Beetle: A Bilingual Model Suite for Modelling Second-Language Processing

Suchir Salhan, Catherine Arnett, James A. Michaelov & Paula Buttery

University of Cambridge EleutherAI MIT

Correspondence: sas245@cam.ac.uk

Abstract

Bilingual language models (LMs) offer a controlled setting for studying how training conditions shape second-language (L2) behaviour, but prior work typically varies exposure structure, scale, and architecture at once, making it difficult to attribute effects to any single factor. We introduce Beetle, a controlled language model pretraining framework in which tokeniser, target language, training budget, and exposure structure are each independently manipulable, enabling systematic and comparable experimentation of training conditions. Using Beetle, we train and release 285 bilingual and 45 monolingual open-source LMs with rich checkpoints across a range of exposure schedules, data scales and first languages (L1s) to study multilingual pretraining and computational modelling of bilingualism and second language learning. Evaluating models on human bilingual and second language reading-time prediction and grammaticality judgement tasks, we find that staged and temporally structured curricula consistently improve alignment with language learner reading time compared to balanced bilingual training, with the largest gains at smaller data scales and for typologically closer language pairs. The Beetle models are well suited tools to help move computational psycholinguistics beyond its prevailing monolingual, English-centric focus toward models of human bilingual processing, to study cross-lingual learning dynamics, while supporting community-based development of controlled model families.

beetlelm.github.io

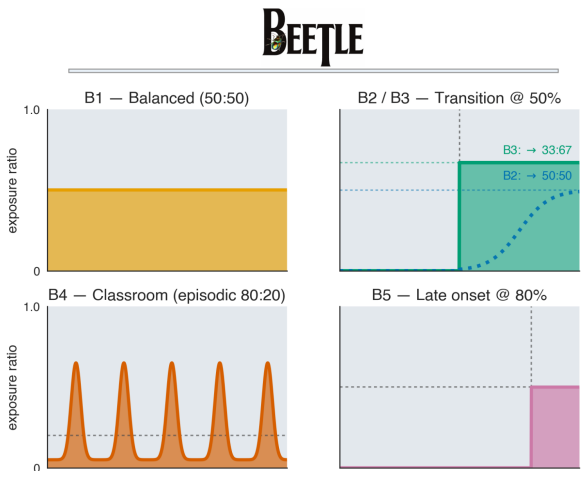

Figure 1: Beetle is an open-source language model pretraining framework. We use it to study bilingual and multilingual language acquisition in language models under tightly controlled experimental conditions. Each facet shows different structured exposure curriculum (B1–B5) used to pretrain bilingual models that vary in the relative per-language data distribution across training steps.

1 Introduction

In contrast to multilingual models, bilingual and second-language (L2) language models (LMs) are models trained on exactly two languages, either jointly or in sequence. These models are a useful middle ground between monolingual and massively multilingual systems for studying multilingual language model pretraining. They also allow us to study bi- and multilingual language acquisition, representations, interpretability, and learner simulation (Frank, 2021; Yadavalli et al., 2023; Oba et al., 2023; Aoyama and Schneider, 2024; Constantinescu et al., 2025; Arnett et al., 2025). However, important questions about the dynamics of bilingual training remain open. When does cross-linguistic transfer emerge during pretraining, how does transfer depend on typological similarity, relative timing and distribution of exposure to each language, and how does joint training compare with learning each language independently (Chang et al., 2022; Blevins et al., 2022; Wang et al., 2024; Rajaee and Monz, 2024; Körner et al., 2026; Zhou et al., 2026)? Existing bilingual models provide useful evidence for transfer, but are often constrained by the available training se-

tups. The only open bilingual family with checkpoints, B-GPT (Arnett et al., 2025), was developed for only a limited set of language pairs and does not provide fine-grained checkpoints over the early stages of training, making it difficult to distinguish effects that emerge during initial acquisition from those that arise later in training. Moreover, the absence of matched monolingual baselines makes it difficult to determine whether apparent bilingual benefits reflect genuine transfer between languages or simply differences in the amount and composition of training data. These limitations motivate controlled bilingual pretraining experiments that vary exposure structure while tracking learning from the earliest stages and comparing bilingual models against matched monolingual controls.

Despite the fact that most people speak at least two languages (Grosjean, 2021), most cognitive science research focuses on monolingual language processing. In computational cognitive science, one limitation has been the lack of available bilingual language models. And the studies that have been done they differ significantly in training scale, exposure schedule, language pair, continual-learning mechanism, and evaluation protocol. As a result, it is hard to attribute behaviour of a model to any single training decision.

To address this, we introduce Beetle, a modular LM pretraining and analysis framework that controls data exposure, alongside the usual choices of data quantity, tokeniser, and architecture (§3), to compare and investigate these confounding factors. We unify previously proposed methods in our training framework and use our models to investigate whether acquisition-inspired language exposure curricula during bilingual LM pretraining leads to better alignment with metrics of human language processing.

Using our training framework, we train a suite of **285 bilingual models** spanning 21 L1s, holding architecture, model size, tokenizer, and overall language mixture fixed. We vary two factors: the L1 (English is always the L2) and the order in which the two languages are presented during pretraining. We train bilingual models with 5 exposure conditions at three FineWeb dataset scales (100M, 2B and 24B tokens), plus a matched human-scale 100M condition. We additionally provide **45 monolingual baselines covering 20 languages**. Every Beetle model is released with its full checkpoint trajectory for studying cross-lingual learning dynamics.

We find that structured exposure improves alignment with human eye-tracking reading times on Multilingual Eye-movement Corpus (MECO) L2 data (Kuperman et al., 2023, 2025) and that ordered curricula tend to outperform fully interleaved training, suggesting that gradual language introduction improves alignment with psycholinguistic data. These effects are not explained by differences in overall proficiency or language balance, and remain consistent across controlled ablations. We additionally evaluate larger massively multilingual models, showing that carefully designed bilingual curricula can match or exceed L2 reading-time alignment under substantially smaller parameter budgets. These results demonstrate that the point during training when languages are introduced is a key determinant of how closely language models reproduce human second-language reading behaviour.

Beyond reading time, we further probe bilingual processing through cross-lingual structural priming (§6.1) and CEFR-stratified L2 learner-error discrimination on BLiSS (§6.2), and find that curriculum effects do not transfer uniformly across these tasks: balanced exposure is often competitive or better on grammaticality and error discrimination even where staged curricula best predict reading times.

The primary contributions of this paper are 1) a flexible training framework to enable controlled multilingual pretraining with learning dynamics 2) a suite of bilingual LMs 3) and modelling efforts for bilingual data on two evaluation domains, which can inform future bilingual cognitive modelling work.

2 Background

Bilingual models. Bilingual and second-language language models (L2LMs) offer a controlled middle ground between monolingual and massively multilingual models, and have been used to study second-language acquisition through measures such as reading times and structural priming (Matusevych et al., 2013; Yadavalli et al., 2023; Oba et al., 2023; Aoyama and Schneider, 2024; Arnett et al., 2025; Constantinescu et al., 2025; Feng et al., 2026). But there is no single recipe for creating an L2LM: prior work varies in what languages are seen, when

they are introduced, and how strongly earlier knowledge is preserved. Some approaches train both languages jointly, while others introduce the L2 after L1 pretraining, using continual-learning methods such as elastic weight consolidation (EWC) to limit forgetting (Constantinescu et al., 2025). Others manipulate the timing and distribution of L1 and L2 exposure directly, as in the structured exposure regimes of B-GPT (Arnett et al., 2025). This heterogeneity makes it difficult to disentangle effects of language pair, model scale, and exposure history. Our framework adopts this structured-exposure approach while systematically controlling these factors, enabling reproducible comparisons across bilingual training curricula.

Modelling human language processing. Within psycholinguistics, LM surprisal is often used to model processing difficulty. Surprisal is incorporated as a regression predictor alongside baseline lexical covariates, and evaluated via the improvement in log-likelihood over this baseline ($\Delta \log L$; Smith and Levy, 2013; Wilcox et al., 2020; Oh et al., 2022; de Varda and Marelli, 2023), making psychometric predictive power dependent on how well a model's probability distribution aligns with human reading behaviour rather than on raw model quality alone. However, while LM–human alignment has been studied extensively in L1 settings, bilingual and L2 reading-time modelling remains comparatively under-represented (Frank, 2014; de Varda and Marelli, 2022; Aoyama and Schneider, 2024), and this gap is especially pronounced in controlled large-scale sweeps of training conditions. In this paper, we hope to help address these limitations in the literature.

3 The Beetle Training Framework

Beetle is a language model pretraining framework that allows controlled manipulation of pretraining conditions. We apply Beetle for controlled pretraining of variables that prior L2LM work has manipulated separately: L2 onset timing, L2 exposure ratio, and continual-learning regularisation. Language exposure in bilingual pretraining is modelled as changing continuously over the course of training, making it possible to represent gradual or abrupt transitions between languages, periodic bursts of exposure, and clustered context-dependent input.

Architecture. The default backbone, **PicoDecoder** (Diehl Martinez et al., 2025), is a 125M-parameter LLaMA-style causal decoder (Touvron et al., 2023): 14 layers, $d_{\text{model}} = 768$, 12 attention heads with one KV head (grouped-query attention), Rotary Position Embeddings, SwiGLU activations, RMSNorm, sequence length 512. This places Beetle on a contemporary LLaMA-class architecture (RoPE, SwiGLU, RMSNorm, grouped-query attention) that mirrors the components used in current open-weight LLMs, rather than the GPT-2-style decoder used in B-GPT (Arnett et al., 2025) and the embedding-and-head adaptation of Aoyama and Schneider (2024). Full training hyperparameters are listed in Table 2.

Beetle also contains modular logic for tokenization, potentially allowing investigation of the benefits of vocabulary overlap, different tokenization schemes (e.g., SentencePiece, UnigramLM, SuperBPE, BPE) and equal compression (Kallini et al., 2025). Our models use a BPE tokenizer based on the `Huggingface` tokenizers library, with a 50K vocabulary. The Beetle models have tokenizers are trained with equal compression, so they have the same compression rate in both languages. This means each model see the amount of information for each language in each sequence.

Exposure curricula. We compare five exposure curricula that manipulate how L2 input is distributed over training. Four curricula differ primarily in when L2 is introduced and how the L1:L2 mixture subsequently changes; B4 instead keeps L2 available throughout training but varies when when the model sees L2 tokens, concentrating them into discrete episodes. The curricula therefore vary in L2 onset, final L1:L2 ratio, and temporal distribution, and are illustrated in Figure 1 (see Appendix C for details). We report the total proportion of training tokens in L2 for each curriculum, since curricula that introduce L2 at different points necessarily provide different amounts of L2 unless their later exposure compensates for the delayed onset.

- **B1 (Balanced):** L1 and L2 are presented in a constant 50:50 mixture throughout training. L2 is therefore available from the beginning and accounts for 50% of all training tokens.
- **B2 (Simultaneous):** Models are trained on L1 only, then halfway through training, L2 is gradually introduced using a sigmoid transi-

tion to a 50:50 mixture for the rest of training. L2 comprises 25% of all training tokens.

- **B3 (Sequential):** Training begins with L1 only. At halfway, L2 is gradually introduced using a sigmoid transition, eventually reaching an L2-dominant 33:67 L1:L2 mixture. L2 accounts for 33.5% of all training tokens.

- **B4 (Classroom / episodic):** Unlike B2, B3, and B5, L2 is available throughout training rather than being introduced at a specific onset. The overall L1:L2 ratio is 80:20, but L2 tokens are temporally clustered into intermittent, high-variance episodes rather than being evenly distributed. Thus, B4 tests the effect of *when L2 tokens occur within training* while keeping L2 available from the outset. L2 is 20% of training tokens.

- **B5 (Late):** Training begins with L1 only, with L2 introduced at 80% of training. A sigmoid transition then gradually increases L2 exposure to a 50:50 L1:L2 mixture for the remainder of training. L2 tokens are 10%.

4 Models & Training

We release **285 bilingual and 45 monolingual Beetle open-source language models** spanning **21 L1s**, with approximately 30 checkpoints per model.

Our experiments hold a fixed architecture, tokeniser, and data source design, as independent controls to measure the effect of L1, exposure curricula (`B1 - B5`) and data scale (100M, 2B, 24B tokens) on L2 performance. All models described in this paper have English as their L2, due to the availability of L2 reading time data (Section 5) and CEFR-stratified morphosyntactic and pedagogical tasks (Section 6.2).

Beetle-Data. The human-scale models are trained on 100M tokens from BabyBabelLM (Jumelet et al., 2025a); the web-scale models (at 100M, 2B, 24B tokens) use FineWeb-2 (Penedo et al., 2025) for L1 and FineWeb-Edu for English. We implement a data decontamination pipeline, removed via 13-gram overlap detection following Brown et al. (2020a), for our 19 evaluation datasets (full list in Appendix B). We release the full data tokenized in the order seen by the model for replicability and studying data memorisation.

