

Pairing Original Text with Meaning-Preserving Rewrites for Sample-Efficient Small-Data Language Model Pretraining

Shaoxiang Wu¹

Alexander Fraser^{1,2,3}

Lukas Edman^{1,2}

¹School of Computation, Information and Technology, TU Munich

²Munich Center for Machine Learning

³Munich Data Science Institute

shaoxiang.wu@tum.de, lukas.edman@tum.de

Abstract

We investigate whether augmenting training data with LLM-generated rewrites can improve performance in small-data language model pretraining. We pair each training sentence with a meaning-preserving rewrite, so that masked tokens can be predicted using both the original and its paired version in the same training window. On FineWeb, we test three rewrite types: *simplification* (shorter and easier), *explanation* (longer and easier), and *paraphrasing* (similar length but harder to read). We also test simplification on OpenStax+BCCampus and Simple Wikipedia. We add a reverse masking curriculum: 7 epochs of whole-word masking (WWM), where entire words are masked instead of individual subword tokens, followed by 3 epochs of standard token-level masking (WWM7→Token3). The masking procedure is not rewrite-aware: it does not use alignments between the original sentence and its rewrite. All three FineWeb rewrite types improve EWoK and GLUE-style fine-tuning over source-matched baselines, and simplification improves Entity Tracking across all three sources. The EWoK and GLUE gains concentrate on concept-swap and sentence-pair tasks, suggesting that rewrite pairing especially helps models link meaning across different surface forms. Despite using only one-tenth as much training data, our best model surpasses the official BabyLM 100M GPT-2 baseline on EWoK, Entity Tracking, and GLUE, suggesting that paired rewrite augmentation can improve sample efficiency under a fixed small-data budget.

1 Introduction

Language models need far more data than humans do to acquire comparable linguistic ability (Warstadt and Bowman, 2022): a 13-year-old has encountered fewer than 100 million words, orders of magnitude less than typical LM pretraining corpora. The BabyLM Challenge and Workshop series

(Warstadt et al., 2023; Choshen et al., 2024; Charpentier et al., 2025; Choshen et al., 2026) studies this gap by constraining pretraining to a developmentally plausible scale, including 10M words in the 2026 STRICT-SMALL track. Edman et al. (2024) recently showed that pairing original sentences with paraphrase and hard-negative rewrites can substantially improve GLUE-style fine-tuning under this budget, providing early evidence that transforming sentences and arranging related variants together can matter under a fixed small-data budget.

In this work, we investigate **meaning-preserving rewrite pairing** as a form of data augmentation under a fixed word budget. We pair each training sentence with a rewrite so that masked tokens can be predicted from both the original and the paired version. For EWoK and GLUE-style fine-tuning, we find gains across all three FineWeb rewrite types, and simplification replicates this pattern across FineWeb, OpenStax+BCCampus, and Simple Wikipedia. These gains concentrate on EWoK concept swap items and GLUE sentence-pair tasks — both of which require linking meaning across different expressions. This suggests that the clearest benefit is not a uniform improvement across all evaluations, but an improved ability to link meaning across different surface forms. We also find that simplification improves Entity Tracking across the same three sources, although this effect is more specific to simplification than the broader EWoK and GLUE gains. Adding a whole-word masking curriculum (WWM7→Token3: 7 epochs masking whole words, 3 epochs masking individual tokens) further strengthens these gains in most conditions.

We make five contributions: (i) a **meaning-preserving rewrite pairing** pretraining strategy, using simplification, explanation, and harder paraphrasing on FineWeb, and simplification on OpenStax+BCCampus and Simple Wikipedia; (ii) evidence that **EWoK and GLUE gains hold across**

Source ^a	Origin	Expected strength
FineWeb	Web sentences from FineWeb-Edu (Penedo et al., 2024)	Broad science-popular and educational web text, covering health/medicine, biology/nature, geography, history/politics, education, and physical science
OpenStax+BCCampus	University-level open textbooks from OpenStax and BCCampus	Education and humanities
Simple Wikipedia	Simple Wikipedia sentences	Simple expressions, broad but shallow topical coverage

Table 1: Three simplification source corpora.

^aSource websites: [FineWeb-Edu](#), [OpenStax](#), [BCCampus](#), and [Simple Wikipedia](#).

Original: Enlightenment thinkers proposed that human reason, coupled with empirical study of the physical world, would lead to progress—the advancement of science and the improvement of the human condition.

Simplified: Enlightenment thinkers believed that using reason and studying the world would lead to scientific progress and better living conditions.

Explanation: Enlightenment thinkers thought that using reason and studying the world could make things better. They believed that reason would help science move forward and improve people’s lives.

Paraphrase: Enlightenment thinkers believed reason and science would lead to progress and a better society.

Table 2: The same FineWeb original rewritten under all three of our rewrite types. Simplification uses simpler wording and a more direct sentence structure; explanation unpacks the content into a more explicit restatement; paraphrase preserves meaning while using different wording and more formal vocabulary, with no instruction to simplify.

three rewrite types on FineWeb and, for simplification, across all three sources; (iii) a more selective **Entity Tracking improvement across three sources of Simplification**, while the other FineWeb rewrite types do not show the same benefit; (iv) a **whole-word masking curriculum** (WWM7→Token3) combined with rewrite pairing, which further strengthens gains in most conditions; and (v) a demonstration of **improved sample efficiency**: our 10M-word model surpasses the official BabyLM 100M GPT-2 baseline on EWoK, Entity Tracking, and GLUE.

2 Data Construction

2.1 Three simplification corpora

We construct sentence-level simplification data from three sources, summarized in Table 1.

For each source, an offline rewrite generator, Qwen2.5-7B-Instruct (Qwen Team, 2024), served with vLLM (Kwon et al., 2023), produces a one-line simplified version of each input sentence, instructed to use simpler vocabulary and shorten the sentence without omitting key information (prompt in Appendix B, Figure 1). We concatenate `original + simplified` for each pair, shuffle with a fixed seed, and truncate to a fixed 10M-word budget. For each source, we construct a raw/original-only 10M corpus and a simplified 10M corpus. This setup is not model-level knowl-

edge distillation: the rewrite generator is used only for offline corpus construction and is absent during pretraining. The pretrained model receives no generator logits, probability distributions, hidden representations, task labels, or online supervision; it is trained from scratch with the standard MLM objective on the resulting fixed-budget corpus. We therefore characterize the intervention as offline text augmentation through meaning-preserving rewriting.

