

BabyConceptLM: Exploring Dynamic Concept Computation under Restricted Language Exposure

Yizhe Xu Ming Song Danni He Jianghao Liu Yitong Wang Qing Cai*

School of Psychology and Cognitive Science, East China Normal University
3663 North Zhongshan Road, Shanghai 200062, China

Abstract

Human language learning takes place under limited exposure and is shaped by structured regularities in the input, whereas standard Transformers apply a largely uniform sequence of computations at every token position. We ask whether dynamic latent segmentation can provide a useful inductive bias for language modeling under BabyLM-scale data constraints. BabyConceptLM causally encodes tokens, groups contextual states into learned contiguous latent segments, processes the shorter sequence with a deeper latent-level backbone, and returns to token resolution through a leakage-safe causal readout. We evaluate two profiles using recipe-matched token-only controls, with analyses of computational cost and segmentation behavior. Across English, multilingual, and Chinese comparisons, BabyConceptLM trails matched token-only controls by 0.60–1.47 points. Estimated padding-aware dense-forward FLOPs are reduced by 26% and 39% for the English and multilingual 5/8/3 models, respectively, but increase for the Chinese 2/4/2 configuration. Segmentation audits show weak alignment with word boundaries and predominantly length-driven behavior, with the larger model frequently reaching its eight-token segment cap. These results show that latent segmentation can yield a favorable compute–accuracy trade-off in some configurations, but does not by itself produce linguistically structured units or improved predictive performance. BabyConceptLM serves as a diagnostic test of hierarchical latent computation under restricted language exposure.

1 Introduction

The BabyLM Challenge provides a controlled setting for asking which inductive biases support language learning under restricted exposure, where training data cannot simply be scaled up (Warstadt

et al., 2023; Hu et al., 2024; Charpentier et al., 2025). This raises an architectural question: how should computation be allocated across a sequence under limited exposure? Standard Transformer language models apply largely uniform computation at every token position (Vaswani et al., 2017), although linguistic information spans multiple scales, from lexical patterns to longer multi-token units. This motivates testing whether computation can be redistributed more efficiently.

A natural alternative is hierarchical computation: local token representations are integrated into larger latent units, allowing deeper computation on a shorter sequence. Related approaches use sentence representations (LCM team et al., 2024), entropy-defined byte patches (Pagnoni et al., 2025), or learned chunks (Hwang et al., 2026); continuous latent reasoning explores a complementary use of model-internal states (Hao et al., 2025). Dynamic Large Concept Models provides the closest architectural basis, combining token-state segmentation, latent processing, and causal readout (Qu et al., 2026). Multilingual interpretability studies also describe cross-language concept alignment and specialized intermediate processing stages (Wendler et al., 2024; Zhao et al., 2024; Brinkmann et al., 2025; Tamo et al., 2026). Whether learned latent segmentation is useful under strongly constrained language exposure remains unclear.

We test this hypothesis through dynamic latent segmentation. The model predicts boundaries between contiguous token representations without requiring an external segmenter at inference time. Such grouping could compress locally predictable information while reserving deeper computation for higher-level representations. We evaluate two complementary dimensions: the performance–compute trade-off and the extent to which learned boundaries reflect structure beyond segment length.

*Corresponding author: qcai@psy.ecnu.edu.cn

BabyConceptLM adapts the Dynamic Large Concept Models design to BabyLM-scale pretraining using a GPT-BERT-derived Hugging Face scaffold (Charpentier and Samuel, 2024). It causally encodes tokens, predicts boundaries, pools states within each segment, processes the shorter sequence with a deeper latent-level backbone, and returns to token resolution through a leakage-safe causal readout. We use *concept* operationally for learned contiguous latent segments, not assumed semantic or linguistic concepts.

We evaluate token-encoder/concept-backbone/readout layer-count profiles 2/4/2 and 5/8/3 against recipe-matched token-only controls, denoted TA1. We ask three questions. **RQ1:** Can latent-segment computation remain competitive while reducing estimated computation? **RQ2:** Do learned boundaries reflect structure beyond segment length? **RQ3:** What multilingual performance and participant-level brain-predictivity patterns does the model exhibit? The first two questions form the primary architectural test. English–Dutch–Chinese evaluation and a separate ChineseBabyLM case study (Song et al., 2026) provide complementary diagnostics, with fMRI analyses motivated by prior model–brain correspondence studies (Caucheteux and King, 2022).

The results reveal a configuration-dependent trade-off. Across three frozen recipe-matched comparisons, BabyConceptLM trails token-only controls by 0.60–1.47 points on shared non-finetuning task aggregates. Padding-aware dense-forward FLOPs are lower for English and multilingual 5/8/3, but slightly higher for Chinese 2/4/2 (Table 2a). The held-out English segmentation audit shows predominantly length-driven behavior: 2/4/2 concentrates near six-token segments, whereas 5/8/3 frequently reaches its eight-token cap, with weak source-word-start alignment (Appendix Table A20). Multilingual effects are heterogeneous (Table 3), and participant-level Chinese fMRI contrasts remain inconclusive (Table 2b).

These findings show why learned latent units require direct audits: sequence compression alone does not establish meaningful higher-level structure. Our contributions are: (1) a causal hierarchical language model adapting learned latent segmentation to BabyLM-scale pretraining; (2) recipe-matched token-only comparisons with padding-aware computational accounting; and (3) segmentation, multilingual, and exploratory brain-

predictivity analyses that characterize the potential and limitations of the learned hierarchy.

2 Method

2.1 Architecture overview

Intuitively, BabyConceptLM encodes tokens causally, predicts boundaries, pools two segment-state sequences, applies the same concept-backbone weights to each, and maps the readout branch back to token resolution. The auxiliary branch provides probe states and optional syntax supervision; the readout branch supplies causal token memory. Figure 1 shows this flow, and Appendix Figure A1 gives the full computation graph.

Given tokens $x_{1:T}$, BabyConceptLM predicts x_{t+1} causally while replacing a uniform token-only stack with a token encoder, a concept backbone, and a token-level readout. We write a profile as token-encoder/concept/readout layer counts, so 2/4/2 has two token, four concept, and two readout layers, whereas 5/8/3 has five, eight, and three. The representation-producing path, with both calls to the shared concept backbone shown explicitly, is

$$\begin{aligned} H &= E(x_{1:T}), & (s, A) &= \Phi(H), \\ Z_{\text{aux}}^{(0)} &= P(H, A), & Z_{\text{aux}} &= M(Z_{\text{aux}}^{(0)}), \\ Z_{\text{rd}}^{(0)} &= P(H, A_{\text{rd}}), & Z_{\text{rd}} &= M(Z_{\text{rd}}^{(0)}), \\ r_t &= D(h_t, Z_{\text{rd}}, A_{\text{rd}}), & u_t &= F(r_t, h_t, p_t). \end{aligned} \quad (1)$$

where E is the causal token encoder, Φ returns boundary scores s and straight-through assignments A , P is the shared pooling/projection module, M is the shared causal concept backbone, D is the leakage-safe readout, and F is optional residual fusion. The reported 2/4/2 profiles set $A_{\text{rd}} = \text{sg}(A)$, whereas the 5/8/3 profiles set $A_{\text{rd}} = A$; sg is forward-pass identity with zero derivative. D uses the corresponding forward hard segment indices to construct its causal memory mask; detachment affects readout-branch pooling gradients, not those indices. The boundary loss acts directly on s . The two calls to M share parameters but are separate executions.

Output logits are computed from u_t ; when fusion is disabled, $F(r_t, h_t, p_t) = r_t$. We refer to each learned contiguous latent segment as a concept. The implementation also executes the pool/projection embedded inside Φ before the two

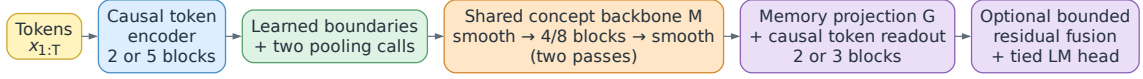


Figure 1: BabyConceptLM overview. Both branches execute the shared backbone, including causal smoothing before and after its blocks. The readout projects completed concept states through G and accesses the causal prefix of the current segment; optional token residual fusion is a separate operation.

named pooling calls; its returned pooled tensors are not consumed downstream. Compute estimates include all three pooling/projection executions and both concept-backbone executions.

2.2 Dynamic segmentation and pooling

The input sums token, position, and, in multilingual training, language embeddings before the token encoder produces $H = (h_1, \dots, h_T)$. For $t > 1$, the content boundary scorer computes

$$\begin{aligned} s_t &= \text{MLP}[h_t; h_{t-1}; |h_t - h_{t-1}|; \\ &\quad h_t \odot h_{t-1}; f_t; f_{t-1}] + b, \\ p_t &= \sigma((s_t - \beta)/\tau), \end{aligned} \quad (2)$$

where f_t is a configuration-dependent structural-label embedding and is zero when that input is disabled. Sequence starts, EOS separators and their following positions, and BOS tokens when present are forced boundaries.

The mass decoder scans left to right. It starts a new segment when a forced boundary occurs, accumulated probability reaches a profile-specific threshold ρ , or the current segment has reached $L_{\max} = 8$ tokens; the accumulator is then reset. Frozen-checkpoint summaries record $\rho = 1.0$ and 0.35 for the audited 2/4/2 and 5/8/3 English profiles, respectively. Both use the same hard mass-decoding rule during training and inference. Let $d_t \in \{0, 1\}$ mark a forced sequence/document boundary, $\bar{p}_t = \max(p_t, d_t)$, and $b_t \in \{0, 1\}$ be the decoder’s hard boundary indicator. Gradients pass through

$$\tilde{b}_t = \bar{p}_t + \text{sg}(b_t - \bar{p}_t), \quad (3)$$

whose forward value is b_t and whose gradient follows $\tilde{p}_t$. For relaxed segment coordinate $\gamma_t = \sum_{i \leq t} b_i - 1$, let C be the number of active segments in the sequence and normalize over their

zero-based indices:

$$\begin{aligned} q_{tc} &= \frac{\exp[-(\gamma_t - c)^2 / (2\sigma_A^2)]}{\mathcal{N}_t}, \\ \mathcal{N}_t &= \sum_{j=0}^{C-1} \exp[-(\gamma_t - j)^2 / (2\sigma_A^2)]. \end{aligned} \quad (4)$$

for $c = 0, \dots, C - 1$, with $\sigma_A = 0.7$. If $c(t)$ is the hard segment index of token t , then $A_{tc}^{\text{hard}} = \mathbf{1}[c = c(t)]$ and $A = q + \text{sg}(A^{\text{hard}} - q)$. Thus the forward assignment is hard while its gradient follows the Gaussian relaxation. The cap changes the forward hard boundary, but the implementation applies no cap-specific gradient mask: away from forced positions, the straight-through derivative still follows p_t . For branch $k \in \{\text{aux}, \text{rd}\}$, take $A^{(\text{aux})} = A$ and $A^{(\text{rd})} = A_{\text{rd}}$. The shared pooling module initializes active segment c as

$$\begin{aligned} \bar{h}_{c,k} &= \frac{\sum_t A_{tc}^{(k)} h_t}{\max\{1, \sum_t A_{tc}^{(k)}\}}, \\ z_{c,k}^{(0)} &= \bar{h}_{c,k} W_{\text{up}} + e_c^{\text{seg}}. \end{aligned} \quad (5)$$

Here W_{up} is the learned projection from token width to concept width, the denominator reproduces the implementation’s lower clamp at one, and e_c^{seg} is the segment-position embedding. In both branches, M applies causal smoothing before and after its shared Transformer stack; Appendix A defines the smoother and architectural settings.

2.3 Leakage-safe token-level readout

Pooling a complete current segment would reveal later tokens within that segment. Let $c(t)$ be the hard segment containing position t . Its causal prefix representation is

$$\begin{aligned} n_{c(t),t} &= \max \left\{ 1, \sum_{i \leq t} \mathbf{1}[c(i) = c(t)] \right\}, \\ \tilde{z}_{c(t),t} &= \frac{\sum_{i \leq t} \mathbf{1}[c(i) = c(t)] h_i}{n_{c(t),t}} W_{\text{up}} + e_{c(t)}^{\text{seg}}. \end{aligned} \quad (6)$$

Within D , $G(z) = \text{Linear}(\text{RMSNorm}(z))$ projects completed backbone states at concept width, independently of optional token residual fusion F . The query at t cross-attends to

$$\mathcal{M}_t = \{G(z_{\text{rd},c}) : 0 \leq c < c(t)\} \cup \{\tilde{z}_{c(t),t}\}, \quad (7)$$

where the current-prefix state bypasses M and G . Consequently, the representation used to predict x_{t+1} contains no position later than t , including within the current segment.

2.4 Bounded token residual fusion

The English 2/4/2 variants and the Chinese Baby-ConceptLM model combine readout state r_t and token-encoder state h_t with a bounded gate,

$$g_t = g_{\max} \sigma(\text{MLP}[r_t; h_t; h_t - r_t; p_t]), \quad (8)$$

$$u_t = \text{RMSNorm}(r_t + g_t \odot (h_t - r_t)). \quad (9)$$

The gate permits a bounded contribution from token-encoder states while retaining the concept-mediated readout as the main prediction path. The English 5/8/3 variants and the shared multilingual model predict directly from the token-level readout; Table A2 records enabled variants, and gate bounds appear in Appendix A.

