Developmentally-plausible Working Memory Shapes a Critical Period for Language Acquisition

Masato Mita and **Ryo Yoshida** and **Yohei Oseki**
The University of Tokyo
{mita, yoshiryo0617, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

Large language models possess general linguistic abilities but acquire language less efficiently than humans. This study proposes a method for integrating the developmental characteristics of working memory during the critical period, a stage when human language acquisition is particularly efficient, into the training process of language models. The proposed method introduces a mechanism that initially constrains *working memory* during the early stages of training and gradually relaxes this constraint in an exponential manner as learning progresses. Targeted syntactic evaluation shows that the proposed method outperforms conventional methods without memory constraints or with static memory constraints. These findings not only provide new directions for designing data-efficient language models but also offer indirect evidence supporting the role of the developmental characteristics of working memory as the underlying mechanism of the critical period in language acquisition.

https://github.com/osekilab/CPLM

1 Introduction

Large language models (LLMs) exhibit general linguistic abilities comparable to those of humans; however, their efficiency in language acquisition remains far inferior. It has been noted that LLMs require data quantities that are three to four orders of magnitude larger than those needed for humans to achieve comparable performance across many evaluation metrics (Warstadt et al., 2023). This disparity in data efficiency reflects the current reliance of LLMs on scaling and suggests not only a significant potential for improving learning efficiency but also the possibility of drawing *insights* from human language processing and acquisition.

An important theoretical framework for understanding the efficiency of human language acquisition is the **Critical Period Hypothesis (CPH)** (Lenneberg, 1967). The CPH posits that there is a specific period during which language can be acquired efficiently and that this ability diminishes thereafter. Various studies, including cases of limited exposure to first language (L_1) during childhood and age-related effects on second language (L_2) acquisition, support the existence of a critical period (CP) (Fromkin et al., 1974; Curtiss, 1977; Johnson and Newport, 1989). However, the reasons why children acquire language more efficiently than adults remain partially unresolved. A compelling explanation is the **Less-is-More Hypothesis** (Newport, 1990), which attributes the superior learning of children to limited cognitive resources such as working memory. According to this hypothesis, children's limited processing capacities enable them to efficiently extract fundamental patterns and structures (e.g., grammatical rules) from linguistic input, whereas adults, with their greater cognitive capacities, are more likely to be distracted by complex information, thereby hindering rule acquisition.

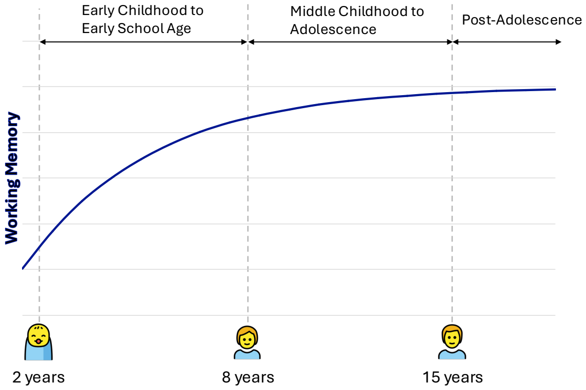


Figure 1: Developmental trajectory of human working memory

This hypothesis not only offers a compelling account of human learning but also has implications for how we design artificial systems. Language is not an arbitrary object of learning but a cultural arti-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9386–9399
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

fact shaped under cognitive constraints. A growing body of work suggests that its structural properties reflect pressures for *efficient communication* —that is, to maximize informational content while minimizing cognitive effort in production and comprehension under human limitations (Zipf, 1949; Jaeger and Tily, 2011; Christiansen and Chater, 2016; Kemp et al., 2018; Fedorenko et al., 2024). Over generations, language has likely evolved to be learnable by agents with limited memory and processing capacity. From this perspective, incorporating such constraints into language models (LMs) is not merely an act of mimicking human limitations, but a theoretically grounded way to introduce an inductive bias that aligns with the nature of the target: language shaped by cognitively bounded agents. Learning under such constraints may help LMs acquire representations better suited to natural language.[1]

Inspired by the *Less-is-More* hypothesis, we use LMs to study the CP for language acquisition, focusing on L_1 acquisition and investigating whether integrating human cognitive developmental characteristics, particularly the developmental properties of *working memory* (Figure 1), into LMs can facilitate efficient language acquisition. Specifically, we propose a method for incorporating the exponential increase in working memory capacity that corresponds to the CP into LMs and analyze its impact on learning efficiency. Using a GPT-2 model (Radford et al., 2019) trained on a Child-Directed Speech (CDS) dataset (Huebner and Willits, 2021), we conduct evaluation experiments with Zorro (Huebner et al., 2021), a targeted syntactic evaluation benchmark specialized for CDS. The results demonstrate that a cognitively plausible model, which initially restricts working memory and gradually relaxes this constraint exponentially as training progresses, outperforms models without memory constraints or with static memory constraints. These findings provide new insights into designing data-efficient LMs, contributing to the field of **natural language processing**, while also offering indirect evidence supporting the role of the developmental characteristics of working memory as the underlying mechanism of the CPH in human language acquisition, contributing to the field of **cognitive science**.

2 Related Work

2.1 Critical Period for Language Acquisition

The CPH posits that language acquisition is most efficient within a specific developmental window, after which it declines. CP effects are observed in both L_1 and L_2 acquisition, suggesting a shared underlying mechanism.

Critical Period for L_1 Acquisition Research in neurolinguistics and cognitive science suggests that there is a biologically determined CP for acquiring an L_1, beyond which full native-like proficiency is unattainable if exposure to language is delayed. Studies on late L_1 learners, such as deaf individuals who acquire sign language after early childhood, indicate severe deficits in grammatical proficiency compared to those exposed to language from birth (Mayberry and Fischer, 1989; Newport, 1990). These findings suggest that neural plasticity, essential for L_1 acquisition, diminishes with age, limiting the ability to develop full linguistic competence. From a theoretical perspective, the existence of the CP for L_1 acquisition is often attributed to biological constraints. Nativist theories propose that L_1 acquisition relies on an innate language faculty that operates most effectively during the CP (Penfield, 1965; Chomsky, 1965; Pinker, 1994). On the other hand, empiricist perspectives argue that the decline in L_1 learning ability may result from environmental factors, such as a reduced need for language learning mechanisms once fundamental linguistic structures have been internalized (Elman et al., 1996; Seidenberg and Zevin, 2006). Despite extensive research, the precise boundary and mechanisms of the CP for L_1 remain a subject of debate.

