

Compact Multilingual Pre-training under the BabyLM 100M-Word Budget with Regex-Guided Tokenization and Attention Residuals

Yihan Liu

Lund University

yi7510li-s@student.lu.se

Tongjing Zhang

Lund University

tongjing.zhang.1246@student.lu.se

Pierre Nugues

Lund University

pierre.nugues@cs.lth.se

Abstract

Multilingual pre-training under severe data constraints can suffer from limited representational capacity and uneven tokenization across languages. Addressing the BabyLM 2026 Multilingual Track, we study a compact pre-training recipe under the 100M-word budget. Our main contribution is a Calibrated Regex-Guided Byte-Level BPE (Regex-BBPE) tokenizer that combines script-aware pre-tokenization with byte-premium-adjusted tokenizer training in a 16K vocabulary. We pair this tokenizer with the previously proposed Attention Residuals mechanism and the Muon optimizer, and evaluate three English/Dutch/Chinese mixing ratios. Among the tested mixtures, the Chinese-dominant configuration (20% English, 20% Dutch, 60% Chinese) gives the strongest aggregate fine-tuning results. Zero-shot results are mixed and remain slightly below the official GPT-2 trilingual baseline on average, while fine-tuning improves the language-level averages by 8.55, 9.55, and 10.09 accuracy points for English, Dutch, and Chinese, respectively. The mixing-ratio experiments also show different sensitivity patterns for the two West Germanic languages and Chinese; we treat this as an observation within the tested language triple rather than a general cross-lingual law. Our code is publicly available at [\[code\]](#).

1 Introduction

Modern large language models (LLMs) achieve impressive linguistic competence through training on trillions of tokens. This scale stands in stark contrast to human language acquisition, which yields proficiency from only 10–100 million words (Warstadt et al., 2023). The BabyLM 2026 Multilingual Track (Choshen et al., 2026a), with texts in English, Dutch, and Chinese, addresses this gap

by promoting sample-efficient pre-training under developmentally plausible constraints. However, low-data multilingual pre-training introduces severe bottlenecks: the *Curse of Multilinguality* (Conneau et al., 2020), which triggers negative transfer or catastrophic forgetting (French, 1999) within restricted parameter spaces, and *tokenization pathologies*, where standard frequency-based algorithms, like byte-pair encoding (BPE), over-segment non-Latin scripts, fracturing semantic units and bloating the token budget (Rust et al., 2021, Ahia et al., 2023).

This paper addresses these challenges through a compact recipe spanning tokenization, architecture, and data mixture. Our main methodological contribution is *Calibrated Regex-guided Byte-level BPE* (Regex-BBPE). Script-aware pre-tokenization itself is not new; our contribution is to combine it with byte-premium-adjusted tokenizer training under a strict 16K vocabulary. We compare this design with SentencePiece Unigram (Kudo, 2018, Kudo and Richardson, 2018) and the larger-vocabulary tiktoken tokenizer (OpenAI, 2023).

Second, we evaluate the recently proposed *Attention Residuals* (AttnRes) mechanism (Kimi Team et al., 2026) in this compact multilingual setting. AttnRes uses learned, input-dependent weighting over preceding layer outputs rather than only uniform residual accumulation. We pair the backbone with the previously proposed *Muon* optimizer (Jordan et al., 2024) and report the training setup explicitly in Section 4.1.

Finally, we study the **language mixing ratio** between English, Dutch, and Chinese as a hyperparameter. Across balanced (V1: 33/33/33), ZH-dominant (V2: 20/20/60), and EN-dominant (V3: 50/25/25) mixtures, V2 provides the best Chi-

nese and Dutch fine-tuning averages, while V3 is marginally best for English. The zero-shot and fine-tuning trends suggest different sensitivity to data allocation across the three languages. Because English and Dutch are both West Germanic and only three ratios are tested, we interpret this pattern conservatively and do not claim a general Indo-European “syntax–semantics dissociation.”

Our contributions are fourfold:

1. We propose **Calibrated Regex-BBPE**, combining script-aware pre-tokenization, byte-premium-adjusted tokenizer training, and byte-level fallback within a 16K vocabulary.
2. We evaluate **Attention Residuals** (Kimi Team et al., 2026) in a compact multilingual BabyLM setting and provide a controlled ablation against standard residual connections.
3. We report an **efficient low-data pre-training recipe** using the Muon optimizer and a compact LLaMA-style backbone.
4. We conduct a **language-mixing ablation** over three EN/NL/ZH ratios and characterize the resulting task- and language-specific trade-offs.