2.5 Depth weighted averaging as an auxiliary variant

We test DWA within 2/4/2 (105.6M parameters) and 5/8/3 (194.6M) token/concept/readout profiles. The noDWA and DWA variants share the corpus, exposure, base architecture, and optimization recipe; DWA adds only 24 and 244 scalar coefficients, so these comparisons are near-parameter-matched. The 2/4/2 implementation uses an unconstrained layer-level combination, while 5/8/3 applies softmax-weighted mixtures after each attention and feed-forward residual update. DWA operates in the token, concept, and readout paths. Cross-profile contrasts also change depth, size, schedule, and output fusion and are not DWA ablations; implementation details appear in Appendix A.

2.6 Matched controls and FLOPs accounting

The 19-layer English and multilingual controls use width 768 and 192.4M parameters, versus 194.6M for 5/8/3; the 10-layer Chinese control has 107.4M parameters, versus 105.6M for 2/4/2. Within each frozen pair, corpus, tokenizer, exposure, optimizer/schedule, batch construction, and

seed are held fixed. Our compute metric is estimated dense-forward FLOPs: counts cover matrix products in every executed path, including both concept-backbone passes, and assign two FLOPs per multiply-accumulate. Padding-aware values allocate the within-microbatch maximum concept length over every dev sequence. Appendix A specifies the formulas, tensor shapes, batching, and exclusions.

2.7 Training objective

All reported pretraining checkpoints use causal next-token prediction only; the retained training arguments set MNTP sampling probability to zero throughout. The enabled loss is

$$\mathcal{L} = \mathcal{L}_{\text{CLM}} + \lambda_{\text{syn}} \mathcal{L}_{\text{syn}} + \lambda_{\text{bnd}} \mathcal{L}_{\text{bnd}} + \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + 10^{-4} \mathcal{L}_z. \quad (10)$$

Boundary supervision is one binary cross-entropy on a composite soft target, not a sum of separately weighted boundary losses. The target combines forced document boundaries, detached adjacent-state cosine change, and enabled tokenizer-derived heuristic-label or phrase-role transitions. In Chinese, an aligned LTP word-start label replaces this composite target before forced document boundaries are restored. The segmentation term regularizes expected and straight-through boundary rates, active-segment lengths, and optionally boundary entropy. Appendix A.1 formally defines every target, coefficient, index, clamp, and loss term and records the profile-specific weights.

3 Pretraining and evaluation

The four English models form two near-parameter-matched DWA ablations. Because the 5/8/3 profile was more competitive in the English NLP and exposure-indexed analyses, we use that profile for the shared MULTILINGUAL entry rather than conducting a second architecture search across languages. The ChineseBabyLM case study instead uses 2/4/2 to retain the smaller 105.6M-parameter budget and bounded token residual, which preserves a direct route for character-level evidence in Hanzi and fMRI evaluation (Song et al., 2026).

3.1 Corpora and pretraining

English STRICT. The English models use the official 100M-word STRICT corpus (Choshen et al., 2026) and a case- and punctuation-preserving 16,384-entry NFKC ByteLevel BPE tokenizer. Tokenized documents are concatenated

with EOS separators and then sliced into fixed-length packed sequences; a sequence may therefore contain portions of multiple documents, while the separator-aware boundary mechanism prevents an unmarked transition between them.

MULTILINGUAL. The multilingual mixture contains 43.095M English, 40M Dutch, and 15M Chinese raw words, corresponding to 100M Byte-Premium-adjusted words under the official factors (Choshen et al., 2026). A shared 16,384-entry tokenizer is trained round-robin; each language is packed separately and identified by a language embedding.

Chinese case study. The Chinese model uses 101.6M Jieba-counted tokens filtered from the baike2018qa encyclopedia-question-answer corpus (hereafter BaikaQA) (Xu, 2019), with 80% used for training. A 16,384-entry NFKC Metaspaces BPE tokenizer is trained from scratch. Offline N-LTP processing (Che et al., 2021) provides word-start and part-of-speech (POS) sidecar labels. The retained environment and cache resolve PyPI ltp==4.2.14 and LTP/small@dea87bf0 (full hash in Appendix B). The original preprocessing run recorded the model name but not its revision in a lock file. No LTP parameters, hidden states, or embeddings initialize or enter BabyConceptLM, and LTP is absent at evaluation. Appendix B gives the filtering and accounting procedure.

Matched controls. Each TA1 model removes segmentation, concept, and readout modules while retaining its paired corpus, tokenizer, exposure sequence, packed batches, AdamW configuration and schedule, and seed 42. English TA1 matches English matched BabyConceptLM (token/concept/readout DWA, initialization 8/8/8), and multilingual TA1 matches Multilingual matched BabyConceptLM (DWA, 6/4/4); the matched references are archived recipes, with public-final submissions reported separately. The Chinese pair also shares the absence of DWA. Its intervention compares the complete concept pathway with token-only depth allocation: boundary prediction, the concept backbone, token-level readout, and layer allocation vary jointly, with a 1.7% parameter-count difference. Appendix Tables A8 and A9 give pinned identities and public-final recipe differences, respectively.

All models are initialized from scratch and

trained with AdamW, mixed precision, gradient clipping, and seed 42. Exposure is accumulated from the actual whitespace word counts of consumed microbatches, with Byte Premiums applied per language for MULTILINGUAL. The four STRICT runs stop at exactly 1B words, and the multilingual run reaches the 1B adjusted-word target within floating-point tolerance. Optimization and checkpoint details appear in Appendix Table A4.

3.2 Evaluation

We pin the official babylm-eval repository at 6f825c2¹ and chinese-babylm-pipeline-final at f4437e1². Source benchmarks include BLiMP (Warstadt et al., 2020), EWok (Ivanova et al., 2025), Entity Tracking (Kim and Schuster, 2023), COMPS (Misra et al., 2023), and GLUE/SuperGLUE (Wang et al., 2019b,a). Challenge-specific variants and aggregation follow the pinned pipeline. STRICT Reading is normalized incremental R^2 , not comprehension accuracy; AoA is 100 times the model-child correlation and is zeroed when $p > .1$. Appendix D gives the complete task/source inventory and metric caveats.

For Chinese fMRI, matched participant IDs give $n = 12$ in six discourse domains and $n = 11$ in Word-fMRI. The focal exploratory Cognition and Language contrasts use paired bootstrap intervals, exact sign-flip tests, Holm correction, and paired effect d_z (Appendix D.1). These tests quantify participant sampling uncertainty conditional on two seed-42 models, not training-seed variance.

Appendix C records checkpoint identities, recipe differences, and score verification; Appendix E gives the English Hippocampus audit protocol (Sap et al., 2020).

4 Results

English STRICT performance, matched controls, and compute accounting address RQ1; the frozen segmentation audit addresses RQ2; and the multilingual and Chinese analyses address RQ3. Scores are percentages unless stated otherwise.

¹Full commit: <https://github.com/babylm-org/babylm-eval/tree/6f825c291e2c4c78ad33b1935fd64d45f52642dc>

²Full commit: <https://github.com/chinese-babylm/chinese-babylm-pipeline-final/tree/f4437e1aaecf18ab5defb6b148857ef7d056e9d6>

4.1 English STRICT performance and estimated forward compute

The primary leaderboard summary is the official *Overall* aggregate (hereafter *Overall*). “Strict-Interaction” reproduces the official baseline label and scores from the pinned evaluation README; it is the 98.4M-parameter public artifact at commit ba5a7aa, not a model trained in this study³.

All four English variants have complete final evaluations. The 5/8/3 variants are the more competitive profile: noDWA is 1.09 Overall points below GPT-2 but only 0.13 lower on NLP, while scoring 1.09 higher on BLiMP and 2.55 higher on GlobalPIQA. Within-profile DWA effects remain mixed. In 2/4/2, DWA changes Overall by -1.41 , NLP by -0.33 , and (Super)GLUE by -0.73 ; in 5/8/3, the corresponding changes are $+0.16$, -0.22 , and -0.35 . The larger 2/4/2 Overall difference is driven primarily by the thresholded Human-like component rather than its NLP aggregate. DWA therefore changes task balance without a consistent aggregate improvement.

Matched token-only control. All main non-finetuning evaluations are complete for the three frozen TA1 pairs. Table 2 separates these within-pair comparisons from the public-final rows above. The English shared macro contains exactly BLiMP, BLiMP Supplement, EWOK, Entity Tracking, and COMPS. Token-only TA1 leads its paired concept reference by 1.47 English points, 0.60 multilingual points, and 0.82 Chinese points. These single-seed pairs estimate non-finetuning performance for the frozen matched recipes; public-final comparisons are reported separately. Because TA1 finetuning is incomplete, Appendix Table A10 reports the shared non-finetuning tasks rather than an *Overall* control score.

Under this FLOPs metric, complete-dev random batching gives 25.94% and 39.33% estimated savings for English and multilingual 5/8/3, whereas Chinese batches make 2/4/2 5.5% more expensive than TA1. Appendix Tables A6–A7 give the distributions and exclusions.

The 2/4/2 profile remains weaker on NLP after complete finetuning. Its Human-like peak is driven by a marginal AoA correlation ($r = .104$, $p = .091$) crossing the pipeline’s $p \leq .1$ threshold; Reading is 0.13–0.22 and has a target-tokenization

mismatch. We therefore retain the official values but do not treat them as robust cross-tokenizer cognitive rankings (Appendix D).

4.2 Exposure-indexed learning

All five BabyLM submissions have fast evaluations at the 28 required revisions. STRICT curves macro-average BLiMP, its Supplement, EWOK, Entity Tracking, GlobalPIQA, and official Reading; tokenizer choice is controlled within each English pair, but the macro inherits the Reading mismatch. MULTILINGUAL curves average official zero-shot families within language and then across languages; panels are not compared because their inventories differ.

Both 5/8/3 variants cross 60 BLiMP between the saved 20M and 30M checkpoints, whereas both 2/4/2 variants cross between 300M and 400M. Thus 30M and 400M are checkpoint upper bounds, not exact acquisition times. Within-profile noDWA/DWA crossings coincide, and their log-exposure AUCs differ by at most 0.26.

Appendix Figure A2 and Table A16 provide the complete trajectories and numerical anchors.

This comparison is recipe-level: relative to 2/4/2, 5/8/3 nearly doubles parameters and blocks, receives about twice as many updates per early token, uses larger learning rates, and adds heuristic syntax/phrase supervision. Isolating the contribution of depth, segmentation, or any other factor requires matched interventions; Appendix Table A5 enumerates the co-varying factors.

4.3 English held-out segmentation behavior

Table A20 quantifies concentration around six-token segments, saturation at the eight-token cap, and alignment with source-word starts. The learned boundary rate excludes window starts and cap-attributed cuts; word-start AUC measures threshold-independent alignment between boundary probabilities and source-word starts.

The 2/4/2 profiles concentrate near six tokens but retain measurable length variation; six-token segments account for 70.16% (noDWA) and 63.32% (DWA), with word-start AUCs of 0.543 and 0.546 indicating weak word-start sensitivity. The 5/8/3 profiles are predominantly cap-driven: DWA raises $P(L = 8)$ by 11.42 points and the cap-attributed boundary share from 58.20% to 69.63%, with word-start AUC 0.448 in both variants. Together, these results characterize length-dominated model-grounded segmentation;

³Hugging Face artifact at the full pinned revision.

Model	Overall	NLP	Human-like	BLiMP	GlobalPIQA	(Super)GLUE
GPT-2 Strict	40.73	53.03	−2.32	74.73	36.62	67.75
Strict-Interaction	39.89	52.42	−3.96	72.89	36.18	67.28
BabyConceptLM 2/4/2 noDWA	38.67	48.21	5.27	62.35	35.67	64.45
BabyConceptLM 2/4/2 DWA	37.26	47.88	0.09	62.44	34.17	63.72
BabyConceptLM 5/8/3 noDWA	39.64	52.90	−6.78	75.82	39.17	67.45
BabyConceptLM 5/8/3 DWA	39.80	52.68	−5.29	76.58	37.59	67.10

Table 1: English STRICT results at 1B word exposures. All four BabyConceptLM rows include complete zero-shot and seven-task finetuning evaluations. Baseline rows are recomputed from the official evaluation README. Human-like is the pipeline-reported mean of Reading and AoA and inherits the metric caveats described in the text. Complete benchmark-family and finetuning results appear in Appendix Tables A11–A12.

