content Word Composition

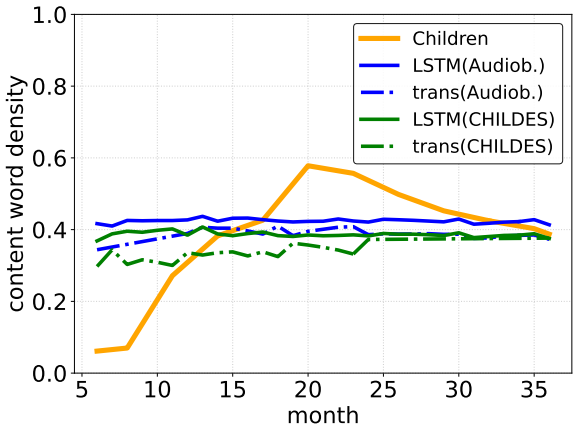

Figure 5: **Proportion of content words in generated text across development.** Children show U-shaped development, while all model configurations maintain stable proportions around 40% across training stages.

The frequency-dependent learning patterns in Section 4.3 may partly stem from differences in functional versus content word production. Function words dominate the highest-frequency bands (Zipf, 2016), whereas our CORPUS-CDI evaluation targets content words (Section 3.4). If models over-rely on function words, this could explain both their elevated lexical diversity (TTR) and their reduced CDI scores.

Children's U-shaped trajectory Figure 5 shows that children exhibit a pronounced U-shaped trajectory. Early productions contain few content words, followed by a sharp rise during the vocabulary spurt, and then a decline as grammatical function words become integrated (Bates et al., 1994; Brown, 1973). Critically, this declining phase coincides with children's steepest CORPUS-CDI growth (Figure 2), showing that children's vocabu-

lary growth is not simply driven by frequency exposure but by a qualitative reorganization of lexical composition.

Models' invariant balance In contrast, all model configurations maintain stable content word proportions of 38-42% across all developmental stages, matching adult written English distributions (Biber et al., 2021). This stability holds regardless of architecture or training corpus—CDS and audiobook models produce nearly identical functional distributions despite different input registers. Models immediately converge to corpus-level statistics and maintain them throughout training.

In contrast, all model configurations maintain stable content word proportions around 40%, which closely matches adult written English distributions (Biber et al., 2021). This invariance holds across architectures and training corpora: both CDS- and audiobook-trained models converge almost immediately to corpus-level statistics and remain fixed throughout training.

This invariance helps explain several earlier findings: (1) Frequency truncation (Section 4.3) disproportionately harms content words, which spread more evenly across frequency bands. (2) CDI underperformance reflects systematic undergeneration of content words, on which CDI norms are based. (3) Elevated TTR may paradoxically arise from over-allocating tokens to a small set of high-frequency function words, forcing broader but shallower sampling of content vocabulary.

Taken together, these results highlight a fundamental mismatch: cross-entropy training aligns models with corpus-level statistics but not with children's developmental trajectory, which reorganizes the functional–content balance over time. Whereas children restructure their lexicon dynamically, models remain tethered to static frequency distributions.

5 Discussion

We introduced CORPUS-CDI, a benchmark that enables direct comparison between children's expressive vocabulary and language model generations by aligning training exposure, production quantity, and frequency distributions. Our analyses yield three main findings that highlight differences between human and model vocabulary acquisition.

First, while both character-level LSTMs and Transformers expand their vocabularies with increased training data, their developmental trajectories diverge largely from those of children. Models consistently produce far greater lexical diversity, maintain stable non-word rates across training, and exhibit slower vocabulary growth without the characteristic acceleration observed in children. These patterns generalize across architectures (LSTM vs. Transformer) and training corpora (audiobooks vs. child-directed speech), suggesting that they reflect fundamental divergence of current learning paradigms rather than dataset or architectures.

Second, these divergences cannot be reconciled through parameter tuning at generation time. Temperature variation experiments show that no single setting simultaneously aligns models with children across TTR, non-word rate, and vocabulary size. In other words, the gap arises from fundamental properties of the learning and production mechanisms rather than from suboptimal decoding choices.

