CogMemLM: Human-Like Memory Mechanisms Improve Performance and Cognitive Plausibility of LLMs

Lukas Thoma[⋆,○,•], **Ivonne Weyers**[○], **Erion Çano**[⋆], **Stefan Schweter**[◇], **Jutta L. Mueller**[○], **Benjamin Roth**[⋆,△]

[⋆]Faculty of Computer Science, University of Vienna, Vienna, Austria
[○]Department of Linguistics, University of Vienna, Vienna, Austria
[•]UniVie Doctoral School Computer Science, Vienna, Austria
[△]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria
{lukas.thoma, ivonne.weyers, erion.cano, jutta.mueller, benjamin.roth}@univie.ac.at
[◇]schweter.ml, stefan@schweter.eu

1 Introduction

Current large language models (LLMs) demonstrate impressive NLP performance, but they require massive amounts of training data. RoBERTa (Liu et al., 2019), for instance, sees 30 billion words during pre-training, which amounts to roughly 300x as many words as a human child hears until the age of 12 (Warstadt and Bowman, 2022). It is one of the explicit aims of the BabyLM challenge (Warstadt et al., 2023) to address this issue by training models on developmentally-plausible quantities and types of data (for similar approaches, see Hosseini et al., 2022; Huebner et al., 2021), in order to ultimately develop more cognitively plausible models that can inform research into human language acquisition (Keller, 2010; Dupoux, 2018).

In the present contribution to the BabyLM STRICT track, we take a threefold approach: firstly, we implement a simple curriculum learning approach and split the provided BabyLM dataset into four sub-datasets by increasing complexity, to broadly structure the data such that it better reflects what kind of input is available to infants and children throughout development (see 2.1). Secondly, we simulate a memory-based vocabulary learning inspired by psycholinguistic work (Perruchet and Vinter, 1998). Starting with a set of single characters, larger linguistics units (sub-words, words, and multi-words) are created based on the core memory mechanisms *activation* and *forgetting*. Possible units are limited in size, imitating working-memory constraints, but become larger across development (see 2.2). Thirdly, we implement redundant text representations to make the compositional aspect of language more salient: The lexicons that emerge from our curriculum learning steps, respectively, shape the (token) encoding of the given input text (see 2.3).

We pre-trained a RoBERTa-base architecture with masked language modeling and our CogMemLM-s model achieves improved results compared to the BabyLM RoBERTa baseline model in 27 out of 39 evaluation tasks. Although the so far integrated mechanisms have been implemented in a simplified form with regard to cognitive plausibility, it is intriguing that our pre-training method already improved performance considerably.

2 Methodology

2.1 Curriculum Learning

Child-directed speech typically consists of shorter and less syntactically complex sentences, more repetitions and limited vocabulary compared to adult-directed speech (Foushee et al., 2016; Kirchhoff and Schimmel, 2005). As the child's language competence increases, the linguistic input received from the environment becomes both more complex and diverse (Kunert et al., 2011). In an attempt to reflect this trajectory, we subdivided the provided 98M word corpus into four approximately equally-sized datasets of increasing linguistic complexity and lexical diversity (for details see Appendix A Table 1). The division was based mainly on the domains which the original corpora stem from and a subjective rating of their linguistic complexity and diversity; i.e. Dataset 1 (least complex) included materials mainly from child-speech contexts, whereas Dataset 4 (most complex) comprised the Wikipedia and Written English corpora. Although this split is rather coarse, it is only a first attempt at a curriculum learning approach, which may be followed-up by more fine-grained analyses and sub-divisions of the available materials.

