Sleep-Consolidated Language Modeling

Peter Shamoun* pshamoun@ucsd.edu

Sai Sri Lasya Yadlapati* syadlapati@ucsd.edu

Tianqi Zhang* tiz019@ucsd.edu

Alex Warstadt awarstadt@ucsd.edu

Abstract

Standard language model training through repeated epochs over the same dataset is both computationally inefficient and cognitively implausible; humans don't learn by reading the same book ten times in a row, but rather consolidating difficult memories during sleep, a process that becomes less flexible as we age. This leads to high resource requirements and limitations when using language models as cognitive models. Existing work on cognitively-plausible and compute-efficient training focuses mostly on curriculum learning, while sleep-inspired consolidation mechanisms normally target catastrophic forgetting for continual learning. In this paper, we propose improving data efficiency in language model pretraining via. sleep-consolidated learning, which replaces the standard n-epoch training loop with a single pass over the data, interspersed with *sleep phases* where the model replays only the samples it struggled with the most. We train *RoBERTa-PreLayerNorm* models on the 100M-token BabyLM corpus and evaluate whether this biologically-motivated approach can match multi-epoch baseline performance, while using significantly less data. We also conduct specific comparisons between different replay criteria and other ablation studies to further examine the effect of every step in the sleep mechanism.

Code: https://github.com/Peter-Shamoun/CLIMB-Sleep

Website: https://lasyayadlapati.github.io/sleep-consolidated-learning

1 Introduction

Language Models (LMs), while powerful, are extremely resource-intensive to train. Effective LMs require hundreds of times more data to build an understanding of language than human children, often more text than a human will ever be exposed to in their entire lifetime. For example, ELMo and BERT were both trained on billions of words, RoBERTa was trained on 30 billion words, and Meta AI's Llama 4 was trained on around 30 trillion words (Warstadt and Bowman 2024; Aaron Grattafiori et al. 2024). In comparison, Hart and Risley (1992) estimate that human children are exposed to about 3 million to 11 million words per year. That means in order to acquire and use language in a meaningful way, LMs see hundreds to millions of years' worth of linguistic content (Warstadt and Bowman 2024). The resource requirements to build an effective language model from random initialization (also known as pre-training) restrict language model research to larger organizations with lots of funding, making academic research of LM pre-training difficult and costly.

Most neural language models are also trained on data over multiple epochs. This means that over the course of a training run, the model will see the same dataset anywhere from hundreds to thousands of times. This is entirely different from how humans acquire language: we live through every day once, and experience everything only once. In addition to not being cognitively plausible, epoch training also potentially wastes computational resources. Different samples of the data will have different difficulty levels that determine how easily the model can learn their patterns. Epoch training means that even after the model has mastered easier samples, it is still re-trained on those same samples in the same contexts as before, leading to overfitting and shortcut learning, where the model picks up on spurious relationships rather than learning real semantic content (Xiao, Hudson and Al Moubayed 2023). This disconnect between the training schedules of LMs and human language acquisition not only could contribute to their inefficiency, but also means that there are strong limitations in the use and interpretation of LMs as cognitive models.

Humans, however, do revisit experiences in a critical stage for cognitive development: sleep. Neuroscience research has shown that sleep is not just a period of rest, but critical for developing memories. While the waking brain is optimized for encoding new experiences into memory, during sleep, the brain undergoes a process called *memory consolidation*, where they are stabilized and integrated into pre-existing synaptic networks (Rasch and Born 2013). This two-stage process for memory formation assumes that, because long-term memory stores take longer to train and short-term memory is easily overwritten by new experiences, sleep provides an essential off-line environment, where recent experiences are revisited repeatedly to gradually integrate into long-term stores without overwriting older memories (Marr 1971). By consolidating abstract representations of our memories in sleeping periods, humans also retain world knowledge and episodic memory of recent events (declarative memory) as well as intuition and unconscious long-term memories that influence their behavior (non-declarative memory), both of which are important for learning complex skills such as language (Rasch and Born 2013).