L1s. Beetle models cover a typologically diverse set of L1s, with coverage continually expanding to support broader psycholinguistic and cross-lingual modelling. At 100M FineWeb, our experiments report results on Danish, German, Greek, Estonian, Basque, Filipino, Finnish, Hebrew, Hindi, Icelandic, Italian, Japanese, Korean, Dutch, Polish, Russian, Spanish, Turkish, and Chinese. At larger data scales, we focus on 9 L1s (German, Hindi, Italian, Dutch, Polish, Russian, Spanish, Turkish, Chinese). For BabyBabelLM, we report 100M-token experiments for German, Chinese, and Dutch.

Checkpointing Checkpoints are saved according to a $\log_2$ schedule following Pythia (Biderman et al., 2023), which re-starts at the beginning of each new exposure phase. This enables granular investigation of learning dynamic near phase boundaries, but with more sparse checkpointing during less critical periods of training. For each model, we release $\approx$ 30 checkpoints. Following Diehl Martinez et al. (2025), checkpoints save model gradients, activations and circuits, facilitating detailed interpretability analyses on circuit formation in bilingual LM pretraining (Inaba et al., 2025). To the best of our knowledge, our models will be the first with checkpoints available for several of these languages, enabling more linguistically diverse training dynamics research.

Baselines. We compare our models against two types of baselines. **Monolingual baselines** are matched-architecture Beetle models pretrained on a single language for German, Dutch and Chinese across all scales and Russian, Italian, Turkish and Basque at 100M/2B scale. **Massively Multilingual LLMs**: we also compare Beetle models to massively multilingual base models, specifically, Apertus 8B (Swiss AI Initiative, 2025), XGLM 4.5B (Lin et al., 2022), Llama 3.1 1B (Grattafiori et al., 2024), Gemma 3 270M (Team et al., 2025), and Qwen 3 0.6B (Yang et al., 2025). Full baseline model details are given in Table 9.

4.1 Analysing Curriculum-Induced Learning Dynamics

Before comparing our models with human behavioural data and metrics of bilingual language, we visualize the training dynamics (Figure 3) and test grammatical learning (Table 8) in order to validate that the exposure curricula demonstrate expected patterns. In order to avoid confounds

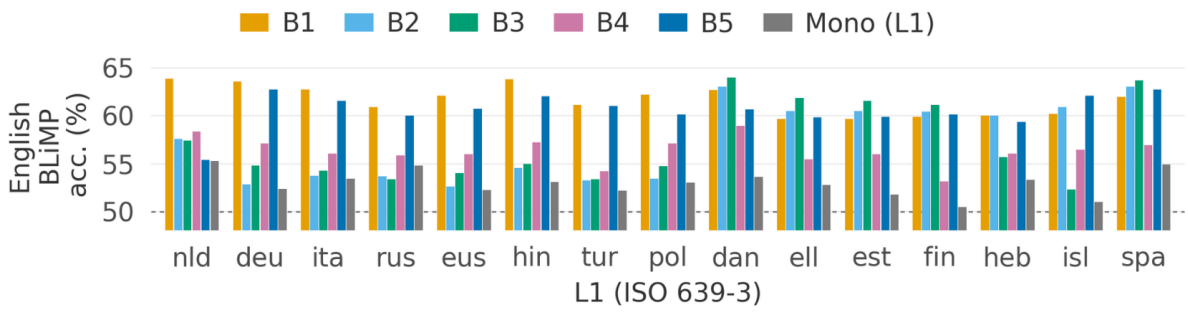

Figure 2: Beetle English BLiMP Accuracy for 100M FineWeb models across 15 L1s (Dutch, German, Italian, Russian, Basque, Hindi, Turkish, Polish, Danish, Greek, Estonian, Finnish, Icelandic and Spanish) and 5 exposure curricula. Full Monolingual and Multi-BLiMP accuracies for the Beetle models are in Appendix G.

from tokenizers or writing systems (Yang et al., 2026), we calculate the sentence-level negative log-likelihood (Sent-NLL) for each parallel item in FLORES-200 and calculate the mean. This provides a comparable metric of model performance across languages and across models.

Cross-Lingual Transfer and L1 Retention The FLORES-200 trajectories in Figure 3 reveal a strikingly asymmetric pattern of bilingual acquisition: introducing English causes an abrupt, curriculum-aligned increase in L1 sentence-level NLL, followed by partial recovery, while English NLL consistently declines after its introduction. Observed alignment of the shock is aligned with the L2 onset and subsequently recovers: L1 NLL increases under B2/B3, compared with the later-onset B5. NLL changes less under B4, where English ultimately accounts for only 20% of the mixture. Chinese and Hindi exhibit large disruptions than other languages, such as Dutch and German. This divergence is consistent with a competition-for-capacity account of cross-lingual interference: acquiring a newly introduced language produces an immediate cost to the language already represented by the model, with the size of the cost depending strongly on the exposure schedule (Whitford and Titone, 2015).

Mono- and Multi-BLiMP. We evaluate the Beetle models on monolingual BLiMPs (BLiMP (Warstadt et al., 2020) for English; BLiMP-NL (Suijkerbuijk et al., 2025b) for Dutch; ZhoBLiMP (Liu et al., 2025)) and MultiBLiMP (Jumelet et al., 2025b). At the human-scale budget of 100M tokens, B1 produces the strongest English grammar: a mean BLiMP score of 61.6 across 15 L1s, versus 56.3 for the strongest staged curriculum, B4. B1 is also the top bilingual curriculum for 9 of 15 L1s (versus 5 for B3 and 1 for B5; Figure 2). Strikingly, every bilingual curriculum beats its L1-monolingual control, whose mean score is only 52.9 and remains close to chance. The advantage is even larger on MultiBLiMP (Eng): B1 reaches 85.2, compared with 67.4 for the monolingual controls. The benefit is language-specific, however: on MultiBLiMP (L1), monolingual models lead with 86.0, narrowly ahead of the best bilingual curriculum, B2 (85.1). Thus, at 100M tokens, balanced bilingual exposure is particularly effective for acquiring English grammar, without requiring a staged curriculum. This result is not driven by the expanded evaluation: within the seven-L1 subset reported in the paper, B1 is again the strongest curriculum for all seven languages (62.6 vs. 54.2 for B2). Full L2 English BLiMP scores across all L1s and curricula are reported in Table 8.

5 Reading Time (MECO L2)

The study of language processing in multilinguals is still an emerging area in computational psycholinguistics. Beetle provides a way to manipulate model training curricula such that we can test hypotheses about how different patterns of learning shape model predictions; and thus may enable research into how different forms of language exposure may impact patterns of reading.

Data scale is also a relevant factor in modelling human language processing. Research suggests that improvements in models' next-word prediction capability (as measured by perplexity) only lead to a better fit to reading times up to point, after which the alignment becomes worse (Kuribayashi et al., 2021; Wilcox et al., 2020; Oh et al., 2022, 2024). This transition has been reported to occur at around 2B training tokens (Oh and

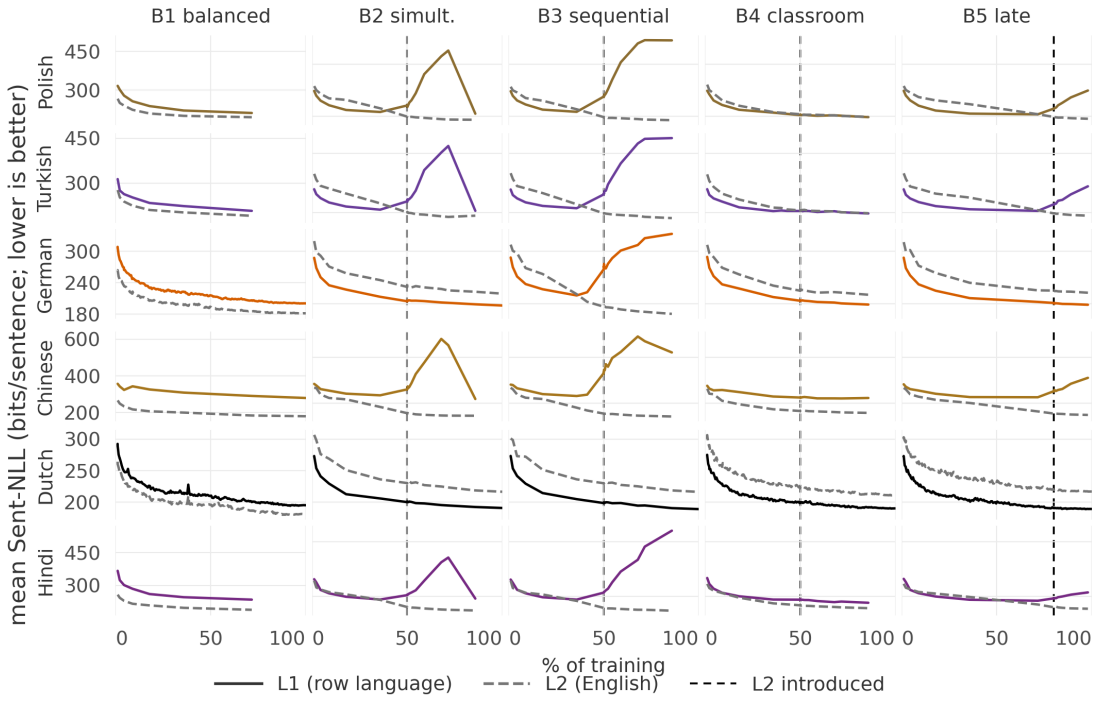

Figure 3: **FLORES Sentence NLL.** Each facet is one model; curves report mean sentence-level NLL in bits on 1,012 parallel FLORES-200 devtest sentences (lower is better), with solid lines for L1 and dashed lines for English. The vertical dashed line shows L2 English introduction.

Schuler, 2023), with some variation depending on reading time measure or dataset (Aoyama and Wilcox, 2025; Michaelov and Bergen, 2026). Beyond this point, the fit to larger model surprisal degrades to a greater extent (Oh and Schuler, 2023; Michaelov and Bergen, 2026), which may explain why smaller models trained on large amounts of data often perform better than larger ones (Oh et al., 2022, 2024). Additionally, there is a movement to use "human-scale" pretraining data (∼100M tokens; Choshen et al., 2026) that is also developmentally plausible, composed of child-oriented data (Jumelet et al., 2025a). We compare both of these data scales against near-Chinchilla-optimal (Hoffmann et al., 2022) web-scale data (24B tokens for 125M parameter models).

MECO eye-tracking dataset. We evaluate how bilingual exposure curricula affect how well Beetle's surprisal predicts human L2 reading times using the Multilingual Eye-Movement Corpus (MECO), a multilingual eye-tracking benchmark covering controlled reading in typologically diverse languages (Siegelman et al., 2022, 2025; Kuperman et al., 2023, 2025). We focus on English L2 reading by Dutch, German, and Chinese L1 participants using the combined Wave 1+2 releases[1]. MECO supports controlled cross-linguistic comparisons across matched reader populations. We use Beetle surprisals to predict *first-pass duration* (also known as *gaze duration)*, a widely-studied psycholinguistic eye-tracking measure that represents the time spent fixated on a word in the first pass before moving to another word, which shows strong surprisal-based prediction effects (e.g., Wilcox et al., 2020), and in the multilingual processing literature has been associated with early lexical access and initial word integration during reading (Dussias, 2010; Luke et al., 2023; Nahatame and Uchida, 2025). We compare the Beetle models to other baseline LMs, selected on the basis of being massively multilingual, and thus confirmed to or likely to be trained on the languages of interest. Given the aforementioned scaling effects, we select the smallest model of each family available. Mixed-effects regressions compare models with and without surprisal predictors, with improvements measured as $\Delta \log L$(see Appendix E for full details).

