EMNLP2026 BUDAPEST

UNIVERSITY OF CAMBRIDGE

BEETLE

Open Bilingual Model Suite for Modelling Second-Language Processing

Building Open "Sovereign" AI Solutions for Equitable Assessment and Simulation

Suchir Salhan
ALTA CST | NLIP Group | Department of Computer Science & Technology
sas245@cam.ac.uk

ROADMAP

Motivation

Why Bilingual Models?

The BEETLE Training Framework

Controlled Bilingual Model Training

Models & Results

Learning Dynamics, Reading Time & Bilingual Processing Effects

Applications

Open “Sovereign” AI Solutions for Equitable Assessment and Simulation

Motivation

Why Bilingual Models?

Educational Machine Learning

The goal of Educational Machine Learning is to develop techniques to use AI to enhance learning environments, personalise education and support teachers.

The ALTA Institute focuses on educational machine learning to support (second) language learning contexts (adaptive learning, responsible and explainable AI) .

Several of our research projects focus on Large Language Models (LLMs) to simulate learners (e.g., learner simulations)

Off-the-Shelf or Train-from-Scratch?

There is a practical demand for student models of bilingual and second language learners across different ages and first language backgrounds (e.g., CUP&A's North Star 2025 → equitable assessment solutions for diverse language learners!).

There are many routes to pedagogical simulation for different purposes!

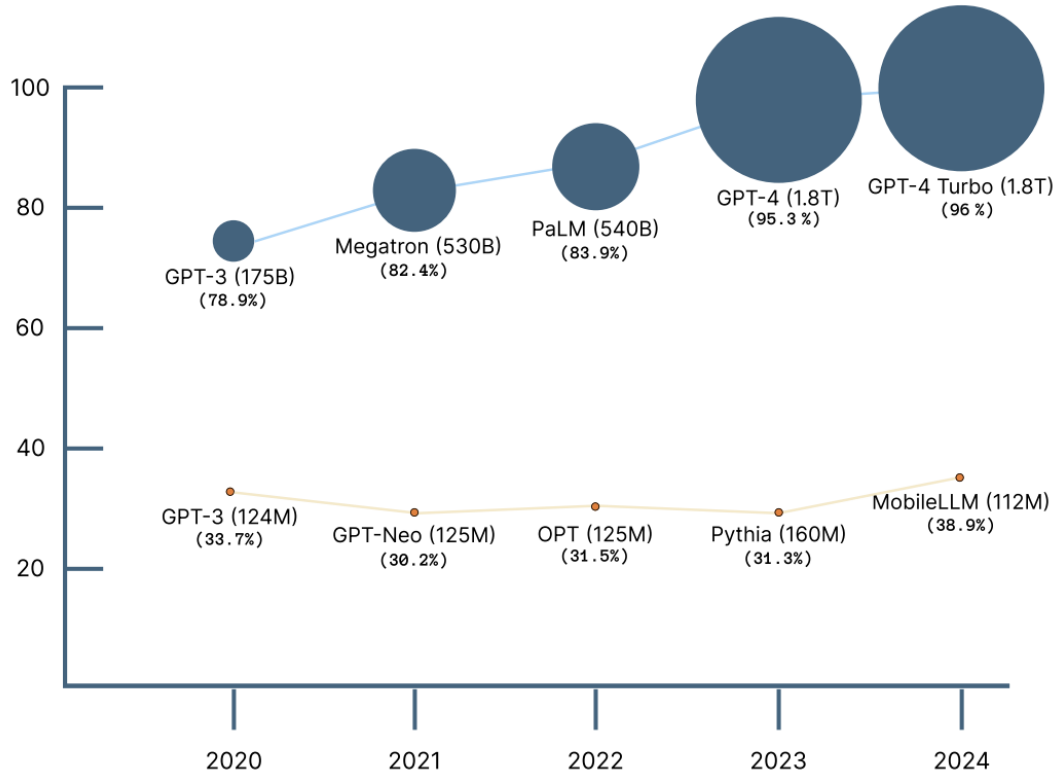
- Why not prompt LLMs?
- Why not use learner corpora?

Goal 1: Sovereign AI & Open-Source Ecosystems

Control and Trust in AI systems is important, especially in high-stakes applications.

- Sovereign AI
- Open Data
- **Open Checkpoints** → learning dynamics (how do models learn? What order do they acquire capabilities?)

Goal 1: Sovereign AI & Open-Source Ecosystems



Diehl Martinez et al (2026)
Pico: A Modular Framework for Hypothesis-Driven Small Language Model Research - ACL Anthology