2 Methods

2.1 Dataset

We use one of the BabyLM Challenge datasets, a curated corpus that is designed to mimic the linguistic input that children receive during early language acquisition. Specifically, we utilize the 100M-word text-only dataset, which roughly represents the amount of word tokens a child encounters by age 13 (Warstadt et al. 2023).

This dataset is made up of a combination of sources from specifically two domains, as shown in Table 1:

1. **Transcribed speech** (56%, ∼55M words): movie and video subtitles, dialogue transcripts, adult-child interactions, and phone conversations

2. **Child-directed language** (44%, ∼43M words): children's books, standard and simplified encyclopedia articles, and literary texts

This composition is meant to reflect the oral and written language input children naturally receive, with a majority coming from spoken or conversational sources to mirror how hearing children acquire language (Warstadt et al. 2023). Table 1 provides a more detailed breakdown of each source.

Table 1: Composition of the BabyLM 100M Word Dataset

Domain	**Source**	**Description**	**Words (M)**	**%**
Transcribed Speech	OpenSubtitles	Movie and TV subtitles	31.28	31%
	QED	Educational video subtitles	10.24	11%
	British National Corpus	Transcribed dialogue	8.16	8%
	CHILDES	Adult-child interactions	4.21	5%
	Switchboard Corpus	Telephone conversations	1.18	1%
	Subtotal		*55.07*	*56%*
Child-Directed Language	Simple Wikipedia	Simplified encyclopedia	14.66	15%
	Wikipedia	Standard encyclopedia	10.08	10%
	Children's Book Test	Children's books collection	5.55	6%
	Children's Stories Text Corpus	Selected children's stories	3.22	3%
	Standard Project Gutenberg	Literary texts	9.46	10%
	Subtotal		42.97	*44%*
Total			**98.04**	**100%**

Note: QED (QCRI Educational Domain Corpus), British National Corpus (only dialogue portions), CHILDES (Child Language Data Exchange System), Switchboard (Dialog Act Corpus telephone conversations).

2.2 Codebase

Due to similarities in goal and implementation, we build upon the code of Diehl Martinez et al. (2023)[1] for our training pipeline. However, we define a novel mechanism and implementation for sleep-consolidated learning.

2.3 Sleep-Consolidated Learning

We replaced the standard multi-epoch training with a single-pass, sleep-consolidated approach consisting of alternating *wake phases* and *sleep phases*, as shown in Figure 1. This is designed to replace the multi-epoch training paradigm of a traditional LM with a cognitively plausible replay system inspired by the process of *memory consolidation*.

Before training, a number of hyperparameters related to the sleep mechanism are assigned:

- *n_phases*: the number of sleep/wake cycles for this run
- *replay_ratio*: the ratio of data to be stored in the *replay buffer*
- *max_steps*: the maximum number of training steps for each phase
- *n_augmentations*: the number of recontextualizations to perform during the *sleep phase*
- *max_seq_length*: the length used to split samples during context-augmented padding

The full dataset is first split into *n_phases* folds. Then, the model goes through *n_phases* cycles of *wake phases* and *sleep phases*.