2.2 Lexicon Creation

Because of computational and memory limitations in humans, any type of input, including language input, has to be "chunked" into units that can be stored and further manipulated (Archibald, 2017;

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 180–185
December 6-7, 2023 ©2023 Association for Computational Linguistics

Baddeley, 2003). For infants, the additional challenge consists in learning to chunk the perceived language input such that the resulting memorized chunks align with word boundaries, which allows for words to be stored in and retrieved from the lexicon. Inspired by the PARSER model for word segmentation (Perruchet and Vinter, 1998), we used a memory-based, variable parsing algorithm for lexicon creation. We start with a set of single characters and from these, larger linguistics units (sub-words, words, and multi-words) are created based on the core memory mechanisms *activation* and *forgetting*. The text data is processed sentence by sentence. Sentences are split into linguistic sub-units (percepts), which vary in size (see Appendix A). If a percept already exists, its activation value is increased by 1, strengthening its representation, if not, an entry is created and receives an activation of 1. After each processed sentence, forgetting is applied by subtracting 1/1000 from all activations. Any percept that is not re-activated within 1000 sentences (activation = 0) is removed from the lexicon. In curriculum 1 (C 1), lexicon creation starts with an empty lexicon, C 2 builds upon the lexicon of C 1 and so on. A 10 % sample of each data set was processed to create the lexicons which resulted in the following number of percepts: 13,444 after C 1, 22,740 after C 2, 25,887 after C 3 and 39,126 after C 4. We used the lexicon information to roughly dimension the vocabulary size of the respective curriculum tokenizers (see A.3) and to re-represent the training data for the perception shaping (see 2.3).

2.3 Perception Shaping

The BabyLM dataset was given in three different representations during pre-training: original text, coarse re-representation, and fine re-representation. For the coarse re-representation, text was processed left-to-right and the lexicon was searched for the longest fitting percept. Following this percept, an additional whitespace was added. For the fine re-representation, the identified percepts were split up further based on smaller units in the lexicon. The representation with the highest activation on average was used to split the coarse percept. Again, whitespaces were added after identified percepts. In the final step, whitespaces were normalized (multiple spaces to one). Usually, an existing token for e.g., the word "ended" would always be encoded with the corresponding token ID. In our training, however, linguistic units would also be encoded in two alternative representations, which increases the likelihood of "ended" also being encoded as "end" and "ed".

3 Results and Conclusion

Building on psycholinguistic work on memory-based word learning, we simulated lexicon creation given the BabyLM dataset as input. We used this information in a four step curriculum learning approach to guide the encoding of text, thereby increasing the cognitive plausibility in the following aspects: the language acquisition trajectory is reflected in the (increasing) number and quality of available linguistic units (percepts), which are not static, as usual in modern NLP, but change over time in our pre-training method. These percepts are further used to create redundant representations of text, based on the assumption that elements in memory shape perception in humans. Our CogMemLM-s model shows increased performance in 27 out of 39 tasks compared to the BabyLM RoBERTa baseline model, which is a significant result ($p = 0.0071$, for details see Appendix A, all results are based on the BabyLM Evaluation Pipeline Warstadt et al. (2023); Gao et al. (2021)). The most striking improvement was archived in the BLiMP and BLiMP Supplement task sets, for which the relative change is 54 % and 46 %, respectively (better performance in 16/17 tasks).

Although these results suggest that implementing human-like cognitive mechanisms in LLMs is a promising avenue for future research and can result in substantial gains in performance also for small training datasets, a few limitations should be addressed. The memory processes as implemented here are relatively simplistic and do not yet consider that forgetting, as observed in humans, is non-linear (Ebbinghaus, 1885; Vlach and Sandhofer, 2012). Nor have we considered interference, which may have a substantial impact in lexicon creation (James et al., 2023). Furthermore, the chunk size of units that infants segment from language input and that subsequently enter the lexicon remains a topic of considerable debate (Grimm et al., 2017). Finally, many aspects of our approach are so far only integrated at text level, however, the lexical information could also be directly implemented in the tokenizer. Planned ablation studies will allow a more detailed evaluation of these first results and provide direction for future extensions of the present implementations.