Third, frequency analyses reveal the mechanistic source of these differences. Children acquire words across frequency bands within a narrow age window (around 35–40 months), whereas models require exponentially more exposure for low-frequency words (equivalent to 300–600 months of input for words around log frequency 1.2–1.4 per million). This extreme frequency sensitivity follows directly from cross-entropy optimization, which down-weights rare words during training. Importantly, the interaction with word categories amplifies the divergence: models preserve adult-like content/function ratios (around 40% content words) throughout training, while children show a U-shaped trajectory driven by the integration of function words. Since function words cluster in high-frequency bands whereas content words (the CDI evaluation focus) distribute more evenly, models' frequency-proportional learning underproduces the vocabulary class children prioritize.

These findings point to imitations of predictive accuracy as the sole training objective for modeling acquisition. Children do not rely solely on distributional statistics; they integrate referential grounding, social-pragmatic inference, and episodic memory of word-learning events to acquire vocabulary from sparse observations. Progress in cognitively grounded language modeling may therefore require learning objectives beyond predictive accuracy, such as memory-augmented architectures for rare word retention, contrastive objectives that emphasize lexical representations, or interactive learning paradigms where production influences subsequent input.

6 Limitations

Our evaluation framework has several limitations. First, we evaluate character-level models rather than speech-based systems. While character input provides an upper bound by preserving invariant word forms and explicit boundaries, it bypasses the acoustic variability and segmentation challenges that infants solve (Dupoux, 2018). Speech-based language models (Lakhotia et al., 2021; Nguyen et al., 2024) must additionally contend with phonetic variability and ambiguous word boundaries. Extending CORPUS-CDI to speech therefore requires addressing the non-trivial task of mapping continuous acoustic output into discrete word forms suitable for CDI scoring. Moreover, our models lack multimodal grounding—the word-object associations and interactive feedback that children exploit. Such grounding likely supports children's frequency-independent learning by providing alternative cues beyond distributional statistics.

Second, the CHILDES transcripts used as reference data introduce possible biases. Transcribers often normalize children's productions by correcting errors or completing fragments, which may underestimate non-word rates and inflate apparent model–child gaps. Conversely, our character-level models inherently generate orthographically valid forms, potentially overstating their production accuracy relative to children's speech.

Third, our comparison of parental CDI reports with model outputs involves methodological misalignment. CDI scores reflect whether parents judge a child to "say" a word, capturing semantic knowledge and contextual usage in addition to production capability. In contrast, our CORPUS-CDI metric counts raw production frequency in unconstrained model text. A child uttering dog once demonstrates robust lexical knowledge, whereas a model generating dog 60 times may not. While our threshold ($\tau = 60$) calibrates vocabulary growth slopes, the construct validity gap between CDI reports and model generation remains.

Finally, our study focuses on English and a single vocabulary assessment tool (CDI). Cross-linguistic validation is essential for establishing universality: languages with richer morphology, less transparent orthography, or different child-directed registers may yield different alignments between model and child learning. Character-level models in particular may be sensitive to orthographic properties, limiting generalizability beyond English.

7 Ethics Statement

All data used in this study is already publicly available. This work focuses on model benchmarking and internal simulations, aiming to advance fundamental research in this area. However, we do not consider this testing sufficient for deployment; any future real-life applications should undergo rigorous validation in deidentifying.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. 2024. Language learning, representation, and processing in humans and machines: Introduction to the special issue.

Elizabeth Bates, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J. Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, 21(1):85–123.

Douglas Biber, Stig Johansson, Edward Finegan, Susan Conrad, and Geoffrey N Leech. 2021. Grammar of spoken and written english.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv e-prints*, pages arXiv–2108.

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bastian Bunzeck, Daniel Duran, Leonie Schade, and Sina Zarrieß. 2024. Graphemes vs. phonemes: Battling it out in character-based language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 54–64, Miami, FL, USA. Association for Computational Linguistics.

Bastian Bunzeck and 1 others. 2025. Child-directed language does not consistently boost syntax learning in language models.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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*.

Jing Feng, Hao Cheng, Roy Schwartz, and Mari Ostendorf. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Larry Fenson and 1 others. 2007. Macarthur-bates communicative development inventories.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

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.

Zébulon Goriely, Richard Diehl Martinez, Andrew Caines, Lisa Beinborn, and Paula Buttery. 2024. From babble to words: Pre-training language models on continuous streams of phonemes. *arXiv preprint arXiv:2410.22906*.