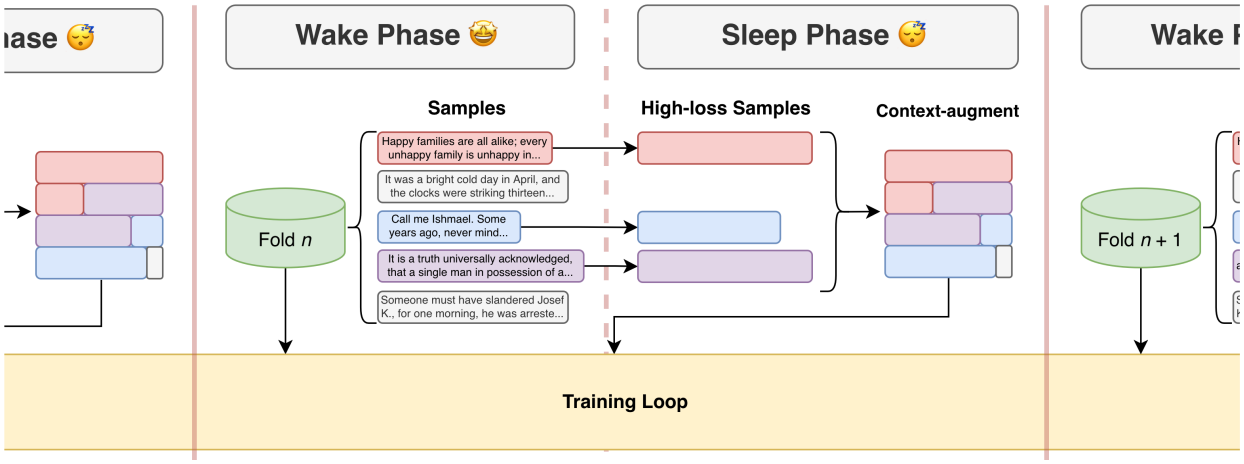

Figure 1: Visualization of one cycle of the sleep mechanism. This demonstrates the selection of high-loss samples and the context-augmented padding process.

[1] https://github.com/codebyzeb/CLIMB

2.3.1 Wake Phase

During each *wake phase*, the model processes a contiguous block of training data comprising of a single fold. It is trained for the masked language modeling task, with the binary cross-entropy loss function. During the *wake phase*, the loss for each sample is tracked and stored for use in the *sleep phase*.

Each *wake phase* ends once the model has seen every sample within the current fold once. If the model reaches *max_wake_steps* steps before this happens, then the *wake phase* ends early. This means that the model will see every sample within the dataset at most once during a *wake phase*.

2.3.2 Sleep Phase

Upon switching from a *wake phase* to a *sleep phase*, a proportion of samples are selected to add to a *replay buffer*, a store of particularly difficult samples. This selection is done by a weighted random sampling from all samples seen during the previous *wake phase*. Samples that had higher loss during the *wake phase* are more likely to be selected. The number of samples selected is determined by the *replay_ratio*.

The samples in this *replay buffer* are then put through context-augmented-padding, as seen in Figure 2 (Xiao, Hudson and Al Moubayed 2023). Samples are shuffled, then combined to form a contiguous block of tokens. This block is then re-split not on the original sample boundaries, but into even samples of length *max_seq_length*. This process effectively places tokens within new contexts, allowing the model to learn patterns more abstractly, as described in Xiao, Hudson and Al Moubayed (2023).

The data is recontextualized *n_augmentations* times, to form *n_augmentations* blocks of recontextualized data from the *replay buffer*. The model is then trained on this recontextualized data for *max_sleep_steps*.

Upon the end of the *sleep phase* and the start of the next *wake phase*, the replay buffer is emptied. This means that after each cycle, the data in the fold of the dataset within that cycle are never again seen by the model.

2.4 Baseline Model

We compare the performance of our Sleep-Consolidated model against a baseline model trained using a standard multi-epoch approach. Due to the nature of the *sleep phase*, it is difficult to measure the number of full passes through the dataset. To ensure a fair comparison of sample efficiency, we instead match the total number of training steps (e.g., 10,000 steps) rather than the number of epochs. This allows us to determine if the selection of data and data replay within the sleep phases allows the model to achieve lower loss rates and better generalization with equivalent gradient updates.

