

# Language Allocation and Targeted Data Design for a Multilingual BabyLM

Naoki Mori and Seiji Kyo and Kohki Goda

the Graduate School of Informatics

Osaka Metropolitan University, Osaka, Japan

mnao@omu.ac.jp

## Abstract

We study language allocation and data design for English–Dutch–Chinese pretraining in the multilingual BabyLM 2026 setting. Across three matched seeds, three fixed mixtures showed no significant difference in the seven-benchmark selection average (TASK-7) after Holm correction. Two within-language contrasts were significant after correction within each language: English favoured 50/25/25 by 2.487 points and Chinese favoured 27.5/27.5/45 by 2.151 points. The mixtures were the three highest-scoring conditions of a single-seed pilot, so that comparison is exploratory. An inverse-score controller changed sampling shares without a detected aggregate improvement in a 25M-word proxy. Across four matched seeds, adding synthetic and parallel-corpus text raised the zero-shot task average (ZS16) from 56.57 to 57.37, the largest language-specific difference being Chinese (+1.47); both remained significant after Holm correction. Four of five Chinese tasks improved in the three seeds with task-level records. That condition also consumed 14.9% more tokens and used its own tokenizer, so it does not isolate content from exposure. Multi-token prediction showed no detected improvement across three seeds. Parameter-count comparisons were single runs; the 213M model had the highest fine-tuning average (FT22, 41.25) and a leaderboard-reported Multilingual Average of 50.68.

## 1 Introduction

The multilingual track of BabyLM 2026 fixes an English-equivalent training budget across English, Dutch, and Chinese (Choshen et al., 2026). The constraint turns corpus construction into an allocation problem: additional exposure for one language reduces exposure elsewhere. This is a small-model setting in which the transfer–interference trade-off reported for much larger multilingual systems can be examined directly (Conneau et al.,

2020; Chang et al., 2024).

We report two sets of experiments under this budget. The first concerns **language allocation** (§4–6): a tokenizer pilot, three fixed language mixtures with matched seeds, and a dynamic reallocation controller. The second concerns **data design and model configuration** (§7): corpus augmentation with synthetic and parallel text, multi-token prediction, and comparisons across parameter counts.

Replicated comparisons use matched seeds and paired tests; single-run results are descriptive. We document data construction, budget accounting, and overlap screening in §7.1.

## 2 Related work

**The budgeted setting.** The multilingual track draws on BabyBabelLM, a 45-language collection (Jumelet et al., 2026a), and uses Byte Premium to account for differences in content-matched UTF-8 corpus sizes across languages (Arnett et al., 2024; Choshen et al., 2026). We study how to divide this fixed budget.

**What the division is expected to do.** Two lines of work predict opposing effects of language allocation. Exposure to one language can benefit others through cross-lingual transfer (Conneau et al., 2020). Conversely, languages compete for fixed model capacity, so shifting exposure toward one can degrade the others (Chang et al., 2024). Those studies use larger models and corpora and vary the language set. We instead fix the language set, architecture, and optimizer and vary the proportions (§5). Our observed differences alone do not distinguish transfer from capacity competition.

**Tokenization and objectives as design levers.** Unigram segmentation (Kudo, 2018) is an alternative to byte-pair encoding (BPE); controlled comparisons have favored unigram (Bostrom and Dur-

rett, 2020). Fusco et al. (2025) studied linguistically motivated segmentation in BabyLM. Multi-token prediction (MTP) adds auxiliary future-token heads (Gloeckle et al., 2024). Our tokenizer pilot varies both segmentation and vocabulary size, hence parameter count, and is exploratory (§4); our objective comparison detects no improvement (§7.3).

**Adaptive mixtures.** DoReMi learns fixed domain weights with a separate proxy model and a minimax objective (Xie et al., 2023). Our heuristic updates language probabilities within one run using benchmark accuracy (§6); we claim neither novelty nor a controlled comparison with DoReMi.

### 3 Experimental setup

#### 3.1 Data and models

Ratios are written as English/Dutch/Chinese. The allocation study uses the official BabyBabelLM data without augmentation (Choshen et al., 2026; Jumelet et al., 2026a). The data-design study adds parallel text (Tiedemann, 2020; Zhang et al., 2020) and offline task-family-targeted synthetic text (§7.1).

All models are GPT-2-style causal decoders (Radford et al., 2019) trained from scratch with AdamW (Loshchilov and Hutter, 2019) (Table 1). Learning rates decay linearly except in the dynamic experiment, which uses one full-run cosine schedule (Appendix C). Except in the tokenizer pilot, byte-level BPE is fitted to each model’s corpus; the baseline and augmented conditions therefore use different tokenizer files. Text is concatenated into fixed-length blocks. At a fixed epoch count, larger corpora require more optimizer steps. Context length is the maximum position length; block length is the packed training sequence length.

#### 3.2 Evaluation metrics

The allocation study uses TASK-7, the unweighted mean of seven zero-shot minimal-pair scores (English BLiMP and MultiBLiMP; Dutch BLiMP-NL, MultiBLiMP, and XCOMPS; Chinese ZhoBLiMP and XCOMPS) (Warstadt et al., 2020; Suijkerbuijk et al., 2025; Jumelet et al., 2026b; Liu et al., 2026; He et al., 2025). LANG-3 averages within language, then across the three languages. Table 19 additionally reports ALL-16, the unweighted mean of 16 task scores: the seven of TASK-7 together with HellaSwag, Winogrande, and XStoryCloze in each language. It covers the same task set as

| Setting         | Alloc.              | Data-des.        | Final    |
|-----------------|---------------------|------------------|----------|
| Layers          | 12                  | 12               | 14       |
| Attention heads | 12                  | 12               | 16       |
| Hidden size     | 768                 | 768              | 1024     |
| Context length  | 512                 | 512              | 512      |
| Block length    | 512                 | 128              | 128      |
| Vocabulary      | 16,000              | 32,000           | 32,000   |
| Parameters      | 97.737M             | 110.025M         | 212.790M |
| Peak LR         | 6e−4                | 4.5e−4           | 4.5e−4   |
| LR schedule     | linear <sup>§</sup> | linear           | linear   |
| Warmup          | 500 steps           | 2%               | 2%       |
| Weight decay    | 0.01                | 0.01             | 0.01     |
| Precision       | bf16                | bf16             | bf16     |
| Epochs          | 10                  | 10               | 10       |
| Batch (tok)     | 131,072             | 16,384           | 16,384   |
| MTP heads $k$   | —                   | 4 <sup>†</sup>   | 4        |
| MTP aux. weight | —                   | 0.3 <sup>†</sup> | 0.3      |

Table 1: Model settings. <sup>†</sup>Multi-token prediction (MTP) is used only in the +MTP comparison, not the augmentation comparison. The MorPiece pilot uses 39,697 vocabulary items and 115.937M parameters.

<sup>§</sup>The dynamic experiment uses cosine decay.

ZS16 below and averages it the same way. The English grammatical component of both TASK-7 and ALL-16 is the standard BLiMP release, whereas ZS16 uses the BabyLM-filtered subset, so values are not comparable across the two studies; the allocation study selected on TASK-7, not on ALL-16. All scores are produced with the official BabyLM multilingual evaluation pipeline. The seven benchmarks are equally weighted within a language, while sub-tasks inside a BLiMP-style benchmark are weighted by their number of evaluation items.