Critical Period for L_2 Acquisition CP effects are also observed in L_2 acquisition, where late learners struggle with pronunciation, morphology, and syntax (Johnson and Newport, 1989; Hartshorne et al., 2018). While biological constraints play a role, entrenchment—where prior exposure to L_1 limits flexibility in learning new linguistic structures—is also a factor (Ellis and Lambon Ralph, 2000; Seidenberg and Zevin, 2006). Although the CP for L_2 acquisition is an important topic, this study focuses on the CP for L_1 acquisition, since our goal is to design data-efficient LMs by exploring the mechanisms of CP in L_1 acquisition.

[1] Futrell and Mahowald (2025) for an alternative view suggesting that, given the empirical success of machine learning, effective learning may not require cognitively inspired constraints or inductive biases.

2.2 The Role of Language Models in Acquisition Theories

In recent years, computational models have played a crucial role in elucidating the mechanisms of language acquisition. These models enable controlled investigations of learning mechanisms and environments, which are difficult to achieve with human participants, and they are used to test theoretical claims such as the "poverty of the stimulus" (Clark and Lappin, 2011). For instance, McCoy et al. (2020), Wilcox et al. (2024), and Warstadt et al. (2023) have employed LMs to directly test hypotheses about language acquisition, demonstrating that such models can provide proof-of-concept evidence for *learnability*. These studies have attracted attention as efforts to deepen theoretical discussions on language acquisition through computational modeling.

While Transformer-based LMs are not cognitive models in a strict sense, they are widely adopted in acquisition research as abstract "model learners" (Warstadt and Bowman, 2022). Rather than replicating the full complexity of human cognition, these models are used to investigate the role of specific biases by selectively adding or removing them. This approach allows researchers to assess whether certain linguistic phenomena can be acquired purely through statistical learning or require inductive constraints, thereby testing the necessity of such biases. When a model fails to acquire a phenomenon in the absence of a particular bias, but succeeds once the bias is introduced, it offers at least weak evidence for the bias's relevance in human language acquisition (McCoy et al., 2020). Our study follows this reverse-engineering paradigm (Dupoux, 2018), using LMs not as literal simulations of human learners, but as controlled testbeds for cognitive hypotheses.

Constantinescu et al. (2025) investigated CP phenomena in L_2 acquisition and L_1 attrition,[2] assuming a shared underlying mechanism for CP effects across L_1 and L_2. They simulated L_2 exposure at varying ages to examine how LMs differ from human learners, finding that LMs do not naturally exhibit CP effects. To artificially induce such effects, they employed Elastic Weight Consolidation (Kirkpatrick et al., 2017), a regularization method for mitigating catastrophic forgetting, thereby mimicking a maturational decline in plasticity. Their findings suggest that CP effects are not an inevitable outcome of statistical learning but may instead involve innate mechanisms. While this study shares the broader objective of enhancing the cognitive plausibility of LMs as models of human language acquisition, it differs from Constantinescu et al. (2025) in both *focus* and *methodology*. Rather than modeling CP effects through dataset manipulation or post-CP plasticity constraints, this study explicitly addresses the **developmental processes unfolding during the CP itself**. Specifically, we integrate a mechanism to simulate the progressive growth of working memory capacity throughout the CP, a factor considered crucial for L_1 acquisition but previously unmodeled in LM-based research. By incorporating developmental constraints, this study aims to provide a more fine-grained computational model of early L_1 acquisition and its cognitive underpinnings, advancing the developmental plausibility of LMs.

3 Critical Period-inspired Language Model

3.1 Modeling Developmental Trajectory of Human Working Memory

Human working memory undergoes substantial developmental changes, progressing through three distinct stages: early childhood to early school age (2–7 years), middle childhood to early adolescence (8–14 years), and post-adolescence (15 years and older). During early childhood, both information retention capacity and processing ability improve rapidly, reflecting a significant expansion of cognitive resources (Cowan et al., 1999; Gathercole et al., 2004). This rapid growth begins to decelerate during middle childhood and early adolescence as the brain approaches maturation (Luna et al., 2004; Gathercole et al., 2004). By post-adolescence, working memory capacity plateaus, reaching adult-level performance (Sowell et al., 2002; Luna et al., 2004).

Based on these observations, we characterized the growth trajectory of working memory, as illustrated in Figure 1, using an exponential model of the form $y = b - a^x$ $(0 < a < 1)$. In this model, b represents the asymptotic upper limit of working memory capacity, corresponding to adult-level performance, while a determines the rate of growth. Specifically, smaller values of a result in steeper early growth, reflecting the rapid cognitive development observed during early childhood, whereas

[2] The phenomenon in which earlier cessation of L_1 exposure increases the likelihood of L_1 forgetting.

larger values of a indicate a slower rate of change.

This modeling approach is justified for several reasons. First, the horizontal asymptote inherent in the exponential function accurately represents the biological ceiling of adult working memory capacity. Second, the rapid initial increase observed during early childhood is consistent with the steep growth predicted by this exponential form. Finally, alternative models, such as logarithmic or linear growth, fail to account for both the early rapid development and the eventual plateau: logarithmic models imply unbounded growth, while linear models oversimplify the deceleration phase. Thus, the exponential model $y = b - a^x$ offers a concise and biologically plausible representation of the developmental trajectory of human working memory, aligning well with observed patterns and theoretical considerations.[3]

3.2 Integrating Human Working Memory into Language Models

In this study, Attention with Linear Biases (ALiBi) (Press et al., 2022) is employed to model the constraints of human working memory. ALiBi is a method for Transformer (Vaswani et al., 2017) models that does not use positional embeddings but instead applies a distance-dependent linear penalty to attention scores. Specifically, the attention score for an input sequence of length L is calculated as follows:

$$\begin{aligned} \text{Attention Score} &= \operatorname{softmax}\left(q_i K^\top + m \cdot B\right), \\ B &= \begin{bmatrix} -(i-1) & -(i-2) & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{1}$$

Here, $q_i \in \mathbb{R}^{1\times d}$, $K \in \mathbb{R}^{L\times d}$, $m \in \mathbb{R}_{[0,1]}$, and $B \in \mathbb{R}^{1\times L}$ represent the query, the key, a scalar slope specific to each attention head, and a bias matrix encoding the relative distances between queries and keys, respectively, where B_i is defined as the negative absolute difference between the query position i and each key position. The values of m are set geometrically for each head. For example, in an 8-head model, the values of m are assigned as follows: $m = 1, \frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{128}$. The slope m takes values in the range $[0, 1]$, ensuring a consistent interpretation of its influence on attention scores. By penalizing attention scores for query-key pairs with greater distances, ALiBi introduces a *recency bias* to the model. Originally, ALiBi was proposed to enhance the extrapolation capability of Transformer models. More recently, Clark et al. (2025) has shown that incorporating it into attention score computation during training allows for the estimation of surprisal patterns resembling human reading times. This suggests its potential for modeling human-like memory decay and cognitive limitations.