Similarly, we also train a baseline-like model, in which we set *sleep phase* hyperparameters

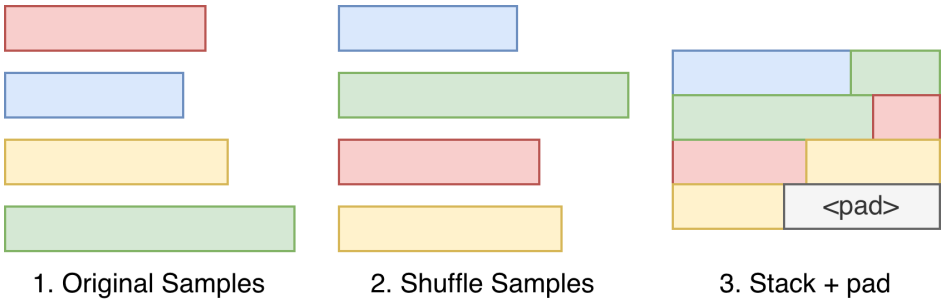

Figure 2: Context-augmented padding process. Samples are shuffled, stacked, truncated to the input length, then padded at the end, to fit samples from different contexts in with each other.

to values that equate to multi-epoch training. This results in one *wake phase* containing the entire dataset, then one long *sleep phase*, in which the data is shuffled without contextualization. The *replay_ratio* is also set to 1.0, so the entire dataset is trained on during the *sleep phase*.

We use a small *RoBERTa-PreLayerNorm* model (Liu et al. 2019) with 8 hidden layers, 8 attention heads, a hidden size of 256, and an intermediate size of 2048 as our model architecture for all experiments.

2.5 Replay Strategy Experiments

During the *sleep phase*, high-loss clips are selected for replay. To test the effect of the selection, we test three selection strategies:

1. Random Replay: Clips from the *wake phase* are selected uniformly at random for the *replay buffer*.
2. Weighted Replay: Clips from the *wake phase* are selected at random, with selection probability weighted by loss.
3. Strict Replay: Clips from the *wake phase* are selected deterministically, with the highest loss clips being stored in the *replay buffer*.

2.6 Hyperparameter Sweeps

In order to examine the effects of specific sleep-related hyperparameters as outlined in subsection 2.3, we perform a hyperparameter sweep over several values for *n_phases*, *max_sleep_steps*, and *replay_ratio*. In an initial bayesian hyperparameter sweep, these hyperparameters had the largest effect on model performance, so we selected them for a full grid-search to more thoroughly test their effects. For our target metric, we aim to minimize the mean perplexity on a hidden test set.

2.7 Evaluation

To validate the efficacy of sleep-consolidated learning, we evaluate our models on two primary fronts: linguistic acceptability and downstream task performance, while also employing specific metrics to analyze the internal dynamics of the sleep phases.

Linguistic Benchmarks. To assess general linguistic capabilities, we utilize the BabyLM evaluation pipeline . We report results on the Benchmark of Linguistic Minimal Pairs (BLiMP) to measure grammatical acceptability and syntactic acquisition without heavy fine-tuning (Warstadt et al. 2020). Additionally, we evaluate derivational morphology with wug-tests for both adjective nominalization and past-tense formation (Hofmann et al. 2025; Weissweiler et al. 2023), as well as entity state tracking (Kim and Schuster 2023).

Sleep-Specific Metrics. To specifically test the hypothesis that replaying difficult samples improves the learning of ”hard” concepts, we move beyond aggregate loss and track granular metrics during training. Specifically, we monitor the standard deviation of perplexity and the maximum perplexity across samples within batches. A decrease in the standard deviation of perplexity would indicate that the model is successfully consolidating difficult samples, narrowing the gap between ”easy” and ”hard” data points.

3 Results

3.1 Hyperparameter Sweeps

Figure 3 shows the importance of each hyperparameter for minimizing mean test perplexity as well as the correlation of each parameter with perplexity. The number of sleep phases (*n_phases*) is the most important hyperparameter by a large margin, and is weakly positively correlated with perplexity. This indicates that the model performs *worse* the more sleep phases there are.

Surprisingly, the max number of steps in the sleep phase is positively correlated with perplexity. This is counterintuitive to conventional understandings in multi-epoch training. More training steps generally leads to better performance up to a point, after which the model will begin to overfit.