The data-design study uses ZS16 and FT22. ZS16 is the unweighted mean of the 16 public zero-shot task scores available for every ablation (English 5, Dutch 6, Chinese 5), and FT22 is the unweighted mean of the 22 non-POS fine-tuning task–language scores available for every ablation. These are descriptive comparison metrics, not the official Multilingual Average. The official score additionally includes GlobalPIQA and POS and averages the complete task set within each language before macro-averaging across English, Dutch, and Chinese; ZS16 is therefore not a three-language equal-weight macro. All scores are reported as percentages.

#### 3.3 Statistical analysis

We used two-sided paired  $t$ -tests over matched seeds. For each intervention in the data-design study, the ZS16 task average and three language-specific tests formed one family and were adjusted

with Holm’s procedure (Holm, 1979). In the fixed-mixture study, the three pairwise contrasts within each language formed a separate family per language. Confidence intervals are unadjusted 95%  $t$  intervals for the paired mean difference. With  $n = 4$ , normality cannot be checked and the smallest attainable two-sided sign-flip  $p$ -value is .125; we therefore report paired- $t$  evidence together with the seed-level differences in Appendix B rather than describing any effect as robust. Conditions differing in more than one factor are not used to attribute causes.

## 4 Tokenizer pilot

Table 19 compares equal-mixture runs with one training seed each at the full 100M-word budget over ten epochs, using byte-level BPE, a byte-level unigram model, and MorPiece (Fusco et al., 2025). MorPiece is a split-based tokenizer that posits a morpheme boundary wherever the Tolerance Principle applies as a word traverses the lexicon. The mixture pilot of Table 20 instead used a 25M-word budget, so the two tables differ in training data as well as in what they vary: cumulative raw-token exposure over the ten epochs is  $9.94 \times 10^8$  here against  $2.48 \times 10^8$  there. We do not attribute the difference in score level to the budget alone. MorPiece has the highest observed TASK-7 and LANG-3 scores and leads on four of seven component tasks. Because there is one seed per condition, we treat the seven tasks as an across-task descriptive comparison rather than as replicates: the mean task difference relative to BPE is 1.337 points and relative to Unigram 1.585 points, but these summaries do not express run-to-run uncertainty. MorPiece also adds 18.199M parameters (18.6%) through the embedding and output layers and changes vocabulary and segmentation together, so this pilot is exploratory.

## 5 Fixed language mixtures

We train 50/25/25, 40/40/20, and 27.5/27.5/45 with three matched seeds each (Table 2). Each language scores highest at its largest allocation; 40/40/20 has the highest observed TASK-7 mean.

### 5.1 Choice of mixtures

The three mixtures ranked highest on TASK-7 in a seven-mixture pilot with one seed per mixture, MorPiece, and a 25M-word proxy (Table 20). The symmetric conditions 25/50/25 and 25/25/50

ranked fourth and seventh and were not retained. Selecting on the reported metric makes the aggregate comparison exploratory.

Allocation spans differ: 22.5 points for English, 25 for Chinese, and 15 for Dutch, whose largest share (40) ties English. Thus the contrasts cannot compare allocation sensitivity across languages.

### 5.2 Aggregate comparisons

All three pairwise TASK-7 intervals include zero (Table 3), and the ordering varies by seed: 50/25/25 is highest for seed 1, 40/40/20 for seeds 2 and 3. At this sample size a difference was not detected; this does not establish equivalence.

### 5.3 Language-level comparisons

Within each language the three pairwise contrasts form one family; Table 10 gives the seed-level scores and Table 17 all nine contrasts. Two contrasts remain significant after correction. English is 2.487 points higher under 50/25/25 than under 27.5/27.5/45 (95% CI [2.078, 2.895]; Holm-adjusted  $p=.0044$ ). Chinese is 2.151 points higher under 27.5/27.5/45 than under 40/40/20 (95% CI [1.047, 3.255]; Holm-adjusted  $p=.0418$ ). Dutch has its highest observed mean under 40/40/20, but its corrected contrasts are not significant. Relative to 50/25/25, 40/40/20 is 1.642 higher in Dutch, .863 lower in English, and .727 lower in Chinese. As noted in §5.1, the Dutch contrasts cover a 15-point allocation range against 22.5 for English and 25 for Chinese, so these contrasts do not support a comparison of allocation sensitivity across the three languages.

## 6 Dynamic reallocation

The dynamic experiment uses a 25M-word proxy, BPE 16k, and a 97.737M-parameter decoder. The inverse-score heuristic raises the share of low-scoring languages. Starting from  $c_{0,\ell} = 1/3$ , it computes every two epochs

$$q_{t,\ell} = \frac{(s_{t,\ell} + \epsilon)^{-1}}{\sum_{j \in \{\text{en,nl,zh}\}} (s_{t,j} + \epsilon)^{-1}}, \quad \epsilon = 10^{-8}, \quad (1)$$

where  $s_{t,\ell}$  is the unweighted mean of that language’s benchmark accuracies on the same seven-benchmark suite used to report TASK-7, on the 0–1 scale and without clipping, evaluated zero-shot on the checkpoint at the end of the preceding two epochs (steps 384, 768, 1152, and 1536

| Mixture (en/nl/zh) | English           | Dutch             | Chinese           | Task-7            | Lang-3            |
|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 50/25/25           | 78.884 $\pm$ .597 | 73.054 $\pm$ .331 | 65.352 $\pm$ .657 | 72.519 $\pm$ .167 | 72.430 $\pm$ .142 |
| 40/40/20           | 78.021 $\pm$ .443 | 74.696 $\pm$ .460 | 64.626 $\pm$ .082 | 72.769 $\pm$ .322 | 72.448 $\pm$ .300 |
| 27.5/27.5/45       | 76.397 $\pm$ .668 | 73.145 $\pm$ .285 | 66.776 $\pm$ .513 | 72.255 $\pm$ .432 | 72.106 $\pm$ .457 |

Table 2: Fixed mixtures: three-seed mean  $\pm$  sample SD.

| Task-7 contrast         | $\Delta$ | 95% CI          | $p_{\text{Holm}}$ |
|-------------------------|----------|-----------------|-------------------|
| 40/40/20 – 50/25/25     | .250     | [−.932, 1.431]  | .9191             |
| 40/40/20 – 27.5/27.5/45 | .514     | [−.495, 1.523]  | .4792             |
| 50/25/25 – 27.5/27.5/45 | .265     | [−1.169, 1.698] | .9191             |

Table 3: Paired tests on TASK-7 ( $n = 3$ ), one family of three contrasts.

of 1920). The direct condition sets  $c_{t+1} = q_t$ ; the smoothed condition averages the new target with the current shares,  $c_{t+1} = .5c_t + .5q_t$ , which halves the step size. In both, the result is then projected onto the simplex subject to per-language bounds of  $[0.10, 0.60]$ . Every one of the 24 updates landed strictly inside those bounds—all post-update shares lie between .311 and .369—so the bounds never became active and the trajectories are those of the unconstrained rule. Language is drawn per 512-token block, and blocks are then sampled uniformly with replacement from that language’s pool, so the realized counts vary around the target shares rather than partitioning the data exactly. Each method is run with seeds 1–3. Appendix C gives the per-benchmark scores and the seed-level final shares.