Goal 2: Sample Efficiency

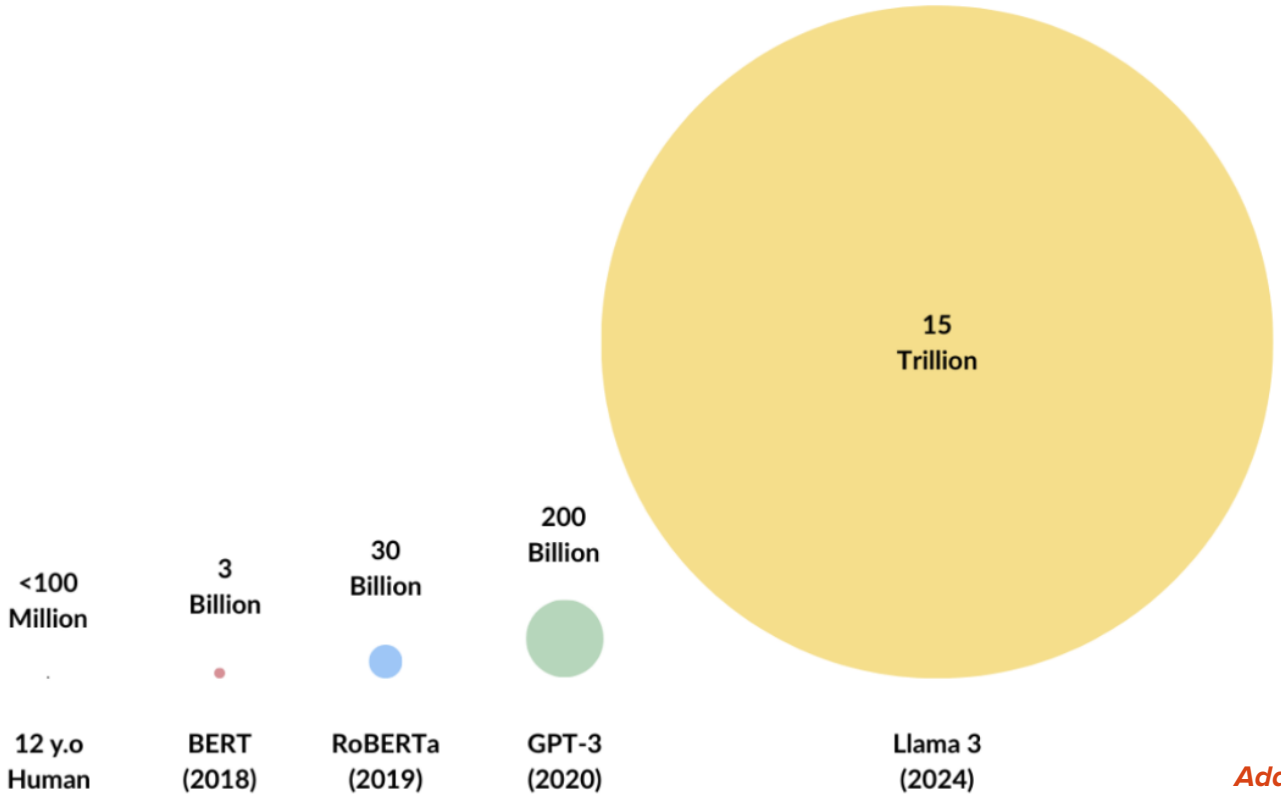
AI systems use 1000xs more data to learn human-like capabilities; do we really need to be prompting closed LLMs for learner-like simulations?

Goal 2 → Small Models & BabyLMs

Small Models = fewer "parameters" (e.g., < 8B)

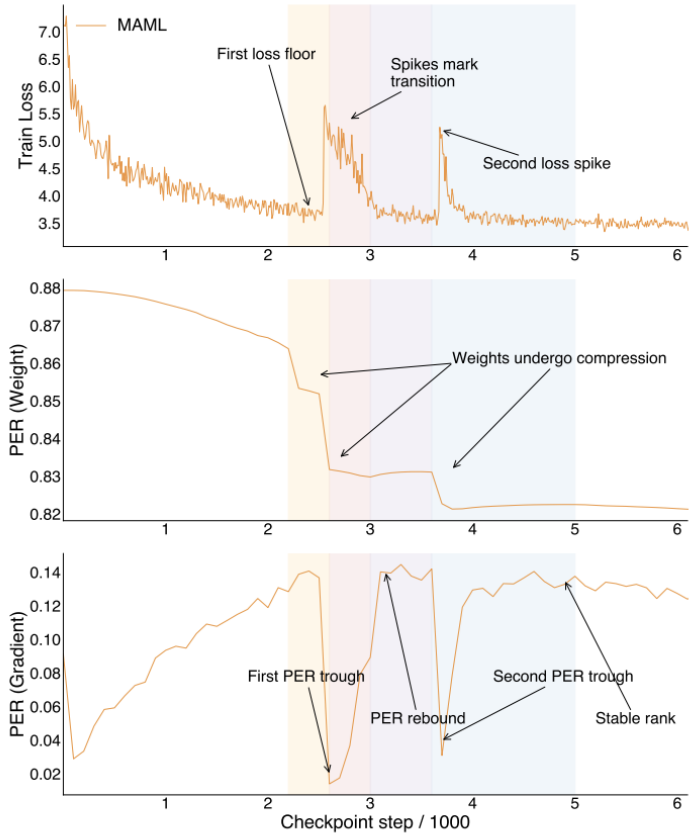
Small Models + Human-Like Data Scales = BabyLMs

Goal 2 → BabyLMs



Adapted from Goriely (2026)

Goal 2 → Small Models with Checkpoints



Diehl Martinez et al (2026)
Pico: A Modular Framework for Hypothesis-Driven Small Language Model Research - ACL Anthology

Introducing BEETLE

Human Language Processing. Most people are multilingual, but most studies using language models to understand human language processing are overwhelmingly monolingual (English!).

Multilinguality & Interpretability (Open-Source AI Research). How does cross-lingual transfer occur? How much does it vary by language pair? To what extent does training curriculum matter?