Table 2 contrasts all three of our rewrite types — simplification, explanation, and paraphrase — on the same FineWeb original.

Why sentence-level. We use a 64-token context window for training, so sentence-level pairs usually fit in a single sequence. This means a masked word in one sentence can be predicted using the other sentence in the pair. Paragraph-level pairs would often be too long to fit in one window, reducing this effect.

Simplification. Simplification is our main rewrite type and is applied to all three source corpora. It preserves the original sentence’s meaning while making the wording shorter and easier to read, mainly through simpler vocabulary and more direct phrasing. This gives the model two aligned expressions of the same content: the original sentence, which retains the source style and domain wording, and a simplified

Source	N	Ratio	Lex. Ov.	FRE		FKGL		DaleC		DiffWd	
				Orig.	Simp.	Orig.	Simp.	Orig.	Simp.	Orig.	Simp.
FineWeb	163,150	0.9	0.5	43.9	58.9	13.1	9.7	11.0	10.0	0.3	0.2
OpenStax+BCCampus	179,165	0.6	0.4	39.6	50.2	13.6	10.1	10.9	10.6	0.3	0.3
Simple Wikipedia	331,074	0.8	0.6	52.8	58.8	11.1	8.9	10.8	10.8	0.3	0.3

Table 3: Difficulty metrics for original vs. simplified text, computed over each source’s pairs (OpenStax and BCCampus pooled together). Ratio and Lex. Ov. compare the simplified sentence to its original; FRE, FKGL, DaleC, and DiffWd are reported separately for the original and the simplified side of each pair. Higher FRE is easier; lower FKGL, DaleC, and DiffWd are easier. Bold marks the easier side for each readability metric.

Rewrite	N	Ratio	Lex. Ov.	FRE		FKGL		DaleC		DiffWd	
				Orig.	Rewr.	Orig.	Rewr.	Orig.	Rewr.	Orig.	Rewr.
Simplification	163,150	0.9	0.5	43.9	58.9	13.1	9.7	11.0	10.0	0.3	0.2
Explanation	162,478	1.1	0.4	43.9	67.7	13.1	7.7	11.0	9.1	0.3	0.2
Paraphrase	159,932	1.0	0.4	43.9	31.4	13.1	14.5	11.0	12.1	0.3	0.4

Table 4: Difficulty metrics for FineWeb original vs. rewrite text, for each of our three rewrite types. Same metric definitions as Table 3. Bold marks the easier side for each readability metric.

version, which exposes the same concepts in a more accessible form. We use this setting to test whether pairing original text with an easier meaning-preserving rewrite improves learning beyond training on the original source text alone.

Explanation. To test whether the benefit requires making text shorter, we also generate *explanation* rewrites on FineWeb. Explanation rewrites are longer than the original (they add context rather than compressing), but still aim to make the text easier to read. If explanation also helps, the benefit is not about shortening but about making the text simpler.

Paraphrasing. To test whether the benefit requires the rewrite to be easier to read, we also generate *paraphrases* on FineWeb. Paraphrases keep the same meaning and similar length, but change the wording and sentence structure. As confirmed in §3, the paraphrases are actually *harder* to read than the originals. If paraphrasing also helps, then the benefit does not require making the text easier — just different.

2.2 Raw source-matched baselines

Each corpus is paired with a source-matched raw Token and WWM curriculum baselines of equal size, which lets us answer: *does any observed gain come from rewriting itself, or from the source domain?*

3 Data Difficulty Analysis

This section checks, using automatic readability metrics, that the rewrites change text difficulty in

the direction we claim.

3.1 Automatic metrics

We use six automatic metrics, computed with the `textstat` library¹: rewrite ratio (rewrite/original word count), lexical overlap (word-set overlap between original and rewrite), Flesch Reading Ease (FRE; higher is easier) and Flesch-Kincaid Grade (FKGL; lower is easier) (Kincaid et al., 1975), and Dale-Chall score and difficult-word ratio (DaleC and DiffWd; both lower is easier) (Chall and Dale, 1995).

3.2 Comparison across the three corpora

We compute these six metrics on both the original and the simplified side of every pair: 163,150 FineWeb pairs, 179,165 OpenStax+BCCampus pairs (the two textbook sources pooled together), and 331,074 Simple Wikipedia pairs (Table 3).

All three sources become shorter after simplification (ratio 0.634–0.893) and keep much of the original vocabulary (overlap 0.38–0.56). FineWeb and OpenStax+BCCampus improve on all four readability metrics. OpenStax+BCCampus is compressed the most but remains the hardest to read after simplification, because its source text is very dense. Simple Wikipedia is a partial exception: two metrics improve, but two move slightly in the wrong direction, likely because the original text is already simple.

¹<https://github.com/textstat/textstat>

3.3 Comparison against explanation and paraphrase

We apply the same six metrics to the FineWeb original+explanation and original+paraphrase pairs (§2.1), to see how each rewrite type compares to simplification on the same source (Table 4).

The three rewrite types differ clearly. Explanation is longer than the original (ratio 1.104) but becomes much easier to read (FRE goes from 43.87 to 67.74, the largest gain). Paraphrasing keeps a similar length (ratio 0.954) but becomes *harder* to read on all four metrics (FRE drops from 43.90 to 31.38). This confirms that paraphrases are not just simplified text with a different name. Both explanation and paraphrasing also change the wording more than simplification does (lower lexical overlap: 0.383 and 0.354 vs. 0.505).

4 Main Experiments

4.1 Baselines

We report two families of baselines: (i) global baselines trained on the standard BabyLM-provided corpus (`mlm_regular`, multiple seeds), and (ii) source-matched baselines for each of the three dataset sources (§2.2). Official challenge baselines are taken from the BabyLM 2026 evaluation repository (BabyLM Team, 2026).