Results First, we consider L2 reading time for Dutch-, German-, and Chinese-L1 models. For these language pairs, we use models trained on all five curricula and four dataset sizes and com-

[1] https://meco-read.com/

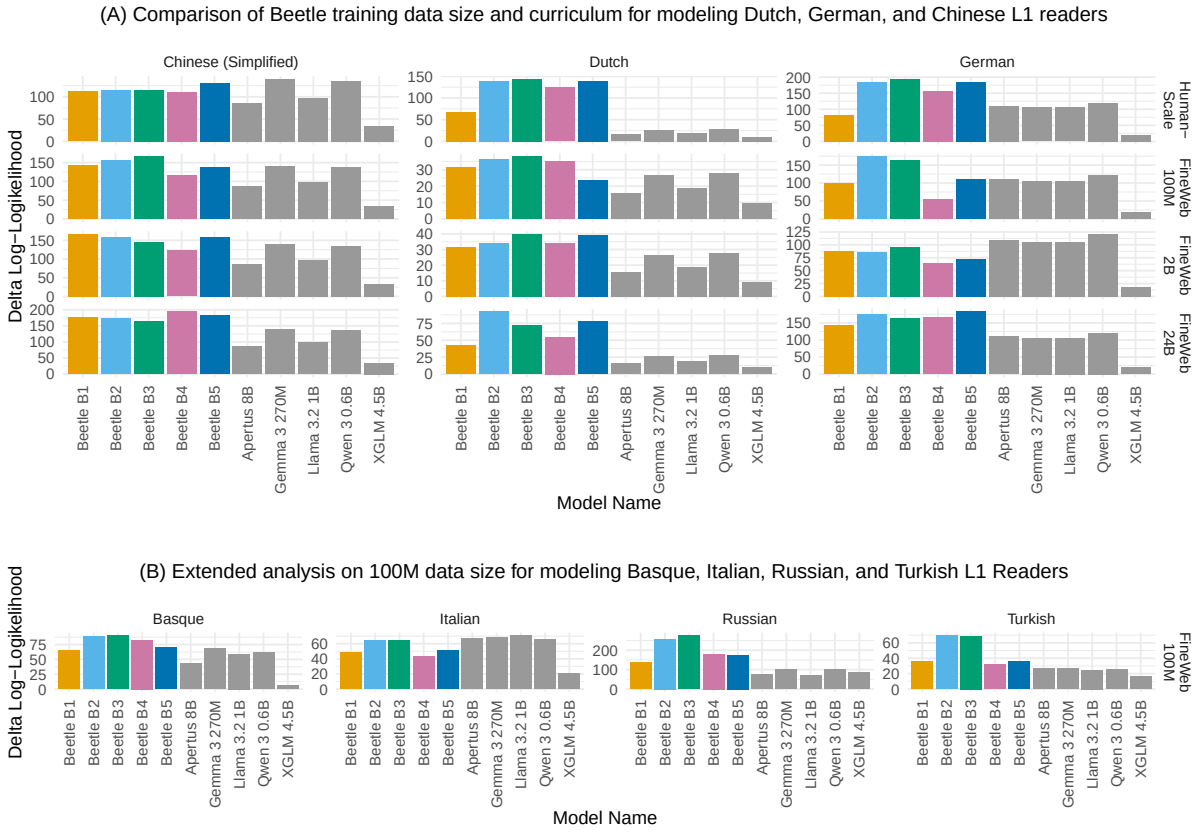

Figure 4: **Comparison of L2 Reading Time Alignment (MECO, (Siegelman et al., 2025; Kuperman, 2025)) of Beetle and multilingual LLMs** : (a) B1−B5 Dutch/Chinese/German-English Beetle (FineWeb 100M, 2B, 24B (Penedo et al., 2025), BabyBabelLM (Jumelet et al., 2025a)) and (b) B1 – B3 Basque/Italian/Russian/Turkish-English Beetle (100M FineWeb)

pare how well they predict reading time (Figure 4A). Figure 4 shows thatthe MECO alignment of the Beetle models is comparable to, or exceeds, multilingual LLMs at a range of parameter counts. Across seven L1s in Figure 4, staged exposure curricula (B2/B3/B5) consistently produce stronger surprisal–reading-time alignment than balanced bilingual exposure (B1); an effect strongest for Russian and German/Dutch L1s. Across training scales, effects of staged exposure are strongest for Dutch and German readers and weaker for Chinese readers, suggesting that curriculum benefits interact with typological proximity to English.

Comparing Human-Scale and FineWeb data at 100M token training budget, child-oriented training data from BabyBabelLM (Jumelet et al., 2025a) provides higher MECO alignment for Dutch and German L1 reading times, though this is not observed for Chinese L1 data. For Chinese, Dutch, and German, the models trained on 24B tokens show a closer fit than than at the smaller data scales, which contrasts with previous research on monolingual reading time showing a decrease in fit beyond 2B tokens (Oh and Schuler, 2023). Figure 4 shows reading time prediction for four additional L1s: Basque, Italian, Russian, and Turkish. For all language pairs except for Italian, surprisal calculated using our models better predict reading time than surprisal calculated with larger multilingual models.

6 Bilingual Processing Effects

6.1 Structural Priming

Previous work has argued that bilingual speakers share syntactic representations between languages, which allows for cross-linguistic syntactic priming (Hartsuiker et al., 2004): people are more likely to re-use the same syntactic structure in one language after hearing it in another language. It has been demonstrated that language models also exhibit crosslingual structural priming, showing that language models also share crosslingual representations (Michaelov et al., 2023; Arnett et al., 2025). We replicate the genitive prim-

ing (Bernolet et al., 2013)experiments in Arnett et al. (2025) with human-scale and web-scale Dutch-English Beetle models. Following Arnett et al. (2025), we quantify structural priming as the difference in normalized probability of a target sentence following congruent versus incongruent primes, fitting a linear mixed-effects model for each model–language combination with prime type as a fixed effect and experimental item as a random intercept, reporting the final-checkpoint results while applying Benjamini–Hochberg false-discovery-rate correction for multiple comparisons. We report the results in Table 1, which show robust priming effects across curricula for both training corpus sizes, with the exception of a marginally significant effect in the web-scale B3 model. This shows that the Beetle models learn to share crosslingual representations like other bilingual models, such as the B-GPT models (Arnett et al., 2025) and larger multilingual models (Michaelov et al., 2023), even when our models have seen small amounts of L2 data and are trained on much smaller datasets.

	FineWeb		HumanScale	
Curr.	diff	p	diff	p
Bernolet (genitive)				
B1	1.051	.0004	1.369	<.0001
B2	0.991	.037	1.095	<.0001
B3	0.969	.051	1.216	<.0001
B4	1.252	.0008	0.565	.005
B5	1.075	.0003	0.571	<.0001

Table 1: Cross-lingual structural priming for the 24B FineWeb and HumanScale models: surprisal difference (nats) between congruent and incongruent primes, with permutation p-values.

6.2 L2 Discrimination & CEFR

We secondly evaluate whether the curriculum effects observed in reading-time prediction also extend to grammatical discrimination. We use BLiSS (Gao et al., 2025), which tests whether models distinguish between a corrected sentence s_{corr}, a human learner error s_{lrn}, and an artificial error s_{art}. We measure sentence plausibility using bits per token (BPT),

$$\text{BPT}(s) = -\frac{1}{|s|}\sum_{t=1}^{|s|}\log_2 p(w_t \mid w_{<t}).$$

where lower BPT means that the model considers the sentence more likely. We report RP@0, the proportion of items for which the model assigns a lower BPT to the human learner error than to the corrected sentence, i.e., whether it prefers the learner error over the correction. On BLiSS, staged curricula (B2, B3-33, B5) generally outperform balanced exposure (B1), although the differences are smaller than those observed for MECO reading-time prediction (Fig. 5). On FineWeb `nld–eng`, the curriculum ranking is B3-33 $>$ B5 $>$ B2 $>$ B4 $>$ B1. The best curriculum also depends on the L1: B3-33 performs best for Germanic L1s, whereas B5 performs best for Chinese. Notably, BLiSS and MECO favour different curricula within Beetle (Spearman $\rho = -0.47$, $p < 0.001$), suggesting that curriculum effects do not transfer uniformly across evaluation tasks. Overall, BLiSS shows smaller curriculum differences than MECO reading-time prediction (App. F.1), with no single curriculum consistently dominating across languages. This suggests that the benefits of staged exposure depend on the aspect of language behaviour being evaluated. Detailed BLiSS results for the Beetle models and for the multilingual baselines are reported in Table 5 and Table 6, and CEFR-graded pedagogical results in Table 7.

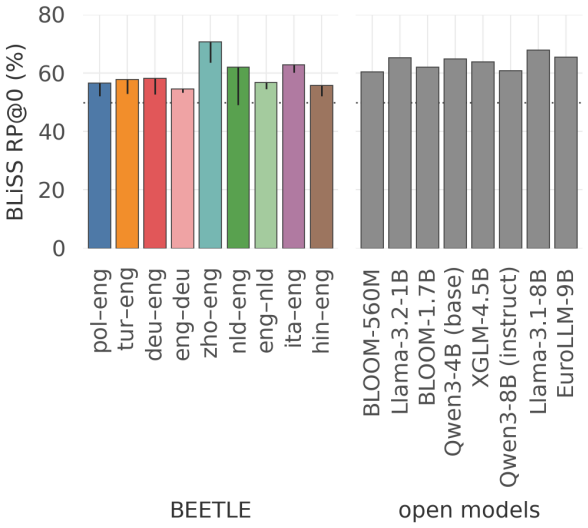

Figure 5: **BLiSS: Beetle vs. multilingual LLM performance.** The best Beetle models (zho–eng HumanScale and nld–eng FineWeb) are compared with the same baselines and a Dutch monolingual anchor (chance = 50). Large multilingual LLMs reach a higher ceiling (RP@0 $\approx$ 66–71). Unlike on MECO, Beetle does not outperform the multilingual baselines on BLiSS, indicating a dissociation between the two measures. Detailed results are reported in subsection F.1.

7 Discussion and Conclusion

Although psycholinguistics and cognitive science have historically focused on monolingual English-speaking, high-resource, and WEIRD populations (Majid and Levinson, 2010; Bylund, 2022), bilingualism and second-language (L2) processing are central to understanding human language processing. Computational approaches using language models inherit this bias, and an over-reliance on English and monolingual participants limits the generalisability of theories of language processing and acquisition (Blasi et al., 2022).

To address this gap, we introduced Beetle, a modular bilingual pretraining framework that disentangles exposure *structure* from training *scale* while holding architecture, tokeniser, and L2 target language fixed. Using Beetle, we pretrained 285 bilingual and 45 monolingual language models spanning 21 L1s and five exposure curricula, and we release checkpoints, training data, and reproducible pipelines to support future work on bilingual language modelling, learning dynamics, and cognitive modelling. Beetle provides interpretable and competitive models of bilingual processing. Small bilingual models trained under carefully controlled exposure conditions frequently matched or exceeded substantially larger massively multilingual systems on cognitively oriented evaluations, narrowing the gap between open-data academic models and closed-data industrial systems. Yet no single curriculum dominated.

Sequential and late-onset schedules produced stronger alignment with human L2 reading times, suggesting that temporally structured exposure better captures incremental bilingual processing, whereas balanced exposure achieved the strongest grammaticality (BLiMP) performance, indicating that more integrated exposure may better support robust grammatical generalisation; on BLiSS error discrimination the curriculum differences were smaller, with no single schedule dominating. That different schedules produce measurably different linguistic and cognitive behaviours, even with architecture and training volume held constant, suggests that carefully structured bilingual pretraining may yield stronger computational models of human language processing and language acquisition.

By releasing a reproducible framework alongside open bilingual models designed for studying bilingualism and L2 acquisition, Beetle lowers the barrier to systematic research at the intersection of NLP, psycholinguistics, and cognitive science. We hope it will support future work using language models as experimentally controllable systems for investigating how language experience shapes processing, prediction, and acquisition across diverse linguistic settings.

7.1 Future Work

There are several features of the Beetle framework that we did not discuss in the paper, due to space limitations. We introduce them here and highlight their potential utility.

Trilingual and Multilingual Models. Beetle is designed for multilingual pretraining, not just for training bilingual LMs. It can be used to develop more complex exposure curricula for multilingual LM pretraining or trilingual language modelling

Alternative Architectures and Additional Manipulations. Following recent work investigating training dynamics across architectures (Michaelov et al., 2025), we implement encoder-decoder, MoE, and SSM architectures (described in App. A). Extending the analyses in this paper to other architectures could help us understand whether these effects are specific to transformer language models. Beetle additionally integrates a range of continual-learning strategies (parameter regularisation, replay, meta-learning, distillation, and architectural variants); the full registry and taxonomy are given in App. C.2 (Table 3, Table 4).

Training Dynamics. We trained models with rich checkpoints, but only did limited analyses of how the models changed over training across different curricula. Up to now, there have been very few models with checkpoints for any language other than English. Future work could investigate how consistent training dynamics are across languages.

Interpretability. PicoDecoder's compatibility with Pico-Analyze (Diehl Martinez et al., 2025) allows mechanistic analysis of how bilingual pretraining manipulations affect weight geometry (e.g., shifts in effective rank or representational similarity) across checkpoints (Diehl Martinez et al., 2025; Inaba et al., 2025).

8 Limitations

L2-English. The Beetle pretraining experiments reported in the paper focus on non-English L1 data to English L2. Non-English L2s are not the focus due to limited evaluation metrics; this makes claims about directional asymmetries in transfer limited.

Matched Controls and Statistical Significance. Comparisons between Beetle and pretrained models, such as BLOOM-7b1, are not fully controlled for differences in parameter count, training data, tokeniser, or compute budget; a fuller control would involve training a similarly sized model (e.g., 1B parameters) on a matched FineWeb-based bilingual dataset with uniform sampling, and main results are reported using a single random seed per configuration. In particular, Beetle models use task-specific BPE tokenisers, whereas baselines such as BLOOM and XGLM use different tokenisation schemes, and we aggregate surprisal to the word level by summing subword contributions, which introduces a known source of noise when comparing across tokenisers (Pimentel and Meister, 2024). Within-Beetle comparisons are unaffected, since all compared models within a curriculum share the same tokeniser.