ROADMAP

The BEETLE Training Framework



Controlled Bilingual Model Training



BEETLE



THE BEETLE MODEL ZOO

21 languages

× monolingual & bilingual

× 5 curricula

× 3 data scales

× 2 data types

one factor varied at a time.

beetlelm.github.io

beetlelm.github.io

Bilingual Language Models

- Models trained on just two languages.
- Bilingual Model Learning Dynamics can be used to study cross-lingual transfer between different languages.
- Sometimes with an explicit first language (L1), with a second language (L2) introduced later in training

Bilingual Language Models

- Often trained in a way that is inspired by bilingual people (known as L2LMs), useful for computational psycholinguistics.
  - Cf. 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)
- Bilingual model training methods differ!
  - e.g., when L2 is introduced in training, ratio of L1:L2 in different training phases and whether methods handle catastrophic forgetting

The BEETLE Training Framework

- A modular language model training & analysis framework
- Systematically manipulate:
  - Training data curriculum
  - Architecture
  - Tokenizer
- Rich checkpointing for studying learning dynamics and catastrophic forgetting

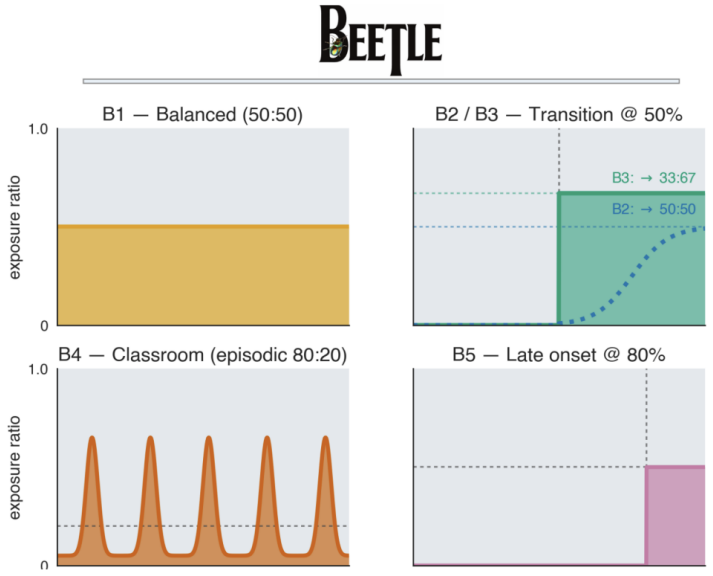
Model Suite

- 285 bilingual language models
  - 21 unique L1s:
    i. German, Chinese, Dutch, Italian, Hindi, Russian, Spanish, Greek, Basque, Tagalog, Polish, Turkish, Japanese, Korean, Finnish, Estonian, Icelandic, Danish, Hebrew, Norwegian, Serbian
  - L2 is always English
    i. Due to availability of downstream evaluation tasks.
- 3 data sources and sizes:
  - FineWeb-2 (100M, 2B, 24B tokens) (Penedo et al 2025)
  - BabyBabelLM (100M tokens) (Jumelet et al 2026)

Model Suite

5 acquisition-inspired data curricula:

- Balanced (B1): 50/50 mix of L1 and L2
- Simultaneous (B2): L1 only for first half, then 50/50 mix
- Sequential (B3): L1 only for first half, L2 only for second half
- Classroom (B4): episodic exposure, 80/20 mix
- Late (B5): L2 onsets 80% through training