Replay ratio and contextualization both do not display significant correlation with perplexity, which is also surprising. The *replay_ratio* governs how much data is included in the replay buffer, and contextualizing the data significantly impacts what the model sees during training. We would expect significant effects from both these methods. The low to zero correlation indicates that these factors actually do not impact model performance as much as we had thought.

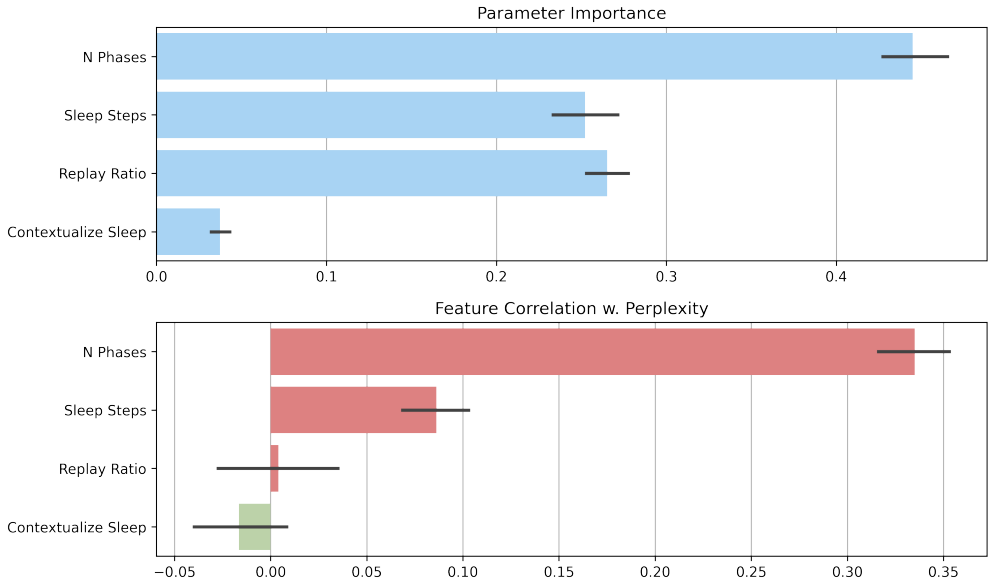

Figure 3: Importance and correlation of each hyperparameter with respect to mean test perplexity (loss). Error bars show bootstrapped 95% CIs.

3.2 Replay Strategy Experiments

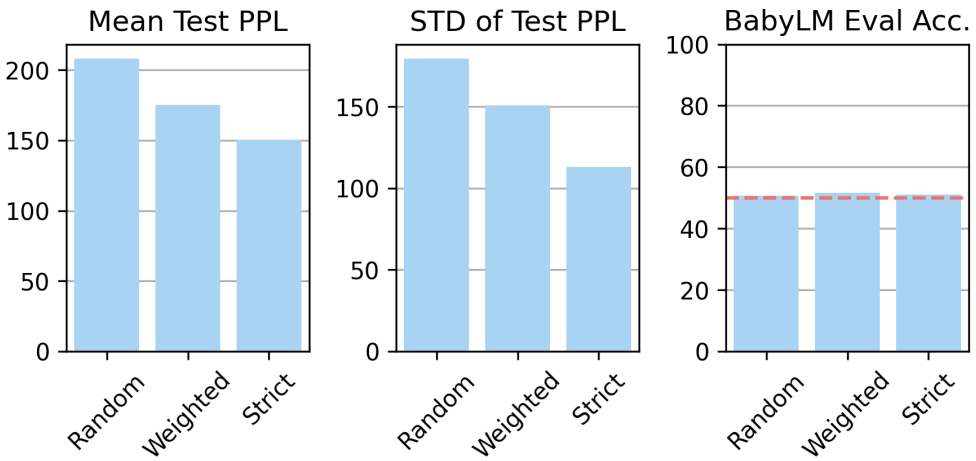

Figure 4: Results from the replay strategy experiments.