However, since the slope m in ALiBi is fixed for each attention head, the approach does not inherently reflect the developmental increase in working memory capacity (i.e., reduced decay) over time (Figure 1). Therefore, this study proposes a method, **DYNAMICLIMIT-EXP**, which replicates the developmental characteristics of working memory during the CP, specifically its exponential growth. This is achieved by exponentially decreasing the slope m in ALiBi as training epochs progress. In this method, the slope m in the ALiBi mechanism is updated at each epoch t as follows:

$$m_t = m_0 \cdot r^t, \tag{2}$$

where m_0 represents the initial slope, $r \in (0, 1)$ is the decay rate, and t denotes the current epoch. In this study, the model's working memory capacity w_t is formulated as follows:

$$w_t := 1 - m_t. \tag{3}$$

This definition links the dynamically decaying slope m_t to the model's working memory capacity w_t: as m_t decreases exponentially, w_t grows, enabling broader contextual retention over time. By simulating this developmental trajectory, the model initially focuses on short-range dependencies and gradually attends to longer ones.

4 Experiments

This study explores whether LMs trained from scratch can achieve more efficient L_1 acquisition by incorporating the developmental characteristics of human working memory. Specifically, we aim to determine whether this approach can replicate the increased efficiency of L_1 acquisition observed during the CP in L_1 acquisition, focusing on the developmental advantages before the end of this period.

[3] ACT-R (Anderson and Milson, 1989) suggests that working memory *decays* exponentially in language processing, while we propose that working memory *grows* exponentially in language acquisition, but whether the shared exponential function between language processing and acquisition is a coincidence remains to be investigated in future.

Model	Overall*	D-N Agr	S-V Agr	Ana. Agr	Arg. Str*	Binding†	Case*	Ellipsis†	Filler. Gap*	Irregular	Island†	Local. Atr*	Quantifiers*	NPI*
NoLimit	56.5	49.8	49.7	49.9	44.8	61.8	70.8	73.3	72.1	51.7	61.7	47.1	47.9	53.9
StaticLimit	56.8	50.2	49.9	49.8	44.4	60.5	70.3	71.4	74.7	52.2	62.9	45.3	52.3	54.4
DynamicLimit-Linear	61.6	51.0	49.6	49.5	64.3	60.3	88.6	47.6	90.8	53.0	57.0	47.9	56.8	84.3
DynamicLimit-Exp	62.2	50.8	50.0	49.6	67.7	58.7	95.2	43.1	93.6	52.2	53.6	51.3	57.6	85.0

Table 1: Accuracy (%) of models trained on AO-CHILDES dataset. * and † indicate items where DynamicLimit-Exp performed significantly better or worse than NoLimit, respectively (z-test for proportions, $p < 0.05$).

Model	Overall*	D-N Agr	S-V Agr	Ana. Agr	Arg. Str*	Binding†	Case*	Ellipsis†	Filler. Gap*	Irregular	Island*	Local. Atr*	Quantifiers	NPI*
NoLimit	54.7	50.3	50.0	47.2	68.4	62.6	73.4	60.8	42.9	53.4	51.1	42.7	41.2	42.6
StaticLimit	54.7	50.4	50.0	47.1	73.7	61.2	87.4	57.3	56.1	52.3	53.0	40.8	42.0	38.9
DynamicLimit-Linear	58.6	50.0	50.5	48.4	71.9	58.8	96.9	38.7	82.7	51.6	57.9	59.6	41.5	53.4
DynamicLimit-Exp	59.1	49.8	50.4	46.0	71.5	59.3	97.7	37.4	86.5	51.1	58.0	60.5	42.2	53.9

Table 2: Accuracy (%) of models trained on Wikipedia dataset. * and † indicate items where DynamicLimit-Exp performed significantly better or worse than NoLimit, respectively (z-test for proportions, $p < 0.05$).

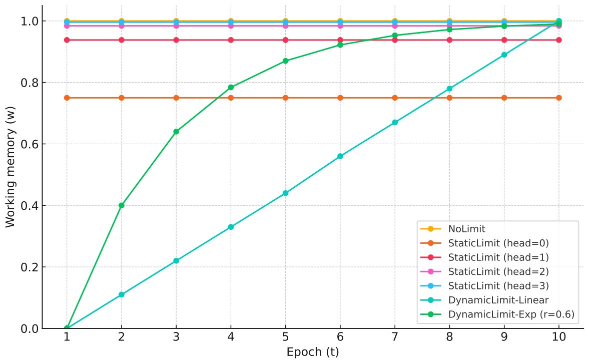


Figure 2: Trajectory of working memory capacity for each model (num. of epochs = 10)

4.1 Configurations

Models We used the `transformers` (Wolf et al., 2020) implementation of the GPT-2 (Radford et al., 2019) as the base LM. While some studies utilize RoBERTa (Liu et al., 2019) as a base model (Huebner et al., 2021; Warstadt et al., 2023), we selected GPT-2 for two primary reasons: (1) its unidirectional (left-to-right) predictions more effectively capture human working memory constraints, and (2) GPT-based architectures dominate modern LLMs (OpenAI, 2023; Touvron et al., 2023b).

Dataset We used AO-CHILDES (Huebner and Willits, 2021)[4] as the training dataset, which is derived from the CHILDES dataset (Macwhinney, 2000) and records CDS from conversations between children and adults. AO-CHILDES contains 5 million words of speech directed at English-speaking children aged 1–6 years and controls for external factors such as age group, speaker variation, and situational context. As a preprocessing step, following Haga et al. (2024), all sentences were converted to lowercase, and sentences shorter than three words were excluded. Since the AO-CHILDES dataset contains only about 5 million words, training a standard GPT-2 model would likely result in overfitting. To mitigate this, we followed existing studies on small language models (SLMs) trained with CDS datasets (Huebner et al., 2021; Haga et al., 2024) and constructed an SLM with 4 layers, 4 attention heads, and 256 embedding dimensions for the base model. Details of the training configuration for the base model are provided in Appendix A.

Furthermore, to determine whether the CP effect stems from exposure to specific linguistic stimuli, such as CDS, or from the model's cognitive developmental properties independent of input, we conducted a complementary experiment using Wikipedia (written language, adult-oriented) as training data. Following Huebner et al. (2021), 500,000 sentences were randomly sampled from the English Wikipedia corpus. We used the latest version of Wikipedia, as of January 2025,[5] and preprocessed it using `WikiExtractor`.[6] We provide the sentence length distribution for the AO-CHILDES and Wikipedia datasets used in this ex-

[4] https://github.com/UIUCLearningLanguageLab/AOCHILDES

[5] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[6] https://github.com/attardi/wikiextractor

periment in Appendix B.