4.2 Training and evaluation procedure

All models share the same encoder architecture and training procedure as our BabyLM STRICT-SMALL baseline, following the architecture and optimization setup of Edman and Fraser (2025) but not their adaptive MLM mechanism: a DeBERTa-style encoder (He et al., 2021) (hidden size 384, intermediate size 1280, dropout 0.1), trained with the LAMB optimizer (You et al., 2020), a cosine learning-rate schedule, and a sequence-length curriculum that starts at 64 tokens and increases to 256 tokens partway through training, with batch size scaled inversely to sequence length. We use a 40k-vocabulary SentencePiece BPE tokenizer (Kudo and Richardson, 2018; Sennrich et al., 2016) with byte fallback, trained separately for each data condition on its own training corpus.

For each corpus, we train two masking variants: (i) **Token**, standard token-level MLM with 15% masking throughout, and (ii) **WWM7→Token3**, which uses whole-word masking for the first 7 epochs and then switches to token-level masking for the last 3 epochs. Whole-word masking

was introduced in Google’s subsequent BERT release (Devlin et al., 2019; Google Research, 2019); our contribution is the 7-epoch WWM to 3-epoch token-level curriculum and its combination with rewrite pairing, rather than WWM itself.

Whole-word masking differs from token masking when a word is split into multiple subword pieces. Because we use SentencePiece BPE, word boundaries are recovered from the tokenizer’s word-start marker, shown in tokenizer output as an underscore-like prefix (`_`). Pieces with this prefix start a new word span, and following pieces without it are treated as part of the same word. For example, if one word is segmented into several pieces, WWM masks the entire span rather than only one piece. With token masking, the model may see part of the word and only need to recover the missing subword; with WWM, all pieces of the selected word are masked together, so the model must recover the full word from context. We chose the 7/3 split as a practical balance: fewer WWM epochs may not be enough to establish word-level context learning, while using only WWM with no token-level phase loses fine-grained subword prediction.

All models are trained for 10 epochs on a fixed 10M-word budget, with 3 random seeds per condition; we report seed means unless noted. Paired bootstrap tests over seeds (10,000 samples) show significant EWoK and GLUE gains, as well as Entity Tracking gains for simplification conditions, over source-matched raw baselines ($p < 0.05$; all but FineWeb Explanation Token EWoK reach $p < 0.01$).

After training, each model is evaluated with the official BabyLM 2026 strict-track evaluation pipeline (BabyLM Team, 2026). BLiMP (Warstadt et al., 2020), BLiMP Supplement (Warstadt et al., 2023), EWoK (Ivanova et al., 2024), COMPS (Misra et al., 2023), Entity Tracking (Kim and Schuster, 2023), and Eye-tracking and Self-Paced Reading (de Varda et al., 2023) are scored via the pipeline’s standard perplexity-comparison protocol, and GLUE- and SuperGLUE-style tasks (Wang et al., 2018, 2019) are evaluated by fine-tuning the pretrained encoder with the organizers’ default fine-tuning hyperparameters. We report both an aggregate zero-shot average and per-task breakdowns where relevant to our claims (§5, §5.4).

Condition	EWoK	±	Entity	±	GLUE	±
FW Raw Token	51.2	1.2	21.2	0.3	65.1	0.3
FW Simp Token	52.9	0.3	26.1	2.3	68.3	0.1
FW Simp WWM	55.4	0.6	28.9	0.4	69.9	0.1
FW Expl Token	52.8	1.6	22.1	1.6	68.7	1.1
FW Expl WWM	54.7	0.5	22.2	2.0	69.8	1.3
FW Para Token	52.7	0.7	19.0	1.0	68.5	0.3
FW Para WWM	54.2	0.6	19.1	0.6	69.2	0.0
OS+BCC Raw Token	50.5	0.9	20.4	1.0	66.0	0.9
OS+BCC Simp Token	53.7	1.7	25.5	2.5	69.2	0.5
OS+BCC Simp WWM	54.9	0.7	25.6	2.4	69.1	0.3
SW Raw Token	50.2	0.1	20.5	0.4	65.0	0.9
SW Simp Token	51.4	0.4	23.2	1.9	68.7	0.8
SW Simp WWM	53.4	1.0	23.8	1.8	69.8	0.5

Table 5: EWoK, Entity Tracking, and GLUE mean±std over 3 seeds for conditions with full seed replication. The FineWeb block includes simplification, explanation, and paraphrasing. EWoK and GLUE gains, and Entity Tracking gains for simplification rows only, are significant over source-matched raw baselines ($p < 0.05$, paired bootstrap, 10,000 samples).

5 Results

The improvement is most pronounced on EWoK and GLUE-style fine-tuning, so §5.1 and §5.3 concentrate on these two, before §5.4 asks why. We also observe a narrower but consistent effect on the Entity Tracking task: simplification improves it across all three sources, while the other FineWeb rewrite types do not show the same benefit (§5.2). Table 5 reports mean±std for EWoK, Entity Tracking, and GLUE across all conditions where 3 seeds are available. Across the rewrite-pairing conditions, WWM7→Token3 improves performance in most cases. Since WWM masking alone (without rewrite pairing) does not produce meaningful gains over the raw Token baseline, we omit the Raw WWM conditions from all sub-item tables for brevity; full aggregate scores, including Raw WWM and all zero-shot metrics, are reported by source in Appendix C.

5.1 EWoK results

Simplification improves EWoK over the raw Token baseline in all three sources. Both FineWeb ablations, including paraphrasing, improve EWoK by a comparable margin despite its rewrites being harder, not easier, to read (§3.3); we return to why this may be the case in §5.4.

Physical-dynamics and *quantitative-properties* both improve substantially under simplification for all three sources, making them the clearest source-general component of simplification’s EWoK ben-

Source	Raw	Tok.	WWM
<i>physical-dynamics</i>			
FineWeb (Simp.)	50.3	61.4	69.2
FineWeb (Expl.)	50.3	58.6	63.3
FineWeb (Para.)	50.3	63.3	66.4
OpenStax+BCCampus	55.0	62.8	61.7
Simple Wikipedia	52.8	60.8	63.9
<i>quantitative-properties</i>			
FineWeb (Simp.)	51.4	66.1	70.4
FineWeb (Expl.)	51.4	63.8	66.7
FineWeb (Para.)	51.4	58.7	60.9
OpenStax+BCCampus	53.5	66.2	69.5
Simple Wikipedia	52.1	64.8	68.0
<i>social-properties</i>			
FineWeb (Simp.)	51.4	55.4	56.9
FineWeb (Expl.)	51.4	48.3	53.2
FineWeb (Para.)	51.4	51.9	56.2
OpenStax+BCCampus	47.7	60.3	60.9
Simple Wikipedia	52.2	48.1	51.3

Table 6: Mean EWoK accuracy (3 seeds) on physical-dynamics, quantitative-properties, and social-properties, for the Raw, and Token/WWM7→Token3 conditions of each source. For FineWeb, all three rewrite types (Simplification, Explanation, paraphrasing) are shown against the same FineWeb Raw baseline.

efit. *Social-properties* tells a different story: it is OpenStax+BCCampus’s single largest gain, plausibly reflecting that source’s humanities and social science content, yet only gains modestly for FineWeb and regresses for Simple Wikipedia. This suggests simplification’s EWoK benefit has both a source-general component (*physical-dynamics*, *quantitative-properties*) and a source-specific one (*social-properties*) tied to a source’s particular world-knowledge content.

