

OLMo2 with Hybrid MLM-CLM Training for the BabyLM 2026 Multilingual Challenge

Zichao Wei

Saarland University

ziwe00001@stud.uni-saarland.de

Abstract

We present our submission to the BabyLM 2026 Multilingual Shared Task. Our system uses OLMo2, a modern decoder-only architecture with grouped-query attention, RoPE, and SwiGLU, trained from scratch on the official BabyLM trilingual corpus (English, Dutch, Chinese). We employ a hybrid training objective following (Charpentier and Samuel, 2024) that combines masked language modeling (87.5%) with causal language modeling (12.5%), enabling the decoder-only model to perform well on both bidirectional and autoregressive evaluation tasks. With a multilingual BPE tokenizer (32K vocabulary) and approximately 79M parameters, our model achieves a multilingual average of 47.12 on the official leaderboard. We release our model and training code.¹

1 Introduction

The BabyLM Challenge (Warstadt et al., 2023) encourages the development of language models trained on limited, developmentally plausible data. The 2026 edition introduces a multilingual track requiring models to handle English, Dutch, and Chinese within a 100M-word budget. The previous edition (Charpentier et al., 2025) demonstrated that hybrid masked-causal training objectives are particularly effective in this data-constrained setting.

We explore whether modern architectural advances—specifically the OLMo2 architecture (Team OLMo et al., 2025)—can be effectively combined with the hybrid training paradigm (Charpentier and Samuel, 2024) in a multilingual setting. Our key contributions are:

- **Architecture choice:** We adopt OLMo2, a decoder-only transformer with grouped-query attention (GQA), rotary position embeddings

(RoPE), SwiGLU activations, and RMSNorm. To our knowledge, this is the first application of OLMo2 to the BabyLM challenge.

- **Hybrid training objective:** Following Charpentier and Samuel (2024), we combine masked language modeling (MLM) with causal language modeling (CLM), using MLM for 87.5% of batches and CLM for 12.5%. This enables the decoder-only model to leverage bidirectional context for BLiMP-style evaluation tasks while retaining autoregressive capabilities.
- **Multilingual BPE tokenizer:** We train a byte-pair encoding tokenizer with Unicode script pre-tokenization, producing a 32K vocabulary that efficiently handles all three languages.

2 Model

2.1 Architecture

Our model uses the OLMo2 configuration with 8 layers, 8 attention heads (4 key-value heads via GQA), a hidden dimension of 512, and an intermediate dimension of 1,376. Word embeddings are tied between input and output. The total parameter count is approximately 79M. The maximum sequence length is 512 tokens.

2.2 Data and Tokenizer

We use the official BabyLM 2026 multilingual datasets with a 40/30/30 English/Dutch/Chinese split by document count, totaling approximately 100M words. We train a BPE tokenizer with NFKC normalization and lowercase folding, using Unicode script-based pre-tokenization to handle the distinct writing systems. The vocabulary size is 32,000 tokens.

¹Model: <https://huggingface.co/exusiaiw/emnlp-babylm-2026-v2>

Category	EN	NL	ZH
Zeroshot Avg	41.33	55.24	49.95
Finetune Avg	48.59	45.80	42.80
Language Avg	45.48	50.21	45.66
Multilingual Avg	47.12		

Table 1: Results on the BabyLM 2026 multilingual evaluation suite. Multilingual Average is the official leaderboard score.

2.3 Training

We train for 10 epochs using AdamW with a peak learning rate of 1×10^{-3} , weight decay of 0.1, warmup ratio of 0.06, and gradient accumulation over 2 steps (effective batch size of 64 sequences $\times$ 512 tokens = 32,768 tokens per step). Training uses FP16 mixed precision and takes approximately 7 hours on a single NVIDIA A100 GPU.

2.4 Hybrid Objective

Following Charpentier and Samuel (2024), for 87.5% of training batches, we apply masked language modeling: 15% of tokens are selected for prediction, of which 80% are replaced with a [MASK] token, 10% with a random token, and 10% left unchanged. The attention mask is set to be fully bidirectional. For the remaining 12.5% of batches, we apply standard causal language modeling with a triangular attention mask. This hybrid approach allows the decoder-only architecture to develop bidirectional representations while maintaining autoregressive generation capability.