Evaluation We evaluate the grammatical abilities of these models using a developmentally inspired targeted syntactic evaluation benchmark, Zorro (Huebner et al., 2021). Zorro is designed for assessing the syntactic and grammatical knowledge of LMs in child-directed language and consists of 13 mid-level categories and 23 subcategories. Each subcategory contains 2,000 sentence pairs, with one grammatically acceptable and one unacceptable sentence per pair.[7] Below is an example of a minimal pair from the "Subject-verb agreement (S-V AGR)" category:[8]

(1) a. The **lie** on the foot is flat.

b. *The **lies** on the foot is flat.

By inputting both the acceptable and unacceptable sentence into the model and calculating the proportion of pairs where the model assigns a higher probability to the acceptable sentence, we obtain the grammaticality judgment score (Accuracy). In this study, we report scores for each mid-level category (henceforth, *grammatical items*) as well as their macro-average.

4.2 Baselines

We prepared the following three baseline models to precisely analyze the learning effects of different working memory limitation strategies:

- **NoLimit**: A model with no memory constraints. Working memory remains constant from the early stages of training, simulating the mature working memory observed post-adolescence. This configuration is equivalent to a vanilla GPT-2 (Radford et al., 2019).
- **StaticLimit**: A model applying standard ALiBi (Press et al., 2022) during attention score calculation, where memory constraints remain fixed throughout training.
- **DynamicLimit-Linear**: A model in which the ALiBi slope m decreases linearly over the course of training.

To ensure a fair comparison between the linear and exponential growth curves of working memory, we controlled the initial and final values of working memory capacity w_t in DynamicLimit-Linear and DynamicLimit-Exp to be as similar as possible. Specifically, we set the number of training epochs to 10 and configured both models with an initial slope of $m = 1.0$ and a final slope of $m = 0.0$. Figure 2 illustrates the trajectory of working memory capacity for each model. All models were trained using three different seeds, and we report the average results across these runs.

4.3 Results

Developmentally-plausible working memory shapes the CP for L_1 acquisition Table 1 presents the accuracy of each model trained on the AO-CHILDES. Compared to NoLimit and StaticLimit, which do not account for developmental changes in working memory, DynamicLimit-Linear and DynamicLimit-Exp, which simulate its gradual growth, achieve significantly higher overall performance. Among them, DynamicLimit-Exp attains the highest overall accuracy, supporting the effectiveness of a cognitively plausible mechanism. The comparable performance of StaticLimit to NoLimit suggests that the gradual introduction of working memory constraints throughout training is crucial, rather than their static application. These results indicate that DynamicLimit-Exp effectively replicates the CP effect observed in human L_1 acquisition.

The CP depends on the child's learning algorithm, not the input stimulus Table 2 presents the accuracy of models trained on Wikipedia, showing trends similar to those observed in Table 1, where the models were trained on AO-CHILDES. Specifically, DynamicLimit-Linear and DynamicLimit-Exp outperform NoLimit and StaticLimit in overall accuracy, with DynamicLimit-Exp achieving the highest performance, further supporting the efficacy of incorporating developmental working memory constraints. These findings suggest that the CP effect does not depend solely on exposure to specific linguistic stimuli (e.g., CDS) but rather on the learning algorithm itself, which mirrors human cognitive development. This result aligns with existing research (Feng et al., 2024; Padovani et al., 2025), which have reported that CDS is not uniquely valuable for training LMs. This finding suggests that our method is applicable to LLM pretraining, as they typically use non-CDS datasets such as Common Crawl and Wikipedia (Touvron et al., 2023a).

[7] While Zorro lacks naturalness, this can aid in isolating syntactic ability from lexical or semantic cues (Gulordava et al., 2018).

[8] See Appendix C for the full list of grammatical categories.

Model	**OVERALL**	D-N AGR	S-V AGR	ANA. AGR	ARG. STR	BINDING	CASE	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	LOCAL. ATR	QUANTIFIERS	NPI
AO-CHILDES														
DYNAMICLIMIT-EXP (↑)	62.2	50.8	50.0	49.6	67.7	58.7	95.2	43.1	93.6	52.2	53.6	51.3	57.6	85.0
DYNAMICLIMIT-EXP (↓)	56.5	49.9	49.7	50.1	44.7	61.9	70.6	73.3	72.0	51.8	61.9	47.0	48.1	54.1
Δ (↑, ↓)	**5.7***	**0.9**	**0.3**	-0.5	**23.0***	-3.2†	**24.6***	-30.1†	**21.6***	**0.4**	-8.3†	**4.4***	**9.5***	**30.8***
Wikipedia														
DYNAMICLIMIT-EXP (↑)	59.1	49.8	50.4	46.0	71.5	59.3	97.7	37.4	86.5	51.1	58.0	60.5	42.2	53.9
DYNAMICLIMIT-EXP (↓)	52.9	50.4	50.1	47.4	68.7	62.3	74.4	60.2	44.2	53.2	51.7	42.7	40.6	42.2
Δ (↑, ↓)	**6.1***	-0.6	**0.3**	-1.4	**2.9**	-3.0	**23.3***	-22.8†	**42.3***	-2.2	**6.3***	**17.8***	**1.7**	**11.7***

Table 3: Performance difference when changing the direction of the cognitive constraints in DYNAMICLIMIT-EXP. * and † indicate items where DYNAMICLIMIT-EXP (↓) performed significantly better or worse than DYNAMICLIMIT-EXP (↑), respectively (z-test for proportions, p < 0.05).

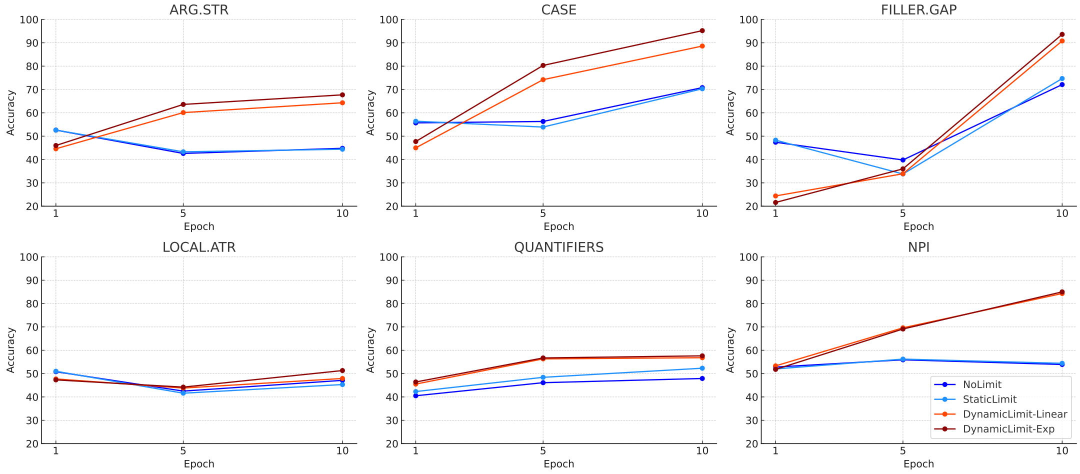


Figure 3: Accuracy trajectories over training epochs (1, 5, 10) for six grammatical categories that showed significant final-stage improvements with developmental constraints.