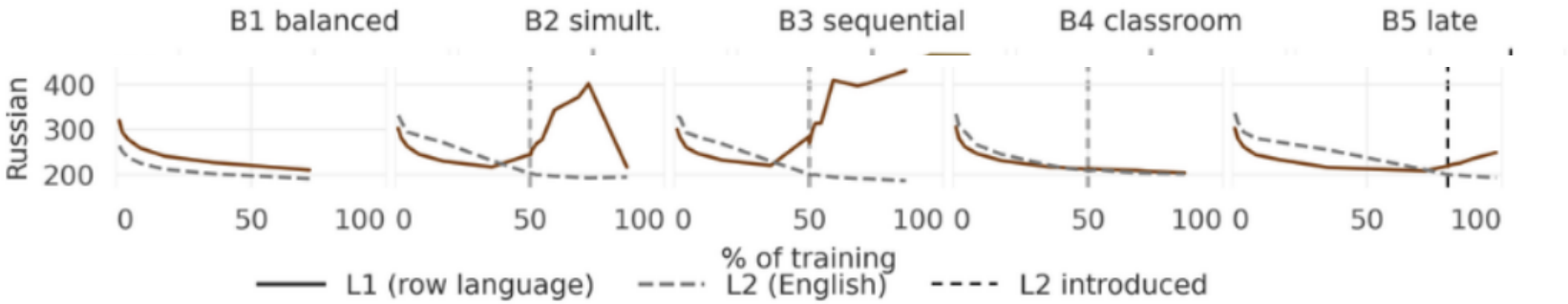


Models & Results



Learning Dynamics, Reading Time & Bilingual Processing Effects

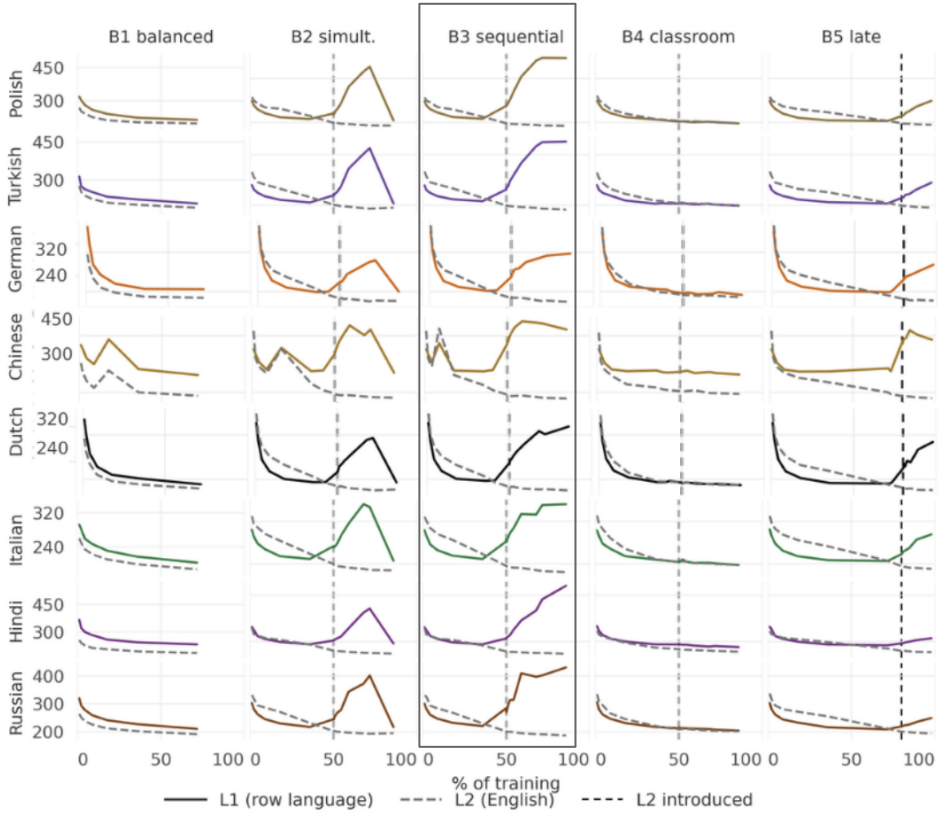
Learning Dynamics



Do bilingual exposure curricula change training consistent with expected patterns?
FLORES-200 is being used to measure how well each Beetle model predicts natural sentences in its first language (L1) and second language (L2).
lower values mean the model assigns higher probability to the observed text

Learning Dynamics

English NLL



- Models with continued/late L1 exposure do not exhibit catastrophic forgetting **(B1, B4, B5)**
- When the model stops seeing L1 it forgets L1 (B3)

L2 Reading Time: Method

- **Data.** First-pass gaze duration from the **Multilingual Eye Tracking Corpus (MECO)**: reading time data from second-language English speakers
- **Models.** Beetle models and 5 massively multilingual baseline models (Apertus, Gemma, Llama, Qwen, XGLM)