Zébulon Goriely and Paula Buttery. 2025. IPA-CHILDES & G2P+: Feature-rich resources for cross-lingual phonology and phonemic language modeling. *arXiv preprint arXiv:2504.03338*.

Akari Haga, Miyu Oba, and Yohei Oseki. 2024. Modeling overregularization in children with small language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14451–14466. Association for Computational Linguistics.

Pierre A Hallé and Bénédicte de Boysson-Bardies. 1994. Emergence of an early receptive lexicon: Infants' recognition of words. *Infant behavior and development*, 17(2):119–129.

Jennifer Hu, Alvin Wei Ming Tan, Steven Y Feng, and Michael C Frank. 2025. Language production is harder than comprehension for children and language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

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 preprint arXiv:2412.05149*.

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.

Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. Mewl: Few-shot multimodal word learning with referential uncertainty. In *International Conference on Machine Learning*, pages 15144–15169. PMLR.

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

Eliza Kosoy, Emily Rose Reagan, Leslie Lai, Alison Gopnik, and Danielle Krettek Cobb. 2023. Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of lamda responses. *arXiv preprint arXiv:2305.11243*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and 1 others. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Marvin Lavechin, Maureen De Seyssel, Lucas Gautheron, Emmanuel Dupoux, and Alejandrina Cristia. 2022. Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*, 8(1):389–407.

Marvin Lavechin, Maureen De Seyssel, Hadrien Titeux, Guillaume Wisniewski, Hervé Bredin, Alejandrina Cristia, and Emmanuel Dupoux. 2025. Simulating early phonetic and word learning without linguistic categories. *Developmental Science*, 28(2):e13606.

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*.

Ori Lavi-Rotbain and Inbal Arnon. 2022. The learnability consequences of Zipfian distributions in language. *Cognition*, 223:105038.

Ori Lavi-Rotbain and Inbal Arnon. 2023. Zipfian distributions in child-directed speech. *Open Mind*, 7:1–21.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2024. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. *arXiv preprint arXiv:2405.13828*.

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

Lori Markson and Paul Bloom. 1997. Evidence against a dedicated system for word learning in children. *Nature*, 385(6619):813–815.

Gina McKee, Dana McDaniel, Sarah Garrett, and Ruth A. Berman. 2022. Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in Psychology*, 13:905789.

Céline Ngon, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat, and Sharon Peperkamp. 2013. (non) words,(non) words,(non) words: evidence for a protolexicon during the first year of life. *Developmental Science*, 16(1):24–34.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, and 1 others. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

Mitja Nikolaus and Abdellah Fourtassi. 2021a. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 200–210.

Mitja Nikolaus and Abdellah Fourtassi. 2021b. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th Conference on computational natural language learning*, pages 391–407.

Eva Portelance, Michael C. Frank, Dan Jurafsky, Gregory Scontras, and Rui Cao. 2023. Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*, 47(8):e13334.

Okko Räsänen and Daniil Kocharov. 2024. Age-dependent analysis and stochastic generation of child-directed speech. *arXiv preprint arXiv:2405.07700*.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of Child Language*, 14(2):201–209.

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*.

Linda B Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Daniel Swingley. 2011. The looking-while-listening procedure. *Research methods in child language: A practical guide*, pages 29–42.

Alvin Wei Ming Tan, Sunny Yu, Bria Long, Wanjing Anya Ma, Tonya Murray, Rebecca D Silverman, Jason D Yeatman, and Michael C Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *arXiv preprint arXiv:2406.10215*.

Michael Tomasello. 2000. The social-pragmatic theory of word learning. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 10(4):401–413.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Alexandre Verine, Benjamin Alamé, Paula Czarnowska, Sylvain Lamprier, Jacopo Staiano, and Nathalie Aussenac-Gilles. 2025. Improving diversity in language models: When temperature fails, change the loss. *arXiv preprint arXiv:2508.09654*.

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 and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

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

Marieke Wolters, Ori Lavi-Rotbain, and Inbal Arnon. 2024. The facilitating effect of Zipfian distributions on word learning in children. *Developmental Science*, 27(3):e13456.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books.

A Appendix

A.1 Training and generation details

Table 1 summarizes the age-appropriate scaling of training data across three exposure scenarios. Linguistic exposure estimates are based on daylong recording studies (Cristia et al., 2019), which document substantial variability in the amount of speech heard by children across environments. Annual exposure ranges from approximately 100 hours in low-input households to 1000 hours in high-input ones, with 500 hours per year representing a mid-range estimate.