Our two FineWeb ablations diverge on *quantitative-properties*: both explanation and paraphrasing improve it less than simplification does, paraphrasing least of all. On physical-dynamics, paraphrasing’s gain is comparable to simplification’s and explanation’s, and on social-properties it falls between the other two. No single rewrite type dominates on all three sub-items, but all three produce a gain on *physical-dynamics* and *quantitative-properties*, consistent with these being the most source-general and rewrite-type-general component of the EWoK benefit (§5.4).

5.2 Entity Tracking results

The Entity Tracking task shows a more selective pattern than EWoK. Simplification improves Entity Tracking over source-matched raw Token baselines for all three sources.

Unlike EWoK and GLUE, this gain does not generalize to all meaning-preserving rewrite types. On FineWeb, explanation gives only a small improvement, while paraphrasing falls below the raw Token baseline. Thus, the Entity Tracking result appears to be tied specifically to simplification. This contrasts with EWoK and GLUE, where simplification, explanation, and paraphrasing all improve performance.

5.3 GLUE results

As shown in Table 5, simplification substantially improves GLUE over the source-matched raw baselines for all three sources. This pattern is not limited to simplification: explanation and paraphrasing also improve GLUE by similar margins, even though their rewrites are not uniformly easier to read.

Table 7 breaks this down by sub-task (full breakdown in Appendix F) for the five that gain consistently across all three sources and all three FineWeb rewrite types.

MNLI, MRPC, and QQP gain the most consistently across all three sources and both FineWeb rewrite types: all three require relating two given sentences rather than answering questions about a single passage. RTE gains as well under both rewrite types, though more modestly for paraphrasing. BoolQ and WSC move inconsistently; MultiRC gains modestly under simplification but little under paraphrasing. The sentence-pair advantage holds for paraphrasing as well as simplification, consistent with the account in §5.4.

5.4 Mechanisms behind the main gains

EWoK: Improved context-to-concept mapping, shared across rewrite types. EWoK (Ivanova et al., 2024) tests world knowledge with minimal pairs scored by perplexity. In *concept swap* items, context and target jointly change concept (e.g., "increasing"/"accelerating" vs. "decreasing"/"slowing down"), requiring the model to map context onto the right concept word. In *variable swap* items, concept words stay fixed and only entity role assignment changes — a different capacity. Our hypothesis concerns concept swap items only. The three specific EWoK sub-items reported in Table 6 — *physical-dynamics*, *quantitative-properties*, and *social-properties* — contain only concept swap items in the evaluation data and therefore fall directly within this analysis.

Under MLM, a masked word in one sentence

Source	Raw	Tok.	WWM
<i>MNLI</i>			
FineWeb (Simp.)	51.2	59.4	62.8
FineWeb (Expl.)	51.2	59.4	61.1
FineWeb (Para.)	51.2	58.7	60.9
OpenStax+BCCampus	51.4	60.2	61.8
Simple Wikipedia	52.1	59.5	62.5
<i>MRPC (F1)</i>			
FineWeb (Simp.)	82.4	87.2	88.6
FineWeb (Expl.)	82.4	87.0	88.7
FineWeb (Para.)	82.4	86.6	88.6
OpenStax+BCCampus	82.6	87.2	88.5
Simple Wikipedia	82.2	86.2	88.8
<i>QQP (F1)</i>			
FineWeb (Simp.)	65.9	71.7	73.3
FineWeb (Expl.)	65.9	71.7	72.9
FineWeb (Para.)	65.9	71.8	72.5
OpenStax+BCCampus	67.3	72.1	72.5
Simple Wikipedia	66.1	71.4	73.1
<i>RTE</i>			
FineWeb (Simp.)	59.0	63.6	65.7
FineWeb (Expl.)	59.0	64.3	66.4
FineWeb (Para.)	59.0	61.2	63.8
OpenStax+BCCampus	62.1	65.5	62.4
Simple Wikipedia	57.3	65.5	66.9
<i>MultiRC</i>			
FineWeb (Simp.)	66.0	67.1	67.1
FineWeb (Expl.)	66.0	65.9	67.4
FineWeb (Para.)	66.0	66.3	66.5
OpenStax+BCCampus	66.8	67.4	67.0
Simple Wikipedia	64.2	66.8	67.0

Table 7: Mean fine-tuning score (3 seeds) on MNLI, MRPC, QQP, RTE, and MultiRC, for the Raw, and Token/WWM7→Token3 conditions of each source, with FineWeb’s explanation and paraphrasing rewrites shown alongside simplification. MRPC and QQP are scored with F1; the remaining tasks with accuracy.

can be predicted using the other sentence in the pair, even though the two sentences use different words to express the same meaning. This may train the model to associate a given context with a range of words that share the same meaning, rather than always predicting the same word it has seen before.