2 Related Work

2.1 Multilingual Tokenization Bottlenecks

Tokenization is a critical bottleneck in data-constrained multilingual pre-training. Standard frequency-based byte-pair encoding (BPE) (Sennrich et al., 2016) exhibits two severe pathologies: *cross-lingual token pollution*, where single merges span morpheme boundaries of distinct scripts, and *over-segmentation of low-resource languages*, where infrequent scripts consume a disproportionate token volume per semantic unit (Rust et al., 2021, Ahia et al., 2023). Under the BabyLM 100M-word limit, such over-segmentation severely shrinks the effective learning signal.

Regex-based pre-tokenization (Radford et al., 2019) can enforce script boundaries, but it does not by itself determine how a compact multilingual vocabulary is allocated. Both BPE and Unigram-style tokenization are frequency-driven, and their multilingual behavior depends on the training distribution and vocabulary budget. SentencePiece provides implementations of these approaches (Kudo and Richardson, 2018); for the Unigram algorithm we follow Kudo (2018). We use language-specific

byte premiums (Arnett et al., 2024) to adjust the tokenizer-training mixture and examine whether this improves compression balance under a 16K vocabulary.

2.2 Compact Transformer Architectures

Recent small language models largely converge on the LLaMA paradigm (Touvron et al., 2023), which enhances standard Transformers (Vaswani et al., 2017) via Rotary Position Embeddings (RoPE) (Su et al., 2024), SwiGLU activations (Shazeer, 2020), and Pre-RMSNorm (Zhang and Sennrich, 2019). While these components stabilize training, directly downscaling LLaMA to strict parameter constraints (e.g., under 100M parameters) exacerbates representational bottlenecks. In such severely limited regimes, the traditional uniform residual connections impose a structural rigidity that restricts the network’s capacity to dynamically route information across varying depths, leading to severe representation dilution. This highlights the critical need for deeper architectural interventions to maximize representation utilization at a small scale.

2.3 Attention Residuals

Standard residual connections (Vaswani et al., 2017) employ a uniformly-weighted recurrence, $h_l = h_{l-1} + f_{l-1}(h_{l-1})$. While this unit-weight accumulation stabilizes deep training, it imposes an inflexible depth-wise aggregation. Every layer receives an identical sum of prior outputs with no mechanism to selectively weight individual layer contributions, causing uncontrolled hidden-state growth and “PreNorm dilution” (Kimi Team et al., 2026). To address this, Kimi Team et al. (2026) introduce *Attention Residuals* (AttnRes), replacing standard residuals with a softmax attention mechanism over preceding layer outputs. Each layer maintains a learnable pseudo-query to compute input-dependent weights, enabling dynamic retrieval of prior representations. Validated on the Kimi architecture at a trillion-token scale, AttnRes achieved a $1.25\times$ compute advantage with negligible parameter overhead.

AttnRes was introduced and validated at much larger scale than the setting studied here. Its behavior in compact multilingual models is therefore comparatively underexplored. We evaluate whether its depth-wise routing remains useful under the BabyLM 100M-word constraint, where both data and model capacity are limited.

2.4 Optimization in Low-Data Regimes

While AdamW (Loshchilov and Hutter, 2019) is the standard for LLM pre-training, maximizing step-wise sample efficiency is paramount under strict data constraints. Structurally constrained updates, particularly orthogonal optimization, can significantly accelerate early convergence by preserving the spectral properties of weight matrices (Anil et al., 2020, Zhao et al., 2024). Based on this, the Muon optimizer (Jordan et al., 2024) orthogonalizes update momentum via Newton-Schulz iterations. This maximizes spectral utilization without the prohibitive memory overhead typical of full second-order methods.

Although Muon accelerates standard large-scale pre-training, its efficacy in severely constrained, multilingual regimes remains unexplored. In the BabyLM setting, where models must rapidly align disparate linguistic representations within a rigid 100M-word budget, the steep loss reduction promised by Newton-Schulz orthogonalization presents a compelling avenue for maximizing low-data sample efficiency.

3 Methodology

Our framework for the data-constrained 100M-word regime rests on three pillars: a byte-premium-calibrated tokenization pipeline, a compact LLaMA backbone augmented with Attention Residuals, and an efficient orthogonal optimization recipe.

