EMNLP2026 BUDAPEST

UNIVERSITY OF CAMBRIDGE

BEETLE

A Bilingual Model Suite for Modelling Second-Language Processing

Suchir Salhan  Catherine Arnett  James A. Michaelov  Paula Buttery 

ROADMAP

Motivation

Why Bilingual Models?

The BEETLE Training Framework

Controlled Bilingual Model Training

Models & Results

Learning Dynamics, Reading Time & Bilingual Processing Effects

Motivation

Why Bilingual Models?

Motivation

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

Multilinguality & Interpretability. 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

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

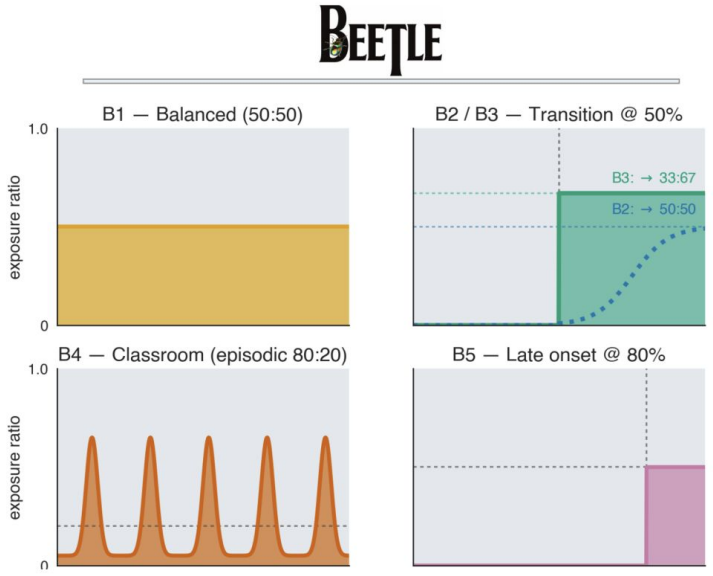
Model Suite

- 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

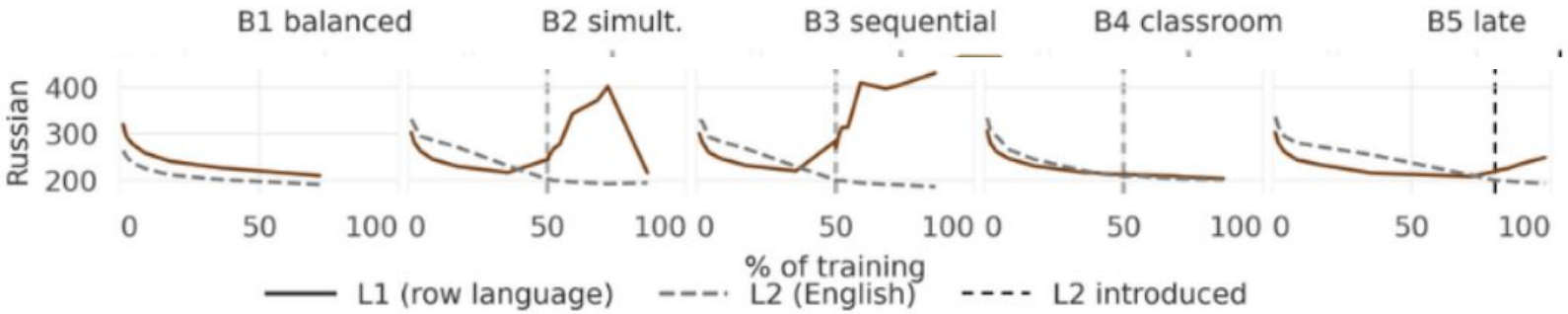




Models & Results

Learning Dynamics, Reading Time & Bilingual Processing Effects

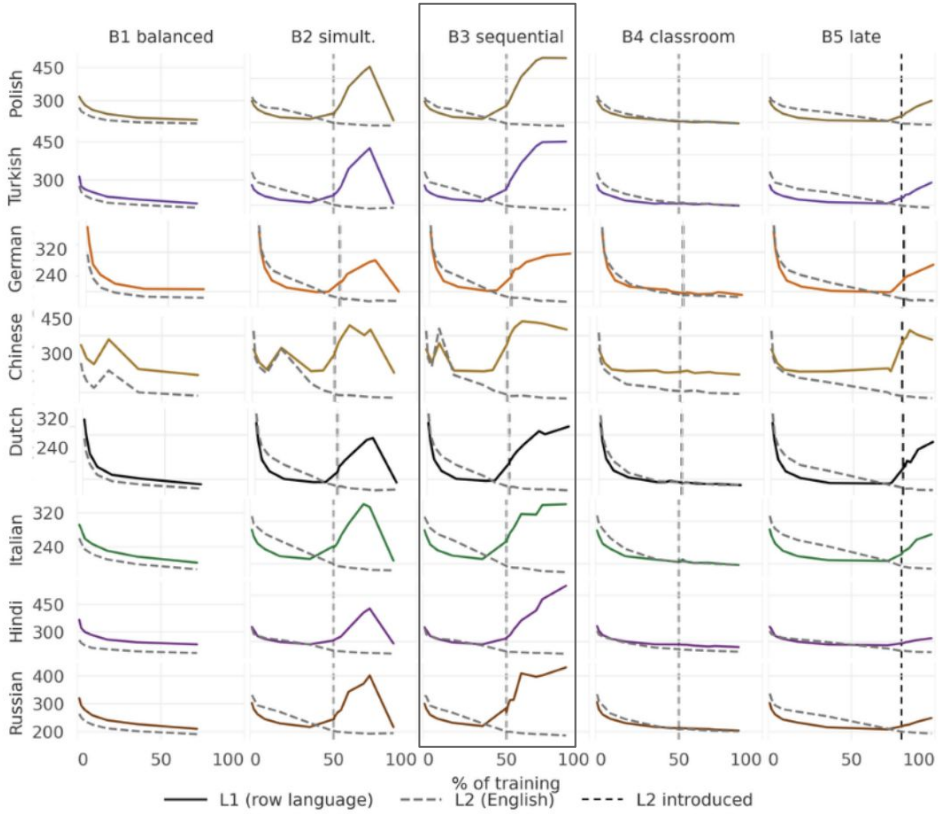
Learning Dynamics



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)