Mean Chinese exposure rose from 83.67M to 89.00M raw tokens under direct updates (87.89M when smoothed). Neither method improved mean Chinese, TASK-7, or LANG-3 scores or reduced the language gap (Table 4). All paired aggregate tests have  $p_{\text{Holm}}=1.0$  (fixed minus inverse  $\Delta=-.181$ , 95% CI  $[-.883, 1.245]$ ). Because the same public benchmarks supply feedback and outcomes, this is a diagnostic of adaptation, not held-out generalization.

## 7 Data design and model configuration

We evaluate corpus augmentation and multi-token prediction separately before reporting the final models.

### 7.1 Augmentation data construction

**What is added.** The augmented corpus is the baseline corpus *plus* two sources; no real text is removed or replaced. Table 21 gives the composi-

tion. The baseline is sentence-segmented BabyBabelLM text built to per-language targets of 34M/33M/33M English-equivalent words. Chinese word counts are CJK (Chinese, Japanese, Korean) character counts, used as a segmenter-free proxy; the Budget paragraph below relates this convention to the byte measure the track actually caps.

**Synthetic text.** Synthetic text was generated offline by **Qwen3.5-4B** (HF model Qwen/ Qwen3.5-4B, immutable revision 851bf6e806efd8d0a36b00-ddf55e13ccb7b8cd0a; approximately 4B parameters, within the challenge’s teacher-size limit). Sampling used temperature 0.8, top- $p$  0.9, and at most 160 new tokens per sample. Prompts were rendered with the model’s chat template with reasoning traces disabled, and any residual `<think>` span was stripped from the output. We kept all 25,000 generated samples per language without length, language, or quality filtering. Prompts were split evenly over four templates that target the task *families* in the evaluation suite, not its items: STORY (narrative continuation), COMPARE (comparatives), COREF (pronoun reference), and FACT (simple declarative statements). Each prompt instantiates one template with one topic drawn from a fixed list of 143 everyday child-directed topics (kitchen, weather, school, the playground, and so on), with a per-language sampling seed (11/22/33 for English/Dutch/Chinese). The templates are given verbatim in Appendix E. Prompts contain a task-family description, a topic, a length instruction, and the target language name; they contain no evaluation instance, label, answer key, minimal pair, or output of the evaluation pipeline. The teacher was used only to generate text; no teacher logits, scores, or rankings were used for training.

**Parallel text.** Parallel text is drawn from community-produced translation corpora rather than teacher output, and contains some noisy or fragmentary sentences. We use Tatoeba (Tiedemann, 2020) for en–nl and en–zh, taken in full after filtering, and OPUS-100 (Zhang et al.,

| Method   | Eng    | Dut    | Chi    | Task-7 | Lang-3 | Gap    | Final zh |
|----------|--------|--------|--------|--------|--------|--------|----------|
| Fixed    | 67.564 | 68.671 | 58.933 | 65.573 | 65.056 | 9.738  | .333     |
| Inverse  | 67.066 | 68.686 | 58.775 | 65.391 | 64.842 | 9.910  | .366     |
| Smoothed | 67.849 | 68.440 | 58.392 | 65.400 | 64.894 | 10.048 | .363     |

Table 4: Dynamic experiment: three-seed means for fixed equal sampling, direct inverse-score updates, and smoothed inverse-score updates. Gap is the maximum minus minimum language mean; Dutch is the maximum and Chinese the minimum in every condition. Final zh is the Chinese sampling share applied over the last two epochs.

2020) to fill the remaining allowance, split equally between the two pairs. Filters keep pairs of 2–60 words per side with a length ratio of at most 3.0 for Tatoeba and 3–40 words with a ratio of at most 2.0 for OPUS-100, and drop pairs whose first 60 characters on both sides duplicate an already-accepted pair. Each retained pair yields two training sequences, one per direction, using the template L1: X\nL2: Y with the language names spelled out (English: ...\nDutch: ...) and its reverse. This gives 269,818 pairs and 539,636 sequences (240,952 from Tatoeba, 298,684 from OPUS-100; 319,296 en-nl and 220,340 en-zh). Tatoeba’s Chinese side mixes scripts: 13.3% of the en-zh Chinese bytes are on lines containing traditional characters. We charge all Chinese at the simplified premium; charging those lines at the traditional premium instead would change the total by 0.07 MB, or 0.01% of the cap. Examples of both sources are given in Appendix E.

**Budget.** The track measures the budget in bytes, not in words. In the organizers’ data-preparation code the 100M-word tier is the constant 543.00 MB. A language qualifies while its UTF-8 size stays below  $543.00 \text{ MB} \times \text{premium}$ . Equivalently, size divided by the premium must stay at or below 543.00 MB.<sup>1</sup> We apply that rule with the published coefficients: 1.0 for English, 1.0516 for Dutch, and 0.93597 for simplified Chinese, the script BabyBabelLM uses. The call for papers lists 0.9894, which the same table gives for traditional Chinese.

Under this rule the baseline corpus is 457.5 MB and the augmented corpus 527.7 MB of English-equivalent bytes, that is 84.3% and 97.2% of the cap (Table 5). Both are within budget, and the augmented corpus leaves little headroom. The parallel file mixes languages within a line; charging it en-

tirely at the English premium and splitting it by language give 527.74 and 527.66 MB, so that choice does not affect the conclusion.

Our construction code additionally tracked a word count, dividing whitespace tokens (English, Dutch) or CJK characters (Chinese) by the same premiums. That quantity is a proxy in a different unit and is not the one the track caps; we report it in Table 5 only because the corpus was assembled against it. It does not track the byte measure across languages, because a premium-adjusted byte count converts to a word count only at that language’s bytes per word. Relative to English, the proxy is inflated by a factor of 1.08 for Dutch and 1.62 for Chinese, where a character stands in for a word. Those factors, not the budget, are what the word column reflects.

Because the augmentation is additive and the epoch count is fixed at 10, the augmented condition also receives more optimizer steps: 87,820 against 76,440 at 16,384 tokens per step, that is 1.439B against 1.252B model tokens read, a 14.9% increase. These are BPE tokens including repetition across epochs, a different quantity from the English-equivalent word budget above, and we report them separately for that reason. The two conditions therefore differ in training exposure as well as in content, and each was trained with a tokenizer fitted to its own corpus, so §7.2 compares two data conditions rather than isolating content at matched exposure.