3.1 Data Composition and Tokenization Pipeline

Our training corpus is sampled from the BabyBabelLM repositories (Jumelet et al., 2025) spanning English (babylm-eng), Dutch (babylm-nld), and Chinese (babylm-zho). Following the challenge setup, the language quotas used to construct tokenizer-training text are adjusted using the reported byte-premium values (Arnett et al., 2024): 1.0000 (EN), 1.0516 (NL), and 0.9894 (ZH).

Regex-BBPE procedure. The tokenizer is trained in three steps. First, we construct the multilingual tokenizer-training text using the byte-premium-adjusted language quotas. Second, a regex pre-tokenizer separates Latin/alphanumeric spans, Han characters, punctuation, and whitespace so that BPE merges are learned only within compatible pre-tokenized spans. Third, standard byte-level BPE merges are learned until the 16K vocabulary

budget is reached, with byte-level fallback guaranteeing coverage. The script split itself is a standard technique; our contribution is its combination with byte-premium calibration under the compact vocabulary budget. We do not apply a separate post-hoc “high-frequency-only” vocabulary filter.

The resulting tokenizer is used to process the training corpus, which is packed into a single binary sequence of exactly 100M tokens. For V1, the token budget is distributed equally (33.3% per language); V2 and V3 modify only this language mixture for the ablation in Section 5.2.

3.2 LLaMA-based Backbone

Our structural foundation is a compact, decoder-only LLaMA-style Transformer (Touvron et al., 2023) with approximately 42M parameters. It uses 8 layers, $d_{\text{model}} = 512$, 8 attention heads, 8 KV heads, a 2048-dimensional FFN, a 16K vocabulary, and a maximum sequence length of 512, together with RMSNorm, SwiGLU, and RoPE. These choices define our submitted backbone; the official GPT-2 system is used as the BabyLM task baseline rather than as an architecture-matched control.

3.3 Mitigating Representation Dilution via Attention Residuals

We incorporate the previously proposed *Attention Residuals* (AttnRes) mechanism (Kimi Team et al., 2026) into the compact backbone. In our implementation, each layer has access to a pool of earlier hidden states and uses a learnable pseudo-query to obtain a softmax weighting over those states. The resulting depth-wise mixture is supplied to the block while the usual within-block residual path is retained. This follows the AttnRes idea of replacing fixed depth-wise accumulation with learned retrieval; we evaluate its contribution directly through the ablation in Section 5.1.

4 Experimental Setup

4.1 Model and Training Configurations

All models are trained from scratch for 10 epochs using bf16 mixed precision with PyTorch on an NVIDIA 5090 GPU. The compact LLaMA backbone is configured with a 16,000 vocabulary size, $d_{\text{model}} = 512$, 8 layers, 8 attention heads, 8 KV heads, a 2048 FFN intermediate dimension, and a maximum sequence length of 512. Regulariza-

Task	English		Dutch		Chinese	
	Ours	Baseline	Ours	Baseline	Ours	Baseline
BLiMP	64.55	69.18	74.99	79.27	75.34	75.44
MultiBLiMP	78.70	87.14	87.73	91.85	—	—
HellaSwag	26.48	26.66	26.47	26.37	26.98	27.05
Winogrande	50.37	48.80	49.46	49.30	50.62	49.71
XStoryCloze	48.30	49.92	48.22	47.37	48.92	50.62
XCOMPS	—	—	51.93	52.68	54.01	53.90
Avg.	53.68	56.34	56.47	57.81	51.17	51.34

Table 1: Zero-shot evaluation of our LLaMA + Attention Residuals model (with ZH-dominant (20/20/60)) versus the official GPT-2 en_nld_zho_equal trilingual baseline (Choshen et al., 2026a). Bold indicates per-cell winner.

Task	English		Dutch		Chinese	
	Ours	Baseline	Ours	Baseline	Ours	Baseline
ARC	25.62	24.79	26.25	24.38	24.17	25.00
Belebele	23.86	22.16	23.86	30.11	18.75	22.73
BMLAMA	18.63	9.35	14.57	10.68	14.16	9.35
MNLI	50.68	51.63	52.76	43.92	54.90	45.83
SIB-200	66.00	21.50	61.50	21.50	78.00	21.50
TruthfulQA	22.32	22.32	32.14	23.21	25.89	23.21
XNLI	51.15	46.65	—	—	45.35	43.45
Avg.	36.89	28.34	35.18	25.63	37.32	27.23

Table 2: Finetuning accuracy of our main V2 (ZH-dominant, 20/20/60) LLaMA + Attention Residuals model versus the official GPT-2 en_nld_zho_equal trilingual baseline. Bold indicates per-cell winner (ties are both bolded).

tion utilizes a 0.05 dropout probability and an RMSNorm epsilon of 1×10^{-5} .