Acknowledgements

The present research was funded by the Go!Digital 3.0 grant program of the Austrian Academy of Sciences (GD3.0 2021-18 CogML). We thank our intern student Célestin Eve for assisting us in this project. Furthermore, we would like to thank Google's TPU Research Cloud (TRC) program for giving us access to TPUs that were used for training our BabyLM models. We would also like to thank Hugging Face for providing the ability to host and perform inferencing of our models on the Hugging Face Model Hub.

References

Lisa MD Archibald. 2017. Slp-educator classroom collaboration: A review to inform reason-based practice. *Autism & Developmental Language Impairments*, 2:2396941516680369.

Alan D. Baddeley. 2003. Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4:829–839.

Nelson Cowan. 2016. Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on Psychological Science*, 11(2):239–264. PMID: 26993277.

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

Hermann Ebbinghaus. 1885. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Ruthe Foushee, Thomas L. Griffiths, and Mahesh Srinivasan. 2016. Lexical complexity of child-directed and overheard speech: Implications for learning. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016*, Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016, pages 1697–1702. The Cognitive Science Society.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Robert Grimm, Giovanni Cassani, Steven Gillis, and Walter Daelemans. 2017. Facilitatory Effects of Multi-Word Units in Lexical Processing and Word Learning: A Computational Investigation. *Frontiers in Psychology*, 8.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv*.

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, Online. Association for Computational Linguistics.

Emma James, M. Gareth Gaskell, Gráinne Murphy, Josie Tulip, and Lisa M. Henderson. 2023. Word learning in the context of semantic prior knowledge: evidence of interference from feature-based neighbours in children and adults. *Language, Cognition and Neuroscience*, 38(2):157–174. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2022.2102198.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden. Association for Computational Linguistics.

Katrin Kirchhoff and Steven Schimmel. 2005. Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, 117(4):2238–2246.

Richard Kunert, Raquel Fernandez, and Willem Zuidema. 2011. Adaptation in Child Directed Speech: Evidence from Corpora.

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. Cite arxiv:1907.11692.

Pierre Perruchet and Annie Vinter. 1998. Parser: A model for word segmentation. *Journal of Memory and Language*, 39:246–263.

Haley Vlach and Catherine Sandhofer. 2012. Fast Mapping Across Time: Memory Processes Support Children's Retention of Learned Words. *Frontiers in Psychology*, 3.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *CoRR*, abs/2208.07998.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

A Appendix

A.1 Schematic Overview of CogMemLM-s

Figure 1 shows the basic concept of our approach: Based on the RoBERTa architecture, CogMemLM-s is first trained on the Curriculum 1 Data Set. In the Tokenizer C 1 only 10 000 elements of the (final) Tokenizer are available, a number that is influenced by the size of the Lexicon C 1. Also based on Lexicon C 1 two alternative representations are created for every original sample in the Curriculum 1 Data Set: a coarse and a fine re-representation, as for the following example sentence:
Original: *She was a beautiful girl.*
Coarse: *She was a be aut if ul girl .*
Fine: *She was a be au t if ul gi rl .*

For curriculum 2, the RoBERTa architecture is initialized based on the resulting ComMemLM-s_c1, and the process descried for C 1 is repeated. The same applies to C 3 and C 4. The number of available elements in the respective tokenizers grows for each curriculum and in C 4 the full model vocabulary is available (see A.3 for further details).

A.2 Percept Lengths in Lexicon Creation

We assume that the mean length of sub-units is three and that initially, there are four working memory slots available for these sub-units. In order to account for cognitive growth throughout infancy and childhood (Cowan, 2016), we increase the number of available working memory slots and thereby the possible length of percepts across curriculum training steps: curriculum 1 (C 1): 4 slots, percepts of length 2-12 characters; C 2: 5, 2-15; C 3: 6, 2-18; C 4: 7, 2-21.