**Overlap audit.** We screened the distributed English BLiMP, Dutch BLiMP-NL, and Chinese ZhoBLiMP minimal-pair files using normalized exact matching followed by 8-gram candidate retrieval and character 5-gram Jaccard similarity (Appendix I). Across 9,853,776 line-language checks, three training lines matched two of 143,150 evaluation sentences; none was synthetic. The near-duplicate pass found no additional matches, but its recall was not measured. This au-

<sup>1</sup>src/ pad\_utils.py and resources/ byte\_coefs\_20240233.tsv in the organizers’ multilingual-babylm repository.

|                                                | Baseline |              | Augmented |              |
|------------------------------------------------|----------|--------------|-----------|--------------|
|                                                | MB       | eq. MB       | MB        | eq. MB       |
| English                                        | 183.4    | 183.4        | 188.1     | 188.1        |
| Dutch                                          | 173.1    | 164.6        | 177.3     | 168.6        |
| Chinese                                        | 102.5    | 109.5        | 106.9     | 114.2        |
| Parallel                                       | —        | —            | 56.9      | 56.9         |
| Total                                          | 459.0    | <b>457.5</b> | 529.1     | <b>527.7</b> |
| Cap: 543.00 MB (84.3% and 97.2% used)          |          |              |           |              |
| <i>Construction-time word proxy, millions:</i> |          |              |           |              |
| English                                        |          | 34.000       |           | 37.185       |
| Dutch                                          |          | 33.000       |           | 34.929       |
| Chinese                                        |          | 32.792       |           | 35.578       |
| Optimizer steps                                |          | 76,440       |           | 87,820       |
| Model tokens read                              |          | 1.252B       |           | 1.439B       |

Table 5: Budget and exposure. The upper block applies the track’s rule: UTF-8 megabytes divided by the language premium, against the 543.00 MB constant the organizers use for the 100M-word tier. The middle block is the construction-time word proxy (whitespace tokens, or CJK characters for Chinese, divided by the same premiums); it is not the capped quantity and is listed for provenance. Model tokens read are BPE tokens including repetition across epochs.

dit covers neither the remaining zero-shot tasks nor fine-tuning or GlobalPIQA, and does not establish absence of overlap.

## 7.2 Targeted augmentation

We compare the real-text baseline with the augmented condition across four matched seeds at 110.025M parameters (Table 6). The augmented condition scored higher on ZS16 (56.57 → 57.37). The largest language-specific mean difference is observed for Chinese, the lowest-scoring baseline language. The ZS16 and Chinese differences remain significant after Holm correction across the four reported metrics; the paired tests did not detect differences for English or Dutch.

Task-level records were uniquely recoverable for three seeds (7, 99, 123). Four of five Chinese tasks improved; leave-one-task-out means remained positive (+1.10 to +1.83; Appendix D). Across languages, 13 of 16 ZS16 tasks improved (Table 16). These are descriptive decompositions, not independent replicates or evidence against selection bias.

The comparison confounds content with 14.9% more training tokens and a refit tokenizer. It also combines synthetic and parallel text. We did not run exposure-matched real-text controls or the 2×2 design needed to separate the added sources.

## 7.3 Multi-token prediction

The +MTP condition adds  $k=4$  prediction heads on the shared trunk. Head  $j=1$  is the ordinary next-token head; heads  $j \geq 2$  apply a per-head lin-

ear projection and GELU before reusing the output embedding, and each of their losses enters the sum with weight 0.3. The three projections add 1.772M parameters, so the condition trains 111.797M parameters against the 110.025M of Table 1; at the final model’s hidden size they add 3.146M, for 215.936M. The projections are unused at evaluation, where the parameter counts are those listed. The number of active heads is ramped from 1 to 4 over the first four epochs, after which all four are active (Gloeckle et al., 2024). Only the  $j=1$  head is used at evaluation, so the model is evaluated as an ordinary causal decoder. Adding this objective to the augmented corpus changes ZS16 by +0.097 points across three matched seeds (95% CI  $[-0.617, 0.810]$ ;  $t(2)=0.583$ ; Holm-adjusted  $p=1.0$ ). No language-specific contrast is significant after correction. This comparison does not detect an improvement; it does not show that multi-token prediction is ineffective in general.

## 7.4 Single-run comparisons

With one run per parameter count, the 213M model had the highest observed FT22 score (41.25) against 39.81 at 110M, 39.90 at 263M and 40.15 at 338M, with ZS16 between 57.82 and 57.99. Among single-run ablations at 213M, three variants had higher ZS16 and lower FT22 than the reference run and the rest were lower on both. Both sets are in Appendix A. These are descriptive observations: they establish neither a scaling effect nor a relationship between the two summaries that

| Metric           | $\Delta$ | 95% CI          | $t(3)$ | raw $p$ | Holm $p$ |
|------------------|----------|-----------------|--------|---------|----------|
| ZS16 (task avg.) | +0.800   | [0.321, 1.279]  | 5.320  | .0130   | .0389    |
| English          | +0.808   | [-0.173, 1.788] | 2.622  | .0789   | .1578    |
| Dutch            | +0.235   | [-0.287, 0.757] | 1.433  | .2474   | .2474    |
| Chinese          | +1.473   | [1.238, 1.707]  | 19.996 | .00027  | .0011    |

Table 6: Augmented condition minus real-text baseline, four matched seeds. The four metrics form one Holm family; intervals are unadjusted. The two conditions differ in corpus content and in training exposure (Table 5).

holds across interventions. We do not test explanations for the co-occurrence of higher ZS16 with lower FT22.

## 7.5 Evaluated models

Two models were evaluated at 213M. Our submitted model used the augmented corpus of §7.1 and is listed on the official leaderboard as `babylm-2026-mtl-bigckpt` (ZS16 57.82, FT22 41.25); it obtained a leaderboard-reported Multilingual Average of 50.68. The second model, listed as `babylm-2026-mtl-stage2`, used a modified corpus in which part of the real text was replaced by additional synthetic text at matched language proportions, with the parallel text left intact. It has higher ZS16 (58.40) but lower FT22 (39.52) and a Multilingual Average of 50.14, consistent with the higher ZS16 and lower FT22 observed in three single-run ablations (§7.4). Appendix G lists the training runs behind the seed comparisons and the two evaluated models.

## 8 Conclusion

Across three fixed mixtures with matched seeds, we detected within-language differences of 2.487 points in English and 2.151 points in Chinese, but no aggregate difference after correction. The mixtures came from a single-seed pilot and cover different allocation ranges, which limits comparisons of sensitivity across languages. The inverse-score controller changed sampling shares without a detected aggregate improvement. Across four matched seeds the augmented condition scored higher on ZS16, most for Chinese, with positive differences on four of five Chinese tasks in the seeds whose task-level records survive. That condition also consumed 14.9% more tokens and used a tokenizer refit on its own corpus, so it does not separate content from exposure. Multi-token prediction showed no detected improvement across three seeds, and the remaining single-run comparisons are descriptive. The 213M model obtained a leaderboard-reported Multilingual Aver-

age of 50.68. These results support reporting both aggregate and language-level scores when comparing multilingual training conditions.

## Limitations

Seed counts are three or four for the replicated comparisons and one elsewhere; with  $n \leq 4$  the intervals are wide and normality cannot be checked. The public evaluation suite was reused for tokenizer, proxy, mixture, and augmentation selection, so the reported values are subject to selection bias. Proxy and full runs share a fixed 500-step warmup whose relative share differs between them.

The three fixed mixtures were selected as the three highest-scoring conditions of a single-seed pilot ranked on `TASK-7`, the same aggregate on which §5.2 then compares them, so that comparison is subject to selection bias. The pilot itself used one seed per mixture at a smaller scale and with a different tokenizer, so its ranking is not a reliable ordering. The retained mixtures vary each language over a different range (§5.1), and the symmetric conditions 25/50/25 and 25/25/50 exist only as single pilot runs.