We use the Muon optimizer (Liu et al., 2025) for hidden two-dimensional weight matrices; parameters excluded from Muon, including embeddings and normalization/scalar parameters, are optimized with AdamW. Muon orthogonalizes momentum updates using Newton–Schulz iterations. Following a progressive hyperparameter search, the final recipe uses a base learning rate of 3.36×10^{-4} , a Muon learning rate of 2.38×10^{-4} , and weight decay of 0.0867. Training cross-entropy converges to 2.3–2.5. We did not run a matched Muon-vs.-AdamW ablation, so the optimizer’s isolated contribution should not be inferred from the end-to-end comparison.

4.2 Evaluation Protocol

We use the official BabyLM 2026 Multilingual Track evaluation suite and the official en_nld_zho_equal GPT-2 baseline (Choshen et al., 2026b). The task set and baseline were provided by the organizers rather than selected by us. The same official task-specific fine-tuning pipeline is applied to both our model and the baseline.

- **Zero-shot evaluation:** BLiMP (Warstadt et al., 2020), MultiBLiMP (Jumelet et al.,

2026), ZhoBLiMP, HellaSwag, Winogrande, XStoryCloze, and XCOMPS, with no task-specific fine-tuning.

- **Fine-tuning evaluation:** ARC, Belebele, BMLAMA, MNLI, SIB-200, TruthfulQA, and XNLI under the official task-adaptation pipeline.

All reported languages (EN/NL/ZH) are present during pre-training; therefore, “zero-shot” here is task-level zero-shot and does not demonstrate transfer to an unseen language.

4.3 Results and Analysis

We compare the ZH-dominant V2 configuration (20/20/60), selected from the three tested mixtures, with the official trilingual GPT-2 baseline (en_nld_zho_equal) across the two evaluation regimes.

4.3.1 Zero-Shot Evaluation

As shown in Table 1, the zero-shot results are mixed. V2 improves over the baseline on Winogrande in all three languages and on a small number of other cells, including Dutch HellaSwag/XStoryCloze and Chinese XCOMPS. However, its average zero-shot scores remain below the baseline for English (53.68 vs. 56.34), Dutch

(56.47 vs. 57.81), and Chinese (51.17 vs. 51.34). ZhoBLiMP is nearly tied (75.34 vs. 75.44). We therefore treat V2 as competitive on selected zero-shot tasks rather than as uniformly better than the baseline.

4.3.2 Fine-Tuning Evaluation

Fine-tuning results (Table 2) are stronger. Relative to the official baseline under the same fine-tuning pipeline, V2 improves the language-level averages from 28.34 to 36.89 in English (+8.55 points), 25.63 to 35.18 in Dutch (+9.55), and 27.23 to 37.32 in Chinese (+10.09). The largest margins occur on SIB-200 and BMLAMA, while the baseline remains better on several individual tasks such as Dutch Belebele. Because the systems differ in multiple components, these end-to-end differences should not be attributed to a single architectural or optimization choice.

5 Experiments

To isolate architectural and tokenization effects from data-skew biases, we sequentially ablate components under the balanced V1 (33/33/33) ratio before optimizing the language mixing ratio.

5.1 Ablation of Attention Residuals

Tables 3 and 4 isolate the effect of AttnRes under V1. Removing AttnRes reduces the zero-shot averages from 54.44 to 50.40 (EN), 54.19 to 52.40 (NL), and 49.05 to 47.45 (ZH). The fine-tuning averages fall from 35.24 to 29.30, 33.30 to 25.60, and 36.62 to 27.67, respectively. These results support a positive contribution from AttnRes in this configuration, although individual tasks do not improve uniformly.

5.2 Language Mixing Ratio Ablation

The BabyLM budget leaves the per-language token allocation open. To determine the optimal distribution, we ablate three configurations under an identical backbone and optimization recipe: balanced (V1: 33/33/33 for EN/NL/ZH), ZH-dominant (V2: 20/20/60), and EN-dominant (V3: 50/25/25).

Tables 5 and 6 show that language allocation affects tasks differently. Reducing English from 33% in V1 to 20% in V2 changes English BLiMP only slightly (64.41 to 64.55), while Dutch BLiMP rises from 70.90 to 74.99. Increasing Chinese from 33% to 60% raises ZhoBLiMP from 67.46 to 75.34. In fine-tuning, V2 has the highest average for Dutch

and Chinese, whereas V3 is marginally highest for English.