5 Analysis

5.1 Testing the "Less-is-more" Hypothesis with Reversed Cognitive Constraints

A key question arising from the results (§4) is whether DYNAMICLIMIT-EXP's superior performance stems from the "Less-is-more" hypothesis (Newport, 1990)—i.e., the gradual growth of working memory—or from unintended side effects. In other words, does the gradual *change* in working memory enhance information capacity, dynamically shifting the model's focus across epochs and ultimately aiding rule generalization? To test this, we introduce a cognitively *implausible* language model, referred to as "DYNAMICLIMIT-EXP (↓)", which shares the same slope trajectory as our proposed DYNAMICLIMIT-EXP (↑) [9] but with its direction reversed, such that working memory capacity decreases over time. Specifically, DynamicLimit-Exp (↑) is set to $m_0 = 1.0, r = 0.6$ (the same setting as in §4), while DynamicLimit-Exp (↓) is set to $m_0 = 0.01, r = 1.668$ to achieve a nearly symmetrical curve.[10]

Table 3 provides evidence supporting the Less-is-more hypothesis, as DYNAMICLIMIT-EXP (↑) consistently outperformed the cognitively implausible DYNAMICLIMIT-EXP (↓). The observed performance gap, particularly in grammatical items requiring both local and non-local dependencies (e.g., CASE, ARG. STR, and FILLER-GAP), suggests that the gradual growth of working memory is crucial for grammatical learning and generalization, as it enables the early extraction of basic patterns followed by the progressive acquisition of complex rules. These findings indicate that the superior performance of DYNAMICLIMIT-EXP (↑) is primarily driven by the developmental trajectory of working memory growth rather than unintended side effects of dynamic shifts in memory focus.

Incidentally, from the series of experimental re-

[9] This section adopts this notation for simplicity.

[10] Since setting the initial slope $m_0 = 0.0$ prevents w_t from being updated in Equation (2), we set it this way for computational reasons.

sults, along with those in §4 (Table 1 and 2), NoLimit and DynamicLimit-Exp (↓) consistently outperform DynamicLimit-Exp (↑) in ELLIPSIS, as exemplified by the following cases:

(2) a. Mark fixed one **worn** canal, and Roger fixed more.

b. *Mark fixed one canal, and Roger fixed more **worn**.

Since resolving ELLIPSIS involves maintaining long-range dependencies, DynamicLimit-Exp (↑) may struggle due to its initial memory constraints. This suggests that grammatical items like ELLIPSIS require substantial memory from the early stages of training, and thus, our proposed method may not be optimal for learning such structures. Alternative workarounds, such as dynamically adjusting memory allocation or hybrid approaches, may be necessary to address this limitation.

5.2 Tracking Developmental Gains Across Training

To more directly support our claim that developmentally guided learning simulates a CP, we examine how model performance unfolds over time—not just at the endpoint but at intermediate stages as well. This addresses the need for stage-by-stage comparisons raised by prior evaluations.

Figure 3 tracks accuracy at Epochs 1, 5, and 10 for six grammatical categories selected based on statistically significant improvements observed in Table 1. At the early stage (Epoch 1), models with larger or fixed memory (NoLimit, StaticLimit) perform better. However, the developmentally constrained model, DynamicLimit-Exp, shows steady gains over time, ultimately surpassing these baselines by Epoch 10 in multiple categories. The improvement is especially pronounced in CASE and FILLER.GAP, highlighting a pattern of late-stage acceleration. These results suggest that incrementally increasing memory capacity over training acts as a beneficial inductive bias, enabling the model to generalize more effectively from limited early experience—consistent with the hypothesized role of a CP in human language acquisition.

5.3 Learning to Represent: The Cognitive Effect of Memory Expansion

We analyze representational change by tracking embedding diversity within epochs and shifts between epochs. Specifically, we consider two complementary aspects of representational change: (i) the diversity or dispersion of embeddings *within* each epoch, which reflects the isotropy and expressiveness of the representation space at a given time; and (ii) the amount of shift in embeddings *between* epochs, which captures the degree of representational update and learning progress over time.

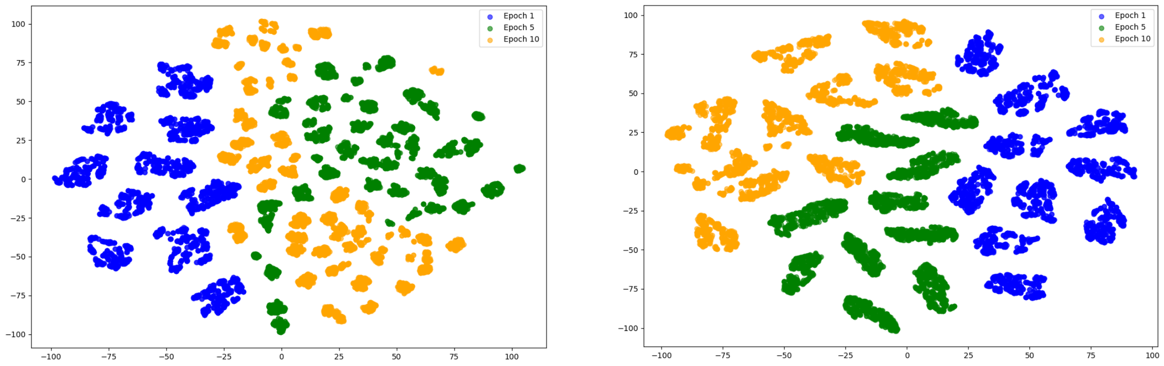
	Entropy			Mean Distance		
Epoch	1	5	10	1-5	5-10	1-10
NoLimit	5.36	5.17	5.19	91.30	28.50	66.28
DynamicLimit-Exp	5.40	5.30	5.39	69.25	70.63	101.92

Table 4: Embedded space analysis of NoLimit and DynamicLimit-Exp at each stage: distribution diversity and distribution distance.