(a) Matched non-finetuning performance and forward FLOPs													
Frozen pair				Concept	TA1	Δ	score	Ideal GF	C/T	Padded GF	C/T	Δ	padded
English	5/8/3	matched	BabyConceptLM–	52.74	54.21	−1.47		72.14/102.07		75.59/102.07		−25.9%	
TA1-19L	†												
MULTI	5/8/3–TA1-19L	†		50.43	51.02	−0.60		61.59/102.07		61.92/102.07		−39.3%	
Chinese	2/4/2–TA1-10L			57.73	58.55	−0.82		47.07/56.77		59.91/56.77		+5.5%	

(b) Paired Chinese CogBench contrasts											
Contrast				n	Δ	(C–T)	95% CI	d_z	Exact p	Holm p	C>T
Cognition				12	+0.288		[−0.274, 0.797]	0.29		.324/.436	9/12
Language				12	+0.281		[−0.118, 0.687]	0.38		.218/.436	7/12
Cognition + Language mean				12	+0.284		[−0.086, 0.645]	0.42		.174/–	9/12
All six discourse domains				12	+0.178		[−0.109, 0.448]	0.35		.257/–	9/12
Word-fMRI				11	−0.159		[−0.566, 0.251]	−0.22		.496/–	4/11

Table 2: Matched-control results. Panel (a): scores are concept minus token. English averages the five explicitly named shared families in the text; MULTILINGUAL is the official fixed-denominator mean of language-level zero-shot averages; Chinese averages ZhoBLiMP, XCOMPS-ZH, and open/hidden Hanzi structure/pinyin. Daggers mark the pinned matched-recipe concept repositories, which are distinct from the public-final submission checkpoints. Ideal FLOPs evaluate the analytic formula at each frozen checkpoint’s full-dev mean concept count. For Chinese, all 89,909 packed dev sequences give $\bar{C} = 86.883$ concepts per sequence and hence 47.07 GF. English/MULTILINGUAL padded values average batch maxima over 200 complete-dev-set random groupings at $T = 256$, $B = 16$; Chinese uses all 89,909 dev sequences at $B = 32$. Negative padded deltas indicate savings. Panel (b): correlations are multiplied by 100; C/T denote concept/token. CIs use paired bootstrap; p -values use exact two-sided sign-flip tests. Holm values in the two focal exploratory rows correct Cognition and Language.

full distributions and paired intervals appear in Appendix E. The nearly closed 2/4/2 fusion gate suppresses the direct token-level residual, although h_t still serves as the readout query.

4.4 Multilingual results

The public multilingual entry is 0.54 points above trilingual GPT-2 and narrows the language spread from 4.29 to 1.61; lower Dutch zero-shot together with higher Chinese *Overall* and Dutch/Chinese finetuning show heterogeneous transfer. This public comparison differs in mixture, size, tokenizer, and architecture; the matched zero-shot pair is reported separately, with TA1 at 51.02 versus 50.43 for 5/8/3. Appendix Tables A13–A15 report all tasks. MECO is excluded from the aggregate; displayed MECO and Hanzi scores were public-

Block	Model	Multi	EN	NL	ZH	Spread
Overall	GPT-2 tri.	45.94	47.60	46.92	43.31	4.29
	Ours	46.48	47.47	46.11	45.86	1.61
Zero-shot	GPT-2 tri.	n/a	53.63	55.05	47.25	n/a
	Ours	n/a	53.75	50.73	49.52	n/a
Finetune	GPT-2 tri.	n/a	43.09	39.82	40.69	n/a
	Ours	n/a	42.77	42.06	43.42	n/a

Table 3: Public MULTILINGUAL aggregates for babyconceptLM-multi (Ours) and the official trilingual GPT-2 reference. Spread is the range across language Overall scores.

leaderboard-verified at manuscript freeze.

4.5 Chinese linguistic and cognitive trade-off

We analyze ChineseBabyLM as a separate case study outside the BabyLM 2026 submission. Ap-

pendix Table A17 separates matched TA1 from published and unmatched baselines (Charpentier and Samuel, 2024; Song et al., 2026). The ChineseBabyLM 2026 final-leaderboard Space is accessible and lists babyconceptLM-baike102M_bpe40; its displayed task scores match the retained outputs produced by pinned pipeline commit f4437e1. fMRI correlations are multiplied by 100.

Across six linguistic/Hanzi measures, TA1 averages 58.55 versus 57.73; BabyConceptLM is higher only on hidden pinyin (+0.05). Word-fMRI also favors TA1, 56.77 versus 56.61, whereas mean discourse fMRI favors BabyConceptLM, 10.92 versus 10.74.

Participant contrasts remain inconclusive: Cognition is +0.288 for BabyConceptLM (exact $p = .324$, Holm $p = .436$), Language is +0.281 ($p = .218/.436$), and their participant-wise mean is +0.284 with 95% CI $[-0.086, 0.645]$ and $p = .174$. No domain survives six-way Holm correction; Reward favors TA1 before correction ($p = .032/.193$). Appendix Figures A3–A5 instead use an unmatched GPT-BERT baseline and are descriptive.

5 Discussion and Conclusion

This work investigates whether learned latent segmentation can provide a useful inductive bias for language modeling under BabyLM-scale exposure. BabyConceptLM groups tokens into latent segments and processes the shorter sequence before returning to token-level prediction. Across three frozen recipe-matched comparisons, it scores 0.60–1.47 points below token-only controls on the evaluated non-finetuning aggregates. Its computational benefit depends on the configuration: padding-aware dense-forward estimates are 26% lower for English and 39% lower for multilingual 5/8/3, but 6% higher for Chinese 2/4/2. Thus, latent compression offers a configuration-dependent trade-off between estimated computation and predictive performance.

The held-out English audit shows that learned segmentation is dominated by segment length and the imposed length cap, with weak alignment to word starts. A learned boundary mechanism can therefore settle into length-driven grouping under the present objectives. Future hierarchical models should test stronger objectives or constraints that encourage linguistically informed segmentation,

together with fixed-length and random-boundary controls to establish the value of learned boundaries.

Overall, BabyConceptLM provides a controlled study of latent segmentation under limited language exposure. Its findings highlight three priorities for evaluating alternative architectures: verify what the learned segments capture, account for the computation actually allocated under padding, and compare predictive performance under matched training conditions. Together, these checks clarify when hierarchical computation offers a useful trade-off for small-scale language learning.

Limitations

First, the learned segmentation remains largely length-driven. Our English narrative audit measures word-start alignment; testing syntactic, semantic, and discourse boundaries and other languages requires broader evaluation. Stronger structural objectives and fixed-length or random-boundary controls would help assess which benefits depend on learned boundaries.

Second, some matched token-only comparisons jointly change segmentation, latent computation, readout, and layer allocation, leaving individual component effects to be isolated. The controls cover shared non-finetuning tasks, with English and multilingual pairs based on archived recipes rather than final submissions. The 30M-versus-400M BLiMP contrast also compares recipes differing in model size, optimization, and supervision, so it identifies a recipe-level learning-speed difference.

Third, experiments use a single training seed and BabyLM-scale models. Multiple seeds and larger models are needed to assess robustness across runs and scales. Participant-level fMRI effects are small and inconclusive, limiting cognitive interpretation.

Finally, our compute analysis estimates dense forward FLOPs, including batch padding. End-to-end speed, peak memory, and energy benefits require direct hardware measurements.

References

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the*

- Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Charlotte Caucheteux and Jean-Rémi King. 2022. [Brains and algorithms partially converge in natural language processing](#). *Communications Biology*, 5:134.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. [Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: Why not both?](#) In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.
- Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. [N-LTP: An open-source neural language technology platform for Chinese](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. 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*.
- Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. [MuBench: Assessment of multilingual capabilities of large language models across 61 languages](#). *arXiv preprint arXiv:2506.19468*.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. [Training large language models to reason in a continuous latent space](#). In *Proceedings of the Second Conference on Language Modeling*.
- 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.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. [Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.
- Sukjun Hwang, Brandon Wang, and Albert Gu. 2026. [Dynamic chunking for end-to-end hierarchical sequence modeling](#). In *International Conference on Learning Representations*.
- Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U. Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian C. Paulun, Maria Ryskina, Ekin Akyürek, Ethan G. Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2025. [Elements of world knowledge \(EWoK\): A cognition-inspired framework for evaluating basic world knowledge in language models](#). *Transactions of the Association for Computational Linguistics*, 13:1245–1270.
- Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. [MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *arXiv preprint arXiv:2504.02768*.
- Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.
- LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, and 2 others. 2024. [Large concept models: Language modeling in a sentence representation space](#). *arXiv preprint arXiv:2412.08821*.
- 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*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

- Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srini Iyer. 2025. [Byte latent transformer: Patches scale better than tokens](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9238–9258, Vienna, Austria. Association for Computational Linguistics.
- Xingwei Qu, Shaowen Wang, Zihao Huang, Kai Hua, Fan Yin, Rui-Jie Zhu, Jundong Zhou, Qiyang Min, Zihao Wang, Yizhi Li, Tianyu Zhang, He Xing, Zheng Zhang, Yuxuan Song, Tianyu Zheng, Zhiyuan Zeng, Chenghua Lin, Ge Zhang, and Wenhao Huang. 2026. [Dynamic large concept models: Latent reasoning in an adaptive semantic space](#). In *The Fourteenth International Conference on Learning Representations*.
- Maarten Sap, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020. [Recollection versus imagination: Exploring human memory and cognition via neural language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1978, Online. Association for Computational Linguistics.
- Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, Sara Maria Da Fonseca, Nicolas Dirix, Wouter Duyck, Argyro Fella, Ram Frost, Carolina A. Gattei, Areti Kalaitzi, Nayoung Kwon, Kaidi Lõo, and 12 others. 2022. [Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus \(MECO\)](#). *Behavior Research Methods*, 54(6):2843–2863.
- Siyuan Song, Zhiheng Qian, Yunhao Zhang, Linyang He, Xiaozhe Ji, Yingxin Lin, Hongao Zhu, Chongtian Shao, Chuhan Lang, Luan Li, Rui Wang, Renfen Hu, Shaonan Wang, and Hai Hu. 2026. [The first ChineseBabyLM challenge: Training data-efficient and cognitively plausible language models for Chinese](#). *arXiv preprint arXiv:2607.10745*.
- J. Ben Tamo, Daniel Carlander-Reuterfelt, Jonathan Rubin, Oleg Poliannikov, Dezhi Hong, and Mingxian Wang. 2026. [LinguaMap: Which layers of LLMs speak your language and how to tune them?](#) In *The Fourteenth International Conference on Learning Representations*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. [SuperGLUE: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *International Conference on Learning Representations*.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. [Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- 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.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Bright Xu. 2019. [NLP Chinese Corpus: Large scale chinese corpus for NLP](#).
- Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. [How do large language models handle multilingualism?](#) In *Advances in Neural Information Processing Systems 37*, pages 15296–15319. Curran Associates, Inc.

A Implementation and optimization details

Token and readout blocks have width 768 with 3,072-dimensional feed-forward layers and 12 attention heads; concept blocks have width 896 with 3,584-dimensional feed-forward layers and 14 heads. All blocks use RMSNorm, GEGLU feed-forward layers, normalized queries and keys, and tied input–output token embeddings. Hidden and attention dropout are 0.1; activation dropout is zero.

Concept smoothing and memory projection. Writing $\mathcal{B}$ for the causal concept Transformer stack, including enabled DWA, the shared backbone is $M = S \circ \mathcal{B} \circ S$. For an active concept position c with input state v_c , the causal smoother is

$$S(v)_c = (1 - \alpha_S)v_c + \frac{\alpha_S}{|\mathcal{J}_c|} \sum_{j \in \mathcal{J}_c} v_j,$$

where $\mathcal{J}_c$ contains the valid concept indices from $\max(0, c - k_S + 1)$ through c ; padded outputs are zero. All reported concept profiles use $k_S = 3$ and $\alpha_S = 0.35$ for both smoothing calls in each branch. The readout applies G , an RMSNorm followed by an affine 896×896 projection, to completed concept memory. Current-prefix states use the token-to-concept pooling projection directly and bypass M and G . The projection G is distinct from the optional token residual fusion F .

Path	Width	FFN	Shared settings
Token encoder	768	3,072	RMSNorm, GEGLU
Concept backbone	896	3,584	normalized Q/K
Token-level read-out	768	3,072	tied embeddings

Table A1: Architectural dimensions shared across profiles.

For path states $v_0, \dots, v_i$, DWA computes $\hat{v}_i = \sum_{j=0}^i \alpha_{ij} v_j$. The 2/4/2 coefficients are layer-level, unnormalized, and initialized as $(0, \dots, 0, 1)$, selecting the current residual state. In 5/8/3, each attention sub-block first updates the residual state, which is appended to the state history; DWA then mixes that history and replaces its latest entry with the mixture. The same sequence follows the feed-forward residual update. These coefficients are softmax normalized; English initializes prior-state logits to 0 and the current-state logit to 4. The multilingual model uses current-state logits 6, 4, and 4

in the token, concept, and readout paths. DWA adds 24 and 244 parameters to the English profiles, respectively. The bounded fusion gate uses $g_{\max} = 0.40$ for the English 2/4/2 models and 0.45 for babyconceptLM-baike102M_bpe40.