Cumulative exposure thus increases from 50–500 hours at 6 months to 300–3000 hours by 36 months. Assuming a standard speech rate of three words per second, this corresponds to training sets ranging from 0.58M words (6 months, low exposure) to 34.71M words (36 months, high exposure).

Age	Hours	Words	Characters	Utterances
Low exposure scenario (100 hours/year)				
6	50	0.58	2.95	58
12	100	1.16	5.89	116
18	150	1.73	8.84	173
24	200	2.31	11.78	231
30	250	2.89	14.73	289
36	300	3.47	17.67	347
Medium exposure scenario (500 hours/year)				
6	250	2.89	14.73	289
12	500	5.78	29.45	578
18	750	8.68	44.18	868
24	1000	11.57	58.90	1157
30	1250	14.46	73.63	1446
36	1500	17.35	88.35	1735
High exposure scenario (1000 hours/year)				
6	500	5.78	29.45	578
12	1000	11.57	58.90	1157
18	1500	17.35	88.35	1735
24	2000	23.14	117.80	2314
30	2500	28.92	147.25	2892
36	3000	34.71	176.70	3471

Table 1: **Age-appropriate scaling of training datasets across three exposure scenarios.** Estimates are derived from three exposure scenarios (100h, 500h, 1000h per year), reflecting observed variability in children's linguistic environments (Cristia et al., 2019). Word counts assume 3 words/second; character counts assume 5.1 characters/word; utterance counts assume 10 words/utterance.

Model training employed carefully tuned hyperparameters (Table 2), selected through systematic validation experiments to ensure training stability and computational efficiency across different data

scales.

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.** Configurations balance model capacity, stability, and efficiency across training scales.

The examples in Table 4 highlight distinct generation behaviors across datasets and model architectures under a fixed sampling temperature ($T = 1.0$). Models trained on child-directed input (Child, CHILDES) tend to produce shorter, phonologically simple, and semantically grounded utterances, resembling early stages of language development. In contrast, audiobook-trained models generate longer, syntactically structured sequences but often lack pragmatic coherence. Comparing architectures, Transformers exhibit more globally consistent syntax and longer dependency spans, while LSTMs display stronger local predictability but higher rates of morphological irregularities. These qualitative patterns align with quantitative trends observed in our lexical and nonword diversity metrics, underscoring how data type and inductive bias jointly shape emergent linguistic structure.

A.2 Result details

To establish reliable vocabulary acquisition metrics, we conducted systematic analyses of different word count thresholds (Figure 6). This calibration process aimed to align our CORPUS-CDI scores with empirical observations of child vocabulary development.

month	sec/hour	estimation	production
6	61.25	55,125	782
7	65.30	58,769	1,323
8	69.35	62,412	1,278
9	73.40	66,056	3,964
10	77.44	69,699	627
11	81.49	73,343	1,148
12	85.54	76,986	1,361
13	89.59	80,630	568
14	93.64	84,273	2,720
15	97.69	87,917	13,765
16	101.73	91,560	3,620
17	105.78	95,204	7,589
18	109.83	98,847	15,197
19	113.88	102,491	17,090
20	117.93	106,134	19,912
21	121.98	109,778	33,458
22	126.02	113,421	33,880
23	130.07	117,065	63,738
24	134.12	120,708	191,062
25	138.17	124,352	130,525
26	142.22	127,995	136,195
27	146.27	131,639	148,741
28	150.31	135,282	158,910
29	154.36	138,926	190,263
30	158.41	142,569	195,549
31	162.46	146,213	196,677
32	166.51	149,856	184,024
33	170.56	153,500	185,534
34	174.60	157,143	185,691
35	178.65	160,787	153,515
36	182.70	164,430	299,537

Table 3: **Monthly progression of speech metrics.** Estimated exposure and production rates over developmental time.