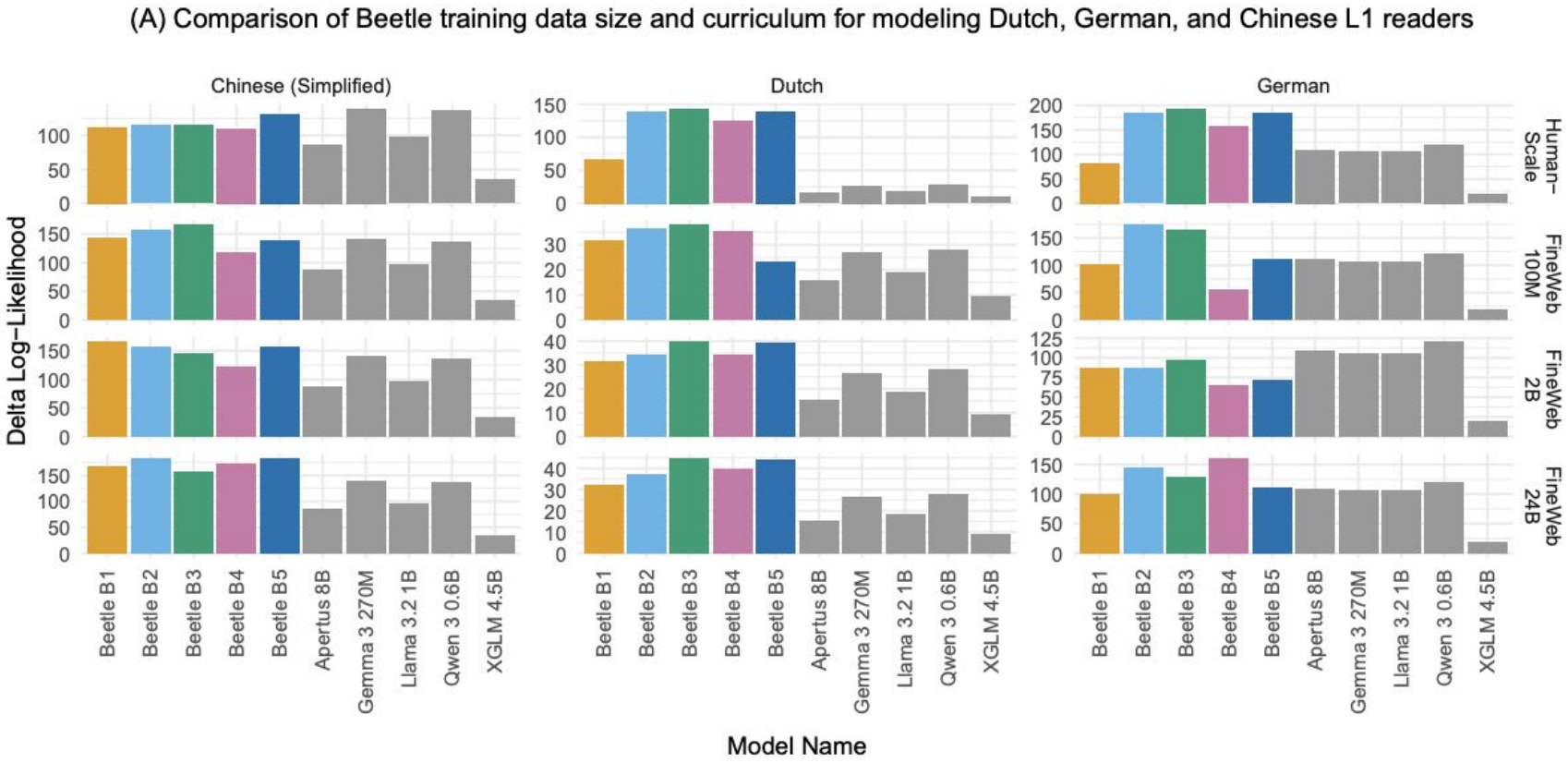
L2 Reading Time: Method

- **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.