The English heuristic syntax lookup maps tokenizer pieces to eight buckets: noun-like suffixes, verb-like forms, modifiers, function words, linkers/punctuation, prepositions, clause markers, and a residual/special bucket. A separate eight-way phrase-role lookup uses other/content, verb, adjective, determiner, preposition, linker, clause, and special roles. Phrase starts are generated deterministically at major heads, object starts, and linker/special transitions, except preposition–object continuations. The two builders contain 324 hard-coded set memberships (218 unique normalized strings) and 25 unique suffix endings. They are literal constants in our source, applied only to the in-budget tokenizer vocabulary; no NLTK resource, dictionary file, external corpus lookup, tagger, or parser is loaded, and they add no training text. Code inspection cannot retrospectively establish what references, if any, informed the authors’ original item selection.

A.1 Objective and compute definitions

The retained training arguments for every reported BabyConceptLM checkpoint set `mntp_final_ratio`, `mntp_late_ratio`, `mntp_start_ratio`, and `mntp_ramp_end_ratio` to zero. A legacy `masked_fraction` field remains serialized but is not consulted once the scheduled MNTTP probability is zero. Every training batch therefore enters `mode=clm`, which enables triangular masks in token and concept self-attention.

Boundary supervision is a single binary cross-entropy. Its composite target is

$$y_t = \max\left\{y_t^{\text{doc}}, \frac{1}{2}d_{\cos}(\text{sg}(h_{t-1}), \text{sg}(h_t)), \frac{3}{5}y_t^{\text{syn}}, \frac{17}{20}y_t^{\text{phr}}\right\},$$

$$\mathcal{L}_{\text{bnd}} = \text{BCEWithLogits}(s_t, y_t), \quad (11)$$

where $d_{\cos}(a, b) = (1 - \cos(a, b))/2$; $y_t^{\text{doc}} = 1$ at forced sequence/document boundaries and is zero elsewhere; and y_t^{syn} and y_t^{phr} indicate an adjacent change in the heuristic-label and phrase-role categories defined above, including the deterministic phrase-start rules. Disabled structural targets are omitted. A valid aligned Chinese LTP word-start label replaces the composite target before forced

Model	Setting / profile	DWA / output fusion	Boundary target / structural input or loss	Size
babyconceptLM-en-noDWA	STRICT / 2/4/2	None / bounded residual	Semantic change / none	105.6M
babyconceptLM-en-DWA	STRICT / 2/4/2	Layer linear / bounded residual	Semantic change / none	105.6M
babyconceptLM-en-noDWA	STRICT / 5/8/3	None / readout only	Semantic + heuristic phrase / heuristic syntax + $\mathcal{L}_{\text{syn}}$	194.6M
babyconceptLM-en-DWA	STRICT / 5/8/3	Sub-block softmax / readout only	Semantic + heuristic phrase / heuristic syntax + $\mathcal{L}_{\text{syn}}$	194.6M
TA1 token-only	STRICT / 19 token layers	Token DWA / token only	None / none	192.4M
babyconceptLM-multi	MULTILINGUAL / 5/8/3	Sub-block softmax / readout only	None / none	194.6M
TA1 token-only	MULTILINGUAL / 19 token layers	Token DWA / token only	None / none	192.4M
babyconceptLM-baike102M_bpe40	ChineseBabyLM / 2/4/2	None / bounded residual	LTP word labels; composite fallback / none	105.6M
TA1 token-only	ChineseBabyLM / 10 token layers	None / token only	None / none	107.4M

Table A2: Model configurations and supervision sources. In the supervision column, the slash separates the boundary target from structural input or loss. For Chinese, valid aligned LTP word labels replace the composite target; the latter is used only as a fallback at invalid or unaligned positions. Forced document boundaries and the eight-token decoding cap apply across concept profiles and are not listed as supervision.

document boundaries are restored. For English 5/8/3, $\mathcal{L}_{\text{syn}}$ predicts the majority heuristic bucket in segment $c + 1$ from auxiliary-branch state c .

Let ℓ^* be the target tokens per segment, $r^* = 1/\ell^*$ its target boundary rate, $\bar{p}$ the valid-token mean of p_t , $\bar{b}$ the valid-token mean of $\tilde{b}_t$, and ℓ_c the length of active segment c . For valid positions V , $N = |V|$, and probabilities clamped within log-arithmetic, define

$$\mathcal{H}(p) = -\frac{1}{N} \sum_{t \in V} [p_t \log p_t + (1 - p_t) \log(1 - p_t)], \quad (12)$$

$$\mathcal{L}_{\text{seg}} = w_{\text{exp}}(\bar{p} - r^*)^2 + w_{\text{hard}}(\bar{b} - r^*)^2 + w_{\text{len}}\mathcal{L}_{\text{len}} + w_{\text{ent}}\mathcal{H}(p), \quad (13)$$

$$\mathcal{L}_{\text{len}} = \frac{1}{C} \sum_{c=0}^{C-1} \left([L_{\min} - \ell_c]_+^2 + [\ell_c - L_{\max}]_+^2 + w_*[(\ell_c - \ell^*)/\ell^*]^2 \right), \quad (14)$$

where $[a]_+ = \max(a, 0)$, C is the number of active segments, $L_{\min}$ and $L_{\max}$ are configured bounds, and the four w coefficients plus w_* are nonnegative profile-specific weights. The implementation averages the final length term over all active segments in the minibatch. For vocabulary logit o_{tv} at a valid target, $\mathcal{L}_z = N^{-1} \sum_t (\log \sum_v e^{o_{tv}})^2$.

For dense-forward accounting, a self-attention block at sequence length L , width d , and GEGLU feed-forward width f contributes

$$F_{\text{self}} = 8Ld^2 + 4L^2d + 6Ldf, \quad (15)$$

using two FLOPs per multiply-accumulate. The dense readout concatenates C_B concept slots and T current-prefix slots, allocating memory length $M_B = C_B + T$, where C_B is the within-microbatch maximum concept count. Its causal mask restricts access without shrinking these dense tensors. With token width d , concept width d_c , and readout GEGLU width f_{rd} , one readout block contributes

$$F_{\text{rd}} = 4Td^2 + 4M_Bd_c d + 4TM_Bd + 6Tdf_{\text{rd}}.$$

This counts query/output projections at length T , key/value projections at length M_B , both $T \times M_B$ attention products, and GEGLU. Evaluated shapes are $T = 256$, $d = 768$, $d_c = 896$, token/readout FFN 3,072, and concept FFN 3,584. Counts also include boundary projections, all three pooling/projection calls, both shared backbone executions, prefix-current pooling, the concept-memory projection G once ($2C_Bd_c^2$), tied vocabulary projection, and the enabled Chinese gate. They exclude embedding lookup, normalization, smoothing averages, nonlinearities, softmax, loss, optimizer operations, segmentation control flow, and hardware utilization. English and multilingual use 200 complete-dev-set random partitions at $B = 16$; Chinese uses all 89,909 dev sequences at $B = 32$. These estimates cover matrix operations rather than profiler-measured runtime.

For the audited English 2/4/2 models, the frozen summaries record $\rho = 1.0$, $\ell^* = 4.0$, and $(w_{\text{exp}}, w_{\text{hard}}, w_{\text{len}}, w_*, w_{\text{ent}}) = (0.3, 0.5, 1.0, 0.1, 0)$. For English 5/8/3, the corresponding values are $\rho = 0.35$, $\ell^* = 3.2$, and $(0.5, 0.5, 0.25, 0.08, 0)$. Both profiles use $L_{\min} = 1$, $L_{\max} = 8$, boundary offset $\beta = 0.18$, probability temperature $\tau = 1.0$, Gaussian

Setting	λ_{syn}	λ_{bnd}	λ_{seg}	Structure / schedule
English 2/4/2	0	0.08 $\rightarrow$ 0.06	0.03	semantic target; boundary decay
English 5/8/3	0.08	0.16 $\rightarrow$ 0.12	0.02	syntax input/loss; phrase target; boundary decay
MULTILINGUAL	0	0	0.02	auxiliary term decays at 30–60%
Chinese 2/4/2	0	0.08 $\rightarrow$ 0.06	0.03	LTP word overwrite; composite fallback; stage decay

Table A3: Auxiliary objectives and schedules. All reported runs use causal language modeling, no masked-objective sampling, and logit regularization $10^{-4}\mathcal{L}_z$.

assignment width $\sigma_A = 0.7$, and the hard mass decoder in training and inference. A sampling temperature of 0.5 is configured but inactive under mass decoding. The same summaries record $\lambda_{\text{syn}} = 0.08$ for both audited English 5/8/3 checkpoints; all other rows use zero. The English schedule decays only the outer boundary weight.

Learning rates use 3% linear warmup followed by cosine decay and a final linear cooldown. Public-final English 5/8/3 uses a 0.1 floor and 1.6% cooldown; English matched BabyConceptLM/TA1 and both multilingual recipes use a 0.03 floor and 3% cooldown. The 2/4/2 μP groups use base width 768: token and unclassified matrix parameters retain 4×10^{-5} and Adam $\epsilon = 10^{-8}$, while width-896 concept matrices use 3.43×10^{-5} and 8.57×10^{-9} ; biases, norms, embeddings, and the tied head are assigned to the unscaled “other” group.

Exposure is counted from original whitespace-delimited words before tokenization; multilingual samples additionally store the language and Byte-Premium-adjusted count. Counts are summed across microbatches and devices. The first state meeting each official target is saved at every 1M words through 10M, every 10M through 100M, and every 100M through 1B. The multilingual terminal checkpoint differs from the nominal 1B adjusted-word target only by floating-point round-off and supplies the 1B evaluation revision. Each checkpoint records target and actual exposure, overshoot, tokenizer, model configuration, and loading code; optimizer state is retained separately for resumption.

B Corpus construction

The English corpus contains 100M whitespace-delimited words in 11,601,896 documents. The multilingual mixture uses official English, Dutch, and Chinese streams with Byte Premiums 1.0, 1.0516, and 0.9894, respectively. One pass con-

tains 43.095M, 40M, and 15M raw words, or exactly 100M adjusted words. Tokenization and packing are language-specific after training the shared tokenizer, so no sequence crosses a language boundary.

For the Chinese self-compiled corpus, only the answer field of BaikeQA (baike2018qa) (Xu, 2019) is retained. Answers are split at newlines; paragraphs shorter than 32 or longer than 4,096 Chinese characters are removed, and only candidates with more than 32 Jieba tokens are retained. Filtering stops at 101,603,104 Jieba tokens, split contiguously into 81,282,497 training and 20,320,607 validation tokens. The tokenizer is trained on the training portion. Offline N-LTP processing (Che et al., 2021) produces word-boundary sidecars and token-aligned Universal POS (UPOS) sidecars; adjacent UPOS changes are derived inside the model when constructing the composite fallback boundary target. At aligned positions, the word-start label overwrites that target, while invalid or unaligned positions retain the composite fallback. The retained environment and cache retrospectively identify PyPI `ltp==4.2.14` and `LTP/small@dea87bf0da3c5d054bbd426a78d7fa9efd347799`; the original metadata stored LTP/small but did not explicitly pin its revision. These sidecars are training labels only: no LTP parameters, hidden states, or embeddings initialize or enter BabyConceptLM, and LTP is not loaded for evaluation. The suffix *bpe40* denotes 40 training epochs, not vocabulary size.

C Matched-control provenance

Table A8 identifies all matched concept and TA1 checkpoints by public repository, immutable commit, and weight hash. English matched BabyConceptLM and Multilingual matched BabyConceptLM are the references used for the paired comparisons; they are

Models	Corpus / exposure	Length / global batch	Token / concept LR
English 2/4/2 public pair	official 100M / 1B words	256 / 128 $\rightarrow$ $\sim$ 512	4×10^{-5} / 3.43×10^{-5}
English 5/8/3 public pair	official 100M / 1B words	128 $\rightarrow$ 256 at 80% / 128 $\rightarrow$ 64	1.5×10^{-4} / 1.2×10^{-4}
English matched BabyConceptLM / TA1	archived 100M / 1B words	128 $\rightarrow$ 256 at 60% / 128 $\rightarrow$ 64	1.2×10^{-4} / 1.0×10^{-4}
babyconceptLM-multi public final	43.095/40/15M raw / 1B adjusted	128 $\rightarrow$ 256 at 60% / 128 $\rightarrow$ 64	1.0×10^{-4} / 8.0×10^{-5}
Multilingual matched BabyConceptLM / TA1	adjusted thirds / 1B adjusted	128 $\rightarrow$ 256 at 60% / 128 $\rightarrow$ 64	1.0×10^{-4} / 8.0×10^{-5}
babyconceptLM-baike102M_bpe40	81.283M training tokens / 40 epochs	256 / 128 $\rightarrow$ $\sim$ 512	4×10^{-5} / 3.43×10^{-5}

Table A4: Optimization settings. All runs use AdamW, gradient clipping at 1.0, mixed precision, weight decay 0.1, $(\beta_1, \beta_2) = (0.9, 0.98)$, and seed 42. The 2/4/2 concept rate is the implemented 768/896 μ P scaling of the nominal 4×10^{-5} rate.