Figure 4 visualizes t-SNE projected embeddings for the FILLER.GAP category, where DynamicLimit-Exp showed clear gains over baselines (§4.3, §5.1). In the NoLimit model (Figure 4a), embedding clusters expand from Epoch 1 to 5 but subsequently contract and overlap by Epoch 10. This pattern reflects a reduction in within-epoch diversity (i.e., lower isotropy) and minimal between-epoch shift, indicating that the representations stagnate and fail to evolve structurally as training progresses. In contrast, DynamicLimit-Exp (Figure 4b) produces more structured trajectories: clusters remain well-separated within epochs and continue to shift meaningfully across epochs. This suggests not only sustained representational plasticity but also a finer-grained encoding of syntactic distinctions over time.

To quantify these trends, Table 4 reports entropy (capturing embedding dispersion) and mean Euclidean distance between clusters (capturing separation).[11] The NoLimit model shows a drop in entropy and a plateau in inter-cluster distance after Epoch 5, consistent with representational collapse. Meanwhile, DynamicLimit-Exp maintains higher entropy and exhibits a steady increase in distance, consistent with ongoing structural refinement. These findings indicate that developmentally guided memory expansion helps preserve expressive, isotropic, and well-separated embedding spaces—properties that support better generalization in language models (Diehl Martinez et al.,

[11] The Appendix D shows how to calculate each measure.



(a) NOLIMIT

(b) DYNAMICLIMIT-EXP

Figure 4: Embedded space at each learning stage for NOLIMIT and DYNAMICLIMIT-EXP (FILLER. GAP)

Dataset	NOLIMIT	DYNAMICLIMIT-EXP
[5,10]	**47.2**	46.8
[11,50]	47.0	**58.7***
[51,100]	40.6	**42.5***
[101, 150]	37.3	**40.8***

Table 5: Accuracy in Zorro when the length of the sentence is changed. * indicates a statistically significant differences ($p < 0.05$).

2024).[12]

5.4 Influence of Input Stimulus Length

We analyze how sentence length affects the performance of NOLIMIT and DYNAMICLIMIT-EXP. To assess their adaptability, we created four Wikipedia-based datasets, each with 500,000 sentences in length ranges: [5,10], [11,50], [51,100], and [101,150].

The results in Table 5 reveal notable differences in model performance. For shorter sentences in the [5,10] range, NOLIMIT achieves slightly higher accuracy compared to DYNAMICLIMIT-EXP. However, in the [11,50] range, DYNAMICLIMIT-EXP significantly outperforms NOLIMIT, achieving 58.7 compared to 47.0. This suggests that DYNAMICLIMIT-EXP excels at handling moderately long sentences, likely due to its ability to dynamically adjust working memory. For longer sentences in the [51,100] and [101,150] ranges, DYNAMICLIMIT-EXP consistently outperforms NOLIMIT. These findings highlight the benefits of dynamic working memory expansion in facilitating rule generalization and contextual adaptation across diverse sentence lengths. While NOLIMIT exhibits competitive performance on short sentences, its stagnation on longer sentences underscores its limited ability to generalize complex patterns. Conversely, DYNAMICLIMIT-EXP's consistent performance across varying sentence lengths supports its suitability for grammatical items requiring the processing of both short and long contexts.

6 Conclusion

This study proposed a method for integrating the developmental trajectory of human working memory into the training process of LMs, inspired by the *Less-is-More* hypothesis. The proposed method, DYNAMICLIMIT-EXP, initially restricts working memory and gradually relaxes it exponentially during training. Experiments on both AO-CHILDES and Wikipedia showed that DYNAMICLIMIT-EXP improves grammatical learning efficiency compared to conventional methods without memory constraints or with static memory constraints. These findings suggest a promising direction for building more data-efficient LMs by leveraging cognitively inspired inductive biases.

Beyond its engineering implications, this study also offers insight into the cognitive mechanisms underlying L_1 acquisition. While our results do not directly demonstrate that working memory development is necessary for human learners, they serve as a computational-level plausibility test (Marr, 1982), showing that the hypothesized link between cognitive constraints and rule learning, central to the *Less-is-More* hypothesis, can be instantiated in artificial learners. Combining the observed learning efficiency gains and the cognitive plausibility of our approach, we support the hypothesis-generating idea that such developmental constraints may plausibly aid human language acquisition as well.

[12] Similar trends were found for CASE; see Appendix E.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. We are also grateful to Akiyo Fukatsu for her insightful comments and the Computational Models of Language (CoLa) Reading Group at UC Irvine for their constructive feedback. This work was supported by JSPS KAKENHI Grant Number 24H00087, JST PRESTO Grant Number JPMJPR21C2, Grant-in-Aid for JSPS Fellows JP24KJ0800, JST SPRING Grant Number JPMJSP2108.

Limitations

Scalability. One limitation of this study is the constrained scale of the experimental setup. The primary goal of this study is to computationally replicate the CP in L_1 acquisition, as discussed in cognitive science (Lenneberg, 1967; Fromkin et al., 1974; Curtiss, 1977; Johnson and Newport, 1989). Following previous studies (Huebner et al., 2021; Haga et al., 2024), we designed the experiment to be as ecologically valid as possible by training an SLM using CDS. While this controlled setting allows for a more precise analysis and simulation of the Less-is-More hypothesis, it remains unclear how our findings contribute to the data efficiency of LLMs. The experimental results with Wikipedia (Table 2, 3, 5) provide a promising outlook in this direction, but further investigation with larger models and datasets is necessary to determine the effectiveness and limitations of the proposed approach.

Language. In this experiment, we investigated the replication of the CP effect in L_1 acquisition using English. However, since the CP effect is observed across various languages (Patkowski, 1980; Johnson and Newport, 1989), it remains to be tested whether the proposed approach is effective in multilingual environments. To our knowledge, there is currently no targeted syntactic evaluation specifically designed for CDS across different languages, such as Zorro. Zorro was developed based on BLiMP (Warstadt et al., 2020), an adult-oriented targeted syntactic evaluation for English, and recent studies have proposed multilingual versions of BLiMP (e.g., JBLiMP (Someya and Oseki, 2023) for Japanese and CLiMP (Xiang et al., 2021) for Chinese). Therefore, developing CDS-specific versions based on these multilingual BLiMPs could help address this limitation.

References

John R. Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Morten H. Christiansen and Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell.

Christian Clark, Byung-Doh Oh, and William Schuler. 2025. Linear recency bias during training improves transformers' fit to reading times. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7735–7747, Abu Dhabi, UAE. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Nelson Cowan, Lara Nugent, Emily M. Elliott, Igor Ponomarev, and John Scott Saults. 1999. The role of attention in the development of short-term memory: age differences in the verbal span of apprehension. *Child development*, 70 5:1082–97.