Notably, the frequency effects observed in the Machine CDI are evaluated on a selected subset matched with the human-CDI set. It is unclear whether this effect is influenced by the randomness of sampling. Also, the selected words are typically in the median frequency bands, making it unknown whether such an effect stems from the limited generation size. To address this concern, we expanded our analysis of the model generations using the same and larger sampling sets. Figure 7 shows the word frequency distributions of the training and generated datasets by the LSTM model trained on 3.4M words, with different generation sizes. Expanding the generation set mitigates some

Setting	Generated Sequences
Child	dad dada mommy looked it missed its wheel i want my plate for it
Audiobook (LSTM)	wounded for woman says no i don't know what i fuller over his week scully a
Audiobook (Trans)	to o he had been and the work of the lord ascend mass of softly rounded head
CHILDES (LSTM)	can bump he said I was wondering what yeah detfive wroggy on drh aulf
CHILDES (Trans)	on helsewhere he doesn't like him does he hafta dragged on doh aunty quite

Table 4: **Examples of generated sequences.** Samples are drawn from models trained on different datasets and architectures at a sampling temperature of 1.0. The boundary marker is replaced with blank space for readability.

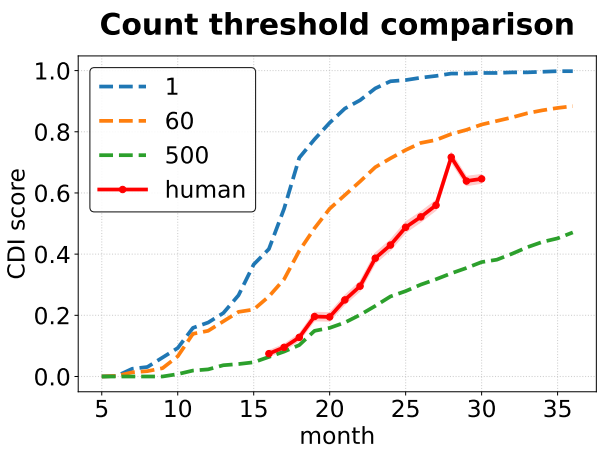

Figure 6: **Count-based calibration** Comparison of different word frequency thresholds for determining word acquisition. The optimal threshold (60 occurrences) was selected based on alignment with parental vocabulary size reports (red line). Each threshold produces distinct vocabulary growth trajectories, with higher thresholds resulting in more conservative estimates.

truncation effects, as more high-frequency words are generated, but low-frequency words remain underrepresented. This pattern reflects the model's limited ability to reproduce realistic word distributions, especially for rare words, even with larger sampling sizes.

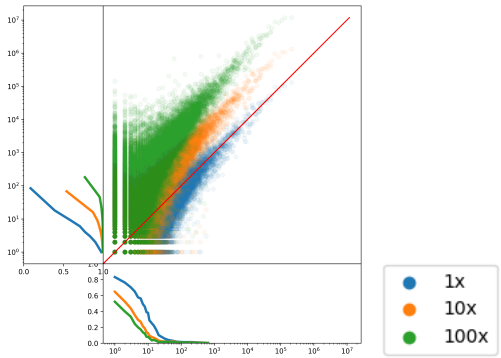

Figure 7: **Word count distribution(temperature = 1.0)**.Word frequency relationships between training and generation sets under different sampling conditions (1x, 10x, 100x generation volume). x-axis: word frequency in the training set; y-axis: word frequency in the generation set. Side plot on the bottom show Missing rates as a function of word count in input corpus.

Although models trained on higher-exposure estimates (1000 h) achieve slightly higher CORPUS-CDI scores than those trained on smaller sets (100 h, 500 h), the overall developmental trajectories remain similar. This pattern indicates that scaling input exposure alone does not yield more human-like developmental dynamics; instead, the gap likely arises from structural differences in learning mechanisms rather than data quantity.

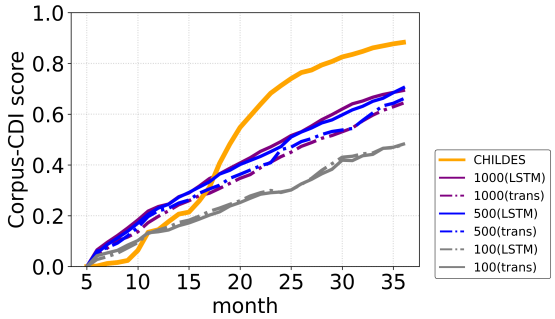

Figure 8: **Model vocabulary growth across estimated annual exposure levels.** Comparison of CORPUS-CDI trajectories for models trained on 100-, 500-, and 1000-hour exposure estimates, alongside CHILDES child production data. Models show roughly linear vocabulary growth that scales modestly with total exposure but remains far below the human developmental curve.