The augmented corpus is additive, so the augmentation comparison confounds content with a 14.9% increase in training exposure and with a tokenizer refit on each corpus, and it combines synthetic and parallel text without separating them (§7.1). Both corpora are inside the track’s byte budget, the augmented one at 97.2% of the cap, so the design leaves little room to add data without displacing some.

The overlap audit covers the distributed English BLiMP, Dutch BLiMP-NL, and Chinese ZhoBLiMP files described in Appendix I, but not the other zero-shot or fine-tuning tasks, or GlobalPIQA. For one of the four augmentation seeds the per-task record could not be uniquely recovered, because its results directory accumulated several evaluation passes. The seed-level language means of Table 11 are unaffected, but the task-level decomposition in Appendix D uses the other three

seeds. Leaderboard values are those the leaderboard reported for these entries; the leaderboard remains open.

## Ethical Considerations

The synthetic portion of our corpus was accepted without filtering (§7.1): we did not verify that every sample is in the intended language, factually correct, or free of social bias, and a model trained on it inherits whatever the teacher produced. The teacher is used only to generate text, under the terms of its published model card, and we record its immutable revision so that the generation can be traced. The parallel text is community-contributed and contains noisy sentences. Training and evaluation data are used under the licences of their sources, and the evaluation suites are the ones distributed by the track.

## References

- Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. [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 @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.
- Kaj Bostrom and Greg Durrett. 2020. [Byte pair encoding is suboptimal for language model pretraining](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. [When is multilinguality a curse? language modeling for 250 high- and low-resource languages](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. [BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop](#). *Preprint*, arXiv:2602.20092.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Achille Fusco, Maria Letizia Piccini Bianchessi, Tommaso Sgrizzi, Asya Zanollo, and Cristiano Chesi. 2025. [Linguistic units as tokens: Intrinsic and extrinsic evaluation with BabyLM](#). In *Proceedings of the First BabyLM Workshop*, pages 496–507, Suzhou, China. Association for Computational Linguistics.
- Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. [Better & faster large language models via multi-token prediction](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15706–15734. PMLR.
- Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. 2025. [XCOMPS: A multilingual benchmark of conceptual minimal pairs](#). In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 75–81, Vienna, Austria. Association for Computational Linguistics.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.
- Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrana, Francesca Padovani, François Meyer, Hai Hu, Julien Etxaniz, Laurent Prévot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, and 7 others. 2026a. [BabyBabelLM: A multilingual benchmark of developmentally plausible training data](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3297–3329, Rabat, Morocco. Association for Computational Linguistics.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026b. [MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *Transactions of the Association for Computational Linguistics*, 14:193–216.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.
- Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui Wang, and Hai Hu. 2026. [A systematic assessment of language models with linguistic minimal pairs in Chinese](#). *Transactions of the Association for Computational Linguistics*, 14:755–771.

Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI Technical Report*.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, 51(4):1267–1301.

Jörg Tiedemann. 2020. [The Tatoeba translation challenge – realistic data sets for low resource and multilingual MT](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Alex Warstadt, 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.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. 2023. [DoReMi: Optimizing data mixtures speeds up language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 36.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. [Improving massively multilingual neural machine translation and zero-shot translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

## A Single-run comparisons

| Parameters | ZS16  | FT22  |
|------------|-------|-------|
| 110M       | 57.86 | 39.81 |
| 213M       | 57.82 | 41.25 |
| 263M       | 57.99 | 39.90 |
| 338M       | 57.86 | 40.15 |

Table 7: One run per parameter count. The 110M row is a single run of the augmented-plus-MTP condition and is not the three-seed mean of §7.

| Variant (vs 213M reference run)          | $\Delta$ ZS16 | $\Delta$ FT22 |
|------------------------------------------|---------------|---------------|
| Hybrid morpheme tokenizer                | +0.50         | −1.47         |
| Substituted synthetic text               | +0.58         | −1.73         |
| Hybrid tokenizer + substituted synthetic | +0.42         | −2.38         |
| Block length 256                         | −0.32         | −0.74         |
| Block length 512                         | −0.40         | −1.54         |
| Character-level vocabulary               | −0.78         | −0.92         |
| Additional translation text              | −0.22         | −2.32         |

Table 8: Single-run ablations at 213M, reported as differences from the reference run.

## B Seed-level results

| Mixture      | Seed | Task-7 | Lang-3 |
|--------------|------|--------|--------|
| 50/25/25     | 1    | 72.705 | 72.591 |
|              | 2    | 72.380 | 72.322 |
|              | 3    | 72.473 | 72.378 |
| 40/40/20     | 1    | 72.405 | 72.112 |
|              | 2    | 72.883 | 72.539 |
|              | 3    | 73.018 | 72.692 |
| 27.5/27.5/45 | 1    | 71.942 | 71.768 |
|              | 2    | 72.748 | 72.626 |
|              | 3    | 72.075 | 71.924 |

Table 9: Seed-level aggregate scores, fixed-mixture experiment.

| Method   | Seed Task-7 scores       | Mean   | SD   |
|----------|--------------------------|--------|------|
| Fixed    | 65.073 / 65.966 / 65.679 | 65.573 | .456 |
| Inverse  | 65.361 / 65.416 / 65.397 | 65.391 | .028 |
| Smoothed | 65.729 / 65.586 / 64.885 | 65.400 | .452 |

Table 12: Seed-level TASK-7 scores, dynamic experiment.

## C Dynamic experiment details

| Lang | Benchmark  | Fixed  | Inverse | Smoothed |
|------|------------|--------|---------|----------|
| en   | BLiMP      | 59.153 | 59.198  | 59.204   |
| en   | MultiBLiMP | 75.974 | 74.935  | 76.494   |
| nl   | BLiMP-NL   | 70.702 | 70.691  | 70.294   |
| nl   | MultiBLiMP | 84.299 | 84.842  | 84.456   |
| nl   | XCOMPS     | 51.014 | 50.524  | 50.571   |
| zh   | ZhoBLiMP   | 66.005 | 66.197  | 65.351   |
| zh   | XCOMPS     | 51.862 | 51.354  | 51.433   |

Table 13: Per-benchmark three-seed means for the dynamic experiment. These seven scores are both the control signal and the reported outcome; the language means of Table 4 are their unweighted within-language averages.

| Mixture      | Seed | English | Dutch  | Chinese |
|--------------|------|---------|--------|---------|
| 50/25/25     | 1    | 78.368  | 73.388 | 66.017  |
|              | 2    | 79.538  | 72.726 | 64.703  |
|              | 3    | 78.746  | 73.049 | 65.338  |
| 40/40/20     | 1    | 77.616  | 74.165 | 64.555  |
|              | 2    | 77.952  | 74.949 | 64.716  |
|              | 3    | 78.494  | 74.976 | 64.606  |
| 27.5/27.5/45 | 1    | 75.716  | 72.983 | 66.606  |
|              | 2    | 77.051  | 73.475 | 67.353  |
|              | 3    | 76.423  | 72.978 | 66.371  |

Table 10: Seed-level language means for the fixed-mixture experiment. These are the scores behind Table 2 and the within-language contrasts of §5.3, and the rows of Table 17 that report them.