Factor	2/4/2	5/8/3	Interpretive consequence
Parameters / blocks	105.6M / 8	194.6M / 16	Capacity and depth are not controlled
Early length / global batch	256 / $\sim$ 128	128 / 128	5/8/3 receives about twice as many updates per token
Nominal token / concept LR	$4.0/3.43 \times 10^{-5}$	$1.5/1.2 \times 10^{-4}$	Optimization scale differs
Early scaled token / concept LR	$1.0/0.857 \times 10^{-5}$	$15/12 \times 10^{-5}$	Approximately 15/14-fold early rate difference
Structural signal	Semantic target only	Syntax input/loss + phrase target	Additional structural targets could affect BLiMP
Output fusion	Bounded residual	Readout only	The prediction path also changes

Table A5: Cross-profile factors that co-vary with the exposure-indexed BLiMP difference. The early scaled 2/4/2 rates follow its linear token-batch scaling at the initial 128-sequence global batch; both recipes use 3% warmup. The table enumerates co-varying factors and plausible contributors; estimating individual effects requires matched interventions.

distinct from the final leaderboard artifacts babyconceptLM-583-en@cb0668f69a87 and babyconceptLM-multi@12d023b28c30. Matched analyses use the pinned reference checkpoints; public-final artifacts are reported separately because their preprocessing, tokenizer, exposure, or schedule differs. The Chinese pair is likewise pinned to the public repositories in Table A8. All multilingual values in Table A10 use the task definitions and fixed-denominator collation pinned at babyLM-eval@6f825c2.

Checkpoint interfaces and score verification.

The five BabyLM submissions use Transformers-compatible model/tokenizer interfaces, support autoregressive string scoring, and retain the 28 exposure checkpoints described in Appendix A. Their English and MULTILINGUAL training uses no LTP model or other pretrained tagger/parser; LTP sidecars belong only to the separate Chinese case study (Appendix B). At manuscript freeze, the BabyLM leaderboard ver-

ified the reported scores, including server-scored MECO and hidden Hanzi. Chinese case-study scores come from frozen local outputs of chinese-babylm-pipeline-final@f4437e1; the currently accessible ChineseBabyLM 2026 final leaderboard displays matching values for babyconceptLM-baike102M_bpe40.

D Evaluation inventory and inference

STRICT zero-shot evaluation covers BLiMP (Warstadt et al., 2020), the BLiMP Supplement, EWoK (Ivanova et al., 2025), Entity Tracking (Kim and Schuster, 2023), COMPS (Misra et al., 2023), Reading, and the challenge-defined GlobalPIQA variant; finetuning covers BoolQ, MultiRC, RTE, WSC, MRPC, QQP, and MNLI drawn from GLUE and SuperGLUE (Wang et al., 2019b,a). Reading averages normalized incremental- R^2 scores from four eye-tracking outcomes and spillover-controlled self-paced reading. The frozen implementation independently to-

Pair	$T/\bar{C}$	Ideal GF C/T	Padded GF C/T	History PF C/T	History saving	Padding evidence
English	4.539	72.14/102.07	75.59/102.07	509.27/631.39	19.3%	2,687 seq.; 200× random $B = 16$
MULTILINGUAL	7.983	61.59/102.07	61.92/102.07	348.10/565.18	38.4%	156,303 seq.; 200× random $B = 16$
Chinese	2.946	47.07/56.77	59.91/56.77	636.00/816.82	22.1%*	89,909 seq.; random $B = 32$

Table A6: Forward FLOPs audit. C/T denote concept/token. For the final-checkpoint columns, $\bar{C}$ is the deterministic concept count averaged over every packed dev sequence, and $T/\bar{C}$ uses $T = 256$; this ratio is not a training-history mean or the mean of per-segment lengths. Ideal GF evaluates the analytic formula at $\bar{C}$, whereas padded GF evaluates it at within-batch maximum concept counts. The Chinese full-dev values are $\bar{C} = 86.883$, $T/\bar{C} = 2.946$, and 47.07 ideal GF. History values separately integrate every logged training interval using its sequence length, effective batch size, and logged mean segment length. A rough forward-plus-backward estimate is three times the history column. The asterisk marks the Chinese history estimate as optimistic: current tensors pad concept sequences to the within-batch maximum, and historical logs do not retain that distribution. Padded final values use the evidence in the last column.

Frozen pair	Dev seq.	Concepts/seq.	Batch max	P5/P50/P95	Saving / repeat 95% interval
English matched	2,687	56.39 ± 4.44	64.32 ± 2.53	61/64/69	25.94% / 25.88–26.02%
BabyConceptLM–TA1					
Multilingual matched	156,303	32.07 ± 0.35	32.84 ± 0.95	32/32/34	39.33% / 39.329–39.336%
BabyConceptLM–TA1					
Chinese BabyConceptLM–TA1	89,909	86.88 ± 33.92	143.64 ± 6.17	134/143/154	−5.52% / [−5.554, −5.469]%

Table A7: Full-dev padding distributions for all matched pairs. Each of 200 repetitions randomly partitions every dev sequence without replacement into groups of $B = 16$ for English/MULTILINGUAL and $B = 32$ for Chinese; incomplete final groups are retained. Batch-max means, SDs, and P5/P50/P95 pool allocated batches. Saving is signed relative to TA1, so the negative Chinese value denotes additional concept-model FLOPs. “Repeat 95% interval” is the central interval of the 200 repetition-level mean savings, not a population confidence interval.

kenizes the target word; our post-hoc audit found a continuation mismatch on 1,726/1,726 items and a different piece count on 462. Fast evaluations are applied to the required exposure checkpoints. AoA fits acquisition points from the full checkpoint sequence, computes Pearson r with CDI-derived child AoA, and retains r only when $p \leq .1$.

MULTILINGUAL evaluation covers language-specific zero-shot suites, the challenge-defined GlobalPIQA variant, MECO reading-time prediction (Siegelman et al., 2022), finetuned classification, and cross-lingual part-of-speech tagging; Chinese additionally includes Hanzi structure and pinyin. MultiBLiMP and XCOMPS follow their released benchmarks (Jumelet et al., 2025; He et al., 2025); HellaSwag, Winogrande, and XStoryCloze use the MuBench versions (Han et al., 2025). MECO and hidden Hanzi predictions use server-side scoring. Task definitions, language subsets, preprocessing, and aggregation are fixed by baby1m-eval@6f825c2; this pinned inventory is the source for challenge-specific variants without a separate archival citation.

ChineseBabyLM contains 14 tasks in three tracks. NLU comprises ZhoBLiMP, XCOMPS-ZH, AFQMC, OCNLI, TNEWS,

CLUEWSC2020, C3, and diagnostic NLI. The Hanzi track contains open and hidden structure and pinyin tests. The cognitive track uses responses from 11 participants reading 672 isolated words and 12 participants listening to 60 stories comprising 52,269 words. The official pipeline fits encoding models from frozen representations to fMRI responses and reports correlation-based scores.

D.1 Participant-level matched contrasts

Pinned encoding implementation. At f4437e1, discourse features average the selected hidden state’s subtokens within each word, apply an HRF convolution, and downsample to a 0.71-second TR with a 19-TR offset. Word-fMRI instead mean-pools non-padding positions for each isolated word. The default layer_index=−1 selects the final exported state (last_hidden_state for Word-fMRI). Discourse ridge regression scores each voxel by the temporal mean of the product of sample-standardized predicted and observed responses. The report takes the NaN-aware voxel mean within each participant/ROI; the nested-CV branch first averages voxel scores over five outer folds. Word-fMRI uses Pearson correlation across valid voxels for each trial and averages the highest

Pair	Matched concept repository@commit (SHA-256)	DWA	TA1 repository@commit (SHA-256)
English	babyconceptLM-English-matched@e82add172267 SHA 55ba4c4cb71d	T/C/R; 8/8/8	babyconceptLM-TA1-token-only-en@49c3d3ac3f1d SHA 4fdac3f1f3e4
MULTI	babyconceptLM-Multilingual-matched@ea60a3df6a34; SHA 82181ab8371a	T/C/R; 6/4/4	babyconceptLM-TA1-token-only-multi@e0296602741 SHA 49da0bb34d85
Chinese	babyconceptLM-baike102M_bpe40@05fda2ae8658; SHA 6345b814d5d2	none	babyconceptLM-TA1-token-only-zh@8bbf037e56a7 SHA 512c6bbef0dc

Table A8: Frozen matched-control identities. Repository names are relative to the zondaxyz/ Hugging Face namespace; commits are immutable revisions and values marked SHA are weight-hash prefixes. T/C/R denotes token/concept/readout paths; numbers are DWA current-state initialization logits.

Recipe field	Matched concept reference	Public-final concept model
English: A-stable reference versus STRICT final (both 5/8/3)		
Corpus/tokenizer	Archived STRICT processing; tokenizer_100M.json	Reprocessed official STRICT; tokenizer_babylm2026_strict.json
Packed train sequences	1,251,114	1,300,682
Training steps	97,750	101,620
Length switch	60% of steps	80% of steps
Token/concept learning rate	$1.2/1.0 \times 10^{-4}$	$1.5/1.2 \times 10^{-4}$
DWA initialization (T/C/R)	8/8/8	4/4/4
Cooldown / LR floor ratio	0.03 / 0.03	0.016 / 0.10
MULTILINGUAL: archived reference versus public final (both 5/8/3)		
Corpus mixture	Approximately equal adjusted thirds	43.095M/40M/15M raw ENG/NLD/ZHO words
Tokenizer directory	tokenizers-multi/	tokenizers-multi-babylm2026/
Packed train sequences	1,119,914	2,019,494
Training steps	87,500	157,780

Table A9: Archived matched versus public-final training recipes. TA1 uses the matched recipe in the middle column. MULTILINGUAL retains DWA initialization 6/4/4, token/concept learning rates $1.0/0.8 \times 10^{-4}$, a 60% length switch, and cooldown/floor ratios 0.03/0.03 in both recipes. Table A8 identifies the matched checkpoints.

10% of trial correlations, averaging these scores across outer folds when using nested CV. The two tasks therefore use different correlation and aggregation procedures. The scorer supports explicit train/dev/test splits or nested CV; without a test directory, the split branch reports dev performance.

We pair identical participant IDs in the frozen BabyConceptLM-TA1 outputs. For each discourse domain, the analysis uses one reported domain score per participant and model; Word-fMRI uses one word-task score per participant and model. Let $s_{i,d}^C$ and $s_{i,d}^T$ denote the reported correlations, multiplied by 100, for participant i and domain d . The paired difference is $\Delta_{i,d} = s_{i,d}^C - s_{i,d}^T$. For the Cognition–Language and six-domain summary contrasts, we first average the relevant domain scores within participant, then compare models. Participants, not domains or voxels, are the resampling units.

The archived analysis uses 200,000 paired

participant-bootstrap samples for each unadjusted 95% interval, preserving each participant’s two model scores together. Exact two-sided sign-flip tests enumerate all 2^n assignments (4,096 for discourse; 2,048 for Word-fMRI). With $T_{\text{obs}} = |n^{-1} \sum_i \Delta_i|$, the exact p -value is the fraction of sign vectors $\epsilon \in \{-1, +1\}^n$ for which $|n^{-1} \sum_i \epsilon_i \Delta_i| \geq T_{\text{obs}}$. The paired standardized effect is

$$d_z = \frac{\bar{\Delta}}{s_{\Delta}}, \quad s_{\Delta} = \sqrt{\frac{\sum_i (\Delta_i - \bar{\Delta})^2}{n-1}}. \quad (16)$$

Holm adjustment is applied separately to the two focal Cognition/Language tests (Table 2) and to all six discourse-domain tests (Table A19). The within-participant averages and Word-fMRI row report unadjusted exact p -values. Cognition and Language are focal exploratory contrasts; this designation makes no claim of preregistration. All six domains and their six-test correction remain re-

ported.

These intervals characterize participant sampling uncertainty conditional on the two fixed seed-42 models and the retained evaluation outputs. They do not include training-seed variance, story sampling, or uncertainty across analysis choices. The records available for this report do not specify the original bootstrap RNG seed, interval construction, or matched-run layer/split settings. The six-participant descriptive sweeps below are a separate analysis.

The detailed inventory is organized as follows. Table A10 reports every shared task entering the matched-control aggregates. Tables A11–A12 decompose English STRICT results; Tables A13–A15 provide the complete MULTILINGUAL outputs; Figure A2 and Table A16 report the exposure trajectories and anchors; and Tables A17–A19 together with Figures A3–A5 document the complete ChineseBabyLM results and descriptive layer-wise profiles.

E Segmentation protocol

The audit covers the four English STRICT models. Its frozen artifact uses the story field of Hippocorpus V3 (Sap et al., 2020), a collection independent of the BabyLM training corpus containing recalled, imagined, and retold narratives. We decode HTML entities, apply NFKC normalization, remove control characters, collapse whitespace, and preserve case and punctuation. Requiring 100–800 whitespace-delimited words yields 6,854 stories, 1,799,978 words, and 2,135,141 BPE tokens under the STRICT submission tokenizer. Exact text round trips are verified for every document, and token-aligned whitespace-word and word-start sidecars are stored with the artifact.