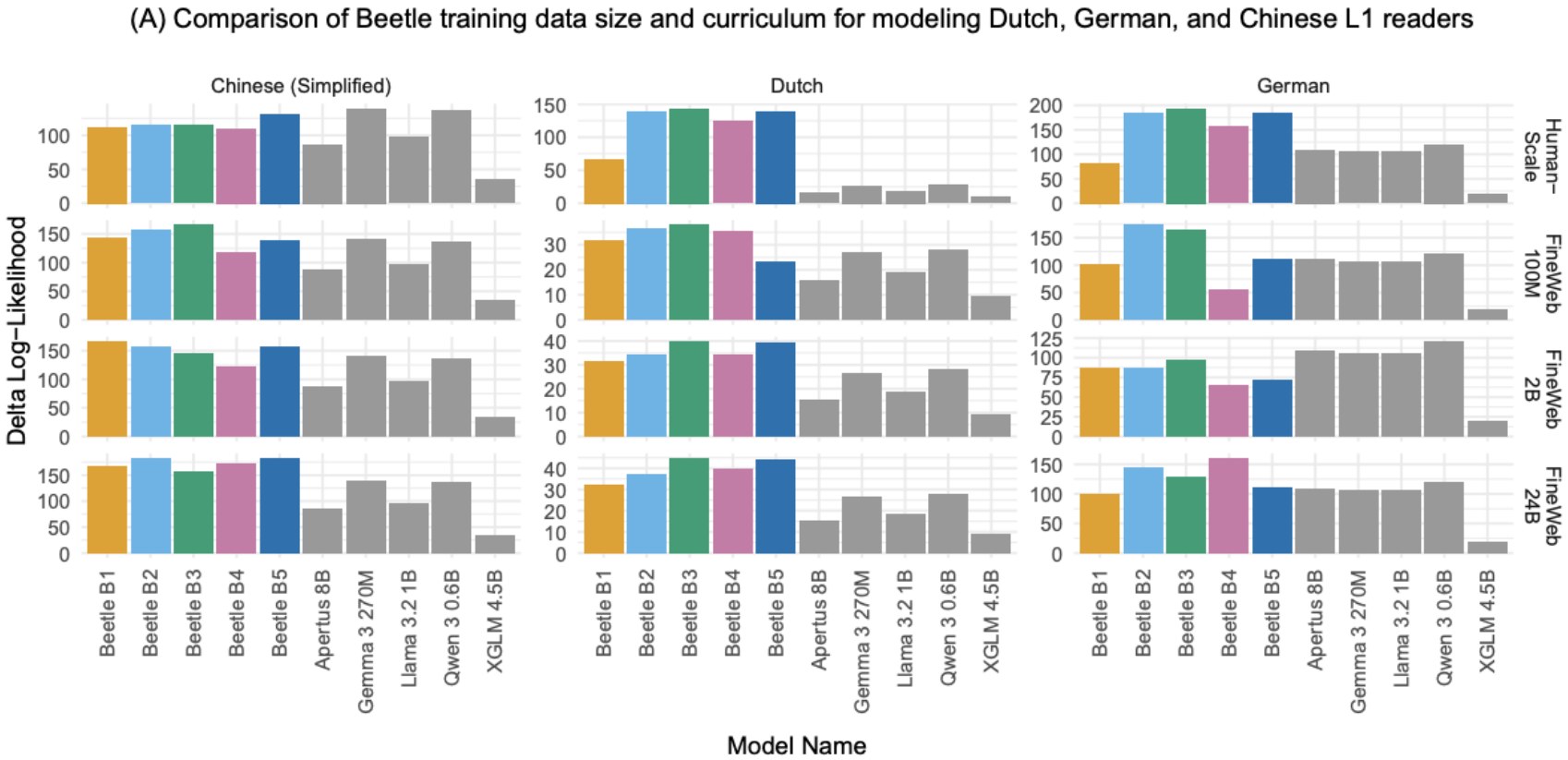
L2 Reading Time: Method

- **Modeling.** We calculate word surprisal from each Beetle model and use it to predict human times, controlling for word length, frequency, and position in the text using linear mixed-effects regression models
- **Comparison.** We compare the fit of the regressions with each language model's surprisal as a predictor.

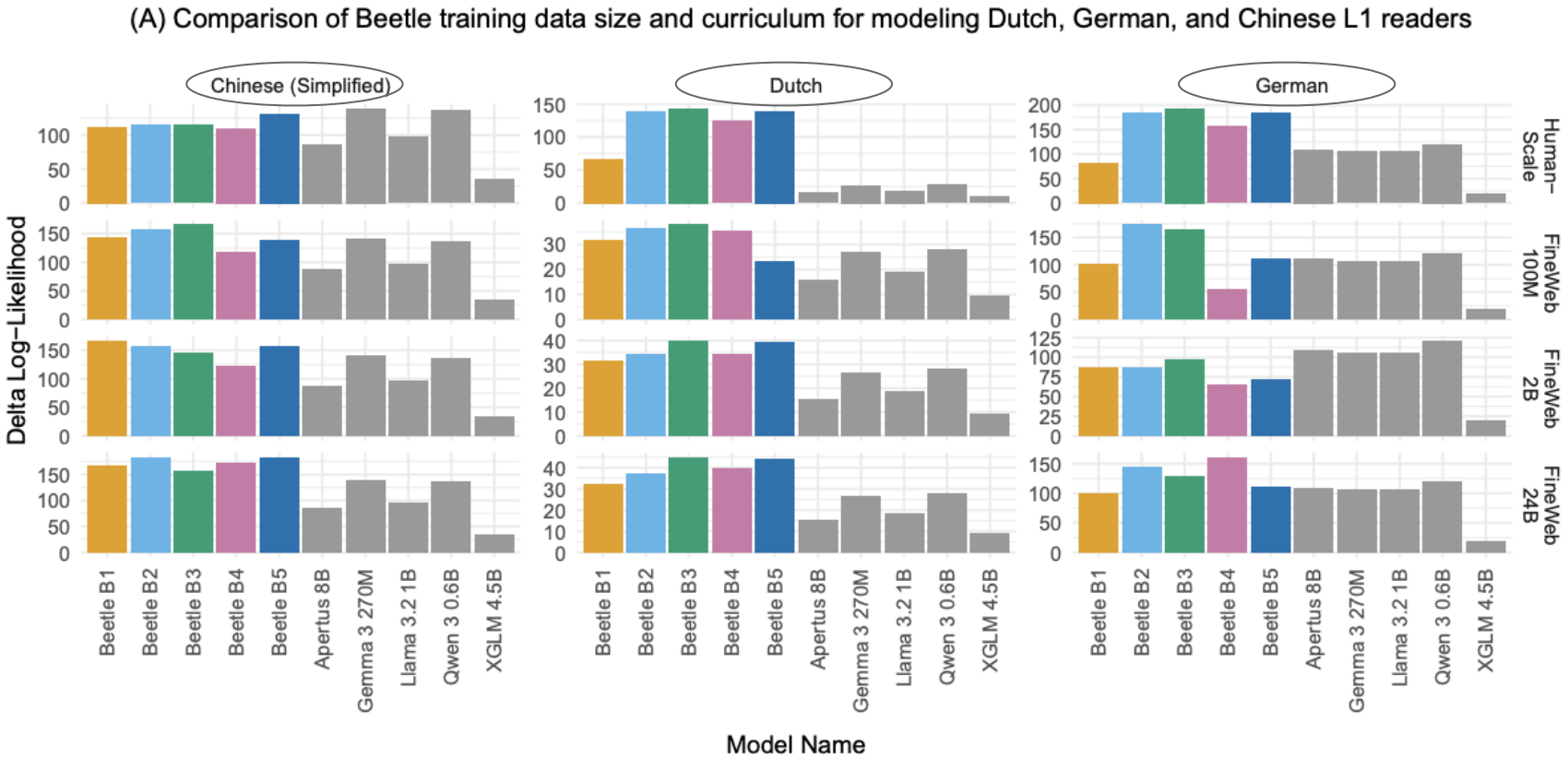
L2 Reading Time: Method

- The contribution of model surprisal is measured by the improvement in regression log-likelihood over a baseline without surprisal (Δ logL).
- **A larger Δ logL indicates closer alignment between a model's predictions and human reading behaviour, beyond what can be explained by basic word-level properties.**