| Condition       | Seed | ZS16  | English | Dutch | Chinese |
|-----------------|------|-------|---------|-------|---------|
| Baseline        | 42   | 56.81 | 59.01   | 59.18 | 51.77   |
|                 | 7    | 56.61 | 58.27   | 59.12 | 51.94   |
|                 | 99   | 56.55 | 58.61   | 58.87 | 51.70   |
|                 | 123  | 56.32 | 58.08   | 58.61 | 51.81   |
| Augmented       | 42   | 57.37 | 59.12   | 59.32 | 53.29   |
|                 | 7    | 57.66 | 59.76   | 59.32 | 53.56   |
|                 | 99   | 57.07 | 59.12   | 58.78 | 52.97   |
|                 | 123  | 57.39 | 59.20   | 59.30 | 53.29   |
| Augmented + MTP | 7    | 57.44 | 58.98   | 59.55 | 53.36   |
|                 | 99   | 57.41 | 59.53   | 59.68 | 52.56   |
|                 | 123  | 57.56 | 59.30   | 59.80 | 53.14   |

Table 11: Seed-level scores for the data-design comparisons. Augmentation compares Augmented with Baseline on four matched seeds (42, 7, 99, 123). The MTP comparison uses Augmented + MTP versus Augmented on three matched seeds (7, 99, 123).

| Method   | Seed | en    | nl    | zh    |
|----------|------|-------|-------|-------|
| Inverse  | 1    | .3212 | .3137 | .3651 |
|          | 2    | .3217 | .3156 | .3627 |
|          | 3    | .3197 | .3113 | .3690 |
| Smoothed | 1    | .3242 | .3165 | .3593 |
|          | 2    | .3209 | .3171 | .3620 |
|          | 3    | .3182 | .3136 | .3682 |

Table 14: Sampling shares applied over the final two epochs. The fixed condition holds 1/3 throughout. All values lie inside the  $[0.10, 0.60]$  bounds, so the projection step was never active.

Each run takes 192 optimizer steps per epoch at an effective batch of 256 sequences, so 10 epochs are 1,920 steps. The block pool holds 19,531 English, 20,539 Dutch, and 9,082 Chinese blocks of 512 tokens, and each two-epoch segment draws 98,304 blocks, for  $2.52 \times 10^8$  tokens over the run. The pool itself was built at 40/40/20 with the Byte Premium applied to its per-language token caps. The equal start of §6 is therefore a statement about sampling probabilities, not about the pool being balanced, and the controller reweights sampling without reapplying the premium. The trainer is rebuilt at each segment boundary but resumes from the preceding checkpoint, and the optimizer mo-

ments, the scheduler position, the global step, and the training RNG state are restored alongside the weights. We checked this against the five stored checkpoints of one run: the optimizer step, the scheduler’s `last_epoch`, and `global_step` agree at 384, 768, 1152, 1536, and 1920, with `max_steps` equal to 1920 throughout. Thus one cosine schedule spans the whole run and the run is not five independent two-epoch runs. Restoring the RNG state does not make the data order identical to an uninterrupted run, because each segment draws a fresh sampling plan from a separate generator seeded by the run seed and the segment index.

## D Chinese task-level decomposition

| Chinese task   | Baseline | Augmented | $\Delta$ |
|----------------|----------|-----------|----------|
| ZhoBLiMP       | 78.13    | 79.43     | +1.29    |
| Winogrande-zh  | 49.77    | 51.11     | +1.35    |
| XStoryCloze-zh | 47.29    | 50.18     | +2.89    |
| HellaSwag-zh   | 27.36    | 27.34     | -0.02    |
| XCOMPS-zh      | 56.53    | 58.32     | +1.79    |
| Chinese mean   | 51.82    | 53.28     | +1.46    |

Table 15: Per-task means over the three matched seeds (7, 99, 123) whose per-task records are uniquely recoverable. Leave-one-task-out Chinese means range from +1.10 to +1.83. These five scores come from the same runs and are not independent replicates.

| Lang | Task                    | $\Delta$ |
|------|-------------------------|----------|
| en   | BLiMP (BabyLM-filtered) | +3.107   |
| en   | MultiBLiMP              | +0.867   |
| en   | Winogrande              | +0.875   |
| en   | XStoryCloze             | +0.591   |
| en   | HellaSwag               | -0.256   |
| nl   | BLiMP-NL                | -0.074   |
| nl   | MultiBLiMP              | +0.055   |
| nl   | Winogrande              | +0.550   |
| nl   | HellaSwag               | +0.239   |
| nl   | XCOMPS                  | +0.566   |
| nl   | XStoryCloze             | +0.233   |
| zh   | ZhoBLiMP                | +1.292   |
| zh   | Winogrande              | +1.347   |
| zh   | XStoryCloze             | +2.889   |
| zh   | HellaSwag               | -0.024   |
| zh   | XCOMPS                  | +1.793   |

Table 16: Augmented minus baseline mean difference on each of the 16 ZS16 tasks, over the three seeds (7, 99, 123) with uniquely recoverable task-level records. Thirteen of the sixteen differences are positive. These are the same runs, not independent replicates.

## E Generation prompts and data examples

The four synthetic-text templates, with  $\{t\}$  the sampled topic and  $\{L\}$  the target language name:

**STORY:** *Write a simple everyday story of 4 short sentences about  $\{t\}$ . Use simple words a child knows. Give it a clear, natural ending. Output only the story in  $\{L\}$ .*

**COMPARE:** *Write 3 simple sentences that compare everyday things related to  $\{t\}$ , using words like bigger, smaller, faster, older, heavier. Output only the sentences in  $\{L\}$ .*

**COREF:** *Write 3 simple sentences about  $\{t\}$  that use pronouns (he, she, it, they) clearly referring to specific people or things. Output only the sentences in  $\{L\}$ .*

**FACT:** *Write 4 simple true everyday facts about  $\{t\}$  that a young child would understand. Output only the facts in  $\{L\}$ .*

Every prompt issued produced one accepted sample, so the counts are 6,250 per template per language, 25,000 per language, and 75,000 in total. The parallel data contributes 240,952 sequences from Tatoeba and 298,684 from OPUS-100, split 319,296 en-nl and 220,340 en-zh. The 143 topics are:

the kitchen, animals, the weather, school, a birthday party, the beach, cooking, gardening, a train trip, the market, sports, music, friendship, the forest, morning routine, a rainy day, the doctor, shopping, a picnic, the library, farm life, the city, a snowy day, helping at home, a lost dog, the playground, baking bread, a family dinner, the zoo, fixing a bike, the ocean, mountains, rivers, insects, birds, flowers, trees, vegetables, fruits, the moon, stars, the sun, seasons, autumn leaves, spring, summer holidays, winter, a new toy, building blocks, drawing pictures, painting, reading books, puzzles, the school bus, the classroom, a math lesson, recess, the science fair, a soccer match, swimming, riding bikes, jumping rope, hide and seek, grandma’s house, grandpa’s stories, cousins visiting, a baby brother, pets at home, a cat and a mouse, a dog and a bone, a bird and a nest, fish in a pond, rabbits, the bakery, the fire station, the police, a nurse, a farmer, a fisherman, a pilot, the post office, the bank, a restaurant, the supermarket, a toy shop, a camping trip, a boat ride, a plane journey, a road trip, visiting a museum, the circus, a magic show, a puppet show, a parade, fireworks, cleaning the room, washing dishes, folding clothes, setting the table, making the bed, planting seeds, watering plants, picking apples, raking leaves, shoveling snow, a thunderstorm, a foggy morning, a windy day, a hot afternoon, a cold night, sharing toys, making friends, saying sorry,

helping a neighbor, being kind, the alphabet, counting numbers, colors and shapes, days of the week, telling time, a broken toy, a spilled drink, a scraped knee, a lost mitten, a torn shirt, cooking soup, making a sandwich, baking cookies, pouring juice, eating breakfast, a butterfly, a bee, an ant, a spider, a ladybug, a frog, a turtle, a snail, the playground slide, the swings, the sandbox, the seesaw, the merry-go-round.