A.3 Tokenizer

We trained byte-level BPE tokenizers on the curriculum datasets as follows: Tokenizer C1 (model vocabulary 10 000) on C 1 dataset, tokenizer C 2 (model vocabulary 20 000) on datasets C 1 and C 2, tokenizer C 3 (model vocabulary 30 000) on datasets C 1, C 2, and C 3, and tokenizer C 4 (model vocabulary 40 000) on the full BabyLM dataset. The intersection of all model vocabularies was used as the final tokenizer (vocabulary size 41 130). In the curriculum training, however, only the tokens of the respective curriculum tokenizer were available (using the IDs of the final tokenizer).

A.4 Model Training

We used the same RoBERTa base model provided by the BabyML organizers for all model instances that we trained. The detailed model parameters are specified at Liu et al. (2019). The training data were organized in four sets of growing complexity, as illustrated in Table 1. The vocabulary size of the full training data is 41130. For each curriculum, the models were trained for 100 epochs, with maximal sequence length 512, learning rate 0.0001 and batch size 256.

A.5 BabyLM Leaderboard Results

Table 2-5 show the results of the official BabyLM model leader board `https://dynabench.org/tasks/baby_strict` for our model and the comparable BabyLM RoBERTa baseline model.

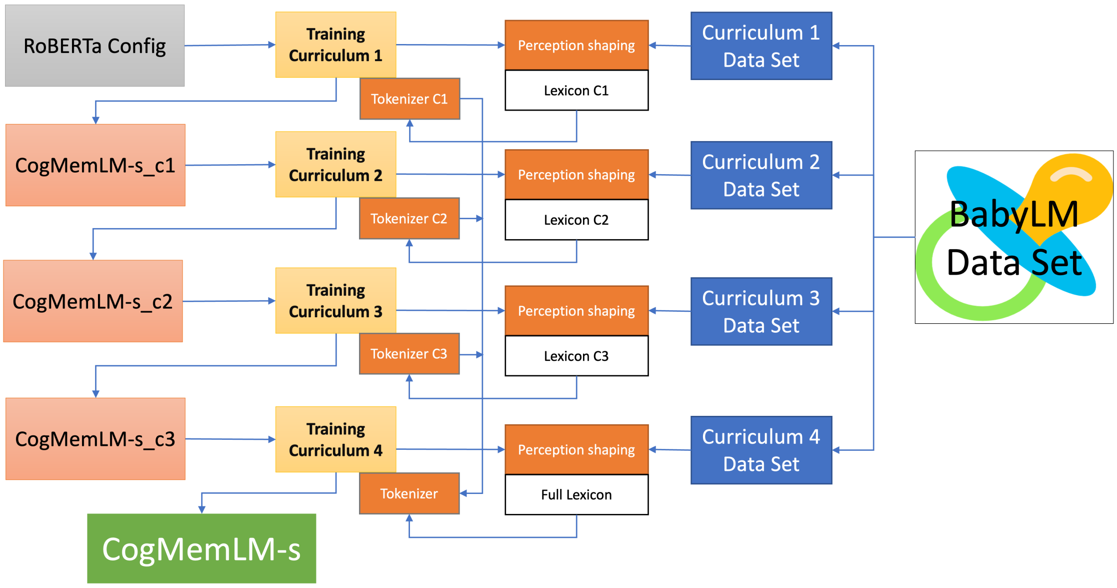

Figure 1: Schematic overview of CogMemLM-s.

	Corpus	**Domain**	**# Words**
C 1	CHILDES (MacWhinney, 2000)	Child-directed speech	4.21 M
	Children's Book Test (Hill et al., 2016)	Children's books	5.55 M
	Children's Stories Text Corpus	Children's books	3.22 M
	OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	31.28 M/4
C 2	Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	1.18 M
	British National Corpus (BNC), dialogue portion	Dialogue	8.16 M
	Simple Wikipedia	Wikipedia (Simple EN)	14.66 M/2
	OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	31.28 M/4
C 3	QCRI Educational Domain Corpus (QED; Abdelali et al., 2014)	Educational video subtitles	10.24 M
	Simple Wikipedia	Wikipedia (Simple EN)	14.66 M/2
	OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	31.28 M/4
C 4	Wikipedia	Wikipedia (English)	10.08 M
	Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Written English	9.46 M
	OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	31.28 M/4

Table 1: Split of the BabyLM-STRICT dataset into curriculum subsets (C 1–C 4). The open subtitles corpus is represented in all curricula, as this type of language input is assumed to be constant across all developmental stages.