The observed pattern offers some support for this explanation. Across all rewrite types and sources, Table 8 shows gains on concept swap items (which test whether the model picks the right meaning given context), while variable swap items (which test something different) stay flat near chance. The gains are already present in the Token-only rewrite conditions and are further strengthened by WWM7→Token3, whereas applying WWM7→Token3 to raw data does not produce consistent aggregate gains over Raw Token (Appendix C). This selective pattern is consistent with rewrite pairing improving context-to-concept map-

Condition	Raw	Tok.	WWM
<i>concept swap</i>			
FineWeb (Simp.)	51.4	52.3	53.7
FineWeb (Expl.)	51.4	52.9	54.0
FineWeb (Para.)	51.4	51.7	52.8
OpenStax+BCCampus	49.9	52.6	53.8
Simple Wikipedia	49.7	50.1	51.8
<i>variable swap</i>			
FineWeb (Simp.)	50.4	49.7	50.0
FineWeb (Expl.)	50.4	49.9	50.0
FineWeb (Para.)	50.4	50.5	49.9
OpenStax+BCCampus	49.9	50.3	50.2
Simple Wikipedia	49.8	50.2	50.5

Table 8: Mean EWoK accuracy (3 seeds) on concept swap vs. variable swap items, for the Raw, and Token/WWM7→Token3 conditions of each source, with FineWeb’s explanation and paraphrasing rewrites shown alongside simplification. Concept swap items test context-to-concept mapping; variable swap items test entity role binding, holding the concept words fixed.

ping rather than producing a broad improvement on all EWoK items. We treat this interpretation as a post-hoc hypothesis supported by the observed decomposition, not as a demonstrated causal mechanism.

This pattern holds for all three rewrite types — including paraphrasing, whose rewrites are harder to read, not easier. This suggests that what matters is not making the text simpler, but pairing it with a different version that expresses the same meaning.

GLUE: Improved performance from rewritten data. The same effect may explain why gains on GLUE concentrate on sentence-pair tasks (MNLI, MRPC, QQP, RTE) — tasks where the model must compare two sentences — rather than single-passage tasks. Training with paired rewrites may generally improve the model’s ability to link meaning across different ways of saying the same thing. [Edman et al. \(2024\)](#) also found large GLUE improvements by training with meaning-related rewrites, though their setup (original + paraphrase + hard-negative triplets) differs from ours and the underlying mechanism may not be the same.

Entity Tracking: Simplification may reduce local referential complexity. The Entity Tracking gain is more selective than the EWoK and GLUE gains, because it appears mainly for simplification. A plausible explanation is that simplified rewrites make entity mentions and their local contexts easier to encode by replacing dense noun phrases, embedded clauses, and domain-specific wording with

more direct expressions, while preserving the same referential structure. This gives the model a simpler second view of the same entities and relations; explanation and paraphrasing also preserve meaning, but they do not consistently simplify this local entity context.

WWM: Harder predictions may encourage contextual learning. The WWM curriculum is a separate masking factor rather than a rewrite-aware mechanism. By masking all subword pieces of a selected word, WWM removes partial-word shortcuts and makes prediction harder: the model must recover the full word from its surrounding context rather than from visible fragments of the same word. This may encourage the model to use broader contextual relations during pretraining, although the lower BLiMP scores suggest a possible trade-off with grammar-focused benchmarks.

Overall, the EWoK and GLUE gains are consistent with the model learning to rely more on meaning and context, and less on memorized word patterns, while the Entity Tracking gain may reflect simplification’s more direct effect on local entity contexts. Future work could test these hypotheses through probing or causal analysis.

5.5 Submission

Our main submission focuses on the STRICT-SMALL track (10M words), where we submit a model trained with WWM7→Token3 simplification pairing.² Table 9 compares this model with the official BabyLM 2026 GPT-2 and interaction baselines ([BabyLM Team, 2026](#)) from both the STRICT-SMALL and STRICT tracks. The most pronounced gains are on EWoK, Entity Tracking, and (Super)GLUE. Notably, our 10M-word model reaches an NLP average of 52.97, exceeding the STRICT-SMALL baselines and even the 100M-word STRICT interaction baseline (52.42). It also surpasses the official 100M-word GPT-2 baseline on EWoK (56.07 vs. 54.37), Entity Tracking (28.45 vs. 16.91), and (Super)GLUE (69.79 vs. 67.75). These results show that better downstream performance can be achieved with one-tenth as much training data, demonstrating a substantial improvement in sample efficiency.

We also make one STRICT-track attempt using span masking. Instead of masking indepen-

²Resources: [code](#); [STRICT-SMALL WWM simplification model](#); [corresponding FineWeb simplification-pair training data](#); and [additional STRICT span-masking model](#).

dently selected tokens or whole words, span masking masks contiguous token spans whose lengths are sampled from a geometric distribution (Joshi et al., 2020). This model is trained on a 100M-word simplification-pair mixture from FineWeb, OpenStax, and BCCampus. Although this setting is not our main ablation, it also improves strongly over the official 100M-word GPT2 baseline on EWoK (57.73 vs. 54.37) and (Super)GLUE (72.60 vs. 67.75), and slightly improves the NLP average (53.10 vs. 53.03).

6 Conclusion

Our results suggest that the rewrite direction does not matter for our two main outcomes: simplification, explanation, and harder paraphrasing all produce consistent gains on EWoK and GLUE-style fine-tuning, indicating that what matters there is semantic-preserving variation itself rather than any particular transformation direction. Simplification also improves Entity Tracking across all three sources, but this effect is more selective: the explanation condition gives only a small gain, and the paraphrase condition does not improve it. Across these outcomes, WWM7→Token3 further strengthens the gains in most conditions. Ablations across corpora of different domain, rewrite difficulty, and rewrite length consistently reproduce the EWoK and GLUE pattern, confirming the robustness of the approach while showing that different evaluation tasks can respond to different rewrite properties.

The Ewok and Glue gains are selective in ways consistent with a single underlying shift toward contextual reliance over surface statistics. On EWoK, the improvement concentrates on concept swap items and is absent on variable swap items, a pattern that holds across all three rewrite types and all three sources. On GLUE, the gain concentrates on sentence-pair tasks (MNLI, MRPC, QQP, RTE), not passage-based ones. This selectivity is consistent with the hypothesis that rewrite pairing encourages the model to exploit contextual meaning rather than memorized surface form, with the particular rewrite type shaping which capacities benefit most. Edman et al. (2024) also found large GLUE improvements from a different meaning-rewrite setup, though the mechanism may differ.

The key practical implication is a substantial improvement in sample efficiency: our 10M-word strict-small model surpasses the official 100M-word GPT-2 baseline on EWoK, Entity Tracking,

and GLUE, suggesting that better organization of limited training data may be as important as increasing its quantity. More broadly, our findings suggest that improving the structure of training data can serve as a complementary scaling axis for small-data language model pretraining — alongside model size, data volume, and compute — one that may be particularly valuable when data is the binding constraint. We leave the design of other targeted data transformations to future work.