Scope of bilingual acquisition modelling. Our curricula attempt to provide computational analogues of several bilingual acquisition scenarios by manipulating the quantity, timing, and distribution of L1 and L2 exposure. B1 approximates relatively balanced bilingual development, with L1 and L2 available in a stable 50:50 mixture throughout training; B2–B4 approximate different forms of L2 learning, varying the onset, eventual proportion, and temporal distribution of L2 exposure; and B5 approximates late L2 learning following prolonged L1 exposure, providing a setting in which L1 attrition may occur. However, these curricula are not holistic models of bilingual acquisition: human bilingual development is shaped by factors beyond exposure history, including interaction, communicative goals, pragmatics, feedback, social context, and active language use, none of which are represented in our training setup. The curricula should therefore be understood as controlled computational analogues that isolate the role of exposure quantity and curriculum structure, rather than as comprehensive models of balanced bilingualism, L2 learning, or language attrition (Barbenel et al., 2026).

Broader Neurolinguistic and Psycholinguistic Alignment. The current evaluation focuses on behavioural proxies such as reading-time measures and CEFR-style tasks. We do not yet evaluate alignment with neural signals such as N400 or P600 responses. Extending the analysis to neurophysiological data is left for future work.

9 Reproducibility Statement

All results are produced using a fully versioned and reproducible analysis pipeline. The MECO L2 evaluation follows the mixed-effects specification in Equation 2 (with the baseline model of Equation 1), implemented in the released analysis scripts. Models are trained on multi-node GPU infrastructure (8× A100 80GB). All checkpoints, tokenisers, and evaluation scripts are publicly released, along with the exact CSV files used to generate figures and tables.

10 Ethics Statement

Beetle releases 285 pretrained bilingual and 45 monolingual L2-English language models, tokenisers, and approximately 30 checkpoints per model. **Dataset provenance**: BabyBabelLM and FineWeb-2 are publicly released, license-compatible multilingual web corpora; we do not introduce new web crawls. We release decontaminated and pretokenized datasets for each L1-English pair across scales. The MECO L2 eye-tracking corpus is used under its public licence with anonymous subject IDs only. **Intended use**: Beetle models are scientific instruments for studying bilingual language acquisition and L2 reading dynamics. They are not deployed L2 tutors, and the CEFR-graded evaluation pipeline is for research-grade scoring rather than placement decisions. **Dual-use considerations**: the CEFR-graded text generation capability of these models could in principle be repurposed to draft L2-graded content for assessment fraud (e.g., automated CEFR-tuned essay generation). This risk is mitigated by the small parameter count of the released models (125M), which underperforms publicly-available frontier LLMs on open-ended generation; and we recommend that any downstream pedagogical-tool deployment include detection-of-machine-generated-text safeguards. **Demographic representativeness**: the

released-model L1 sweep covers typologically diverse L1s, but has limited coverage of lower-resourced L1s.

11 Acknowledgements

We would like to thank CoreWeave for providing GPU resources for model training and evaluation. Suchir Salhan is supported by Cambridge University Press & Assessment. With thanks to Laura Barbenel, Lily Goulder, Aoife O'Driscoll, Filip Trhlik, Bianca Ganescu, Diana Galván-Sosa, Gabrielle Gaudeau; Andrew Caines, and other members of the ALTA Institute and Cambridge Press & Assessment; and Denise Löfflad and Detmar Meurers for their support and constructive feedback on earlier stages of this work.

References

David Demitri Africa, Suchir Salhan, Yuval Weiss, Paula Buttery, and Richard Diehl Martinez. 2025a. Meta-pretraining for zero-shot cross-lingual named entity recognition in low-resource Philippine languages. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 106–127, Suzhuo, China. Association for Computational Linguistics.

David Demitri Africa, Yuval Weiss, Paula Buttery, and Richard Diehl Martinez. 2025b. Learning dynamics of meta-learning in small model pretraining. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 10–23, Mumbai, India. Association for Computational Linguistics.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901.

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Tatsuya Aoyama and Ethan Wilcox. 2025. Language models grow less humanlike beyond phase transition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24938–24958, Vienna, Austria. Association for Computational Linguistics.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Catherine Arnett, Tyler A. Chang, James A. Michaelov, and Ben Bergen. 2025. On the acquisition of shared grammatical representations in bilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria. Association for Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Laura Barbenel, Lily Goulder, Aoife O'Driscoll, Suchir Salhan, Catherine Arnett, Andrew Caines, and Paula Buttery. 2026. L1 influence in L2 language models: A human-centric approach. In *Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL)*, pages 92–116, Grand Hyatt Manchester San Diego, 1 Market Pl, San Diego, CA 92101. Association for Computational Linguistics.

Sarah Bernolet, Robert J. Hartsuiker, and Martin J. Pickering. 2013. From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, 127(3):287–306.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, U. S. Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. *arXiv preprint arXiv:2205.11758*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon

Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Emanuel Bylund. 2022. Weird psycholinguistics. *Struggles for Multilingualism and Linguistic Citizenship*, 173.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. 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: Call for papers for the 2026 babylm workshop. *arXiv preprint arXiv:2602.20092*.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144, Online only. Association for Computational Linguistics.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Richard Diehl Martinez, David Demitri Africa, Yuval Weiss, Suchir Salhan, Ryan Daniels, and Paula Buttery. 2025. Pico: A modular framework for hypothesis-driven small language model research. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 295–306, Suzhou, China. Association for Computational Linguistics.

Ton Dijkstra and Walter J. B. Van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175–197.

Paola E. Dussias. 2010. Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, 30:149–166.

Steven Y. Feng, Aditya Yadavalli, Chenhao Wang, and Tyler A. Chang. 2026. Bilingual BabyLM: Matched mono- and bi-lingual pretraining at 100m words. Forthcoming.

Stefan Frank. 2014. Modelling reading times in bilingual sentence comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Stefan L Frank. 2021. Toward computational models of multilingual sentence processing. *Language Learning*, 71(S1):193–218.

Yuan Gao, Suchir Salhan, Andrew Caines, Paula Buttery, and Weiwei Sun. 2025. Bliss: Evaluating bilingual learner competence in second language small language models. In *Proceedings of the First BabyLM Workshop*, pages 160–174.

Tamar H. Gollan and Victor S. Ferreira. 2009. Should i stay or should i switch? a cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):640–665.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

François Grosjean. 2021. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.

Robert J. Hartsuiker, Martin J. Pickering, and Eline Veltkamp. 2004. Is syntax separate or shared between languages? cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, 15(6):409–414.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals,

Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Tatsuro Inaba, Go Kamoda, Kentaro Inui, Masaru Isonuma, Yusuke Miyao, Yohei Oseki, Yu Takagi, and Benjamin Heinzerling. 2025. How a bilingual LM becomes bilingual: Tracing internal representations with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13458–13470, Suzhou, China. Association for Computational Linguistics.

Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrama, Francesca Padovani, Francois Meyer, Hai Hu, Julen Etxaniz, Laurent Prévot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, Suchir Salhan, Susana Zhou, Yurii Paniv, Ziyin Zhang, Arianna Bisazza, Alex Warstadt, and Leshem Choshen. 2025a. Babybabellm: A multilingual benchmark of developmentally plausible training data.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025b. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 2026. When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3169, Rabat, Morocco. Association for Computational Linguistics.

Victor Kuperman. 2025. How does language distance affect reading fluency and comprehension in English as second language? *Studies in Second Language Acquisition*, 47(3):757–773.

Victor Kuperman, Sascha Schroeder, Cengiz Acartürk, Niket Agrawal, Dominick M Alexandre, Lena S Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, et al. 2025. New data on text reading in english as a second language: The wave 2 expansion of the multilingual eye-movement corpus (meco). *Studies in Second Language Acquisition*, 47(2):677–695.

Victor Kuperman, Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2023. Text reading in english as a second language: Evidence from the multilingual eye-movements corpus. *Studies in second language acquisition*, 45(1):3–37.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, Rui-Jie Tu, and Hai Liu. 2025. ZhoBLiMP: A systematic assessment of language models with linguistic minimal pairs in Chinese.

Steven G. Luke, Rachel Yu Liu, Kyle Nelson, Jared Denton, and Michael W. Child. 2023. An ex-Gaussian analysis of eye movements in L2 reading. *Bilingualism: Language and Cognition*, 26(2):330–344.

Asifa Majid and Stephen C Levinson. 2010. Weird languages have misled us, too. *Behavioral and Brain Sciences*, 33(2-3).

Yevgen Matusevych, Afra Alishahi, and Ad Backus. 2013. Computational simulations of second language construction learning. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 47–56, Sofia, Bulgaria. Association for Computational Linguistics.

Renata F. I. Meuter and Alan Allport. 1999. Bilingual language switching in naming: Asymmetrical costs

of language selection. *Journal of Memory and Language*, 40(1):25–40.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

James A Michaelov and Benjamin K Bergen. 2026. Better language models better model the n400, but not reading time. *Journal of Memory and Language*, 149:104762.

James A Michaelov, Roger P Levy, and Benjamin K Bergen. 2025. Language model behavioral phases are consistent across architecture, training data, and scale. *arXiv preprint arXiv:2510.24963*.

John Mullooly. 2023. CUP&A5: A cefr-stratified multiple-choice question-answering dataset for pedagogical evaluation. Cambridge University Press & Assessment internal dataset.

Shingo Nahatame and Satoru Uchida. 2025. Relative importance of lexical features in word processing during L2 English reading. *Studies in Second Language Acquisition*, 47(5):1383–1406.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1915–1921.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics (volume 1: long papers)*, pages 2644–2663.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

Sara Rajaee and Christof Monz. 2024. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2895–2914.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 54(6):2843–2863.

Noam Siegelman, Sascha Schroeder, Yaqian Borogjoon Bao, Cengiz Acartürk, Niket Agrawal, Lena S Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, et al. 2025. Wave 2 of the multilingual eye-movement corpus (meco): New text reading data across languages. *Scientific data*, 12(1):1183.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Michelle Suijkerbuijk, Zoe Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025a. BLiMP-NL: A corpus of Dutch minimal pairs and grammaticality judgements for language model evaluation.

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

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2019. Lamol: Language modeling for lifelong language learning. *arXiv preprint arXiv:1909.03329*.

Swiss AI Initiative. 2025. Apertus: Democratizing open and compliant LLMs for global language environments.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during llm training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173.

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.

Veronica Whitford and Debra Titone. 2015. Second-language experience modulates eye movements during first- and second-language sentence reading: Evidence from a gaze-contingent moving window paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(4):1118–1129.

Ethan G. Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiulin Yang, Ethan Gotlieb Wilcox, and Catherine Arnett. 2026. Apples to Apples? Towards Comparable Crosslingual Language Model Evaluation.

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

Ej Zhou, Suchir Salhan, Catherine Arnett, and Anna Korhonen. 2026. Cross-lingual alignment without joint training: Do monolingual language models converge on universal representations?

Appendix

A Model Architecture

BeetleLM is built on a custom PICO decoder stack (Diehl Martinez et al., 2025) with curriculum-aware checkpointing, a registry of fifteen continual-learning strategies, and a tokeniser layer supporting per-language, shared-bilingual, and shared-trilingual vocabularies.

By default, Beetle uses the `PicoDecoder`, a LLaMA-style (Touvron et al., 2023) decoder with $L = 14$ blocks, RMSNorm (Zhang and Sennrich, 2019), grouped-query attention (Ainslie et al., 2023) (12 query heads, 1 key/value head), rotary position embeddings (Su et al., 2024), and SwiGLU feed-forward (Shazeer, 2020) expanding to $4d$ and projecting back to d. The shared bilingual vocabulary used in the main runs has size 50,304; per-language and shared-trilingual tokenisers are configured separately. Sequence length is 512 under BF16-mixed precision. Width is the only scale-dependent hyperparameter: $d \in \{96, 384, 768, 1536\}$ for the tiny, small, medium, and large variants. The 125M-parameter "medium" variant ($d = 768$, 12 heads, 1 KV head, 14 layers) is used for every reported run.

A.1 Alternative Architectures

Complementary architectures enable controlled comparisons across inductive biases: **BGPT**, a GPT-2-style decoder with learned positional embeddings used in Arnett et al. (2025); **BeetleMoE**, a top-k mixture-of-experts model with ST-MoE auxiliary routing losses; and **BeetleSSM**, which replaces attention blocks with Mamba-style state-space modules.

Pre-pretraining (Shuffle-Dyck) and time-varying ALiBi slopes can be used with different architectures.