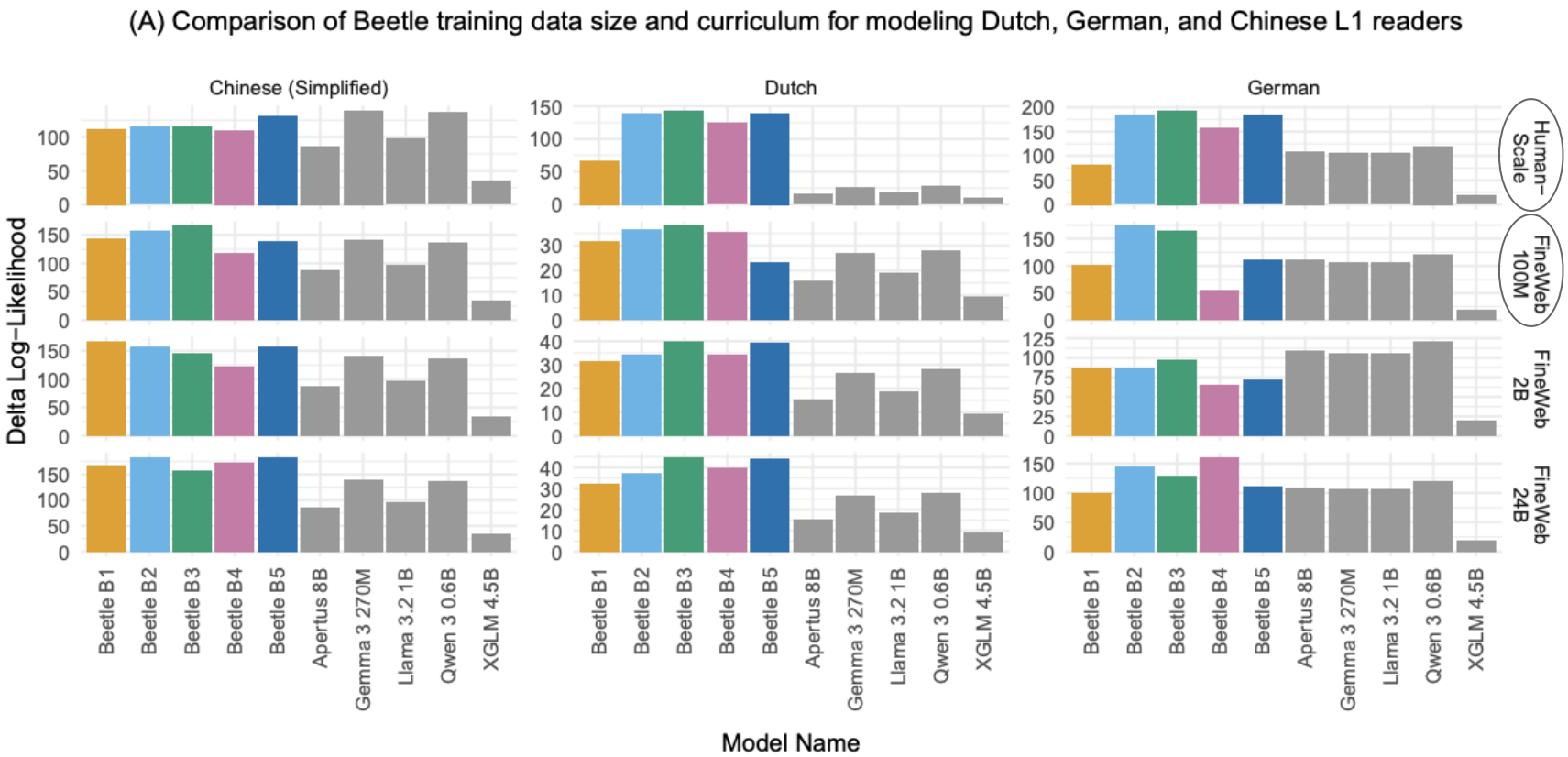
Staged exposure (B2 - B5) generally produces better alignment with human second-language reading times than balanced bilingual training (B1)



L2 Reading Time Detailed Results



L2 Reading Time Detailed Results



Bilingual Processing Tasks

- **Structural Priming:** structural priming is measured as the increase in the probability of producing a target sentence after a matching (congruent) prime compared with a mismatching (incongruent) prime (Arnett et al., 2025)
- **L2 Discrimination and CEFR Simulation:** We use BLiSS (Gao et al., 2025), a benchmark which tests whether models distinguish have learner-like grammatical discrimination between a corrected sentence, a human learner error (from L2 Learner Corpora), and an artificial error.

Bilingual Processing Results

Structural Priming

We replicate a Dutch–English genitive priming evaluation and find that Dutch-English Beetle models exhibit cross-lingual priming across exposure curricula, including models trained with just 100 million tokens.

Difference in surprisal (nats) between congruent and incongruent primes.