Figure 4 displays the evaluation results from the replay strategy experiments. For both mean and standard deviation of test perplexity, strict replay yields the best results, and some weighting by loss is better than uniformly randomly sampling. Looking at the accuracy on linguistic tasks, however, all three strategies seem to be relatively similar.

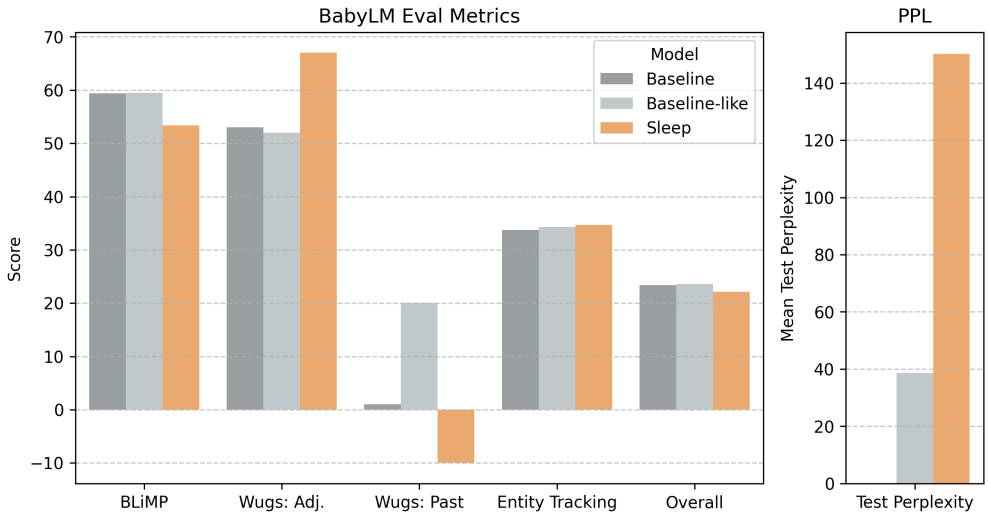

Figure 5: Comparison of evaluation metrics for baseline models and a sleep model.

3.3 Baseline Comparison

Figure 5 shows both BabyLM evaluation accuracy and perplexity for the two baseline models and a sleep model. Both baseline models show similar performance, meaning our hyperparameter configuration successfully mimics multi-epoch training. Comparing sleep and multi-epoch models, interestingly, performance is mixed across different tasks. While the sleep model falls behind in BLiMP accuracy, it outperforms significantly in adjective nominalization (indicating much more human-like behavior) and slightly outperforms the baseline in entity tracking. In a past-tense formation, the sleep model seems actually negatively correlated with humans, indicating the opposite behavior. In terms of perplexity, baseline models far outperform sleep models.

4 Conclusions

Overall, it is clear that *sleep models currently fall short of standard multi-epoch models*, both in terms of next-token prediction and linguistic understanding. However, our results still indicate that this new, cognitively-plausible training paradigm has promise. Firstly, in our hyperparameter sweep, several results were counter-intuitive, showing a divergence from multi-epoch training. Of particular interest is the *replay_ratio*, as our results from the replay strategy experiments show that training a language model on a smaller proportion of the highest-loss clips leads to improved performance. This supports our hypothesis that in a multi-epoch training schedule, steps are wasted by training over easier samples. However, our hyperparameter sweep results seem to suggest a more complex relationship with perplexity loss, as while correlation was low, importance was high (Figure 3).

Additionally, in terms of linguistic understanding, results are inconclusive. While sleep-like training improved performance on some tasks, multi-epoch models fared better on

others. This may have to do with the model overfitting on data in the replay buffer. If samples containing examples of adjective nominalization are more likely to be selected, then naturally, the sleep model would perform better on that task. Conversely, if high-loss samples with past-tense formations happened to exhibit non-native behaviors, this would lead to the negative correlation seen in Figure 5.