For each frozen checkpoint, the analyzer selects up to 2,000 documents by reservoir sampling with seed 42 and extracts one 256-token window per document. It micro-averages tokenizer tokens and source words per concept. Learned boundary rate excludes window starts and cap-attributed cuts. The complete length distribution includes P50, P90, standard deviation, coefficient of variation, Shannon entropy, and $P(L = k)$ for $k = 1, \dots, 8$; sampled-window edge segments are excluded. Word-start alignment includes boundary-probability AUC and the hard-boundary-rate difference between word starts and within-word positions. Bounded-gate statistics ap-

ply only to the 2/4/2 profile. Confidence intervals use 1,000 document-level bootstrap resamples. Matched noDWA/DWA comparisons use identical windows and resample aligned document pairs to form intervals over differences. All values are computed offline from frozen checkpoints rather than reconstructed from batch-level training summaries; each artifact retains document-level metrics plus model, configuration, and tokenizer fingerprints for audit.

Illustrative 2/4/2 cuts. For one shared Hippocorpus window, noDWA produces “the show, near the top | of the theater, and laid | down a blanket. The | opener came on, and,” whereas DWA produces “the show, near the | top of the theater, and | laid down a blanket. | The opener came on.” These examples illustrate realized length variability; corpus-level interpretation relies on the aggregate audit above.

Evaluation	Shared family/task	Concept	TA1	Δ (C-T)
English	BLiMP	75.90	77.37	-1.47
English	BLiMP Supplement	60.28	65.27	-4.99
English	EWoK	55.62	54.89	+0.73
English	Entity Tracking	16.19	16.47	-0.28
English	COMPS	55.72	57.04	-1.32
MULTI-EN	BLiMP	69.06	70.05	-0.99
MULTI-EN	HellaSwag	26.53	26.51	+0.01
MULTI-EN	MultiBLiMP	84.16	85.71	-1.56
MULTI-EN	Winogrande	51.61	49.30	+2.31
MULTI-EN	XStoryCloze	51.78	51.70	+0.08
MULTI-EN	GlobalPIQA	38.14	41.09	-2.96
MULTI-NL	BLiMP-NL	67.26	70.63	-3.37
MULTI-NL	HellaSwag	26.04	26.13	-0.09
MULTI-NL	MultiBLiMP	76.15	71.86	+4.29
MULTI-NL	Winogrande	51.11	49.71	+1.40
MULTI-NL	XCOMPS	51.16	53.50	-2.35
MULTI-NL	XStoryCloze	47.60	46.98	+0.62
MULTI-NL	GlobalPIQA	37.12	37.14	-0.01
MULTI-ZH	HellaSwag	27.17	26.85	+0.32
MULTI-ZH	Winogrande	48.39	49.79	-1.40
MULTI-ZH	XCOMPS	51.71	51.70	+0.01
MULTI-ZH	XStoryCloze	49.38	48.61	+0.77
MULTI-ZH	ZhoBLiMP	66.60	71.93	-5.33
MULTI-ZH	GlobalPIQA	37.65	40.06	-2.41

Table A10: Every task entering the English and multilingual matched-control macros. English compares English matched BabyConceptLM with English TA1; GlobalPIQA is excluded because it was not shared by both frozen outputs. MULTILINGUAL compares Multilingual matched BabyConceptLM with multilingual TA1 and follows the official fixed-denominator collator, including equal weighting of GlobalPIQA parallel and nonparallel subsets. Scores are percentages; minor delta discrepancies reflect calculation before display rounding. Chinese task-level values appear in Table A18.

Model	BLiMP	Supp.	EWoK	Entity	COMPS	GPIQA	SGLUE	Reading	AoA
GPT-2 Strict	74.73	65.00	54.37	16.91	55.85	36.62	67.75	6.93	-11.58
Strict-Interaction	72.89	65.09	54.23	15.89	55.37	36.18	67.28	3.43	-11.36
babyconceptLM-en-noDWA (2/4/2)	62.35	56.63	49.51	17.67	51.18	35.67	64.45	0.18	10.36
babyconceptLM-en-DWA (2/4/2)	62.44	56.41	49.83	17.54	51.10	34.17	63.72	0.18	0.00
babyconceptLM-en-noDWA (5/8/3)	75.82	63.22	53.96	15.07	55.61	39.17	67.45	0.22	-13.79
babyconceptLM-en-DWA (5/8/3)	76.58	63.52	53.30	14.79	55.88	37.59	67.10	0.13	-10.71

Table A11: English STRICT benchmark-family scores. GPIQA and SGLUE abbreviate GlobalPIQA and (Super)GLUE. Reading and AoA are official pipeline outputs retained with the tokenizer and threshold caveats in the text.

Model	BoolQ	MNLI	MRPC	MultiRC	QQP	RTE	WSC	Avg.
GPT-2 Strict	69.66	60.76	85.34	65.92	71.56	57.55	63.46	67.75
Strict-Interaction	67.46	60.68	86.27	66.21	70.51	58.27	61.54	67.28
babyconceptLM-en-noDWA (2/4/2)	68.07	48.72	82.99	61.34	63.02	59.71	67.31	64.45
babyconceptLM-en-DWA (2/4/2)	68.38	48.35	82.49	61.01	61.67	56.83	67.31	63.72
babyconceptLM-en-noDWA (5/8/3)	70.76	57.99	85.90	64.81	70.95	58.27	63.46	67.45
babyconceptLM-en-DWA (5/8/3)	70.28	57.86	83.87	65.80	71.56	56.83	63.46	67.10

Table A12: English (Super)GLUE results. Accuracy is reported except MRPC and QQP, which use F1. All BabyConceptLM entries contain the complete seven-task suite.

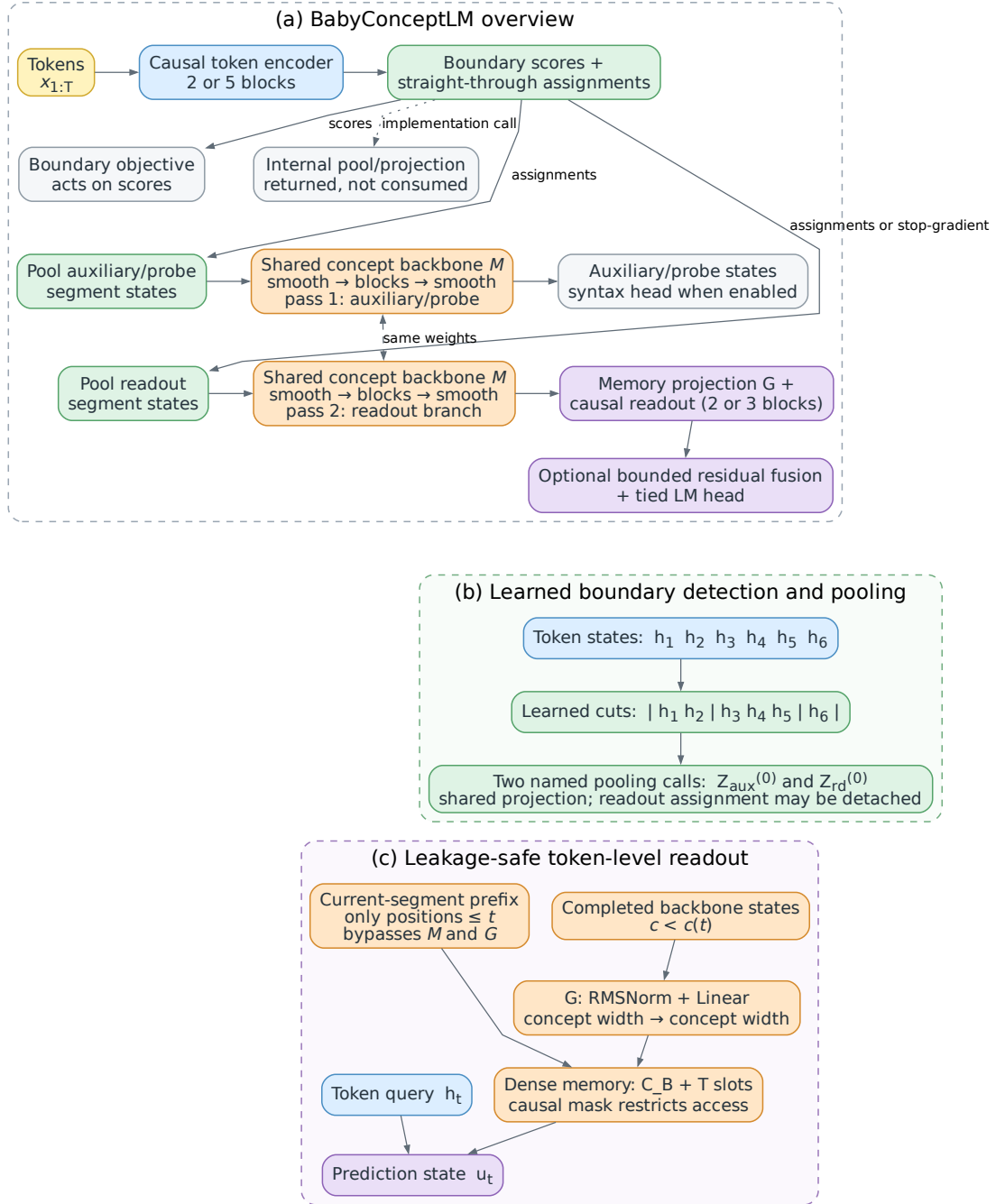


Figure A1: Detailed BabyConceptLM computation graph. Panel (a) separates boundary supervision from the two shared-backbone executions, each with pre- and post-stack causal smoothing, and shows the unused internal pooled tensors. Panels (b) and (c) expand pooling and readout, including completed-memory projection G and the current-prefix bypass.

Language	Zero-shot task	GPT-2 tri.	babyconceptLM-multi
English	BLiMP	70.49	71.24
English	HellaSwag	26.51	26.35
English	MultiBLiMP	85.97	86.88
English	Winogrande	50.12	49.30
English	XStoryCloze	49.54	50.08
English	GlobalPIQA	39.12	38.64
Dutch	BLiMP-NL	77.04	66.34
Dutch	HellaSwag	26.65	26.17
Dutch	MultiBLiMP	92.11	75.42
Dutch	Winogrande	49.38	51.61
Dutch	XCOMPS	53.04	50.93
Dutch	XStoryCloze	48.99	48.07
Dutch	GlobalPIQA	38.11	36.61
Chinese	HellaSwag	26.74	26.96
Chinese	Winogrande	49.63	51.03
Chinese	XCOMPS	53.47	53.46
Chinese	XStoryCloze	47.91	51.63
Chinese	ZhoBLiMP	74.03	75.90
Chinese	GlobalPIQA	31.74	38.15

Table A13: Complete MULTILINGUAL zero-shot tasks used in the aggregate. GPT-2 values reproduce the rounded official README entries.

Language	Finetuning task	GPT-2 tri.	babyconceptLM-multi
English	ARC / Belebele / BMLAMA	24.79 / 26.14 / 12.00	21.88 / 17.61 / 11.67
English	MNLI / part-of-speech / SIB200	50.00 / 93.40 / 69.50	55.35 / 93.97 / 66.50
English	TruthfulQA / XNLI	23.21 / 45.65	22.32 / 52.85
Dutch	ARC / Belebele / BMLAMA	24.38 / 26.14 / 9.93	25.62 / 19.89 / 12.09
Dutch	INCLUDE / MNLI / part-of-speech	19.64 / 49.27 / 94.97	33.04 / 55.35 / 95.25
Dutch	SIB200 / TruthfulQA	71.00 / 23.21	72.00 / 23.21
Chinese	ARC / Belebele / BMLAMA	25.21 / 22.73 / 10.93	25.21 / 19.32 / 10.84
Chinese	INCLUDE / MNLI / part-of-speech	28.57 / 47.58 / 90.40	25.89 / 56.87 / 91.97
Chinese	SIB200 / TruthfulQA / XNLI	72.50 / 23.21 / 45.05	84.00 / 25.89 / 50.75

Table A14: Complete MULTILINGUAL finetuning tasks used in the aggregate. Slash-separated values follow the task order shown in each cell.

Non-aggregate output	GPT-2 tri.	babyconceptLM-multi
Hanzi structure	55.05	53.40
Hanzi pinyin	50.05	52.05
MECO English L1	5157.10	5206.26
MECO Dutch L1	388.27	173.85
MECO Chinese L1	632.02	627.38
MECO Dutch L2	55.92	71.43
MECO Chinese L2	75.47	166.91

Table A15: MULTILINGUAL Hanzi and MECO outputs verified against the public leaderboard at manuscript freeze. MECO delta-log-likelihood values are not normalized across reader groups; both output types are excluded from the aggregate.