Limitations

All three source corpora used in this work consist of factual, expository text — web-based educational content, open textbooks, and encyclopedic articles. It remains an open question whether meaning-preserving rewrite pairing produces similar gains on other text types, such as conversational or spoken language, fictional narratives, or informal social media text. In such genres, sentences may be shorter, more context-dependent, or less self-contained, which could reduce the effectiveness of sentence-level pairing.

We also do not test all rewrite types on all sources. Explanation and harder paraphrasing are evaluated only on FineWeb, while OpenStax+BCCampus and Simple Wikipedia are tested only with simplification. Therefore, our cross-source evidence is strongest for simplification; whether explanation and harder paraphrasing generalize similarly across sources remains open.

Ethics Statement

Our simplification, explanation and paraphrase corpora are generated by prompting an offline rewrite generator (Qwen2.5-7B-Instruct) to rewrite existing text; this introduces stylistic and content biases specific to that generator into the resulting training data, which may not generalize to rewrites produced by other models or pipelines. Our source corpora (FineWeb, OpenStax, BCCampus, Simple Wikipedia) are all openly licensed for research use, and we do not redistribute the generated rewrites beyond what is needed to reproduce our results. All experiments are conducted on English-language data under the BabyLM STRICT-SMALL track and do not involve human subjects.

Acknowledgments

The authors gratefully acknowledge the scientific support and HPC resources provided by the Er-

Model	Track	Overall	NLP	Human	BLiMP	Supp.	EWoK	Entity	COMPS	GPIQA	GLUE	Read.
Ours (WWM simp.)	strict-small	41.80	52.97	2.71	67.20	56.01	56.07	28.45	53.57	39.67	69.79	5.42
Ours (span mask. simp.)	strict	41.93	53.10	2.83	72.38	59.86	57.73	18.43	53.98	36.69	72.60	5.65
GPT2 baseline	strict-small	37.38	48.99	-3.26	65.23	57.25	50.63	19.10	51.81	35.09	63.80	5.63
Interaction baseline	strict-small	38.71	49.04	2.55	63.07	58.13	51.45	19.62	51.29	36.14	63.62	5.10
GPT2 baseline	strict	40.73	53.03	-2.32	74.73	65.00	54.37	16.91	55.85	36.62	67.75	6.93
Interaction baseline	strict	39.89	52.42	-3.96	72.89	65.09	54.23	15.89	55.37	36.18	67.28	3.43

Table 9: Leaderboard submission results for our STRICT-SMALL model, our additional STRICT span-masking attempt, and official BabyLM 2026 baselines. Overall, NLP, and Human are leaderboard aggregate scores; Reading is the mean of Eye-tracking and SPR. GPIQA denotes GlobalPIQA. Bold indicates the best score in each column. Our 10M-word STRICT-SMALL model exceeds the 100M-word STRICT interaction baseline in NLP average and exceeds the 100M-word GPT2 baseline on EWoK, Entity Tracking, and (Super)GLUE; our STRICT span-masking model further improves EWoK and (Super)GLUE.

langen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project b279bb. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by the German Research Foundation (DFG) – 440719683.

This work was co-funded by the European Union (ERC, EPICAL, 101141712). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

- BabyLM Team. 2026. [BabyLM Evaluation 2026: Official evaluation pipeline and baselines](#). GitHub repository.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, and 1 others. 2025. [Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures](#). *arXiv preprint arXiv:2510.24081*.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. [Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.
- 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](#). *arXiv preprint arXiv:2602.20092*.
- Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd babyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.
- Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2023. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, pages 1–24.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- Lukas Edman, Lisa Bylinina, Faeze Ghorbanpour, and Alexander Fraser. 2024. Are babyLMs second language learners? In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 166–173.
- Lukas Edman and Alexander Fraser. 2025. [Mask and you shall receive: Optimizing masked language modeling for pretraining BabyLMs](#). In *Proceedings of the First BabyLM Workshop*, pages 445–453, Suzhou, China. Association for Computational Linguistics.
- Google Research. 2019. [BERT: Whole word masking models](#). Official BERT repository release, May 31, 2019.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.

- Anna Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, Pramod RT, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Josh Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09665*.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Naval Technical Training Command Research Branch Report*.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Alal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pages 17–60.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.
- 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.
- Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. [Large batch optimization for deep learning: Training BERT in 76 minutes](#). In *International Conference on Learning Representations*.

A GlobalPIQA Hidden Task Analysis

GlobalPIQA (Chang et al., 2025) was released as a hidden evaluation task late in the BabyLM evaluation cycle. Table 10 reports the 3-seed mean and standard deviation of the GlobalPIQA average score, computed as the simple average of the parallel and non-parallel splits.

Overall, the GlobalPIQA results show relatively high variance, making it difficult to draw a firm conclusion from individual differences. Most of our source-specific conditions have higher mean scores than the official GPT2 and interaction baselines in Table 9, whose GlobalPIQA scores range from 35.09 to 36.62. At the same time, the BB26 Raw Token baseline reaches 39.08 in one seed despite being trained on the official BB26 strict-small data rather than on any of our constructed corpora, suggesting that the task may be sensitive to random seed or evaluation instability. We therefore remain

Source	Condition	GlobalPIQA avg.
BabyLM	Raw Tok.	36.93 $\pm$ 2.00
FineWeb	Raw Tok.	37.34 $\pm$ 1.59
FineWeb	Raw WWM	40.79 $\pm$ 2.22
FineWeb	Simp. Tok.	37.16 $\pm$ 1.32
FineWeb	Simp. WWM	38.66 $\pm$ 2.59
FineWeb	Expl. Tok.	39.29 $\pm$ 2.01
FineWeb	Expl. WWM	39.28 $\pm$ 3.12
FineWeb	Para. Tok.	39.41 $\pm$ 3.21
FineWeb	Para. WWM	38.77 $\pm$ 2.07
OS+BCC	Raw Tok.	38.04 $\pm$ 2.58
OS+BCC	Raw WWM	39.59 $\pm$ 1.44
OS+BCC	Simp. Tok.	40.39 $\pm$ 1.63
OS+BCC	Simp. WWM	38.45 $\pm$ 0.55
SimpleWiki	Raw Tok.	38.31 $\pm$ 1.03
SimpleWiki	Raw WWM	37.49 $\pm$ 1.99
SimpleWiki	Simp. Tok.	35.99 $\pm$ 2.50
SimpleWiki	Simp. WWM	34.67 $\pm$ 2.27

Table 10: GlobalPIQA hidden-task results for the main ablations. Scores are 3-seed mean $\pm$ standard deviation. The BB26 Raw Token baseline uses our DeBERTa-style architecture trained on the official BabyLM STRICT-SMALL data.

cautious about interpreting GlobalPIQA as clear evidence for any single intervention. At most, the pattern suggests that our three source corpora may be mildly favorable for this task because they are educational or knowledge-oriented texts, but this remains a hypothesis rather than a stable finding.