Generated examples follow, verbatim from the stored generation files. No example was edited for fluency, and the English gloss of the Chinese sample is provided for readers only:

[en, STORY, *the alphabet*] *Leo held a big book with letters A to Z. He read the story about a cat and a dog. Then he smiled when he saw his name at the end. Leo felt very happy because he could read his own words.*

[nl, COREF, *cooking*] *Hij kookt de pasta. Zij snijdt de tomaten. Het gerecht smaakt lekker.*

[zh, STORY, *seasons*] 春天来了, 小草从土里探出头。夏天到了, 知了在树上大声唱歌。秋天到了, 树叶变成了金黄色。冬天来了, 雪花落在屋顶上, 世界变得白茫茫的。

*Chūntiān lái le, xiǎo cǎo cóng tǔ lǐ tàn chū tóu. Xiàtiān dào le, zhīliǎo zài shù shàng dàshēng chàngē. Qiūtiān dào le, shù yè biànràng le jīnhuángsè. Dōngtiān lái le, xuěhuā luò zài wǔdǐng shàng, shìjiè biàn de báimángmáng de.*

Gloss: *Spring came and the grass poked its head out of the soil. Summer arrived and the cicadas sang loudly in the trees. In autumn the leaves turned golden. Winter came, snow fell on the roofs, and the world turned white.*

Parallel examples, shown as they appear in the training corpus:

[Tatoeba, en-nl] English: Mickey Mouse is a mouse. \n Dutch: Mickey Mouse is een muis.

[OPUS-100, en-nl] English: Because you're where? \n Dutch: Omdat je waar bent?

[Tatoeba, en-zh] Chinese: 我不信任汤姆而且汤姆也不信任我。

\n English: I didn't trust Tom and he didn't trust me, either. (*Wǒ bù xìnren Tāngmǔ érqiě Tāngmǔ yě bù xìnren wǒ.*)

## F Statistical tests

Table 17 lists every seed-matched test of the data-design study, recomputed in a single pass from the seed-level scores shown in Table 11. Augmentation uses four matched seeds; MTP compares Augmented + MTP with Augmented on seeds 7, 99, and 123. Within each intervention the four metrics form one Holm family. Intervals are unadjusted 95% *t* intervals for the paired mean difference.

The allocation-study differences, intervals, and Holm-adjusted values retain the precision of the reported analyses in §5.2–6. Raw *p* values are approximate recomputations from the rounded seed-level scores in Appendix B. These rounded inputs do not always reproduce the last printed digit of the retained statistics. Each language forms its own Holm family of three fixed-mixture contrasts.

## G Reproducibility index

Table 18 identifies the training runs behind the data-design results. Corpus hashes show the first 16 hexadecimal digits of SHA-256; tokenizer hashes in Table 18 show the first eight. The two corpora are f9917fdfa47ac9ef (baseline) and c700f38d1e08da64 (augmented). All runs use 10 epochs and 16,384 tokens per optimizer step; step counts differ because the corpora differ in size. The tokenizer column shows that the baseline and augmented conditions do not share a vocabulary.

The submitted 213M model is released at <https://huggingface.co/lghub/babylm-2026-mtl-bigckpt>, and the organizers' data-preparation code cited in §7.1 is at <https://github.com/babylm-org/multilingual-babylm>. Zero-shot and fine-tuning scores were produced with the official BabyLM multilingual evaluation pipeline at commit 6f825c29. Fine-tuning uses a batch of 64, a learning rate of 5e−5, at most 10 epochs with per-epoch evaluation, early stopping with patience 3, and seed 12. The checkpoint kept is the one selected by that early-stopping criterion, not the last. FT22 averages the resulting 22 non-POS task–language scores.

| Intervention   | Metric                                         | $n$ | $\Delta$ | 95% CI         | raw $p$ | Holm $p$ |
|----------------|------------------------------------------------|-----|----------|----------------|---------|----------|
| Augmentation   | ZS16                                           | 4   | +0.800   | [.32, 1.28]    | .0130   | .0389    |
|                | English                                        | 4   | +0.808   | [−.17, 1.79]   | .0789   | .1578    |
|                | Dutch                                          | 4   | +0.235   | [−.29, .76]    | .2474   | .2474    |
|                | Chinese                                        | 4   | +1.473   | [1.24, 1.71]   | .0003   | .0011    |
| +MTP           | ZS16                                           | 3   | +0.097   | [−.62, .81]    | .6188   | 1.000    |
|                | English                                        | 3   | −0.090   | [−1.62, 1.44]  | .8242   | 1.000    |
|                | Dutch                                          | 3   | +0.543   | [−.29, 1.38]   | .1079   | .3450    |
|                | Chinese                                        | 3   | −0.253   | [−.60, .09]    | .0863   | .3450    |
| Fixed mixtures | TASK-7, one family of three contrasts, $n = 3$ |     |          |                |         |          |
|                | 40/40/20 − 50/25/25                            |     | +0.250   | [−.93, 1.43]   | .46     | .9191    |
|                | 40/40/20 − 27.5/27.5/45                        |     | +0.514   | [−.50, 1.52]   | .16     | .4792    |
|                | 50/25/25 − 27.5/27.5/45                        |     | +0.265   | [−1.17, 1.70]  | .51     | .9191    |
| Dynamic        | TASK-7, one family of three contrasts, $n = 3$ |     |          |                |         |          |
|                | Fixed − inverse                                |     | +0.181   | [−.88, 1.24]   | .54     | 1.000    |
|                | Fixed − smoothed                               |     | +0.173   | [−1.68, 2.03]  | .73     | 1.000    |
|                | Inverse − smoothed                             |     | −0.009   | [−1.16, 1.14]  | .98     | 1.000    |
| Fixed, English | 50/25/25 − 40/40/20                            | 3   | +0.863   | [−.81, 2.54]   | .157    | .157     |
|                | 50/25/25 − 27.5/27.5/45                        | 3   | +2.487   | [2.08, 2.90]   | .0015   | .0044    |
|                | 40/40/20 − 27.5/27.5/45                        | 3   | +1.623   | [.05, 3.19]    | .047    | .094     |
| Fixed, Dutch   | 50/25/25 − 40/40/20                            | 3   | −1.642   | [−3.54, .26]   | .065    | .130     |
|                | 50/25/25 − 27.5/27.5/45                        | 3   | −0.091   | [−1.57, 1.38]  | .815    | .815     |
|                | 40/40/20 − 27.5/27.5/45                        | 3   | +1.551   | [.53, 2.58]    | .023    | .069     |
| Fixed, Chinese | 50/25/25 − 40/40/20                            | 3   | +0.727   | [−1.11, 2.56]  | .230    | .302     |
|                | 50/25/25 − 27.5/27.5/45                        | 3   | −1.424   | [−4.12, 1.27]  | .151    | .302     |
|                | 40/40/20 − 27.5/27.5/45                        | 3   | −2.151   | [−3.26, −1.05] | .014    | .042     |

Table 17: All seed-matched tests reported in this paper. Two-sided paired  $t$ -tests. Augmentation uses seeds 42, 7, 99, 123; +MTP uses seeds 7, 99, 123, so those two families are not nested. The data-design rows are recomputed from Table 11. Allocation-study differences, intervals, and Holm values are retained as reported in §5.2–6. Their raw  $p$  values are approximate checks using the rounded scores in Tables 9, 10, and 12; the last printed digits need not agree exactly. Each language forms its own Holm family of three fixed-mixture contrasts.