S. Curtiss. 1977. *Genie: A Psycholinguistic Study of a Modern-day "wild Child"*. Mathematics in Science and Engineering. Academic Press.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5999–6011, Miami, Florida, USA. Association for Computational Linguistics.

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

Andrew W. Ellis and Matthew A. Lambon Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1103–1123.

Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press.

Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. 2024. Language is primarily a tool for communication rather than thought. *Nature*, 630:575–586.

Steven Y. Feng, Noah Goodman, and Michael Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Victoria Fromkin, Stephen Krashen, Susan Curtiss, David Rigler, and Marilyn Rigler. 1974. The development of language in genie: a case of language acquisition beyond the "critical period". *Brain and Language*, 1(1):81–107.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Preprint*, arXiv:2501.17047.

S. E. Gathercole, S. J. Pickering, B. Ambridge, and H. Wearing. 2004. The structure of working memory from 4 to 15 years of age. *Developmental psychology*, 40(2):177–190. Gathercole, Susan E Pickering, Susan J Ambridge, Benjamin Wearing, Hannah 2004/2/26.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Akari Haga, Saku Sugawara, Akiyo Fukatsu, Miyu Oba, Hiroki Ouchi, Taro Watanabe, and Yohei Oseki. 2024. Modeling overregularization in children with small language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14532–14550, Bangkok, Thailand. Association for Computational Linguistics.

Joshua K. Hartshorne, Joshua B. Tenenbaum, and Steven Pinker. 2018. A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177:263–277.

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. Association for Computational Linguistics.

Philip A. Huebner and Jon A. Willits. 2021. *Using lexical context to discover the noun category: Younger children have it easier*, pages 279–331. Psychology of Learning and Motivation - Advances in Research and Theory. Academic Press Inc.

T. Florian Jaeger and Harry Tily. 2011. On language 'utility': processing complexity and communicative efficiency. *WIREs Cognitive Science*, 2(3):323–335.

Jacqueline S Johnson and Elissa L Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, 21(1):60–99.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

E.H. Lenneberg. 1967. *Biological Foundations of Language*. Wiley.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Beatriz Luna, Krista E. Garver, Trinity A. Urban, Nicole A. Lazar, and John A. Sweeney. 2004. Maturation of cognitive processes from late childhood to adulthood. *Child Development*, 75(5):1357–1372.

Brian Macwhinney. 2000. The childes project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., USA.

Rachel I. Mayberry and Susan D. Fischer. 1989. Looking through phonological shape to lexical meaning: The bottleneck of non-native sign language processing. *Memory & Cognition*, 17(6):740–754.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Elissa L. Newport. 1990. Maturational constraints on language learning. *Cognitive Science*, 14(1).

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. *Preprint*, arXiv:2505.23689.

Mark S. Patkowski. 1980. The sensitive period for the acquisition of syntax in a second language. *Language Learning*, 30(2):449–468.

Wilder Penfield. 1965. Conditioning the uncommitted cortex for language learning. *Brain*, 88(4):787–798.

Steven Pinker. 1994. *The Language Instinct: How the Mind Creates Language*. William Morrow and Company.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Mark S. Seidenberg and Jason D. Zevin. 2006. Connectionist models in developmental cognitive neuroscience: Critical periods and the paradox of success. In Yuko Munakata and Mark H Johnson, editors, *Processes of Change in Brain and Cognitive Development*, pages 585–612. Oxford University Press.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Elizabeth R. Sowell, Doris A. Trauner, Anthony Collins Gamst, and Terry L. Jernigan. 2002. Development of cortical and subcortical brain structures in childhood and adolescence: a structural mri study. *Developmental Medicine & Child Neurology*, 44.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 Ryan Cotterell. 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*, pages 1–34. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

George K. Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley.

A Details of the Training Configuration for the Base Models

Table 6 shows the training settings of the base model. For the experiment, a single NVIDIA RTX A5000 (24GB) GPU was used, and the training time for each run was approximately one hour.

Hyperparameter	Value
Model Architecture	GPT-2
Number of Layers	4
Number of Attention Heads	4
Embedding Dimension	256
Dropout Rate	0.1
Learning Rate (η)	5×10^{-6}
Weight Decay	0.01
Batch Size	512
Gradient Accumulation Steps	2
Total Epochs	20
Maximum Sequence Length	32
Learning Rate Scheduler	Cosine with Restarts
Warm-up Steps	10% of Total Steps
Optimizer	AdamW
Optimizer Parameters	$\beta = (0.9, 0.999)$, $\epsilon = 1e`08$
Tokenizer	Trained on CHILDES
Early Stopping Tolerance	1 Epoch
Evaluation Metric	Perplexity

Table 6: Training Configuration (Hyperparameters) for the GPT-2 Model.

B Distribution of the Datasets

Figure 5 shows the sentence length distribution for the AO-CHILDES and Wikipedia datasets used in this experiment. As can be seen from the figure, AO-CHILDES, by its nature, contains more short sentences than Wikipedia.

C Details of Grammatical Items in Zorro

Table 8 shows the full list of grammatical categories in Zorro. Examples are taken from Table 5 in the original paper (Huebner et al., 2021).

D Analysis of Distributional Changes in t-SNE Space Across Training Epochs

This section explains in detail the analysis of the entropy and average distance of embeddings projected into the t-SNE space for different learning epochs.

D.1 Entropy Calculation

To quantify the distribution of embeddings, a 2D histogram is constructed using a fixed grid (50×50 bins). The probability distribution P is obtained by normalizing the histogram. The entropy is then computed as:

$$H(P) = -\sum_i P_i \log P_i, \tag{4}$$

where P_i is the probability of each bin. Higher entropy suggests a more uniform distribution, whereas lower entropy indicates clustering.

	Entropy			Mean Distance		
Epoch	1	5	10	1-5	5-10	1-10
NoLimit	5.30	5.23	5.30	75.47	12.26	87.62
DynamicLimit-Exp	5.29	5.30	5.34	59.91	37.68	97.59

Table 7: Embedded space analysis of NoLimit and DynamicLimit-Exp at each stage: cluster expansion, distribution diversity, and distribution distance.