Hyperparameter	100M	2B	24B
Architecture	PicoDecoder	PicoDecoder	PicoDecoder
Parameters	125M	125M	125M
Layers	14	14	14
d_{model}	768	768	768
Heads (Q/KV)	12 / 1	12 / 1	12 / 1
FFN size	3072	3072	3072
Sequence length	512	512	512
Vocabulary	50K	50K	50K
Languages	9	20	20
Dataset	BabyBabel FineWeb-2	FineWeb-2	FineWeb-2
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}
Weight decay	0.01	0.01	0.01
Warmup steps	500	381	625
Micro-batch	64	64	64
Effective batch	64	512	512
Tokens / step	32768	262144	262144
Max steps	10000	7629	91552
Precision	bf16	bf16	bf16
Strategy	DDP	DDP	DDP
Devices	8×A100	8×A100	8–64×A100

Table 2: Training hyperparameters across data scales.

A.2 Training Hyperparameters

Training hyperparameters for reported experiments are in Table 2. Training was conducted on 8× A100s (64GB). 100M and 2B tokens 125M models can be trained in under an hour on 1 A100. 24B token modles take around 5 hours on 8× A100.

B Beetle-Data

Beetle-Data is a module that prepares pretokenized data for language model training, streaming data from default, or user-specified, training datasets. When streaming raw documents, Beetle-Data applies a 13-gram benchmark-overlap decontaminator (Brown et al., 2020b). Reported experiments perform 13-gram decontamination for 19 evaluation metrics: BLiMP, MultiBLiMP, BLiMP-NL, ZhoBLiMP, FLORES-200, XNLI, UD Treebanks, MECO-L2 stimuli. Finally, the library pretokenises decontaminated data to user-specified training budgets and tokenization.

C Curriculum Parameterisation in Beetle

Beetle represents bilingual and multilingual curricula as continuous exposure schedules rather than as discrete training phases with fixed boundaries. Formally, each curriculum defines a time-varying language mixture $w(t) \in \Delta^{|L|}$ over normalised training progress $t \in [0, 1]$, where $\Delta^{|L|}$ denotes the simplex over the set of languages L. Progress t may be measured either in optimisation steps or in tokens seen. At every point during training, $w_l(t)$ specifies the probability mass assigned to language l, with $\sum_l w_l(t) = 1$.

All curriculum conditions are constructed compositionally from four schedule primitives, summarised in Table 4.

SigmoidRamp. The basic transition mechanism is a smooth sigmoid interpolation between two language-mixture vectors. The transition is parameterised by a midpoint $m \in [0, 1]$, which determines the onset timing, and a steepness parameter k, which controls transition sharpness. Larger values of k approximate an abrupt step function, while smaller values yield gradual transitions. Unless otherwise stated, experiments use $k = 12$. All sigmoid-based curricula (including B2, B3, and B5) share the same implementation, meaning that transition sharpness is controlled by a single scalar parameter rather than by curriculum-specific logic.

MultiSigmoid. More complex staged introductions are implemented by chaining multiple sigmoid ramps. This permits schedules such as $L_1 \rightarrow L_2 \rightarrow L_3$ with independently parameterised transition points. The trilingual curricula (T2 and T3) use this mechanism.

BurstSchedule. Burst schedules superimpose periodic square-wave increases in exposure for a target language on top of an existing base schedule. Each burst process is parameterised by an amplitude (the additive increase in exposure weight), a period (cycle length as a fraction of total training), a duty cycle (the fraction of each cycle during which the burst is active), and a start time t_{start}. During an active burst, the target language receives an increased allocation and the remaining language weights are proportionally renormalised so that $\sum_l w_l(t) = 1$. Burst schedules are intended to model concentrated episodes of exposure, such as intensive classroom interaction or temporary re-exposure to the first language. They are used in conditions such as B2_burst and the L1 re-exposure manipulation in B6.

CyclicalSchedule. Where smoother recurring fluctuations are theoretically motivated, **BEETLELM** uses sinusoidal oscillations rather than discrete square-wave bursts. The cyclical schedule is parameterised by an amplitude and frequency and produces continuous oscillatory variation around a base exposure schedule.

C.1 Episodic Clustering

In addition to continuous exposure scheduling, **BEETLELM** implements episodic clustering through a separate mechanism termed the `EpisodicSampler`. This mechanism is orthogonal to the global exposure schedule $w(t)$. Whereas bursts modify the overall language proportions through time, episodic clustering changes the local temporal structure of sampling.

The episodic sampler operates over a finite set of named contexts $C = \{c_1, \ldots, c_n\}$, each associated with a characteristic language distribution and a marginal sampling probability π_c. Example contexts include `home` with distribution $\{L_1 : 0.95, L_2 : 0.05\}$ and `classroom` with distribution $\{L_1 : 0.3, L_2 : 0.7\}$.

Training examples are sampled in contiguous blocks of size B. Within a block, all examples are drawn from the same context distribution. This breaks the IID assumption typically used in standard language-model training and approximates the contextual non-stationarity characteristic of naturalistic second-language exposure.

When both a global schedule and episodic sampling are active, the effective per-step language weights are computed as the geometric mean of the scheduled distribution $w(t)$ and the currently active context distribution, followed by renormalisation.

Condition B4 (*Classroom*) combines both mechanisms: an underlying sigmoid transition toward an 80:20 language mixture, periodic L2 bursts, and a two-context episodic sampler consisting of `home` (80%) and `classroom` (20%) contexts with block size $B = 64$. Condition B1b is a corresponding non-classroom control using a balanced 50:50 base schedule with episodic clustering only. B4a and B4b are ablations removing episodic clustering and EWC respectively.

A bilingual curriculum is a time-dependent

Family	Strategy	Mechanism	Key hyperparameters
Parameter regularisation			
	EWC (Kirkpatrick et al., 2017)	Diagonal-Fisher quadratic penalty at phase change	$\lambda \in [0, 40]$, $K_{\text{Fisher}} = 10$
	SI	Online importance accumulation	c, damping
	MAS	Gradient-magnitude importance	λ
Replay			
	LAMOL (Sun et al., 2019)	Pseudo-sample generation, nucleus sampling	$\lambda_{\text{replay}} \in [0, 0.5]$, top-$p$
	InsCL similarity replay	Centroid-similarity weighted replay	cosine/Wasserstein, buffer size
	Episodic replay	Stored exemplars, uniform sampling	buffer size, mix ratio
Meta-learning			
	MAML / SMLMT	Inner-loop adaptation on (S, Q) splits	hybrid ratio $\in [0, 0.4]$, N-way, k-shot
Distillation			
	Hinton-KD	Soft-target cross-entropy against pre-L2 anchor	T, λ_{KD}
	Self-distillation	Same model as teacher at earlier checkpoint	checkpoint offset
	FitNets feature distill.	Hidden-state ℓ_2 alignment	layer-pair list
Architectural / auxiliary loss			
	TILT freeze-and-extend	Freeze backbone; train embeddings + LM head	frozen modules
	Exposure-LR plasticity	Token-exposure LR with phase-transition boost	boost $\in [1, 3]$, warmup tokens
	ALiBi scheduler	Time-varying ALiBi slopes across phases	initial/final slope, decay shape
	Shuffle-Dyck pre-pretraining	Formal-language warmup before NL	steps, depth
	EWC + MAML stack	Composite of regularisation + meta-learning	paired hyperparameters

Table 3: The continual-learning strategies implemented in the Beetle training library. The experiments reported here use EWC, LAMOL, and exposure-LR plasticity overlays.

	B1	B2	B3	B4	B5
L2 onset	0	50%	50%	0	80%
Final L2	50%	50%	67%	20%	50%
Transition	none	sig.	sig.	bursts	sig.
Bursting	no	no	no	yes	no
Episodic	no	no	no	yes	no
CL overlay	none	none	none	none	none

Table 4: Controlled curriculum factors for bilingual conditions B1–B5. B4 uses bursty episodic exposure with the same mean L2 ratio as B2 and no EWC.

mixture $\mathcal{D}_{\text{mix}}(t) = \alpha(t)\,\mathcal{D}_1 \cup (1-\alpha(t))\,\mathcal{D}_2$, where $\alpha(t) \in [0, 1]$ is the per-step L2 sampling probability and $t \in [0, T]$ indexes training. The polyglot generalisation is a language-exposure vector $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_N(t))$ with $\sum_i \alpha_i(t) = 1$. Each schedule maps a normalised progress value $p \in [0, 1]$ to a per-language sampling weight; phase boundaries are discrete events that trigger EWC snapshots in continual-learning runs. Rather than specifying $\boldsymbol{\alpha}(t)$ as an arbitrary continuous function, we represent curricula as a sequence of discrete phases $\mathcal{C} = \{(\tau_k^{\text{start}}, \tau_k^{\text{end}}, \mathbf{w}_k)\}_{k=1}^{K}$, where $\mathbf{w}_k \in \mathbb{R}^N$ satisfies $\sum_i w_{k,i} = 1$; for $\tau_k^{\text{start}} \leq t/T < \tau_k^{\text{end}}$ we set $\boldsymbol{\alpha}(t) = \mathbf{w}_k$. This declarative phase representation is used in every reported run.

C.2 Beetle Continual-Learning Strategies

The BeetleLM library contains **fifteen** modular continual-learning strategies, which are grouped in five families, summarised in Tab. 3.

Parameter regularisation: EWC (Kirkpatrick et al., 2017) with $\lambda \in [0, 40]$ and a Fisher snapshot at phase change, SI (synaptic intelligence), and MAS (memory-aware synapses).

Generative and similarity replay: LAMOL (Sun et al., 2019) with $\lambda_{\text{replay}} \in [0, 0.5]$, InsCL-style similarity replay (cosine and Wasserstein centroids), and episodic-sample replay buffers.

Meta-learning: first-order MAML / SMLMT (Bansal et al., 2020; Africa et al., 2025b,a) with hybrid ratio $\in [0, 0.4]$ and N-way/k-shot inner episodes.

Distillation: feature-level (FitNets), output-level (Hinton-KD), and self-distillation against the pre-L2 anchor. *Architectural and auxiliary-loss*: TILT freeze-and-extend, exposure-LR plasticity boost (`plasticity_boost`$\in [1, 3]$), ALiBi-scheduler decay, and Shuffle-Dyck pre-pretraining.

D Bilingual and Second Language Evaluation Methodology

We select different three evaluation families probe distinct facets of second-language competence. MECO measures graded processing alignment through reading-time prediction. BLiSS (Gao et al., 2025) tests categorical grammatical knowledge, and its triplet design—contrasting a corrected form, a genuine learner error, and a syn-

thetic error—additionally asks whether a model internalises an L2-specific interlanguage error structure rather than a generic notion of ungrammaticality. JFLEG (Napoles et al., 2017) tests whether a model prefers corrected forms over learner productions, connecting the framework to grammatical error correction. Because Beetle employs a shared vocabulary under controlled exposure schedules, it offers a platform for studying bilingual-specific phenomena that lie beyond monolingual processing. Candidate directions include lexical facilitation and cognate effects of the kind modelled by Dijkstra and Van Heuven (2002), the resolution of interlingual homographs, code-switching and the switch costs documented in the bilingual production and comprehension literature (Meuter and Allport, 1999; Gollan and Ferreira, 2009), and cross-linguistic syntactic interference (Hartsuiker et al., 2004). These phenomena are noted here as future directions that the controlled-exposure design of the framework is well placed to support.

E Details of Beetle MECO Analyses

For each target word w_i, we compute autoregressive surprisal $S(w_i) = -\log P(w_i \mid w_{<i})$ on the relevant Beetle and baseline models. Following previous work, (e.g., Wilcox et al., 2023), we fit mixed-effects regressions predicting first-pass reading time from current-word and spillover surprisal while controlling for word length, frequency, and sentence position, and evaluate language model alignment with reading time based on the fit of the regressions to the data.

We fit a linear mixed-effects regression models to MECO first-pass duration (also known as *gaze duration*) data from MECO waves 1 + 2 based on our predictors. We first fit baseline regressions with each word's position in the sentence, length in characters, and frequency as main effects, as well as each the preceding two words' length and frequency; and include by-subject uncorrelated random slopes for each of these as well as random intercepts for each subject:

$$\begin{aligned}\text{first\_pass} \sim\ & \text{position}(w_i) + \text{len}(w_i) \\ & + \text{len}(w_{i-1}) + \text{len}(w_{i-2}) \\ & + \text{freq}(w_i) + \text{freq}(w_{i-1}) \\ & + \text{freq}(w_{i-2}) \\ & + (1 + \text{position}(w_i) + \text{len}(w_i) \\ & \quad + \text{len}(w_{i-1}) + \text{len}(w_{i-2}) \\ & \quad + \text{freq}(w_i) + \text{freq}(w_{i-1}) \\ & \quad + \text{freq}(w_{i-2}) \mid\mid \text{subject})\end{aligned} \tag{1}$$

We then fit linear mixed-effects regressions that included all the baseline predictors in addition to the surprisal of the current and previous two words, which were also included as uncorrelated random slopes:

$$\begin{aligned}\text{first\_pass} \sim\ & \text{Surprisal}(w_i) \\ & + \text{Surprisal}(w_{i-1}) \\ & + \text{Surprisal}(w_{i-w}) \\ & + (1 + \text{Surprisal}(w_i) \\ & \quad + \text{Surprisal}(w_{i-1}) \\ & \quad + \text{Surprisal}(w_{i-w}) \mid\mid \text{subject}) \\ & + \text{baseline\_predictors}\end{aligned} \tag{2}$$

Since we hold all regression predictors constant except the surprisal terms, we consider language models with surprisal terms that lead to better fits (i.e., higher log-likelihoods) to first-pass duration to be models that better align with reading time. For easier comparison, we subtract the log-likelihood of the baseline regressions to get $\Delta \log L$:

$$\Delta \log L = \log L(\text{full}) - \log L(\text{baseline}). \tag{3}$$

We thus compare language model alignment with reading time based on the $\Delta \log L$ of regressions that include surprisal calculated using them.