Model	***Ana. Agr***	***Agr. Str***	***Binding***	***Ctrl. Rais.***	***D-N Agr***	***Ellipsis***	***Filler. Gap***	***Irreg. Forms***	***Island***	***NPI***	***Quantifiers***	***S-V Agr***
BLM RoBERTa	59.2	62.05	48.03	54.90	49.76	41.05	56.26	51.76	40.21	38.79	49.10	51.49
CogMemLM-s	88.75	73.31	73.24	71.06	93.65	89.09	73.09	85.24	61.81	70.15	69.65	78.28
change	*49.92*	*18.15*	*52.49*	*29.44*	*88.20*	*117.03*	*29.91*	*64.68*	*53.72*	*80.85*	*41.85*	*52.03*

Table 2: Results of our model compared with BabyLM RoBERTa-base on the BLiMP benchmark. The accuracy of the two models and the relative change between them are reported in percent.
Avg. BLiMP: baseline 50.22, ours 77.27.

Model	***Hypernym***	***QA Congr. (easy)***	***QA Congr. (tricky)***	***Subj.-Aux. Inversion***	***Turn Taking***
BabyLM RoBERTa	50.81	34.38	34.55	45.60	46.79
CogMemLM-s	50.12	67.19	46.06	80.63	65.71
change	*-1.36*	*95.43*	*33.31*	*76.82*	*40.44*

Table 3: Results of our model compared with BabyLM RoBERTa-base on the BLiMP Supplement benchmark. The accuracy of the two models and the relative change between them are reported in percent.
Avg. BLiMP Suppl.: baseline 42.43, ours 61.94.

Model	***CoLA***	***SST-2***	***MRPC (F1)***	***QQP (F1)***	***MNLI***	***MNLI-mm***	***QNLI***	***RTE***	***BoolQ***	***MultiRC***	***WSC***
BabyLM RoBERTa	45.30	87.80	82.00	84.54	77.10	77.94	84.08	54.55	59.89	67.58	61.45
CogMemLM-s	44.93	89.57	82.52	85.84	78.16	79.34	85.39	53.54	68.33	66.59	60.24
change	*-0.82*	*2.02*	*0.63*	*1.54*	*1.37*	*1.80*	*1.56*	*-1.85*	*14.09*	*-1.46*	*-1.97*

Table 4: Results of our model compared with BabyLM RoBERTa-base on the SuperGLUE benchmark. The accuracy and F1 score of the two models and the relative change between them are reported in percent.
Avg. (Super)GLUE: baseline 71.11, ours 72.22.

Model	***CR (Control)***	***LC (Control)***	***MV (Control)***	***RP (Control)***	***SC (Control)***	***CR_LC***	***CR_RTP***	***MV_LC***	***MV_RTP***	***SC_LC***	***SC_RP***
BabyLM RoBERTa	74.68	100.00	99.93	99.98	59.23	-89.04	-91.24	-99.84	-15.30	-57.74	-39.17
CogMemLM-s	91.30	100.00	99.88	86.84	65.81	-68.19	-75.12	-99.97	-86.83	-65.29	-49.54
change	*22.25*	*0.00*	*-0.05*	*-13.14*	*11.11*	*23.42*	*17.67*	*-0.13*	*-467.52*	*-13.08*	*-26.47*

Table 5: Results of our model compared with BabyLM RoBERTa-base on the MSGS benchmark. The Matthew correlation coefficients of the two models and the relative change between them are reported in percent (negative correlation scores indicate surface generalisations, positive correlation scores linguistic generalizations).
Avg. MSGS: baseline 3.77, ours -0.10.