Curve	1M	10M	100M	200M	500M	1B	Log-AUC
STRICT noDWA (2/4/2)	32.64	33.93	35.15	36.65	37.02	37.41	35.06
STRICT DWA (2/4/2)	32.63	33.91	35.64	36.59	37.45	37.35	35.04
STRICT noDWA (5/8/3)	34.73	34.80	39.71	40.17	41.51	41.18	37.72
STRICT DWA (5/8/3)	34.99	35.16	39.13	40.16	40.84	41.14	37.46
MULTILINGUAL mean	44.67	45.04	48.38	49.64	50.68	51.33	47.35

Table A16: Learning-curve anchors. Log-AUC is the trapezoidal mean over $\log_{10}$ exposure from 1M to 1B. The corresponding public Hugging Face repositories expose all 28 evaluated checkpoint revisions.

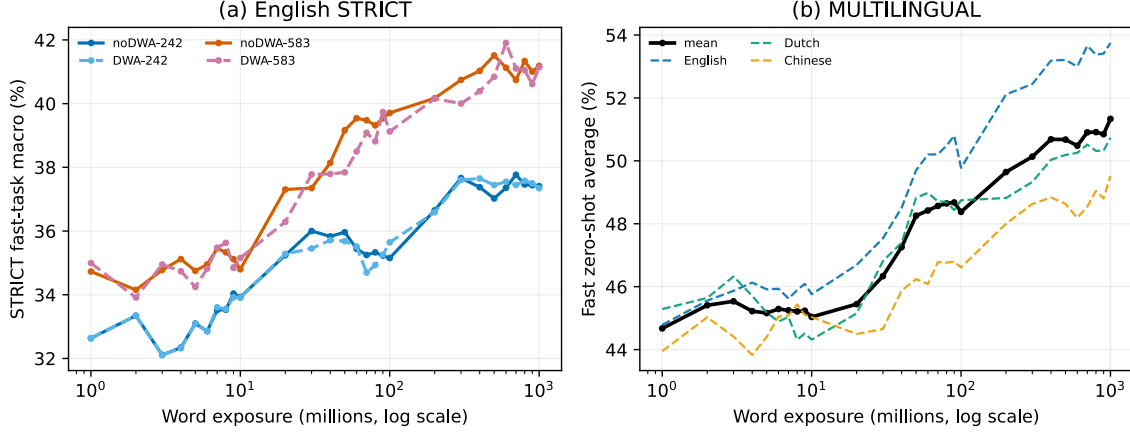


Figure A2: Fast evaluation at all 28 official exposure checkpoints. Left: within-profile English DWA/noDWA pairs. Right: the multilingual mean and language decomposition. Lines connect observed checkpoints and do not imply monotonic learning. Anchor values and log-exposure AUCs appear in Table A16.

Model	ZhoBLiMP	XCOMPS-ZH	Hanzi struct.	Hanzi pinyin	Word fMRI	Discourse fMRI
T5 (14M)	75.82	56.56	51.10	51.40	55.01	8.24
BERT (14M)	75.82	62.66	55.80	45.85	56.02	10.74
Mamba (14M)	74.23	58.88	52.05	52.90	47.18	9.37
Pythia (14M)	73.95	57.02	53.70	50.10	49.89	8.67
Token-level GPT-BERT	76.88	58.98	54.60	50.40	56.45	9.92
Matched TA1	72.25	63.29	56.70	52.85	56.77	10.74
BabyConceptLM	70.08	61.82	56.45	52.15	56.61	10.92

Table A17: Selected ChineseBabyLM results. Official-baseline values reproduce the aligned columns of Table 4 in Song et al. (2026); token-level GPT-BERT is an unmatched recipe. The displayed BabyConceptLM values match the accessible ChineseBabyLM 2026 final leaderboard. Matched TA1 shares the BabyConceptLM data, tokenizer, exposure, optimizer/schedule, batches, seed, and absence of DWA, with a 1.7% parameter difference. The contrast evaluates the complete concept pathway, jointly varying segmentation, concept processing, readout, and layer allocation. Bold marks the largest displayed value.

Model	ZhoB	XComp	H-Str.	H-Pin.	H-Str.-H	H-Pin.-H	Word	Disc.
Matched TA1	72.25	63.29	56.70	52.85	55.95	50.25	56.77	10.74
babyconceptLM-baike102M_bpe40	70.08	61.82	56.45	52.15	55.55	50.30	56.61	10.92

Model	AFQMC	OCNLI	TNEWS	WSC	C3	Diag.	NLI
babyconceptLM-baike102M_bpe40	69.14	60.92	53.56	63.49	30.19		45.10

Table A18: Completed non-finetuning ChineseBabyLM inventory for the matched pair, followed by BabyConceptLM finetuning tasks. H and Disc. denote hidden and discourse; correlations are multiplied by 100. TA1 finetuning was not run.

Discourse domain	<i>n</i>	Δ (C-T)	95% CI	d_z	Exact <i>p</i>	Holm-6 <i>p</i>
Cognition	12	+0.288	[-0.274, 0.797]	0.29	.324	.871
Language	12	+0.281	[-0.118, 0.687]	0.38	.218	.871
Manipulation	12	+0.406	[-0.024, 0.789]	0.54	.088	.442
Memory	12	-0.073	[-0.545, 0.311]	-0.09	.802	.871
Reward	12	-0.153	[-0.257, -0.029]	-0.72	.032	.193
Vision	12	+0.319	[-0.136, 0.850]	0.35	.270	.871

Table A19: Participant-level matched TA1 contrasts for every CogBench discourse domain. Correlations are multiplied by 100 and positive differences favor BabyConceptLM. Intervals are unadjusted paired bootstrap intervals; exact two-sided sign-flip tests use all 2¹² assignments, and Holm-6 corrects the six exploratory domain tests. The focal exploratory two-domain correction for Cognition and Language is reported in Table 2.

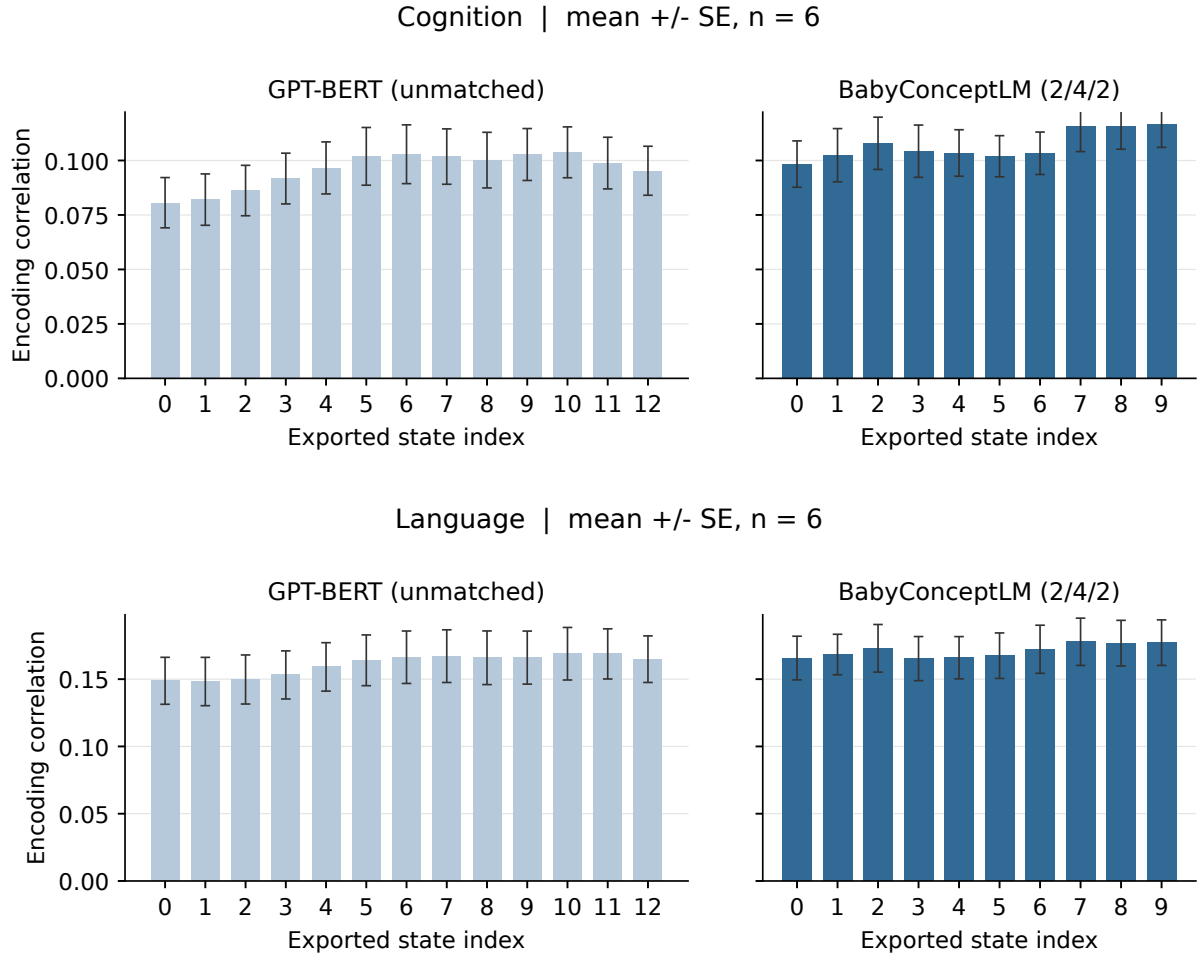


Figure A3: Descriptive discourse-fMRI layer sweeps in Cognition (top) and Language (bottom). Each bar averages the six participant-domain correlations in the retained layer-sweep export; error bars are ± 1 SE (sample $SD/\sqrt{6}$), not confidence intervals. Scores are averaged across participants within each model, state, and domain, without pooling other domains or layers. These unmatched GPT-BERT–BabyConceptLM sweeps are separate from the matched TA1 $n = 12$ tests in Table 2. Index 0 is the embedding output; subsequent indices enumerate within-model exported states, not homologous depths. GPT-BERT has token-level Transformer blocks, not concept/readout stages.

English model	Tok./con.	Word/con.	Learned bnd.	P50/P90	$P(L = 6)$	$P(L = 8)$	Word AUC	Gate
babyconceptLM-en-noDWA (2/4/2)	5.807	4.899	16.74	6/6	70.16	0.57	0.543	0.0003
babyconceptLM-en-DWA (2/4/2)	5.706	4.813	17.09	6/6	63.32	0.43	0.546	0.0003
babyconceptLM-en-noDWA (5/8/3)	5.911	4.987	6.93	8/8	3.61	57.02	0.448	n/a
babyconceptLM-en-DWA (5/8/3)	6.405	5.403	4.64	8/8	2.48	68.44	0.448	n/a

Table A20: Hippocampus segmentation statistics under each model’s native decoder. Boundary and length probabilities are percentages; P50/P90 exclude window-edge-censored segments. Complete histograms and dispersion statistics appear in Table A21 and Figure A6.

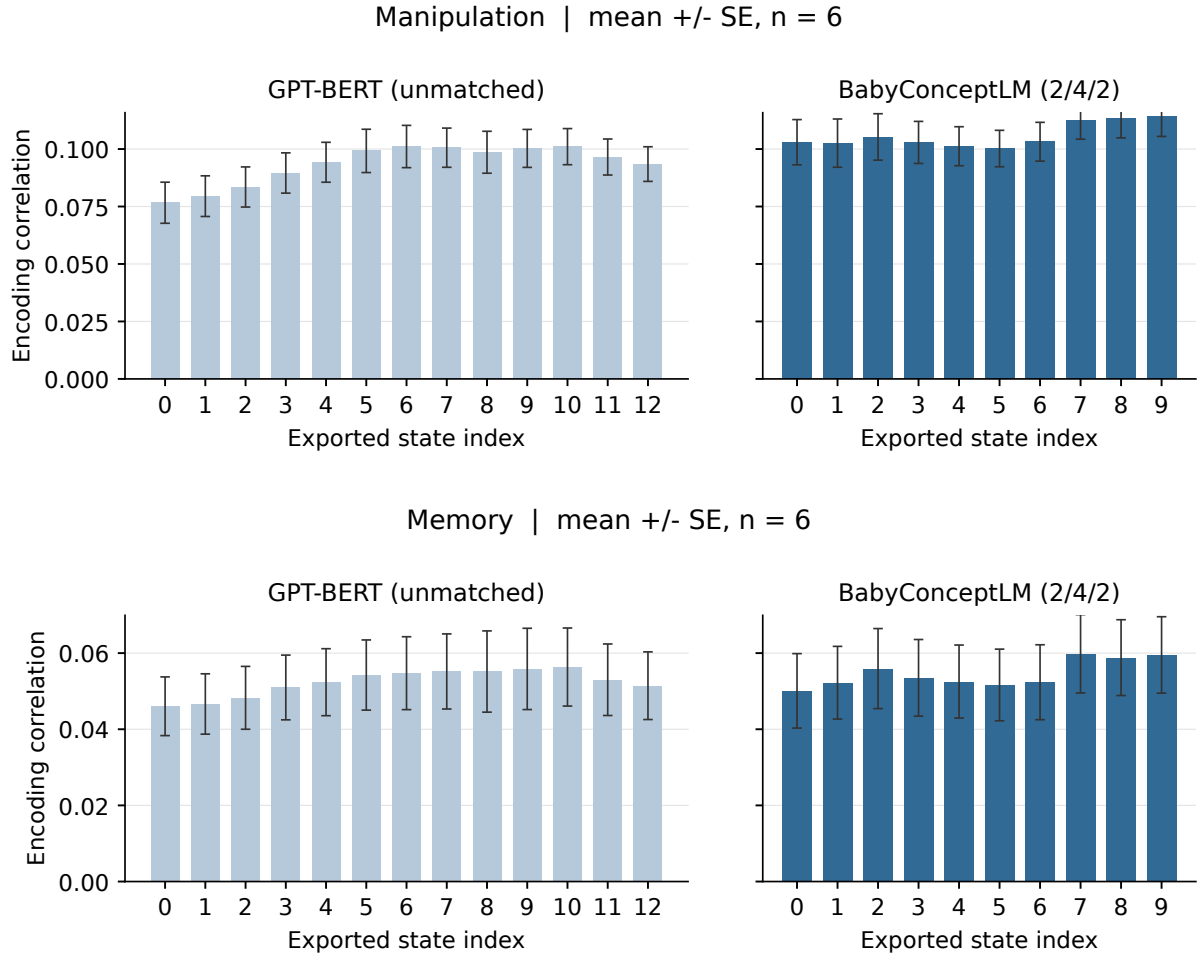


Figure A4: Descriptive discourse-fMRI layer sweeps for unmatched GPT-BERT and BabyConceptLM in Manipulation (top) and Memory (bottom). Bars average six participant-domain correlations separately for each model/state/domain; whiskers are ± 1 SE (sample $SD/\sqrt{6}$), with no pooling across domains or layers. Index 0 denotes embeddings; subsequent indices are within-model exported states and do not imply matched depth or shared concept/readout stages. These six-participant exports are distinct from the matched TA1 $n = 12$ analysis in Table 2.