The script used to fit the regressions is released with the code. Sample sizes are on the order of tens of subjects and tens of thousands of word-level observations.

F BLiSS and Pedagogical Evaluation

We evaluate whether curriculum effects observed in reading-time prediction extend to grammatical discrimination and pedagogical benchmarks.

First, we consider BLiSS (Gao et al., 2025), which evaluates preference behaviour over L2 error triplets, consisting of a corrected sentence s_{corr}, a human learner error s_{lrn}, and an artificial error s_{art}. We compute token-normalized surprisal:

$$\text{BPT}(s) = -\frac{1}{|s|}\sum_{t=1}^{|s|}\log_2 p(w_t \mid w_{<t}),$$

where lower values indicate higher model plausibility.

We report four metrics following (Gao et al., 2025). **Learner Preference (LP)** measures whether s_{lrn} is preferred over s_{corr} ($\text{BPT}(s_{\text{lrn}}) < \text{BPT}(s_{\text{corr}})$). This metric is informative but does not distinguish between grammatical competence and learner simulation behavior. **Human vs. Artificial Preference (HAP)** measures whether s_{lrn} is preferred over s_{art}, and **HAP-τ** applies a margin constraint for robustness. **Strict Order (SO)** requires the full ordering $s_{\text{corr}} \prec s_{\text{lrn}} \prec s_{\text{art}}$. We report all metrics separately as they capture different aspects of model behavior.

F.1 Detailed Beetle BLiSS Results

Table 5 and Table 6 show BLiSS accuracy of Beetle and LLMs. The overall performance is summarised in *Figure* 6. Beyond this headline accuracy, the detailed BLiSS table (Table 5) decomposes performance into RP@0 and RP@τ (random-over-human preference, with and without a margin), NGS (normalised gap score), and CPS (correct-preferred sanity check); reporting these additional metrics separates whether models merely prefer human over random productions (RP@0) from whether they do so by a robust margin (RP@τ), and confirms the effect is not a scoring artefact (CPS/NGS), giving a finer-grained picture than the headline accuracy in Figure 6. We briefly summarise the main results.

Curriculum effects on BLiSS On BLiSS, staged curricula (B2, B3-33, B5) outperform B1, but effect sizes are smaller than those observed on MECO (Fig. 5). On FineWeb `nld–eng`, the ordering is B3-33 > B5 > B2 > B4 > B1 , compared to a larger separation in MECO (approximately 78 $\Delta \log L$). The best-performing curriculum varies by L1: B3-33 ranks highest for Germanic L1s, while B5 ranks highest for Chinese L1. Within Beetle, BLiSS and MECO induce different curriculum rankings (Spearman $\rho = -0.47$, $p < 0.001$).

Summary across discrimination tasks Across BLiSS, curriculum differences are smaller than those observed for MECO reading-time prediction (subsection F.1). Best-performing curricula vary by L1 (B3-33 for Germanic, B5 for Chinese), with no single schedule dominating across settings. Across pedagogical evaluations (Table 7), balanced exposure (B1) consistently performs best on cloze and CEFR-style metrics. In contrast, staged curricula only dominate on MECO reading-time alignment, indicating a clear separation between fluency/correction-based judgments and online processing measures.

F.2 CEFR Task Performance

This section reports Beetle's performance on the CEFR-aligned and other pedagogical evaluations.

Fig. 7 shows performance of Beetle on different pedagogical tasks. Tab. 7 summarises the downstream performance of Beetle 125M models on CLOTH closed-cloze accuracy, Cambridge open-cloze accuracy, JFLEG learner-fluency margin, false-friend disambiguation margin, and the CEFR perplexity gradient, by curriculum and training scale. The individual benchmarks measure the following:

- **CLOTH** (Xie et al., 2018): a large cloze test of English built from middle/high-school exam passages; closed multiple-choice gap-filling measuring grammatical and lexical competence.
- **Cambridge cloze**: open-cloze items from Cambridge English exams graded by CEFR level; the model must supply the missing word, testing productive grammatical knowledge (Mullooly, 2023).
- **JFLEG grammatical correction** (Napoles et al., 2017): a fluency/grammatical-error-correction benchmark; we score the model's preference (bits/token margin) for the corrected sentence over the learner-error version (positive = prefers the correction).
- **False-friend disambiguation**: cognate/false-friend minimal pairs; the margin (bits/token) by which the model prefers the contextually correct sense over the interlingual-homograph distractor.

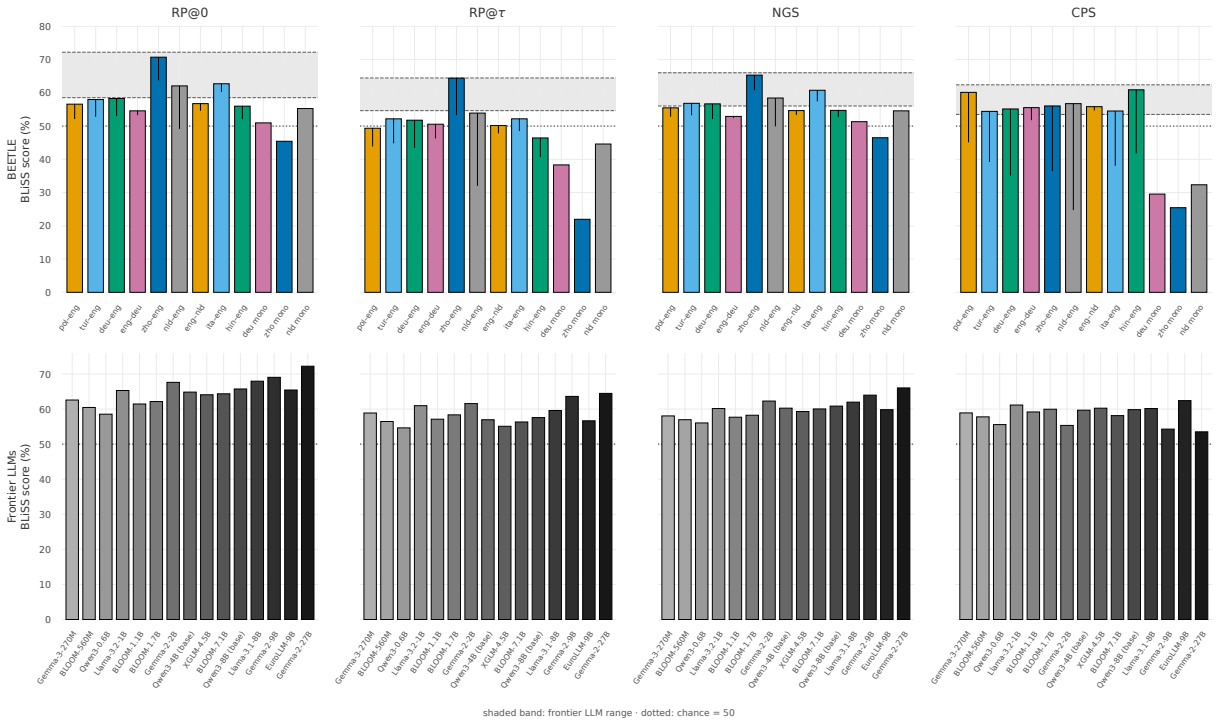

Figure 6: Overall BLiSS learner-minimal-pair accuracy for Beetle models against baseline-matched multilingual LLMs, aggregated across the matched-L1 cohorts. Curriculum differences on BLiSS are smaller than the corresponding differences in MECO reading-time alignment.

- **CEFR perplexity gradient** (Arase et al., 2022): the monotonic change in bits-per-token from CEFR-A1 to CEFR-C2 graded texts; a well-calibrated model should assign lower perplexity to easier (A1) than harder (C2) material, so the A1→C2 gradient indexes proficiency sensitivity.

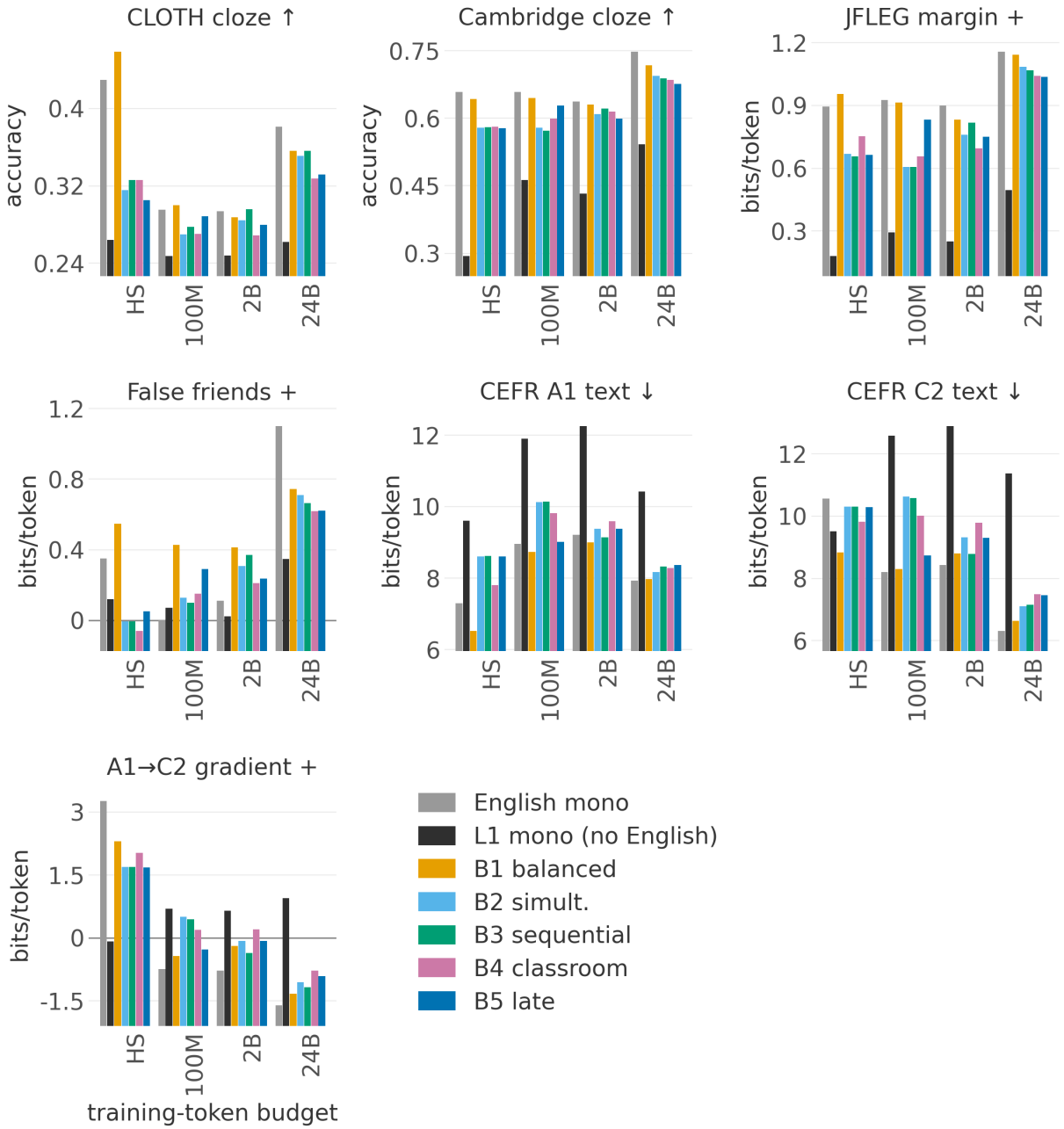

Figure 7: Performance of Beetle 125M models across pedagogical evaluations (CLOTH and Cambridge cloze, JFLEG grammatical correction, false-friend disambiguation, and the CEFR perplexity gradient), by curriculum and training scale. Balanced exposure (B1) is consistently strongest on these fluency- and correction-based measures.