D.2 Mean Distance Between Epochs

To analyze shifts in embedding distributions across epochs, we compute the Euclidean distance between the mean embedding vectors of different epochs:

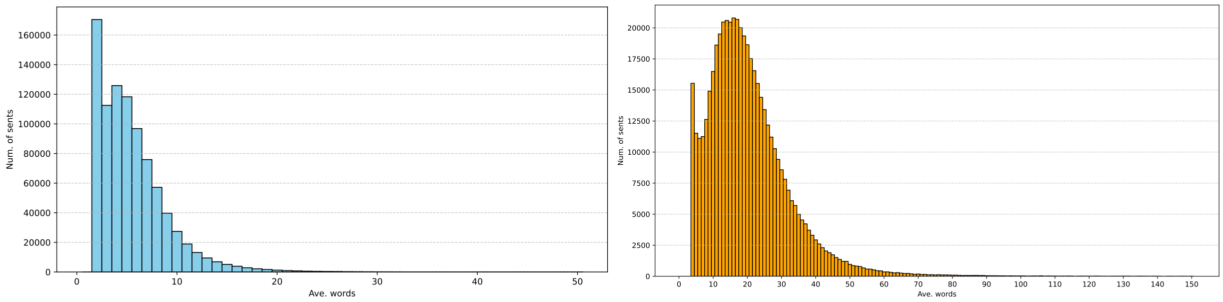
$$D(X, Y) = \|\mu_X - \mu_Y\|, \tag{5}$$

where μ_X and μ_Y are the mean vectors at different epochs. Larger distances imply greater shifts in the learned representation.

E Development of Feature Extraction Capabilities in CASE

Figure 6 visualizes the clustering structure of final layer embeddings using t-SNE for CASE. The embedding space visualizations reveal distinct patterns between NoLimit and DynamicLimit-Exp across training epochs. In NoLimit, the embedding clusters expand between Epoch 1 and Epoch 5 but contract significantly by Epoch 10, suggesting stagnation in representation learning. In contrast, DynamicLimit-Exp maintains structured evolution throughout training, with well-separated clusters that reflect progressive refinement.

Table 7 shows the embedded space analysis. Regarding **entropy**, NoLimit shows a slight decrease over time (Epoch 1 → Epoch 5), reflecting reduced distribution diversity as training progresses. In contrast, DynamicLimit-Exp maintains or slightly increases entropy, suggesting a balanced emphasis on both basic patterns and diverse features, even in later training stages. For **mean Euclidean distances** between clusters, NoLimit exhibits large distances between Epoch 1 and Epoch 5 but demonstrates minimal evolution



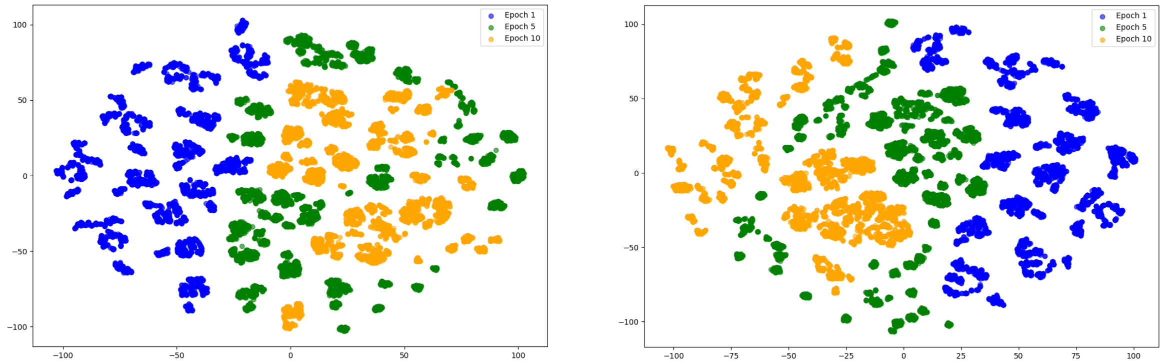
(a) AO-CHILDES

(b) Wikipedia

Figure 5: Word distribution of the AO-CHILDES and Wikipedia datasets used in the experiment

Category	Subcategory	Acceptable Sentence	Unacceptable Sentence
D-N AGR	noun-across_1_adjective noun-between_neighbors	*look at this purple **thing** .* *this **color** must be white .*	*look at this purple **things** .* *this **colors** must be white .*
S-V AGR	verb-across_prepositional_phrase verb-across_relative_clause verb-in_question_with_aux verb-in_simple_question	*the **lie** on the foot is flat .* *the **book** that i like is poor .* *where does the **horse** go ?* *where is the **way** ?*	*the **lies** on the foot is flat .* *the **books** that i like is poor .* *where does the **horses** go ?* *where is the **ways** ?*
ANA.AGR	pronoun_gender	*will Mark want **himself** ?*	*will Mark want **herself** ?*
ARG.STR	dropped_argument swapped_arguments transitive	***give me the poor boat** .* ***he made the slave her label** .* *Philip **thinks** .*	***the poor boat gives me** .* ***the slave made her label he** .* *Philip **affected** .*
BINDING	principle_a	*Ben thinks about himself **calling** this fuel .*	*Ben thinks about himself **called** this fuel .*
CASE	subjective_pronoun	***i brought the wolf** my hill .*	***the wolf brought i** my hill .*
ELLIPSIS	n_bar	*Mark fixed one **worn** canal and Roger fixed more .*	*Mark fixed one canal and Roger fixed more **worn** .*
FILLER.GAP	wh_question_object wh_question_subject	*Laura married the dinner **that the wolf could close** .* *Laura ended the finger **that** can make boats .*	*Laura married **what** the dinner **could close the wolf** .* *Laura ended **who** the finger can make boats .*
IRREGULAR	verb	*Michael **chose** the good one some time ago .*	*Michael **chosen** the good one some time ago .*
ISLAND	adjunct_island coordinate_structure_constraint	*who should William have **without watching the baby** ?* *who must Philip **and the dinosaur turn** ?*	*who should William have **the baby without watching** ?* *who must Philip **turn and the dinosaur** ?*
LOCAL.ATR	in_question_with_aux	*is the whale **getting** the person ?*	*is the whale **gets** the person ?*
NPI	matrix_question only_npi_licensor	***does her boat ever play with the growth ?*** ***only Mark ever finds some suit .***	***her boat does ever play with the growth ?*** ***even Mark ever finds some suit .***
QUANTIFIERS	existential_there superlative	*there are **many** books about soft birds .* *no pig could stand on top of **more than** six days .*	*there are **most** books about soft birds .* *no pig could stand on top of **at least** six days .*

Table 8: Explanation of each grammatical category in Zorro.



(a) NoLimit

(b) DynamicLimit-Exp

Figure 6: Embedded space at each learning stage for NoLimit and DynamicLimit-Exp (CASE)

between Epoch 5 and Epoch 10. This stagnation may highlight the model's failure to effectively generalize new rules. DynamicLimit-Exp, on the other hand, maintains substantial distances across epochs, indicating continuous embedding evolution and refinement throughout training.