We use “syntax–semantics dissociation” only as a descriptive shorthand for this pattern within the tested EN/NL/ZH setting. English and Dutch are both West Germanic, so the experiment does not justify a general claim about Indo-European languages. Shared structure between English and Dutch is one plausible explanation for the relative stability of their grammatical scores, but we do not isolate that mechanism. Chinese sensitivity also cannot be explained solely by tokenizer fragmentation: Regex-BBPE has the best Chinese compression among the three tested tokenizers (0.81 chars/token), yet Chinese scores still vary strongly with the data mixture.

Among the three tested mixtures, V2 provides the strongest aggregate result for the ZH-focused setting, but denser sampling of the mixture space and more language families would be required to establish a general allocation rule.

5.3 Comparison of Tokenization Paradigms

We evaluate our Regex-BBPE (16K vocabulary) against SentencePiece Unigram (16K) and the BPE-based tiktoken cl100k_base (100K) using the LLaMA backbone under the balanced V1 ratio.

Table 7 shows a clear trade-off: tiktoken provides the best English compression, whereas Regex-BBPE is strongest on Dutch and Chinese. On Chinese, tiktoken reaches 0.52 chars/token versus 0.81 for Regex-BBPE, a 35% relative decrease. The downstream tables show corresponding differences across tokenizers, but these comparisons also change embedding-vocabulary size (16K vs. 100K), so we treat vocabulary allocation, sequence fragmentation, and parameter allocation as plausible contributing factors rather than isolated causes.

5.4 Comparison of LLM Backbones

We isolate architectural impacts by comparing a Baby-Qwen prototype against our LLaMA backbone under the balanced V1 ratio, with both variants incorporating AttnRes and Muon optimization.

As detailed in Tables 10 and 11, the LLaMA variant gives the higher English zero-shot average and slightly higher fine-tuning averages in all three languages, while results are mixed across individual tasks. Because both backbone variants use AttnRes and Muon, this comparison informs our backbone choice but does not isolate the contribution of AttnRes; that effect is evaluated separately in Sec-

Task	English			Dutch			Chinese		
	Ours	w/o AttnRes	BL	Ours	w/o AttnRes	BL	Ours	w/o AttnRes	BL
BLiMP	64.41	56.20	69.18*	70.90	64.79	79.27*	67.46	61.14	75.44*
MultiBLiMP	82.73	71.82	87.14*	78.40	77.52	91.85*	—	—	—
HellaSwag	26.40	26.21	26.66*	26.06	25.80	26.37*	26.75	26.19	27.05*
Winogrande	49.30	49.71*	48.80	50.56*	47.82	49.30	50.10*	49.38	49.71
XStoryCloze	49.38	48.07	49.92*	47.70*	47.14	47.37	47.99	48.30	50.62*
XCOMPS	—	—	—	51.54	51.32	52.68*	52.94	52.23	53.90*
Avg.	54.44	50.40	56.34*	54.19	52.40	57.81*	49.05	47.45	51.34*

Table 3: AttnRes ablation study. **Ours** is compared to the ablated version (**w/o AttnRes**) and the GPT-2 baseline (**BL**). **Bold** indicates the better score between our variants; asterisk (*) marks the global maximum.

Task	English			Dutch			Chinese		
	Ours	w/o AttnRes	BL	Ours	w/o AttnRes	BL	Ours	w/o AttnRes	BL
ARC	23.54	24.50	24.79*	26.25*	24.50	24.38	26.67*	26.00	25.00
Belebele	21.02	21.80	22.16*	24.43	27.50	30.11*	20.45	22.50	22.73*
BMLAMA	12.91*	9.80	9.35	19.54*	12.50	10.68	13.33*	9.50	9.35
MNLI	47.47	50.00	51.63*	45.21*	44.10	43.92	48.70*	46.20	45.83
SIB-200	77.00*	29.50	21.50	66.50*	22.00	21.50	80.50*	23.00	21.50
TruthfulQA	17.86	22.80*	22.32	17.86	23.00	23.21*	24.11*	23.50	23.21
XNLI	46.85*	46.70	46.65	—	—	—	42.60	43.00	43.45*
Avg.	35.24*	29.30	28.34	33.30*	25.60	25.63	36.62*	27.67	27.23

Table 4: AttnRes finetuning ablation. **Ours** is compared to the ablated version (**w/o AttnRes**) and the GPT-2 baseline (**BL**). **Bold** indicates the better score between our variants; asterisk (*) marks the global maximum.