Pair	Data	Curr.	BLiSS (matched-L1 cohort) RP@0	RP@τ	NGS	CPS
deu–eng	Human-scale	Mono	51.0	38.3	51.3	29.5
	Human-scale	B1 balanced	**57.2**	**48.2**	55.5	46.8
	Human-scale	B2 simult.	54.8	44.5	54.9	35.1
	Human-scale	B3 sequential	55.5	47.0	55.5	36.7
	Human-scale	B4 classroom	54.8	44.0	54.7	39.3
	Human-scale	B5 late	55.2	44.9	54.6	35.4
	100M tok	B1 balanced	54.6	45.8	54.4	50.1
	100M tok	B2 simult.	55.6	47.1	**56.2**	40.1
	100M tok	B3 sequential	55.6	45.3	55.0	38.2
	100M tok	B4 classroom	54.9	45.5	55.7	44.0
	100M tok	B5 late	56.1	47.1	55.3	48.1
	2B tok	B1 balanced	53.5	43.5	53.4	48.7
	2B tok	B2 simult.	55.5	46.1	55.0	47.0
	2B tok	B3 sequential	53.5	45.8	53.7	49.9
	2B tok	B4 classroom	55.7	45.2	55.6	44.0
	2B tok	B5 late	56.0	44.7	55.2	48.0
	24B tok	B1 balanced	53.1	46.8	53.0	**52.0**
	24B tok	B2 simult.	54.9	46.3	55.6	46.1
nld–eng	Human-scale	Mono	55.3	44.6	54.6	32.3
	Human-scale	B1 balanced	56.6	48.9	55.2	48.4
	Human-scale	B2 simult.	57.5	50.1	57.4	38.6
	Human-scale	B3 sequential	57.7	49.1	57.0	40.3
	Human-scale	B4 classroom	57.3	47.6	55.8	42.3
	Human-scale	B5 late	57.1	48.6	56.7	39.2
	100M tok	B1 balanced	57.1	50.1	57.2	48.9
	100M tok	B2 simult.	59.5	50.2	**58.3**	40.7
	100M tok	B3 sequential	59.3	50.4	57.3	40.3
	100M tok	B4 classroom	57.9	49.8	57.2	41.5
	100M tok	B5 late	57.1	49.0	56.9	48.8
	2B tok	B1 balanced	58.5	50.6	57.0	47.7
	2B tok	B2 simult.	57.4	50.2	57.7	39.6
	2B tok	B3 sequential	58.1	49.1	57.8	37.3
	2B tok	B4 classroom	58.9	49.9	57.7	39.2
	2B tok	B5 late	58.7	50.4	57.6	36.5
	24B tok	B1 balanced	56.6	48.6	54.9	**56.8**
	24B tok	B2 simult.	60.0	51.9	57.9	51.5
	24B tok	B3 sequential	**62.1**	**52.2**	58.1	50.1
zho–eng	Human-scale	Mono	45.4	22.0	46.5	25.5
	Human-scale	B2 simult.	63.8	54.6	60.8	36.5
	Human-scale	B3 sequential	64.0	53.3	61.5	36.4
	Human-scale	B5 late	64.3	55.0	61.9	36.4
	2B tok	B1 balanced	68.5	60.1	64.6	49.5
	2B tok	B2 simult.	68.4	59.0	64.1	48.5
	2B tok	B3 sequential	69.1	60.2	64.7	49.0
	2B tok	B4 classroom	64.8	56.1	63.4	41.4
	2B tok	B5 late	67.6	57.9	64.5	45.4
	24B tok (FW-24B)	B1 balanced	69.8	61.7	64.8	54.8
	24B tok (FW-24B)	B2 simult.	70.2	62.4	64.8	54.6
	24B tok (FW-24B)	B3 sequential	68.3	61.2	63.6	55.2
	24B tok (FW-24B)	B3 sequential *ewc*	69.8	61.3	64.2	**56.1**
	24B tok (FW-24B)	B4 classroom	69.0	62.1	64.7	51.9
	24B tok (FW-24B)	B5 late	**70.7**	**63.3**	**65.0**	53.8
hin–eng	100M tok	B1 balanced	**54.8**	**44.3**	**54.3**	**53.8**
	100M tok	B2 simult.	52.4	40.7	53.6	42.5
	100M tok	B3 sequential	52.1	41.0	53.6	44.2
	100M tok	B4 classroom	54.3	41.5	53.9	46.7
	100M tok	B5 late	54.5	43.7	53.7	53.0
	2B tok	B2 simult.	53.0	40.7	53.0	41.8
ita–eng	100M tok	B1 balanced	**62.6**	**51.3**	59.7	**49.3**
	100M tok	B2 simult.	61.8	49.4	**60.8**	38.1
	100M tok	B3 sequential	62.1	48.5	59.8	38.3
	100M tok	B4 classroom	**62.6**	50.3	60.7	43.5
	100M tok	B5 late	61.5	50.6	59.6	47.5
	2B tok	B3 sequential	61.2	50.3	58.8	48.1
pol–eng	100M tok	B1 balanced	**55.0**	**47.1**	**54.3**	**54.6**
	100M tok	B2 simult.	52.2	43.9	53.1	45.1
	100M tok	B3 sequential	52.1	43.9	52.8	46.4
tur–eng	100M tok	B1 balanced	**57.5**	**50.2**	**55.7**	**50.6**
	100M tok	B2 simult.	53.2	44.9	53.3	39.8
	100M tok	B3 sequential	52.8	44.8	53.2	39.2
	100M tok	B4 classroom	56.3	48.3	55.6	43.0
	100M tok	B5 late	56.1	49.1	55.4	48.7
eng–deu	24B tok	B1 balanced	**54.6**	**46.3**	**52.9**	**51.8**
eng–nld	24B tok	B1 balanced	**56.8**	**50.2**	**54.7**	54.7
	24B tok	B2 simult.	56.6	48.6	54.1	54.7
	24B tok	B3 sequential	54.6	47.8	53.4	**55.8**

Table 5: BLiSS learner-minimal-pair scores (%, 0–100) for Beetle models on the *matched-L1 cohort*, grouped by L1 then token budget. Cell shading encodes score on a fixed scale (darker green = higher). Per L1, the highest value in each metric column is shown in **bold**. Metrics: RP@0 / RP@τ (random-over-human preference, with and without a margin), NGS (normalised gap score), CPS (correct-preferred sanity check).

			BLiSS (matched-L1 cohort)			
Model	**Params**	**L1**	**RP@0**	**RP@τ**	**NGS**	**CPS**
BLOOM-560M	0.56B	Dutch	59.0	52.7	55.4	55.8
		German	58.9	50.7	54.9	52.3
		Chinese	**70.2**	**62.1**	**65.0**	**56.1**
BLOOM-1.1B	1.1B	Dutch	59.7	51.6	55.8	**56.6**
		German	57.6	49.4	54.7	53.1
		Chinese	**70.3**	**63.6**	**65.7**	**56.6**
BLOOM-1.7B	1.7B	Dutch	59.0	51.3	55.9	56.6
		German	57.3	50.2	54.7	55.7
		Chinese	**70.5**	**63.7**	**65.7**	**58.4**
BLOOM-7.1B	7.1B	Dutch	63.1	57.2	58.3	**58.8**
		German	59.3	50.0	55.7	56.7
		Chinese	**73.2**	**65.6**	**67.4**	57.5
XGLM-4.5B	4.5B	Dutch	62.0	54.0	57.5	**63.2**
		German	59.3	50.6	55.3	58.1
		Chinese	**73.7**	**65.7**	**67.3**	59.8
EuroLLM-9B	9B	Dutch	64.4	56.3	57.5	**61.6**
		German	60.3	50.8	55.1	61.3
		Chinese	**73.7**	**65.4**	**67.5**	61.1
Gemma-3-270M	0.27B	Dutch	60.8	55.1	57.2	**56.9**
		German	57.2	51.9	55.3	53.6
		Chinese	**72.0**	**66.5**	**66.7**	56.1
Gemma-2-2B	2B	Dutch	64.7	57.4	58.1	**56.1**
		German	63.0	56.8	58.5	55.2
		Chinese	**75.9**	**71.3**	**69.6**	54.5
Gemma-2-9B	9B	Dutch	65.5	60.4	59.7	54.5
		German	66.0	60.4	60.0	**55.4**
		Chinese	**75.3**	**71.4**	**70.4**	55.2
Gemma-2-27B	27B	Dutch	69.0	61.3	63.0	**55.7**
		German	68.1	59.3	62.0	52.1
		Chinese	**75.2**	**70.0**	**70.3**	54.2
Qwen3-0.6B	0.6B	Dutch	59.5	54.1	55.7	60.3
		German	57.8	49.9	55.4	57.4
		Chinese	**70.9**	**62.2**	**65.4**	**60.9**
Qwen3-4B	4B	Dutch	63.2	55.7	58.9	**61.2**
		German	60.3	53.4	55.7	59.4
		Chinese	**72.8**	**66.1**	**67.3**	56.3
Qwen3-8B	8B	Dutch	65.4	57.0	59.5	**62.0**
		German	59.9	52.6	56.3	57.6
		Chinese	**75.2**	**67.7**	**68.9**	57.7
Llama-3.2-1B	1B	Dutch	63.5	54.9	57.9	**60.8**
		German	61.2	53.5	56.5	59.2
		Chinese	**73.9**	**66.1**	**67.7**	60.1
Llama-3.1-8B	8B	Dutch	67.0	58.7	60.7	**60.6**
		German	62.4	54.8	58.0	57.3
		Chinese	**75.8**	**68.7**	**69.2**	59.3

Table 6: BLiSS learner-minimal-pair scores (%, 0–100) for frontier multilingual-LLM baselines, scored per L1 cohort. *Params* gives the parameter count and *L1* the cohort. Cell shading encodes score on a fixed scale (darker green = higher). Per model, the highest value in each metric column is in **bold**. Metrics: RP@0 / RP@τ (random-over-human preference, with and without a margin), NGS (normalised gap score), CPS (correct-preferred sanity check).

Table 7: **Downstream pedagogical performance of Beetle models.** CLOTH closed-cloze accuracy, Cambridge open-cloze accuracy, JFLEG learner-fluency margin (bits/token; positive = model prefers corrected over learner-error sentence), false-friend disambiguation margin (bits/token), CEFR-A1 / CEFR-C2 BPT, and CEFR gradient. Bold = best per column within the matched-L1 nld–eng family.

Model	Family	CLOTH	Camb. OC	JFLEG	FF	A1 BPT	C2 BPT	A1→C2
B1 nld–eng	HS	**0.380**	**0.644**	**0.949**	**0.585**	6.572	8.818	2.247
B2 nld–eng	HS	0.302	0.596	0.750	0.156	8.051	10.213	2.163
B3-33 nld–eng	HS	0.310	0.587	0.751	0.174	8.106	10.275	2.169
B4 nld–eng	HS	0.304	0.582	0.801	−0.055	7.719	9.755	2.036
B5 nld–eng	HS	0.302	0.584	0.716	0.238	8.127	10.193	2.066
B1 nld–eng	FW-2B	0.369	0.638	0.833	0.457	8.851	8.614	−0.237
B2 nld–eng	FW-2B	0.322	0.567	0.502	0.238	10.317	11.047	0.730
B3-33 nld–eng	FW-2B	0.329	0.614	0.511	0.235	10.354	11.056	0.702
B5 nld–eng	FW-2B	0.323	0.570	0.509	0.318	10.364	11.040	0.676
mono-eng	HS	0.369	0.659	0.888	0.351	7.297	10.558	**3.261**
mono-nld	HS	0.267	0.310	0.306	−0.034	9.875	10.671	0.796

G Detailed Beetle BLiMP and MultiBLiMP results

Table 8 shows BLiMP and MultiBLiMP performance of Beetle models by L1 and scale (Table 8). Figure 8 and Figure 9 give the per-phenomenon BLiMP accuracy grids for the 24B and 2B Beetle models, respectively, and Table 9 reports frontier multilingual-LLM baselines for comparison.

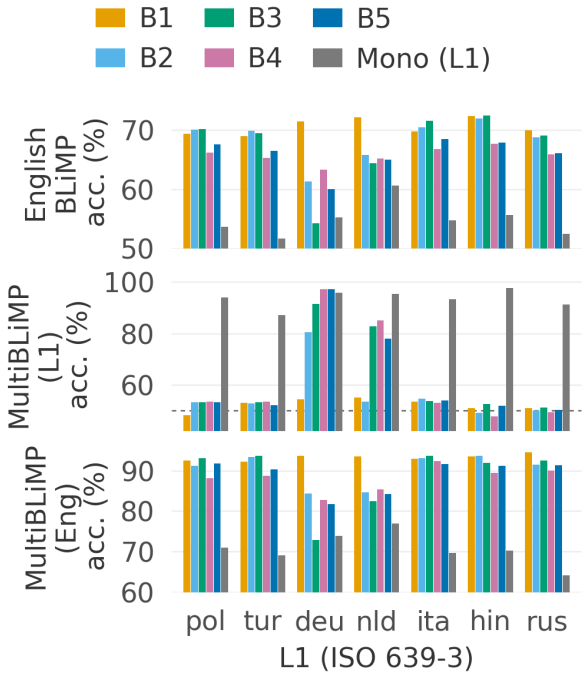

Figure 8: Per-phenomenon BLiMP accuracy for the 24B-token Beetle models, broken down by curriculum and L1. At this scale, differences between balanced and staged curricula are localised to specific phenomena rather than uniform across the benchmark, with the balanced and simultaneous curricula dominating on most syntactic and agreement phenomena.