B Generation Prompts

We give the system prompts used to generate our simplification, explanation, and paraphrase corpora (§2) with the offline rewrite generator (Qwen2.5-7B-Instruct).

You are simplifying academic sentences for a language model training dataset. For each input sentence, output exactly one line: a simplified version.

Simplification rules:

- Keep the core meaning intact.
- Use simpler, more common vocabulary. Replace jargon with plain words.
- Shorten where possible, but do not omit key information.
- Write in a clear, direct style accessible to a general audience.
- Do NOT add explanations, labels, or meta-commentary.
- Output exactly one line of text, nothing else.

Figure 1: System prompt used for simplification generation.

You are rewriting text passages so that a middle school student (age 12-14) can easily understand them.

Rewriting rules:

- Keep ALL the information and key details from the original.
- Replace difficult vocabulary and technical jargon with common words.
- If a key term must be kept, briefly explain what it means in plain language.
- Break up long or complicated sentences into shorter, clearer ones.
- The output should be roughly the same length as the input.
- Write in a clear, natural tone.
- Do NOT add explanations of what you are doing, labels, headers, or commentary.
- Output only the rewritten text,

nothing else.

Figure 2: System prompt used for explanation generation.

You are paraphrasing web text for a language model training dataset. For each input passage, output exactly one line: a paraphrase.

Paraphrase rules:

- Preserve the full meaning. Change structure and vocabulary significantly.
- Match the original register and tense.
- Match the original length as closely as possible -- do not compress or expand.
- Do NOT add explanations, labels, or meta-commentary.
- Output exactly one line of text, nothing else.

Figure 3: System prompt used for paraphrasing’s paraphrase generation.

C Full Aggregate Results

Tables 11–13 report the full aggregate zero-shot, hidden-task, and fine-tuning results, split by source. Raw WWM refers to training on the original (non-rewritten) corpus with WWM7 $\rightarrow$ Token3 masking. Reading is the mean of Eye-tracking and Self-Paced Reading scores. GlobalPIQA is the average of the parallel and non-parallel splits. GLUE is the mean fine-tuning score across the seven GLUE-style tasks.

Metric	Global	FineWeb Raw/Simp.			
	Base (bb26)	Raw Tok.	Raw WWM	Simp. Tok.	Simp. WWM
BLiMP	66.2	67.1	67.9	65.7	67.6
Supplement	58.4	54.8	55.0	54.1	54.9
EWoK	51.0	51.2	51.4	52.9	55.4
Ent. Tracking	20.8	21.2	20.5	26.1	28.9
COMPS	51.3	52.4	52.9	52.0	52.8
Reading	6.2	5.1	5.2	4.7	5.6
GlobalPIQA	36.9	37.3	40.8	37.2	38.7
GLUE	64.4	65.1	64.9	68.3	69.9

Metric	Global	FineWeb Additional Rewrites				
	Base (bb26)	Raw Tok.	Expl. Tok.	Expl. WWM	Para. Tok.	Para. WWM
BLiMP	66.2	67.1	68.4	68.6	67.2	67.2
Supplement	58.4	54.8	56.5	55.5	53.6	54.0
EWoK	51.0	51.2	52.8	54.7	52.7	54.2
Ent. Tracking	20.8	21.2	22.1	22.2	19.0	19.1
COMPS	51.3	52.4	52.5	53.0	51.9	52.1
Reading	6.2	5.1	5.7	5.5	5.4	4.6
GlobalPIQA	36.9	37.3	39.3	39.3	39.4	38.8
GLUE	64.4	65.1	68.7	69.8	68.5	69.2

Table 11: Full aggregate results for FineWeb conditions, mean over 3 seeds. The global baseline is mlm_regular_bb26_40k, trained on the standard BabyLM-provided corpus. Entity Tracking uses the updated hidden-task evaluation, and GlobalPIQA is the hidden-task split average. FineWeb explanation and paraphrasing variants have no separate raw condition; both compare to the FineWeb Raw baseline.

Metric	Global	OpenStax+BCCampus			
	Baseline (bb26)	Raw Tok.	Raw WWM	Simp. Tok.	Simp. WWM
BLiMP	66.2	67.9	68.1	67.2	67.7
Supplement	58.4	55.2	54.6	56.9	57.5
EWoK	51.0	50.5	50.9	53.7	54.9
Ent. Tracking	20.8	20.4	21.4	25.5	25.6
COMPS	51.3	51.0	50.9	51.0	51.4
Reading	6.2	5.7	5.7	5.0	6.1
GlobalPIQA	36.9	38.0	39.6	40.4	38.5
GLUE	64.4	66.0	65.6	69.2	69.1

Table 12: Full aggregate results for OpenStax+BCCampus conditions, mean over 3 seeds. The global baseline is `mlm_regular_bb26_40k`. Entity Tracking uses the updated hidden-task evaluation, and GlobalPIQA is the hidden-task split average.

Metric	Global	Simple Wikipedia			
	Baseline (bb26)	Raw Tok.	Raw WWM	Simp. Tok.	Simp. WWM
BLiMP	66.2	64.8	66.3	66.1	66.2
Supplement	58.4	47.5	46.8	48.2	49.0
EWoK	51.0	50.2	50.0	51.4	53.4
Ent. Tracking	20.8	20.5	20.0	23.2	23.8
COMPS	51.3	52.1	53.1	52.0	52.5
Reading	6.2	5.3	4.9	4.7	5.3
GlobalPIQA	36.9	38.3	37.5	36.0	34.7
GLUE	64.4	65.0	65.1	68.7	69.8

Table 13: Full aggregate results for Simple Wikipedia conditions, mean over 3 seeds. The global baseline is `mlm_regular_bb26_40k`. Entity Tracking uses the updated hidden-task evaluation, and GlobalPIQA is the hidden-task split average.