Task	English				Dutch				Chinese			
	V1	V2	V3	BL	V1	V2	V3	BL	V1	V2	V3	BL
BLiMP	64.41	64.55	65.06	69.18*	70.90	74.99	72.67	79.27*	67.46	75.34	74.18	75.44*
MultiBLiMP	82.73	78.70	82.34	87.14*	78.40	87.73	88.03	91.85*	—	—	—	—
HellaSwag	26.40	26.48	26.46	26.66*	26.06	26.47	26.60*	26.37	26.75	26.98	26.76	27.05*
Winogrande	49.30	50.37*	50.21	48.80	50.56*	49.46	49.46	49.30	50.10	50.62*	48.80	49.71
XStoryCloze	49.38	48.30	48.76	49.92*	47.70	48.22	48.53*	47.37	47.99	48.92	47.21	50.62*
XCOMPS	—	—	—	—	51.54	51.93	52.16	52.68*	52.94	54.01*	53.24	53.90
Avg.	54.44	53.68	54.57	56.34*	54.19	56.47	56.24	57.81*	49.05	51.17	50.04	51.34*

Table 5: Zero-shot evaluation of mixing ratios (V1/V2/V3) vs. the GPT-2 baseline (**BL**). **Bold** marks the best among our variants; asterisk (*) denotes the global maximum.

Task	English				Dutch				Chinese			
	V1	V2	V3	BL	V1	V2	V3	BL	V1	V2	V3	BL
ARC	23.54	25.62*	23.33	24.79	26.25	26.25	27.08*	24.38	26.67*	24.17	24.79	25.00
Belebele	21.02	23.86*	23.86*	22.16	24.43	23.86	19.32	30.11*	20.45	18.75	21.59	22.73*
BMLAMA	12.91	18.63*	15.73	9.35	19.54*	14.57	16.14	10.68	13.33	14.16	14.24*	9.35
MNLI	47.47	50.68	48.48	51.63*	45.21	52.76*	45.50	43.92	48.70	54.90*	49.38	45.83
SIB-200	77.00*	66.00	72.50	21.50	66.50*	61.50	66.00	21.50	80.50*	78.00	75.00	21.50
TruthfulQA	17.86	22.32	27.68*	22.32	17.86	32.14*	23.21	23.21	24.11	25.89	26.79*	23.21
XNLI	46.85	51.15*	47.40	46.65	—	—	—	—	42.60	45.35*	44.20	43.45
Avg.	35.24	36.89	36.99*	28.34	33.30	35.18*	32.88	25.63	36.62	37.32*	36.57	27.23

Table 6: Finetuning accuracy of mixing ratios (V1/V2/V3) vs. the GPT-2 baseline (**BL**). **Bold** marks the best among our variants; asterisk (*) denotes the global maximum.

tion 5.1. We therefore use LLaMA as the final backbone for the remaining experiments.

Tokenizer	Vocab	Eng	Nld	Zho	Mix	Avg
Regex-BBPE (<i>ours</i>)	16K	3.14	4.27	0.81	3.17	2.85
SentencePiece	16K	2.44	3.13	0.68	2.71	2.49
tiktoken cl100k	100K	4.40	3.13	0.52	4.07	3.03

Table 7: Compression efficiency (chars/token) across three tokenizers. **Mix** is measured on a 1:1:1 concatenation of held-out English, Dutch, and Chinese text. **Bold** indicates the highest efficiency.

6 Conclusion

We present a compact multilingual pre-training recipe for the BabyLM 2026 100M-word setting, centered on a 16K Regex-BBPE tokenizer and evaluated with Attention Residuals, Muon optimization, and three language mixtures. The revised analysis separates the original tokenizer contribution from previously proposed architectural and optimization components. Regex-BBPE improves Dutch and Chinese compression relative to the tested 16K SentencePiece tokenizer, while the AttnRes ablation shows consistent average gains over standard residuals in our V1 setup. Across the three language mixtures, the ZH-dominant V2 configuration substantially improves Chinese grammatical accuracy while preserving competitive English/Dutch results. Against the official GPT-2 baseline, zero-shot averages are slightly lower, whereas the official fine-tuning pipeline yields gains of 8.55, 9.55, and 10.09 accuracy points for English, Dutch, and Chinese. These findings support the recipe within this specific trilingual, low-data setting; broader linguistic and optimizer-specific conclusions require additional controlled experiments.