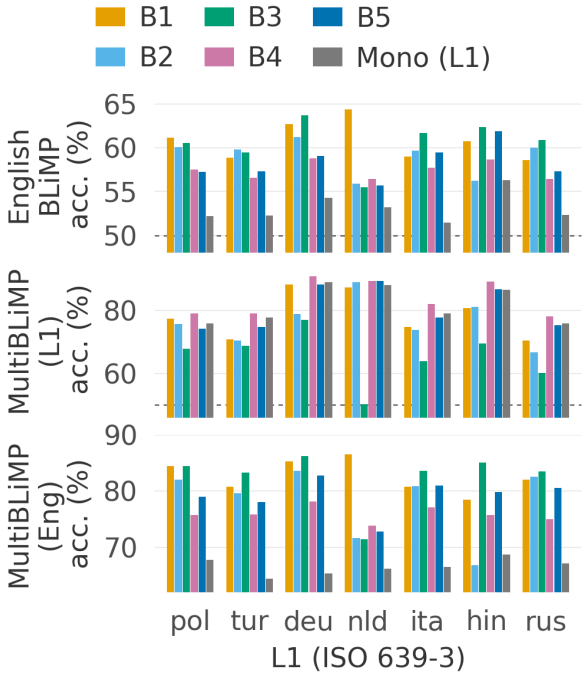

Figure 9: Per-phenomenon BLiMP accuracy for the 2B-token Beetle models, broken down by curriculum and L1. The overall curriculum ordering mirrors the 24B grid in Figure 8, but accuracies are lower and the per-phenomenon gaps between curricula are wider, indicating that the additional data at 24B mainly narrows the harder phenomena rather than reshuffling which curriculum wins.

Pair	Data	Curr.	MultiBLiMP L1	MultiBLiMP Eng	MonoBLiMP L1	MonoBLiMP Eng
deu-eng	Human-scale	Mono	89.9	64.7	-	51.2
	Human-scale	B1 balanced	86.9	80.5	-	65.5
	Human-scale	B2 simult.	89.5	70.9	-	56.1
	Human-scale	B3 sequential	89.9	71.3	-	57.4
	Human-scale	B4 classroom	89.7	72.2	-	57.0
	Human-scale	B5 late	89.6	71.2	-	55.8
	100M tok	B1 balanced	88.6	87.7	-	63.6
	100M tok	B2 simult.	92.1	71.3	-	52.8
	100M tok	B3 sequential	91.7	72.3	-	54.8
	100M tok	B4 classroom	87.6	75.6	-	57.1
	100M tok	B5 late	75.2	83.0	-	62.7
	2B tok	B1 balanced	88.2	85.3	-	62.7
	2B tok	B2 simult.	78.9	83.6	-	61.3
	2B tok	B3 sequential	76.9	86.2	-	63.7
	2B tok	B4 classroom	90.8	78.2	-	58.8
	2B tok	B5 late	88.3	82.7	-	59.1
	24B tok	B1 balanced	95.7	**94.0**	-	**71.5**
	24B tok	B2 simult.	**96.6**	84.4	-	61.3
eng	Human-scale	Mono	-	**79.0**	-	**64.7**
nld-eng	Human-scale	Mono	88.9	63.5	79.0	52.5
	Human-scale	B1 balanced	87.6	80.8	76.8	69.1
	Human-scale	B2 simult.	88.8	74.0	79.1	57.3
	Human-scale	B3 sequential	90.3	73.8	78.3	58.3
	Human-scale	B4 classroom	89.8	74.0	78.7	58.4
	Human-scale	B5 late	88.5	72.3	79.5	56.6
	100M tok	B1 balanced	88.4	88.1	74.5	63.9
	100M tok	B2 simult.	91.1	73.9	76.1	57.6
	100M tok	B3 sequential	92.7	73.9	76.5	57.4
	100M tok	B4 classroom	91.6	73.6	75.8	58.4
	100M tok	B5 late	77.3	84.9	63.2	62.8
	2B tok	B1 balanced	87.3	86.5	72.7	64.4
	2B tok	B2 simult.	88.9	71.7	76.8	55.9
	2B tok	B3 sequential	89.9	73.1	76.8	55.5
	2B tok	B4 classroom	89.3	73.9	77.2	56.4
	2B tok	B5 late	89.3	73.0	76.9	55.7
	24B tok	B1 balanced	94.4	**94.4**	78.3	**72.8**
	24B tok	B2 simult.	**96.0**	84.7	**81.6**	65.8
	24B tok	B3 sequential	95.5	82.6	81.1	64.4
zho-eng	Human-scale	Mono	-	61.7	74.6	51.5
	Human-scale	B1 balanced	-	76.6	70.4	66.0
	Human-scale	B2 simult.	-	68.8	69.8	54.2
	Human-scale	B3 sequential	-	70.3	70.1	54.7
	Human-scale	B4 classroom	-	71.4	69.1	55.7
	Human-scale	B5 late	-	69.7	71.4	54.9
	100M tok	B1 balanced	-	80.8	60.9	62.2
	100M tok	B2 simult.	-	84.0	61.1	63.2
	100M tok	B3 sequential	-	86.2	53.1	64.1
	100M tok	B4 classroom	-	69.4	62.6	57.4
	100M tok	B5 late	-	81.6	58.7	61.8
	2B tok	B1 balanced	-	86.9	63.5	63.5
	2B tok	B2 simult.	-	84.8	59.8	61.4
	2B tok	B3 sequential	-	84.0	56.4	63.0
	2B tok	B4 classroom	-	73.4	63.7	58.2
	2B tok	B5 late	-	79.2	59.4	60.8
	24B tok (FW-24B)	B1 balanced	-	91.7	**75.9**	70.9
	24B tok (FW-24B)	B2 simult.	-	92.9	73.7	71.6
	24B tok (FW-24B)	B3 sequential	-	**93.9**	55.9	**72.7**
	24B tok (FW-24B)	B3 sequential *ewc*	-	93.2	57.5	70.4
	24B tok (FW-24B)	B4 classroom	-	89.6	76.2	67.3
	24B tok (FW-24B)	B5 late	-	89.5	67.5	69.1

Pair	Data	Curr.	MultiBLiMP L1	MultiBLiMP Eng	MonoBLiMP L1	MonoBLiMP Eng
eus-eng	100M tok	B1 balanced	**97.4**	**85.5**	-	**62.1**
	100M tok	B2 simult.	96.7	69.9	-	52.6
	100M tok	B3 sequential	**97.4**	70.3	-	54.0
	100M tok	B4 classroom	96.0	75.8	-	56.0
	100M tok	B5 late	93.0	81.3	-	60.7
	2B tok	B1 balanced	95.2	82.2	-	61.3
	2B tok	B2 simult.	96.7	70.9	-	51.8
	2B tok	B3 sequential	96.7	67.8	-	53.0
	2B tok	B4 classroom	96.7	71.3	-	53.4
	2B tok	B5 late	96.3	70.1	-	52.3
fil-eng	100M tok	B1 balanced	-	**86.1**	-	**65.0**
	100M tok	B2 simult.	-	75.3	-	59.0
	100M tok	B3 sequential	-	75.8	-	58.6
hin-eng	100M tok	B1 balanced	87.6	**86.4**	-	**63.8**
	100M tok	B2 simult.	91.6	67.1	-	54.6
	100M tok	B3 sequential	**91.7**	68.4	-	55.0
	100M tok	B4 classroom	87.6	75.8	-	57.3
	100M tok	B5 late	80.6	79.9	-	62.1
	2B tok	B2 simult.	91.6	66.9	-	56.2
ita-eng	100M tok	B1 balanced	81.7	**87.8**	-	**62.7**
	100M tok	B2 simult.	**84.8**	70.3	-	53.7
	100M tok	B3 sequential	**84.8**	72.6	-	54.3
	100M tok	B4 classroom	79.5	73.5	-	56.1
	100M tok	B5 late	66.8	82.6	-	61.6
pol-eng	100M tok	B1 balanced	77.8	**86.6**	-	**62.2**
	100M tok	B2 simult.	**82.5**	70.9	-	53.4
	100M tok	B3 sequential	81.3	69.1	-	54.8
rus-eng	100M tok	B1 balanced	77.5	**85.1**	-	**60.9**
	100M tok	B2 simult.	81.3	68.6	-	53.7
	100M tok	B3 sequential	**82.1**	70.1	-	53.4
	100M tok	B4 classroom	77.7	76.8	-	55.9
	100M tok	B5 late	66.2	81.8	-	60.0
tur-eng	100M tok	B1 balanced	76.6	**84.4**	74.6	**61.1**
	100M tok	B2 simult.	**81.2**	71.0	**81.5**	53.3
	100M tok	B3 sequential	80.6	69.7	80.8	53.3
	100M tok	B4 classroom	72.5	73.6	73.1	54.2
	100M tok	B5 late	72.2	81.3	61.5	61.0
eng-deu	24B tok	B1 balanced	**93.6**	**93.6**	-	**71.6**
eng-nld	24B tok	B1 balanced	93.5	93.5	-	72.2
	24B tok	B2 simult.	**95.5**	**95.5**	-	**74.2**
	24B tok	B3 sequential	94.0	94.0	-	74.1

Table 8: Accuracy of Beetle models across curricula and L1s and data exposure (100, 2B, 24B) on L1 and L2 MultiBLiMP (Jumelet et al., 2025b) and Monolingual BLiMPs (`English`: BLiMP (Warstadt et al., 2020)); `Dutch`: BLiMP-NL (Suijkerbuijk et al., 2025a); `Chinese`: ZhoBLiMP (Liu et al., 2025)

			MultiBLiMP		MonoBLiMP	
Model	**Par.**	**Cohort**	**L1**	**Eng**	**L1**	**Eng**
BLOOM-560M	0.56B	Dutch	59.8	81.7	52.6	66.7
		German	65.9	81.7		66.7
		Chinese		81.7	72.5	66.7
BLOOM-1.1B	1.1B	Dutch	63.0	92.5	48.7	76.7
		German	76.5	92.5		76.7
		Chinese		92.5	84.1	76.7
BLOOM-1.7B	1.7B	Dutch	66.4	93.0	48.7	76.4
		German	75.2	93.0		76.4
		Chinese		93.0	84.8	76.4
BLOOM-3B	3B	Dutch	66.3	94.8	51.1	74.2
		German	80.7	94.8		74.2
		Chinese		94.8	84.0	74.2
BLOOM-7.1B	7.1B	Dutch	64.0	94.0	55.5	77.5
		German	82.2	94.0		77.5
		Chinese		94.0	84.3	77.5
XGLM-4.5B	4.5B	Dutch	96.9	98.7	83.6	80.9
		German	98.8	98.7		80.9
		Chinese		98.7	84.6	80.9
EuroLLM-9B	9B	Dutch	76.4	84.4	88.0	79.6
		German	77.3	84.4		79.6
		Chinese		84.4	83.5	79.6
Gemma-3-270M	0.27B	Dutch	83.7	97.0	69.9	78.8
		German	90.4	97.0		78.8
		Chinese		97.0	80.7	78.8
Gemma-2-2B	2B	Dutch	91.7	98.7	80.5	76.5
		German	95.6	98.7		76.5
		Chinese		98.7	82.3	76.5
Gemma-2-9B	9B	Dutch	91.2	98.8	84.4	77.3
		German	96.4	98.8		77.3
		Chinese		98.8	82.9	77.3
Gemma-2-27B	27B	Dutch	78.6	82.6	74.0	70.7
		German	81.1	82.6		70.7
		Chinese		82.6	71.5	70.7
Qwen3-0.6B	0.6B	Dutch	81.5	97.0	71.7	81.6
		German	92.6	97.0		81.6
		Chinese		97.0	82.0	81.6
Qwen3-4B	4B	Dutch	93.9	99.5	82.6	82.8
		German	97.2	99.5		82.8
		Chinese		99.5	85.9	82.8
Qwen3-8B	8B	Dutch	94.6	98.7	82.1	82.5
		German	98.5	98.7		82.5
		Chinese		98.7	85.9	82.5
Llama-3.2-1B	1B	Dutch	90.8	99.0	76.0	81.8
		German	96.1	99.0		81.8
		Chinese		99.0	82.9	81.8
Llama-3.1-8B	8B	Dutch	96.1	99.2	85.1	81.2
		German	98.0	99.2		81.2
		Chinese		99.2	84.5	81.2

Table 9: Frontier multilingual-LLM baselines. *Par.* is the parameter count; *Cohort* is the evaluation language.