D Hyperparameters

Table 14 lists the training hyperparameters for both submissions.

Parameter	Strict-small (10M)	Strict (100M)
<i>Model</i>		
Architecture	DeBERTa-v2	DeBERTa-v2
Hidden size	384	768
Intermediate size	1280	2560
Dropout	0.1	0.1
Vocabulary size	40k	40k
<i>Training</i>		
Sequence length	64→256	64→256
Batch size (tokens)	16,384	131,072
Learning rate	7e-3	7e-3
Epochs	10	10
Scheduler	Cosine	Cosine
Mask ratio	0.15	0.15
Random ratio	0.1	0.1
Keep ratio	0.1	0.1
Weight decay	0.01	0.01
Optimizer	LAMB	LAMB
Optimizer ϵ	1e-8	1e-8
Optimizer β_1	0.9	0.9
Optimizer β_2	0.95	0.95
Gradient clipping	1.0	1.0
Masking strategy	WWM7→Token3	Span geo.
Span geo. p	—	0.5

Table 14: Training hyperparameters for the two submissions. Batch size in tokens = sequences $\times$ sequence length. The strict-small model uses WWM7→Token3 simplification pairing; the strict model uses span geometric masking (Joshi et al., 2020) with $p=0.5$.

E Full EWoK Sub-item Breakdown

Table 15 is split into raw/simplification conditions and a continued panel for FineWeb explanation/paraphrasing.

Sub-item	FineWeb			OS+BCC			SimpleWiki		
	Raw	Tok.	WWM	Raw	Tok.	WWM	Raw	Tok.	WWM
social-int.	55.9	53.5	57.4	55.7	56.6	60.8	49.3	51.7	55.3
material-dyn.	55.8	48.4	54.7	47.9	51.0	53.7	49.7	42.5	47.6
physical-rel.	51.2	51.0	50.5	49.4	50.9	49.4	47.5	52.8	51.6
social-rel.	49.6	50.7	50.3	49.8	50.5	51.4	50.3	51.2	50.1
physical-dyn.	50.3	61.4	69.2	55.0	62.8	61.7	52.8	60.8	63.9
physical-int.	50.5	52.3	53.9	50.5	52.2	53.7	50.7	49.5	52.0
material-prop.	48.0	47.8	51.8	48.0	43.7	48.4	50.6	50.0	52.2
social-prop.	51.4	55.4	56.9	47.7	60.3	60.9	52.2	48.1	51.3
agent-prop.	50.5	50.9	50.3	50.1	50.4	50.5	49.6	50.0	51.1
quant.-prop.	51.4	66.1	70.4	53.5	66.2	69.5	52.1	64.8	68.0
spatial-rel.	48.7	44.2	43.8	48.4	45.9	44.0	48.0	44.3	44.1
AVG	51.2	52.9	55.4	50.5	53.7	54.9	50.2	51.4	53.4

Table 15: Full EWoK sub-item breakdown for raw and simplified conditions (UID accuracy, mean over 3 seeds).

F Full GLUE Sub-task Breakdown

Table 16 is split into raw/simplification conditions and a continued panel for FineWeb explanation/paraphrasing.

FineWeb Additional Rewrites					
Sub-item	Raw	E.Tok.	E.WWM	P.Tok.	P.WWM
social-int.	55.9	56.7	58.3	54.5	59.9
material-dyn.	55.8	53.5	52.1	49.0	52.2
physical-rel.	51.2	51.5	52.5	53.8	52.0
social-rel.	49.6	51.3	51.8	50.9	50.3
physical-dyn.	50.3	58.6	63.3	63.3	66.4
physical-int.	50.5	50.8	52.8	50.6	49.3
material-prop.	48.0	50.6	53.3	52.2	53.5
social-prop.	51.4	48.3	53.1	51.9	56.2
agent-prop.	50.5	51.3	51.0	50.9	50.6
quant.-prop.	51.4	63.8	66.7	58.7	60.9
spatial-rel.	48.7	44.5	46.7	43.9	44.4
AVG	51.2	52.8	54.7	52.7	54.2

Table 15: Full EWoK sub-item breakdown for FineWeb explanation and paraphrasing variants. These variants have no separate raw condition; both compare to FineWeb Raw.

Task	FineWeb			OS+BCC			SimpleWiki		
	Raw	Tok.	WWM	Raw	Tok.	WWM	Raw	Tok.	WWM
MNLI	51.2	59.4	62.8	51.4	60.2	61.8	52.1	59.5	62.5
MRPC	82.4	87.2	88.6	82.6	87.2	88.5	82.2	86.2	88.8
QQP	65.9	71.7	73.3	67.3	72.1	72.5	66.1	71.4	73.1
RTE	59.0	63.6	65.7	62.1	65.5	62.4	57.3	65.5	66.9
BoolQ	68.0	66.5	68.0	67.3	66.1	67.0	68.0	67.0	67.7
MultiRC	66.0	67.1	67.1	66.8	67.4	67.0	64.2	66.8	67.0
WSC	63.5	62.8	63.5	64.1	66.0	64.7	65.4	64.7	62.8
AVG	65.1	68.3	69.9	66.0	69.2	69.1	65.0	68.7	69.8

Table 16: Full fine-tuning sub-task breakdown for raw and simplified conditions, mean over 3 seeds. MRPC and QQP are scored with F1; the remaining tasks are scored with accuracy.

FineWeb Additional Rewrites					
Task	Raw	E.Tok.	E.WWM	P.Tok.	P.WWM
MNLI	51.2	59.4	61.1	58.7	60.9
MRPC	82.4	87.0	88.7	86.6	88.6
QQP	65.9	71.7	72.9	71.8	72.5
RTE	59.0	64.3	66.4	61.2	63.8
BoolQ	68.0	67.3	68.3	67.5	68.7
MultiRC	66.0	65.9	67.4	66.3	66.5
WSC	63.5	65.4	64.1	67.3	63.5
AVG	65.1	68.7	69.8	68.5	69.2

Table 16: Full fine-tuning sub-task breakdown for FineWeb explanation and paraphrasing variants. These variants have no separate raw condition; both compare to FineWeb Raw.