3 Results

We evaluate using the official BabyLM 2026 multilingual evaluation pipeline (Biderman et al., 2026), comprising zero-shot tasks (BLiMP, MultiBLiMP, ZhoBLiMP, HellaSwag, Winogrande, XStoryCloze, XCOMPS, GlobalPIQA), fine-tuning tasks (ARC, Belebele, BMLama, MNLI, SIB-200, TruthfulQA, XNLI, INCLUDE, POS tagging), and MECO reading time prediction.

Our model achieves a multilingual average of 47.12 on the official leaderboard. Table 2 shows per-task results across languages.

3.1 Learning Curve

Figure 1 shows the training and evaluation loss over the course of training. The train loss decreases steadily from 10.4 to 3.7, while the eval loss plateaus around 3.9 after epoch 5, suggesting

Task	EN	NL	ZH
<i>Zero-shot</i>			
BLiMP / ZhoBLiMP	71.01	77.91	78.12
MultiBLiMP	85.32	92.32	–
HellaSwag	26.80	26.56	27.78
Winogrande	51.28	50.87	50.70
XStoryCloze	49.92	49.23	49.69
XCOMPS	–	53.71	56.25
GlobalPIQA (par.)	21.36	25.24	23.30
GlobalPIQA (non-par.)	48.00	47.00	51.00
<i>Fine-tuning</i>			
ARC	27.50	26.67	20.83
Belebele	25.57	28.41	22.73
BMLama	24.67	24.42	16.06
MNLI	55.69	58.56	57.77
SIB-200	81.00	80.50	79.00
TruthfulQA	25.89	25.89	21.43
XNLI	55.75	–	52.10
INCLUDE	–	28.57	24.11
POS	92.62	93.42	91.20

Table 2: Per-task accuracy (%) across languages.

that additional training epochs yield diminishing returns at this model scale.

3.2 Checkpoint Evaluation

Figure 2 shows zero-shot task performance at intermediate checkpoints during training. Linguistic tasks (BLiMP, MultiBLiMP, ZhoBLiMP) show rapid improvement in the first 10M words and continue to improve throughout training, reaching 78–92% accuracy. Commonsense reasoning tasks (HellaSwag) plateau early and remain near chance level (27%), consistent with the limited world knowledge available in child-directed speech. XCOMPS and BLiMP-NL show steady improvement, with Dutch benefiting from vocabulary overlap with English.

3.3 Analysis

The model shows consistent performance across languages, with Dutch slightly outperforming English and Chinese on zero-shot tasks, likely due to favorable tokenizer fertility (Dutch shares many subwords with English in our BPE vocabulary). Chinese zero-shot performance benefits from the dedicated ZhoBLiMP evaluation (78.12%), suggesting that the hybrid training objective effectively captures Chinese syntactic patterns.

Fine-tuning performance varies by task type. Tasks like SIB-200 (topic classification) and POS tagging achieve strong results (81–93%), while knowledge-intensive tasks (BMLama, TruthfulQA) are near random, as expected for a model trained only on child-directed speech data. The Glob-