Model	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	SD	CV	H
noDWA-242	0.17	0.16	0.15	0.12	22.54	70.16	6.13	0.57	0.604	0.104	1.189
DWA-242	0.15	0.15	0.16	0.15	30.95	63.32	4.68	0.43	0.617	0.108	1.239
noDWA-583	9.94	8.92	7.10	5.66	4.48	3.61	3.26	57.02	2.671	0.449	2.145
DWA-583	7.95	6.39	5.42	4.03	3.12	2.48	2.18	68.44	2.497	0.386	1.741

Table A21: Complete-segment length distributions. P_k is the percentage of uncensored segments with k tokenizer tokens; SD is in tokens, CV is SD divided by mean length, and entropy H is in bits. The 2/4/2 rows exclude a literal fixed-six decoder but show a narrow near-fixed-six regime.

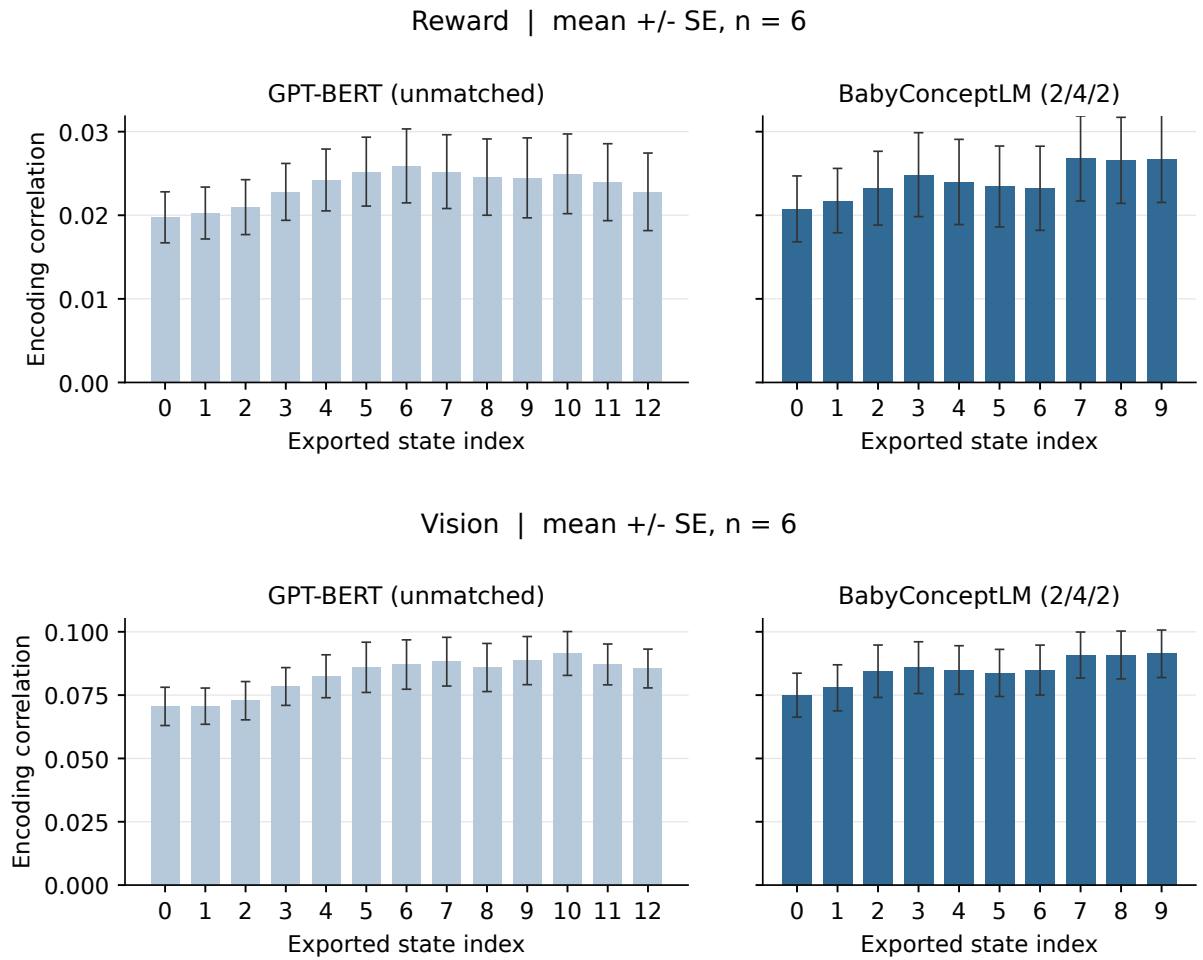


Figure A5: Descriptive discourse-fMRI layer sweeps for unmatched GPT-BERT and BabyConceptLM in Reward (top) and Vision (bottom). Bars average six participant-domain correlations separately for each model/state/domain; whiskers are ± 1 SE (sample $SD/\sqrt{6}$), with no pooling across domains or layers. Index 0 denotes embeddings; subsequent indices are within-model exported states and do not imply matched depth or shared concept/readout stages. These six-participant exports are distinct from the matched TA1 $n = 12$ analysis in Table 2.

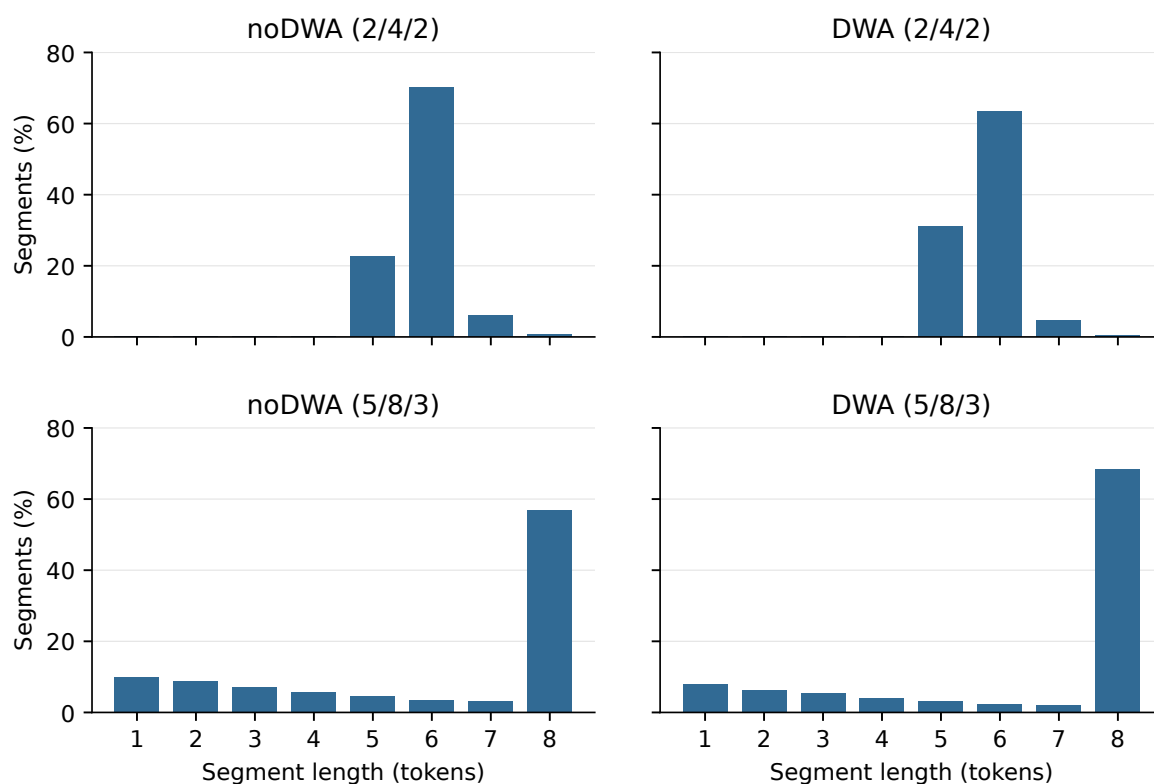


Figure A6: Uncensored Hippocampus segment-length distributions for the four English models, with shared bins and percentage axes. Each panel divides the retained segment counts by its own total: 80,328 (noDWA-242), 81,801 (DWA-242), 78,870 (noDWA-583), and 72,604 (DWA-583). The shallow profile concentrates at lengths five and six; the deep profile concentrates at the eight-token cap while retaining a lower-length tail.

Model / contrast	Quantity	Estimate	95% CI	Resampling unit
noDWA-242	Tokens/concept	5.807	5.801–5.813	document
noDWA-242	Words/concept	4.899	4.887–4.911	document
noDWA-242	Learned boundary (%)	16.74	16.70–16.78	document
noDWA-242	$P(L = 8)$ (%)	0.57	0.43–0.73	document
noDWA-242	Gate mean	0.000295	0.000285–0.000307	document
noDWA-242	Word-start rate difference (pp)	0.26	0.00–0.53	document
DWA-242	Tokens/concept	5.706	5.700–5.712	document
DWA-242	Words/concept	4.813	4.802–4.825	document
DWA-242	Learned boundary (%)	17.09	17.04–17.12	document
DWA-242	$P(L = 8)$ (%)	0.43	0.30–0.59	document
DWA–noDWA (242)	Tokens/concept	–0.101	[–0.105, –0.098]	paired documents
DWA–noDWA (242)	Words/concept	–0.086	[–0.088, –0.083]	paired documents
DWA–noDWA (242)	Learned boundary (pp)	+0.34	0.33–0.35	paired documents
DWA–noDWA (242)	$P(L = 8)$ (pp)	–0.14	[–0.17, –0.11]	paired documents
noDWA-583	Tokens/concept	5.911	5.881–5.942	document
noDWA-583	Words/concept	4.987	4.956–5.017	document
noDWA-583	Learned boundary (%)	6.93	6.80–7.06	document
noDWA-583	$P(L = 8)$ (%)	57.02	56.46–57.59	document
noDWA-583	$P(L = 1)$ (%)	9.94	9.67–10.21	document
noDWA-583	Word-start rate difference (pp)	–1.23	[–1.50, –0.95]	document
DWA-583	Tokens/concept	6.405	6.376–6.432	document
DWA-583	Words/concept	5.403	5.376–5.430	document
DWA-583	Learned boundary (%)	4.64	4.54–4.74	document
DWA-583	$P(L = 8)$ (%)	68.44	67.92–68.91	document
DWA-583	Word-start rate difference (pp)	–0.82	[–1.09, –0.55]	document
DWA–noDWA (583)	Tokens/concept	+0.493	0.483–0.505	paired documents
DWA–noDWA (583)	Words/concept	+0.416	0.407–0.425	paired documents
DWA–noDWA (583)	Learned boundary (pp)	–2.29	[–2.35, –2.24]	paired documents
DWA–noDWA (583)	$P(L = 8)$ (pp)	+11.42	11.19–11.64	paired documents
DWA–noDWA (583)	Word-start rate difference (pp)	+0.41	0.19–0.62	paired documents

Table A22: Document-bootstrap intervals from the English segmentation artifacts. Paired rows use identical Hippocorpus windows for DWA and noDWA; pp denotes percentage points. Exported P50/P90 intervals collapse to their point values, while word-start AUC has no exported interval; these point estimates appear in Table A20.