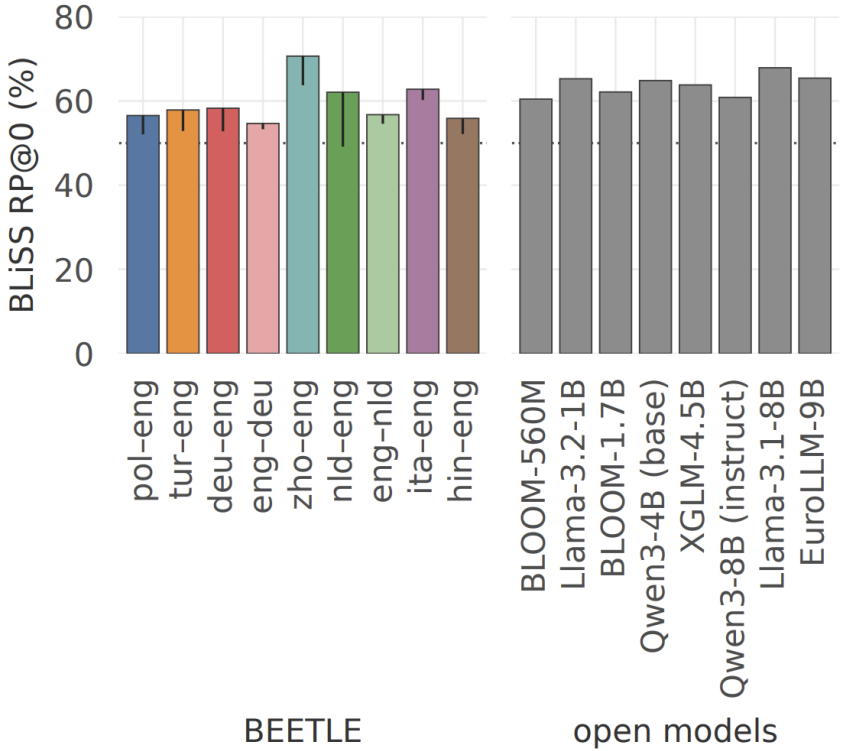
+ve = the model finds the target sentence less surprising after a matching prime → evidence of structural priming..

	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

Bilingual Processing Results

L2 Discrimination and CEFR Simulation

Beetle has comparable performance to multilingual models.





Future Directions

Use Beetle + How to Collaborate!

Use Beetle Models Off the Shelf



Beetle models trained with acquisition-inspired curricula are not just better on psycholinguistic benchmarks … they also produce error profiles closer to real learner populations.

This opens a direct path from more academic evaluations (BLiSS, MECO) to practical pedagogical tools!

Use Beetle Models Off the Shelf

1. Head over to The Beetle Website https://beetlelm.github.io/ 
2. Try out my Tutorial! https://github.com/beetlelm/beetle-explorer 

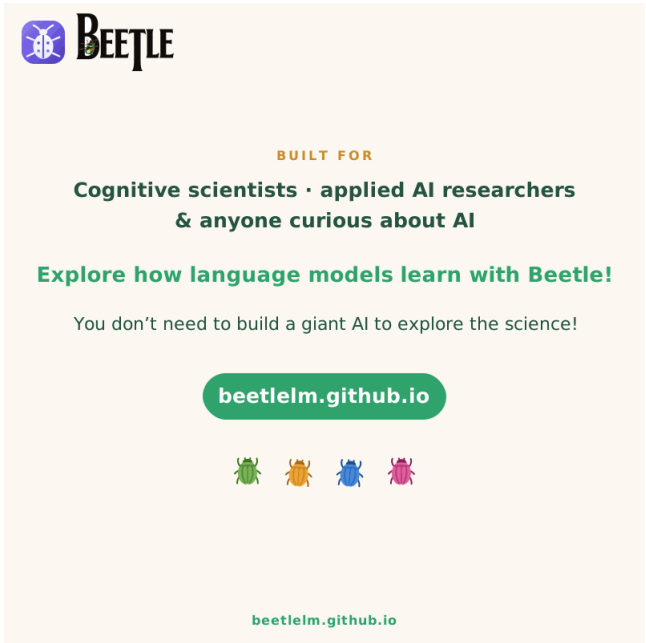
Use Beetle Models Off the Shelf

1. Head over to The Beetle Website https://beetlelm.github.io/
2. Try out my Tutorial! https://github.com/beetlelm/beetle-explorer



Use Beetle Models Off the Shelf

1. Head over to The Beetle Website https://beetlelm.github.io/
2. Try out my Tutorial! https://github.com/beetlelm/beetle-explorer



Train your own 🪲Beetle🪲 Model?

1. Head over to The Beetle Website
2. Check out three repositories!

beetlelm · Train

Train language models with different architectures, tokenisers and language-exposure strategies. Compare five curricula (B1–B5) to investigate how the order and timing of language learning affect the result. GitHub · Guide

beetle-data · Prepare

Turn raw text into training data. Build corpora, train tokenisers, remove benchmark overlap and prepare data for model training, using the same pipeline as the released models. GitHub · Guide

beetle-analyze · Evaluate

Measure what models learn. Evaluate grammatical knowledge, multilingual performance and alignment with human reading times, or investigate how internal representations change during training. GitHub · Guide

More Practical Questions!

Learner-Like Text Generation (Barbenel et al 2026)

- Native Language Identification (NLI)
- NLI of L2LMs using various **instruction-tuning and prompting conditions**

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.

More Practical Questions!

Learner-Like Text Generation (Barbenel et al 2026)

- Instruction Tuning of Beetle Models
  - Three Conditions
    - (1) English-only
      - 30,000 generic QA pairs from Alpaca English
    - (2) English and L1 (1:1)
      - split equally between Alpaca English and its translation for each L1
    - **(3) L1-influenced L2 English → ONLY condition that leads to more L2 learner like text generation**

Seven L1-specific L2 English datasets from the EFCAMDAT (Geertzen et al., 2013) and Write & Improve (Nicholls et al., 2024) learner corpora, each containing 10,000 prompt-essay pairs, except Polish (n=370) and Turkish (n=4,847) due to limited corpus data

More Practical Questions!