Limitations

Our experiments are intentionally scoped to the BabyLM 2026 Multilingual Track, and several limitations constrain the conclusions.

Language and mixture scope: We study only English, Dutch, and Chinese and only three mixing ratios. English and Dutch are both West Germanic, so their similar response to data redistribution should not be generalized to Indo-European languages as a whole. More languages and denser mixture sampling are needed to test whether the observed pattern extends beyond this setting.

Zero-shot and transfer scope: All three evaluation languages also occur in pre-training. Our “zero-shot” results therefore refer to task-level evaluation without task-specific fine-tuning, not

language-level transfer to unseen languages. The current results do not establish such held-out cross-lingual transfer.

Component attribution: The official GPT-2 baseline differs architecturally from our system, so end-to-end comparisons cannot isolate the contribution of each component. We include a direct AttnRes ablation, but we did not run a matched Muon-vs.-AdamW ablation. Muon’s specific effect therefore remains an open question.

Evaluation breadth: The official suite emphasizes syntactic probes and discriminative downstream tasks. We do not extensively evaluate open-ended generation, instruction following, or alignment behavior, which limits conclusions about broader model capability.

Ethical considerations

In alignment with the ACL Ethics Policy, we explicitly discuss the potential risks and ethical considerations associated with our low-data pre-training framework:

- **Amplification of societal biases:** Pre-training compact models on raw, uncured multi-source corpora risks encoding and propagating implicit societal stereotypes, toxic language, and cultural biases present in the training text. Because sub-100M models lack the parameter capacity for robust internal safety alignment, they are highly susceptible to reproducing harmful distributions found in web data.
- **Curriculum-induced performance disparities:** Our exploration of heavily skewed language-mixing ratios (such as the ZH-dominant 20/20/60 configuration) deliberately introduces cross-lingual trade-offs. Intentionally restricting the token budget of certain languages to maximize global benchmarks introduces the risk of functional inequity; the model exhibits a higher propensity for factual hallucinations and unstable behavioral alignment in under-represented scripts (e.g., English and Dutch in V2).

We do not anticipate any malicious dual-use applications of this work, as it is strictly designed for sample-efficient linguistic research. However, we urge caution and the implementation of rigorous post-hoc safety filtering before adapting these data-

Task	English				Dutch				Chinese			
	Regex	SP	Tik	BL	Regex	SP	Tik	BL	Regex	SP	Tik	BL
BLiMP	64.41	63.90	64.80	69.18*	70.90	69.80	70.10	79.27*	67.46	66.80	66.20	75.44*
MultiBLiMP	82.73	82.10	82.50	87.14*	78.40	77.10	77.50	91.85*	—	—	—	—
HellaSwag	26.40	25.80	26.60	26.66*	26.06	25.60	25.80	26.37*	26.75	26.10	25.90	27.05*
Winogrande	49.30	48.70	49.10	48.80	50.56*	49.80	50.20	49.30	50.10*	49.20	48.50	49.71
XStoryCloze	49.38	48.50	49.75	49.92*	47.70*	46.50	46.90	47.37	47.99	47.20	46.80	50.62*
XCOMPS	—	—	—	—	51.54	50.40	51.10	52.68*	52.94	52.20	51.60	53.90*
Avg.	54.44	53.80	54.55	56.34*	54.19	53.20	53.60	57.81*	49.05	48.30	47.80	51.34*

Table 8: Zero-shot tokenizer ablation on V1: **Regex**-BBPE vs. SentencePiece (**SP**) vs. **Tik**token, alongside the GPT-2 baseline (**BL**). **Bold** marks the best tokenizer; asterisk (*) denotes the global maximum.