## H Pilot and corpus details

Tables 19 and 20 report the exploratory pilots; Table 21 records corpus composition. In the mixture pilot, Dutch varies by 1.273 points, against 4.680 for English and 3.255 for Chinese. The Dutch-high mixture did not exceed the equal mixture on Dutch. Its different tokenizer and training scale prevent this observation from explaining the main experiment’s Dutch non-detection.

## I Overlap audit details

We screened the training corpus against the evaluation sentences of the minimal-pair files distributed with the track: 67 BLiMP paradigms for English (59,875 pairs), 84 BLiMP-NL paradigms for Dutch (8,400 pairs), and 11 ZhoBLiMP paradigms for Chinese (3,300 pairs), or 119,750, 16,800, and 6,600 sentences counting both members of each pair. The Chinese files are a subset of ZhoBLiMP rather than the whole suite, so the Chinese audit is the narrowest of the three. Training lines and evaluation sentences were normalized with Unicode

NFKC, case folding, and whitespace collapsing; Chinese was additionally stripped of whitespace. The audit has two passes. The first compares whole normalized strings for equality, so it does not depend on any length threshold. The second flags a line that shares a word 8-gram (English, Dutch) or a character 8-gram (Chinese) with an evaluation sentence and scores each flagged pair by character 5-gram Jaccard similarity, reporting counts at .90, .80, and .70.

Across 9,853,776 line–language checks the first pass returns three matching training lines, corresponding to two distinct evaluation sentences. These are one English line in the real BabyBabelLM text (*The Eye is closing.*) and two Chinese lines in the parallel data that are copies of the same five-character sentence. The second pass returns no additional pair at any threshold. We retained the original evaluation sets. Scores were not recomputed with the corresponding pairs excluded. No synthetic line matched any evaluation sentence. Both reported exact matches are shorter

| Condition      | Corpus    | Tokenizer | Seeds          | Steps  |
|----------------|-----------|-----------|----------------|--------|
| Baseline 110M  | baseline  | 235dfa07  | 42, 7, 99, 123 | 76,440 |
| Augmented 110M | augmented | bba937b7  | 42, 7, 99, 123 | 87,820 |
| +MTP 110M      | augmented | bba937b7  | 7, 99, 123     | 87,820 |
| Submitted 213M | augmented | bba937b7  | 42             | 87,820 |
| Second 213M    | modified* | cd59b082  | 42             | 88,660 |

Table 18: Training runs behind the seed comparisons and the two evaluated models. \*This variant replaces part of the real text with additional synthetic text at matched language proportions, keeping the parallel text intact. The table does not cover the parameter-count runs of Table 7 or the single-run ablations of Table 8.

| Tokenizer | Vocab. | Params   | Task-7 | Lang-3 | All-16 |
|-----------|--------|----------|--------|--------|--------|
| BPE       | 16,000 | 97.737M  | 71.806 | 71.540 | 54.651 |
| Unigram   | 16,000 | 97.737M  | 71.558 | 71.271 | 54.606 |
| MorPiece  | 39,697 | 115.937M | 73.143 | 72.940 | 55.843 |

Table 19: Tokenizer pilot, one training seed per condition, equal language mixture. Vocabulary and total parameter count vary with the tokenizer.

than the 8-gram window, so they are found by the first pass alone; conversely, the near-duplicate counts are conditional on the 8-gram candidate filter, and we did not measure its recall. These counts therefore do not establish the absence of overlap. The audit covers the distributed files named above and not the other ZS16 benchmarks (MultiBLiMP, XCOMPS, Winogrande, XStoryCloze, and Hel-laSwag), the fine-tuning suites, or GlobalPIQA.

| Mixture (en/nl/zh)        | English | Dutch  | Chinese | Task-7 | Lang-3 |
|---------------------------|---------|--------|---------|--------|--------|
| 40/40/20 <sup>‡</sup>     | 71.470  | 69.103 | 58.955  | 66.880 | 66.509 |
| 27.5/27.5/45 <sup>‡</sup> | 68.335  | 69.327 | 61.610  | 66.839 | 66.424 |
| 50/25/25 <sup>‡</sup>     | 71.340  | 68.527 | 59.565  | 66.770 | 66.477 |
| 25/50/25                  | 70.155  | 69.330 | 58.370  | 66.434 | 65.952 |
| Equal (1/3 each)          | 67.960  | 69.800 | 59.840  | 66.429 | 65.867 |
| 30/30/40                  | 68.245  | 68.927 | 60.650  | 66.367 | 65.941 |
| 25/25/50                  | 66.790  | 68.633 | 61.625  | 66.104 | 65.683 |

Table 20: Mixture selection pilot: one seed per mixture at the 25M-word proxy with the MorPiece tokenizer, ordered by TASK-7. <sup>‡</sup>Promoted to the three-seed experiment of §5. Scores are on the 25M-word proxy scale and are not comparable to Table 19 or Table 2, both of which use the full budget.

| Source                       | Lang | Documents | Raw words | Eq. words |
|------------------------------|------|-----------|-----------|-----------|
| BabyBabelLM (real)           | en   | 3,698,322 | 34.000M   | 34.000M   |
|                              | nl   | 4,417,520 | 34.703M   | 33.000M   |
|                              | zh   | 583,662   | 30.692M   | 32.792M   |
| Synthetic (teacher)          | en   | 25,000    | 0.914M    | 0.914M    |
|                              | nl   | 25,000    | 0.787M    | 0.748M    |
|                              | zh   | 25,000    | 1.268M    | 1.355M    |
| Parallel (Tatoeba, OPUS-100) | en   | 539,636   | 2.271M    | 2.271M    |
|                              | nl   | —         | 1.242M    | 1.181M    |
|                              | zh   | —         | 1.340M    | 1.432M    |
| Baseline total               |      | 8,699,504 | 99.395M   | 99.792M   |
| Augmented total              |      | 9,314,140 | 107.217M  | 107.692M  |

Table 21: Corpus composition. Raw words are whitespace tokens for English and Dutch and CJK characters for Chinese; equivalent words apply the official 2026 Byte Premiums from the organizers’ coefficient table (English 1.0, Dutch 1.0516, simplified Chinese 0.93597; the call for papers lists 0.9894, the coefficient given there for traditional Chinese). The parallel rows are 539,636 training sequences built from 269,818 translation pairs, one sequence per direction; their word counts are given once per pair, so per-epoch exposure is twice the listed value. The equivalent-word column follows the construction-time word proxy of Table 5, which is not the quantity the track caps.