5 Limitations and Future Work

We have several hypotheses as to why performance of sleep models falls so short of baselines. For one, the model could be overwriting information learned in past folds with new data from a later fold. This would lead to *catastrophic forgetting*, causing the model to become worse at generalizing to unseen data and explaining the high test perplexities. To fix this, we could take inspiration from another process that happens during sleep as a part of *memory consolidation*: *Synaptic Homeostasis*. During waking hours, synapses are constantly strengthening and firing strongly, making it easier to learn and encode new information. However, this comes with a high resource cost; stronger synaptic connections need more energy and place extra stress on neurons (Tononi and Cirelli 2014). Additionally, neurons with stronger synaptic connections may fire for random chance, effectively wasting resources (Balduzzi and Tononi 2013). Thus, during sleep, important synapses with strong signals are strengthened and consolidated, while less important connections are reduced or even pruned (Tononi and Cirelli 2014). In the future, we will attempt to implement such a mechanism into the sleep mechanism by freezing weights in the model and manipulating learning rates. This would simulate reduced plasticity and serve to ”lock in” information learned in earlier folds of the dataset.

Another explanation could be in the selection of samples for the replay buffer. Currently, we use model training loss to measure difficulty, and assume that higher difficulty samples will be most helpful for learning. However, it's possible that the highest-loss samples contain random noise from the dataset that hinders learning rather than help it. Future studies could explore more nuanced ways to measure how helpful a sample would be for learning.

Multi-epoch learning, while effective for LM training, is not cognitively-plausible or comparable to human learning. While our project has so far been unsuccessful in providing a cognitively-plausible, sleep-inspired alternative, it has produced promising results. We hope that in continuing this project, we may develop such an alternative and forward the democratization of linguistic research.

References

Aaron Grattafiori et al.. 2024. "The Llama 3 Herd of Models." [Link]

Balduzzi, D., and G. Tononi. 2013. "What can neurons do for their brain? Communicate selectivity with bursts." *Theory Biosci* 132 (1): 27–39. [Link]

Diehl Martinez, Richard, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. "CLIMB – Curriculum Learning for Infant-inspired Model Building." In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Singapore Association for Computational Linguistics. [Link]

Hart, Betty, and Todd R. Risley. 1992. "American Parenting of Language-Learning Children: Persisting Differences in Family-Child Interactions Observed in Natural Home Environments.." *Developmental Psychology* 28: 1096–1105. [Link]

Hofmann, Valentin, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. "Derivational morphology reveals analogical generalization in large language models." *Proceedings of the National Academy of Sciences* 122 (19), p. e2423232122. [Link]

Kim, Najoung, and Sebastian Schuster. 2023. "Entity Tracking in Language Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada Association for Computational Linguistics. [Link]

Liu, Yinhan, 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." [Link]

Marr, D. 1971. "Simple memory: a theory for archicortex." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 262 (841): 23–81. [Link]

Rasch, Björn, and Jan Born. 2013. "About Sleep's Role in Memory." *Physiological Reviews* 93 (2): 681–766. [Link]

Tononi, Giulio, and Chiara Cirelli. 2014. "Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration." *Neuron* 81 (1): 12–34. [Link]

Warstadt, Alex, and Samuel R. Bowman. 2024. "What Artificial Neural Networks Can Tell Us About Human Language Acquisition." [Link]

Warstadt, Alex, 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*. Association for Computational Linguistics. [Link]

Warstadt, Alex, 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. [Link]

Weissweiler, Leonie, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. "Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore Association for Computational Linguistics. [Link]

Xiao, Chenghao, G Thomas Hudson, and Noura Al Moubayed. 2023. "Towards more Human-like Language Models based on Contextualizer Pretraining Strategy." In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Singapore Association for Computational Linguistics. [Link]