Dataset (Model)	Temp	TTR (%)		NWType (%)		NWToken (%)		CDI(%)	
		Mean	Last	Mean	Last	Mean	Last	Mean	Last
CHILDES	-	18.3	20.7	9.6	0.7	4.7	0.2	48.6	88.4
Audiobook (LSTM)	**0.3**	43.8 [41.3, 46.1]	44.2	13.5 [12.0, 14.8]	13.8	10.4 [9.2, 11.6]	10.6	62.5 [59.2, 65.8]	63.1
	0.6	45.9 [43.4, 48.2]	46.3	14.8 [13.3, 16.1]	15.1	11.7 [10.5, 12.9]	11.9	65.2 [61.9, 68.5]	65.8
	1.0	48.1 [45.6, 50.4]	48.5	16.4 [14.9, 17.7]	16.7	13.1 [11.9, 14.3]	13.3	68.8 [65.5, 72.1]	69.4
	1.5	**50.3** [47.8, 52.6]	**50.7**	**18.0** [16.5, 19.3]	**18.3**	**14.8** [13.6, 16.0]	**15.0**	**71.9** [68.6, 75.2]	**72.5**
Audiobook (Trans)	**0.3**	44.5 [42.0, 46.8]	44.9	13.8 [12.3, 15.1]	14.1	10.7 [9.5, 11.9]	10.9	63.2 [59.9, 66.5]	63.8
	0.6	46.6 [44.1, 48.9]	47.0	15.1 [13.6, 16.4]	15.4	12.0 [10.8, 13.2]	12.2	65.9 [62.6, 69.2]	66.5
	1.0	48.8 [46.3, 51.1]	49.2	16.7 [15.2, 18.0]	17.0	13.4 [12.2, 14.6]	13.6	69.5 [66.2, 72.8]	70.1
	1.5	**51.0** [48.5, 53.3]	**51.4**	**18.3** [16.8, 19.6]	**18.6**	**15.1** [13.9, 16.3]	**15.3**	**72.6** [69.3, 75.9]	**73.2**
CHILDES (LSTM)	**0.3**	43.8 [41.3, 46.1]	44.2	13.5 [12.0, 14.8]	13.8	10.4 [9.2, 11.6]	10.6	62.5 [59.2, 65.8]	63.1
	0.6	45.9 [43.4, 48.2]	46.3	14.8 [13.3, 16.1]	15.1	11.7 [10.5, 12.9]	11.9	65.2 [61.9, 68.5]	65.8
	1.0	48.1 [45.6, 50.4]	48.5	16.4 [14.9, 17.7]	16.7	13.1 [11.9, 14.3]	13.3	68.8 [65.5, 72.1]	69.4
	1.5	**50.3** [47.8, 52.6]	**50.7**	**18.0** [16.5, 19.3]	**18.3**	**14.8** [13.6, 16.0]	**15.0**	**71.9** [68.6, 75.2]	**72.5**
CHILDES (Trans)	**0.3**	44.5 [42.0, 46.8]	44.9	13.8 [12.3, 15.1]	14.1	10.7 [9.5, 11.9]	10.9	63.2 [59.9, 66.5]	63.8
	0.6	46.6 [44.1, 48.9]	47.0	15.1 [13.6, 16.4]	15.4	12.0 [10.8, 13.2]	12.2	65.9 [62.6, 69.2]	66.5
	1.0	48.8 [46.3, 51.1]	49.2	16.7 [15.2, 18.0]	17.0	13.4 [12.2, 14.6]	13.6	69.5 [66.2, 72.8]	70.1
	1.5	**51.0** [48.5, 53.3]	**51.4**	**18.3** [16.8, 19.6]	**18.6**	**15.1** [13.9, 16.3]	**15.3**	**72.6** [69.3, 75.9]	**73.2**

Table 5: Metric scores across different model settings presented in percentages (%). Mean values are averaged across different months, followed by range across different models. We also reported the score of the 36th month. TTR: Type-token ratio, NWType: Non-word type rate, NWToken: Non-word token rate. CHILDES child production. Best scores for each setting are in **bold**.