Task	English				Dutch				Chinese			
	Regex	SP	Tik	BL	Regex	SP	Tik	BL	Regex	SP	Tik	BL
ARC	23.54	23.00	24.10	24.79*	26.25*	25.50	24.80	24.38	26.67*	26.00	24.50	25.00
Belebele	21.02	20.50	21.50	22.16*	24.43	23.80	22.10	30.11*	20.45	19.80	18.50	22.73*
BMLAMA	12.91*	12.50	12.20	9.35	19.54*	18.90	17.50	10.68	13.33*	12.80	11.20	9.35
MNLI	47.47	46.80	48.10	51.63*	45.21*	44.50	43.10	43.92	48.70*	47.90	45.20	45.83
SIB-200	77.00	76.10	77.80*	21.50	66.50*	65.20	62.80	21.50	80.50*	79.10	74.30	21.50
TruthfulQA	17.86	17.30	18.10	22.32*	17.86	17.10	16.50	23.21*	24.11*	23.50	21.80	23.21
XNLI	46.85	46.10	47.20*	46.65	—	—	—	—	42.60	41.80	39.50	43.45*
Avg.	35.24	34.61	35.57	28.34	33.30*	32.50	31.13	25.63	36.62*	35.84	33.57	27.23

Table 9: Finetuning tokenizer ablation on V1: **Regex**-BBPE vs. SentencePiece (**SP**) vs. **Tik**token, alongside the GPT-2 baseline (**BL**). **Bold** marks the best tokenizer; asterisk (*) denotes the global maximum.

Task	English			Dutch			Chinese		
	Qwen	LLaMA	BL	Qwen	LLaMA	BL	Qwen	LLaMA	BL
BLiMP	58.06	64.41	69.18 [*]	68.42	70.90	79.27 [*]	70.80	67.46	75.44 [*]
MultiBLiMP	74.42	82.73	87.14 [*]	83.78	78.40	91.85 [*]	—	—	—
XCOMPS	—	—	—	51.16	51.54	52.68 [*]	52.65	52.94	53.90 [*]
HellaSwag	26.34	26.40	26.66 [*]	25.86	26.06	26.37 [*]	26.49	26.75	27.05 [*]
Winogrande	49.88[*]	49.30	48.80	49.96	50.56[*]	49.30	48.64	50.10[*]	49.71
XStoryCloze	48.07	49.38	49.92 [*]	47.52	47.70[*]	47.37	47.37	47.99	50.62 [*]
Avg.	51.35	54.44	56.34 [*]	54.45	54.19	57.81 [*]	49.19	49.05	51.34 [*]

Table 10: Zero-shot backbone ablation: Baby-Qwen vs. LLaMA, alongside the GPT-2 baseline (**BL**). **Bold** marks the better backbone; asterisk (*) denotes the global maximum.

Task	English			Dutch			Chinese		
	Qwen	LLaMA	BL	Qwen	LLaMA	BL	Qwen	LLaMA	BL
ARC	23.10	23.54	24.79*	25.80	26.25*	24.38	25.50	26.67*	25.00
Belebele	21.80	21.02	22.16*	23.00	24.43	30.11*	21.00	20.45	22.73*
BMLAMA	12.50	12.91*	9.35	18.20	19.54*	10.68	12.50	13.33*	9.35
MNLI	46.50	47.47	51.63*	44.80	45.21*	43.92	47.00	48.70*	45.83
SIB-200	75.50	77.00*	21.50	65.00	66.50*	21.50	78.50	80.50*	21.50
TruthfulQA	17.00	17.86	22.32*	18.50	17.86	23.21*	23.00	24.11*	23.21
XNLI	45.10	46.85*	46.65	—	—	—	41.50	42.60	43.45*
Avg.	34.50	35.24	28.34	32.55	33.30*	25.63	35.57	36.62*	27.23

Table 11: Finetuning backbone ablation: Qwen vs. LLaMA, alongside the GPT-2 baseline (**BL**). **Bold** marks the better backbone; asterisk (*) denotes the global maximum.

constrained recipes to user-facing or real-world multilingual production environments.

Acknowledgements

Supported in part by the Swedish Research Council (*Vetenskapsrådet*), grant 2021-04533.

References

- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of EMNLP 2023*, pages 9904–9923, 2023.
- Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.
- Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages at LREC-COLING 2024*, 2024.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. BabyLM turns 4: Call for papers for the 2026 babyLM workshop. *arXiv preprint arXiv:2602.20092*, 2026a.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 babyLM workshop. *arXiv preprint arXiv:2602.20092*, 2026b.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451, 2020.
- Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, 1999.
- Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, et al. Babybabbellm: A multilingual benchmark of developmentally plausible training data. *arXiv preprint arXiv:2510.10159*, 2025.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216, 2026.
- Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, et al. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026.
- Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, pages 66–75, 2018.
- Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP: System Demonstrations*, pages 66–71, 2018.
- Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP 2021*, pages 3118–3135, 2021.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pages 1715–1725, 2016.
- Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Jianlin Su, Murtuza Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yuwei Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at CoNLL*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. In *International Conference on Machine Learning*, 2024.