What is the Best Sequence Length for BABYLM?

Suchir Salhan*, Richard Diehl Martinez*, Zébulon Goriely & Paula Buttery

*equal contribution.
{sas245, rd654}@cam.ac.uk

https://huggingface.co/babylm-seqlen

https://github.com/rdiehlmartinez/babylm-seqlen

Many BabyLM submissions use a shorter sequence length. But are shorter sequence lengths actually optimal for training on limited data?

The Case for ...

Long Sequences

Short Sequences

Training Efficiency

More updates

More Cognitively-Plausible? (Psychometric Fit)

~ mimic human working memory limitations

("starting small hypothesis"; Elman 1990)

We train BabyLMs with OPT and Mamba architectures with eight different sequence lengths on Strict Corpus (100M words)

8 distinct BabyLM datasets – shuffled at document level, split into fixed-length chunks with different seq lens.

Why Mamba? SSMs, unlike Transformers, don't impose a hard-cap on sequence length.
We want to investigate whether in data-limited contexts Mamba benefit from longer sequence lengths.

L*, the Training-Optimal Sequence Length for a BabyLM Evaluation Task

The shortest L that yields competitive accuracy relative to other lengths while offering a measurable training-time benefit

Task	OPT				Mamba			
	L^*	% (Longest)	L_{best}	% (Longest)	L^*	% (Longest)	L_{best}	% (Longest)
BLiMP	1024	34.8	64	100.0	512	37.3	2048	33.3
BLiMP Suppl.	256	43.9	64	100.0	64	100.0	64	100.0
Entity Tracking	4096	34.5	8192	38.8	1024	35.2	128	58.4
Wug	4096	34.5	4096	34.5	128	58.4	128	58.4
EWoK	4096	34.5	2048	34.3	1024	35.2	512	37.3
Reading	8192	38.8	8192	38.8	512	37.3	64	100

Training time is expressed as a proportion of the longest run within the same model family to facilitate comparison under setup variance and without exhaustive hyperparameter sweeps.

OPT checkpoints

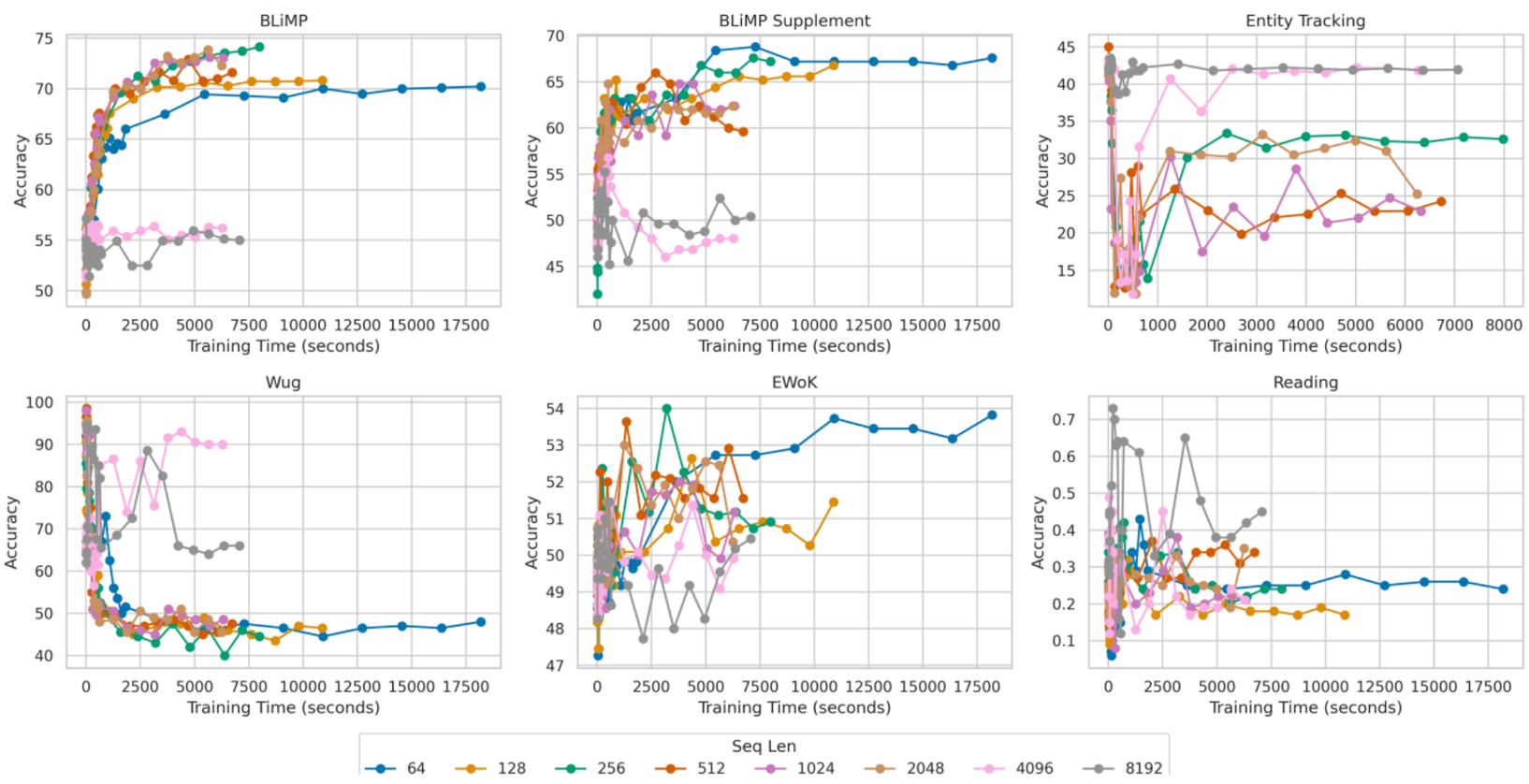
TL;DR: Effect of Sequence Length is Task-Dependent

Shorter Sequence Length: BLiMP, BLiMP Suppl

Longer Sequence Lengths: Wug (and Entity Tracking, Reading Evaluation)

Mamba < OPT

Mamba consistently prefers shorter or mid-range sequence lengths



Limitation we do not vary the mini-batch size or gradient accumulation strategy in conjunction with sequence length.

UNIVERSITY OF CAMBRIDGE

NATURAL LANGUAGE PROCESSING

EMNLP 2025 Suzhou, China 中国苏州