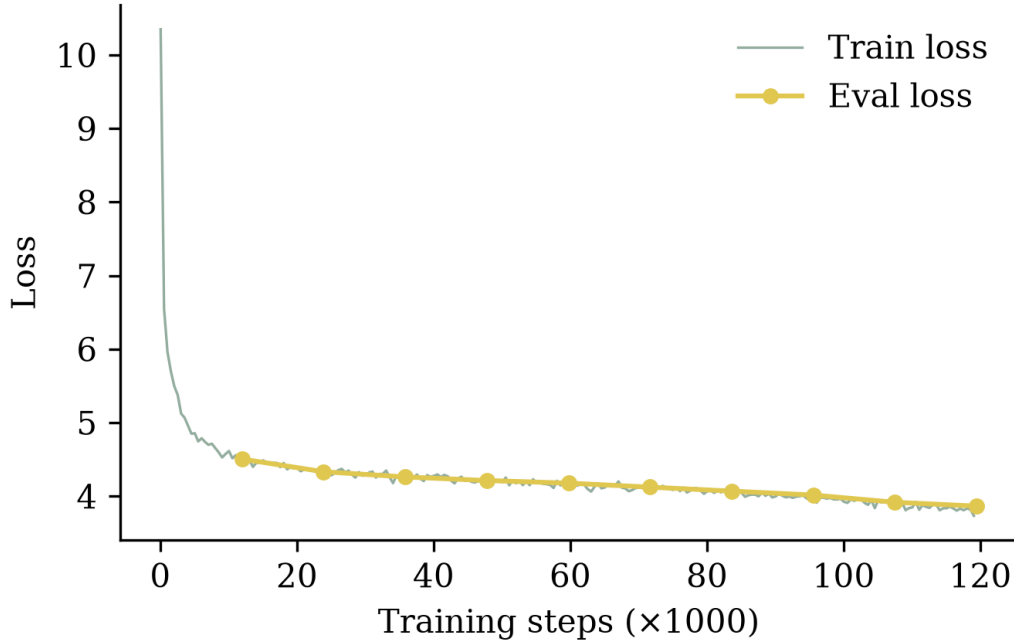


Figure 1: Training and evaluation loss over 10 epochs (119K steps).

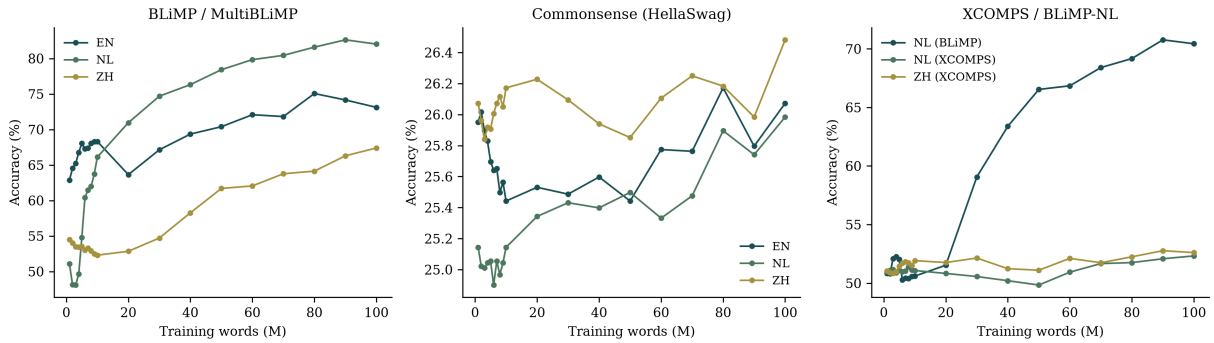


Figure 2: Zero-shot task accuracy at intermediate checkpoints (1M–100M training words). Left: grammatical acceptability. Center: commonsense reasoning. Right: semantic compatibility.

alPIQA parallel task is challenging for all models, while the non-parallel variant shows reasonable performance (47–51%).

The hybrid MLM-CLM objective is motivated by the evaluation setup: BLiMP-style tasks use log-probability comparisons that benefit from bidirectional context, while text generation tasks require autoregressive modeling. Our 87.5/12.5 split prioritizes the more numerous bidirectional evaluation tasks.

4 Conclusion

We presented a simple but effective approach for the BabyLM 2026 multilingual challenge based on OLMo2 with hybrid training. Our results suggest that modern architectural choices (GQA, RoPE, SwiGLU) combined with a hybrid training objec-

tive can achieve competitive performance in data-constrained multilingual settings. Future work could explore curriculum learning across languages and more sophisticated data mixing strategies.

5 Limitations

Our model is trained on a fixed 40/30/30 language split without tuning the mixing ratio. The hybrid objective ratio (87.5% MLM / 12.5% CLM) was also not systematically optimized. Additionally, our model has 79M parameters, which may be sub-optimal for the 100M-word budget—smaller or larger models might achieve better data efficiency.

References

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri

- Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2026. [Lessons from the trenches on reproducible evaluation of language models](#). *Preprint*, arXiv:2405.14782.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. [Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: why not both?](#) In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. [2 OLMo 2 furious](#). *Preprint*, arXiv:2501.00656.
- 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, Singapore. Association for Computational Linguistics.