Learner-Like Text Generation (Barbenel et al 2026)

- A novel dataset of learner English essays by collecting responses to nine prompts from seven L1 backgrounds across three proficiency levels, controlling for text length.
  - 174 participants (M age=30; 42.5% female)
- This design enabled direct comparison with model-generated text by matching prompts between human participants and LMs (RQ2). For each proficiency level and L1, at least six participants were recruited. In total, the dataset comprised 466 essays

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.

More Practical Questions!

Learner-Like Text Generation (Barbenel et al 2026)

- GPT4o-as-Judge for NLI; more fine-grained analysis possible

Low-Level Learner Simulation using LLMs (Goulder et al f.c.)

LLMs generally struggle to generate genuine low-level learner simulation. Coupled with a general sparsity of low-level learner corpora (esp when split by L1), this remains a promising open research direction!

Other Dimensions of Cognitive Simulation

Learner Behaviour in Tasks

Linguistic Capabilities and Generation is only a subset of a much wider range of cognitive behaviours of language learners. *E.g., Simulating Young Learners? Motivation?*

Other Dimensions of Cognitive Simulation

Human-like learning also does not happen in a “vacuum” of text only supervision

The Transformer is not by default the most sample-efficient architecture

Open SLMs offer alternative routes to simulation

- Small Language Models as Judges for Assessment (Salhan, Zou & Buttery 2026/f.c.) - stay tuned!
- **Multi-Agent Systems for Calibration using Beetle + Open SLMs**

Takeaways

- **Language Exposure Matters**
- **Cognitive modelling needs more diverse models and more diverse data**
- **We need more multilingual models with open checkpoints!**

Acknowledgements

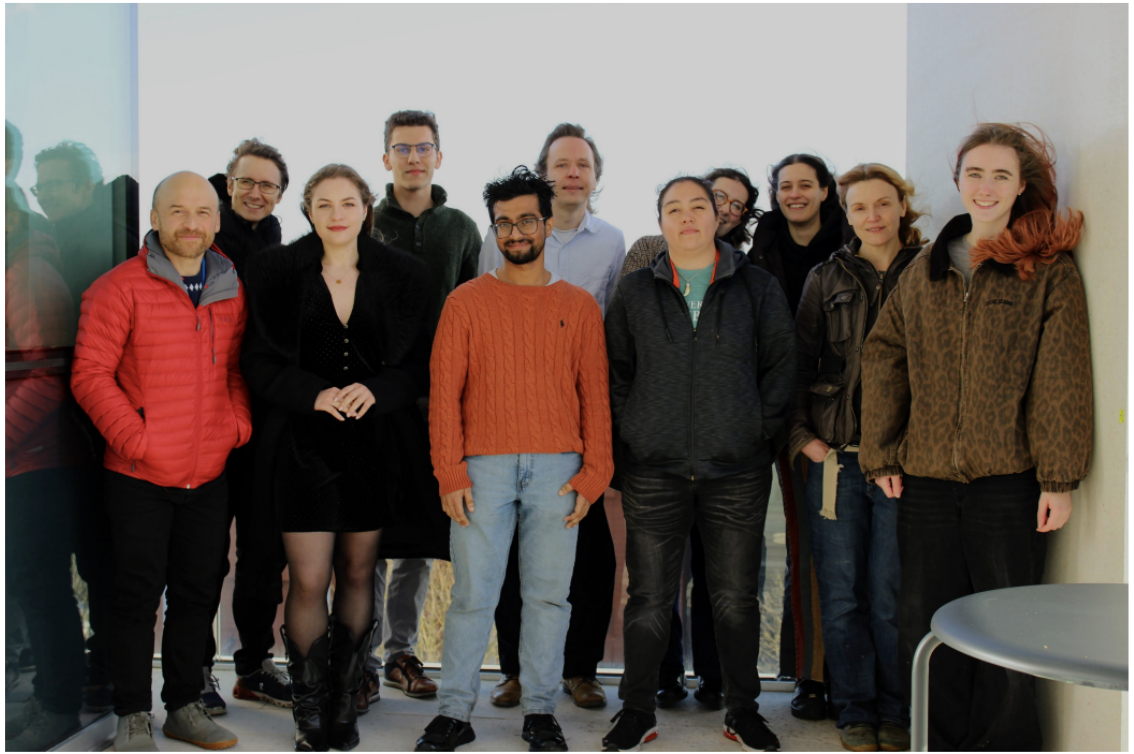
My work has been supported by Cambridge University Press & Assessment.

Beetle Training has been supported by EleutherAI, a non-profit supporting open-source AI.

Thank you to my wonderful collaborators and mentors in the Natural Language & Information Processing (NLIP) Group and ALTA CST!

In particular, thanks to my supervisor, Prof Paula Buttery. Thanks to my research mentors, Catherine Arnett (EleutherAI), Prof Andrew Caines, Richard Diehl Martinez and Zeb Goriely. Thanks to various people in The Buttery Group: Dr Diana Galvan Sosa, Gabrielle Gaudeau, Laura Barbenel, Aoife O'Driscoll, Lily Goulder, Yuan Gao, Filip Trhlik, Bianca Ganescu, Andrzej Szablewski; Donya Rooein (MilaNLP, Bocconi University); and the BabyLM 2026 Co-organisers!

Acknowledgements



EMNLP2026 BUDAPEST

UNIVERSITY OF CAMBRIDGE

Thank you!



Suchir Salhan Catherine Arnett James A. Michaelov Paula Buttery

beetlelm.github.io

EMNLP2026 BUDAPEST

UNIVERSITY OF CAMBRIDGE

BEETLE



beetlelm.github.io

beetlelm.github.io