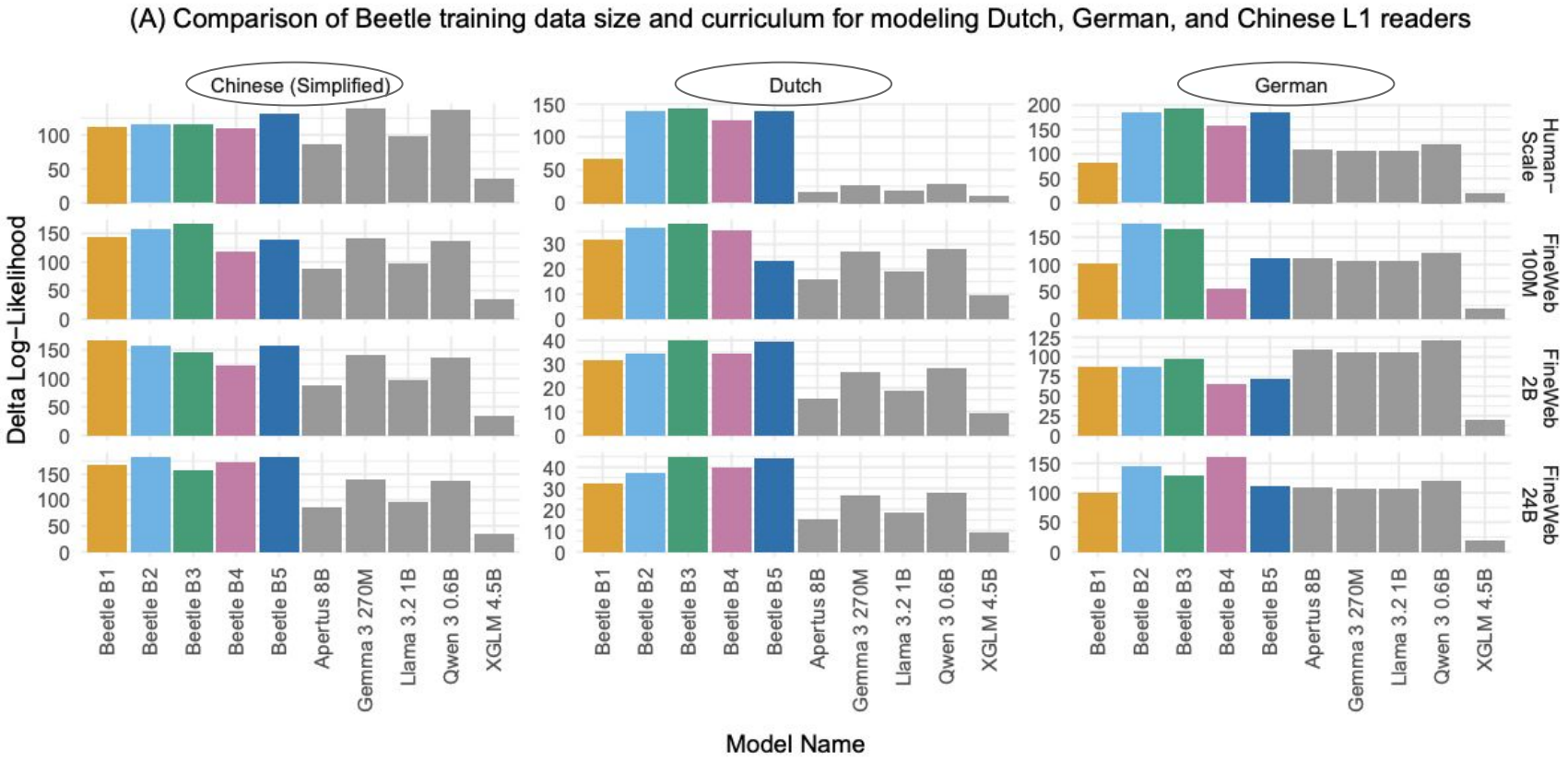
L2 Reading Time: Method

- 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.**

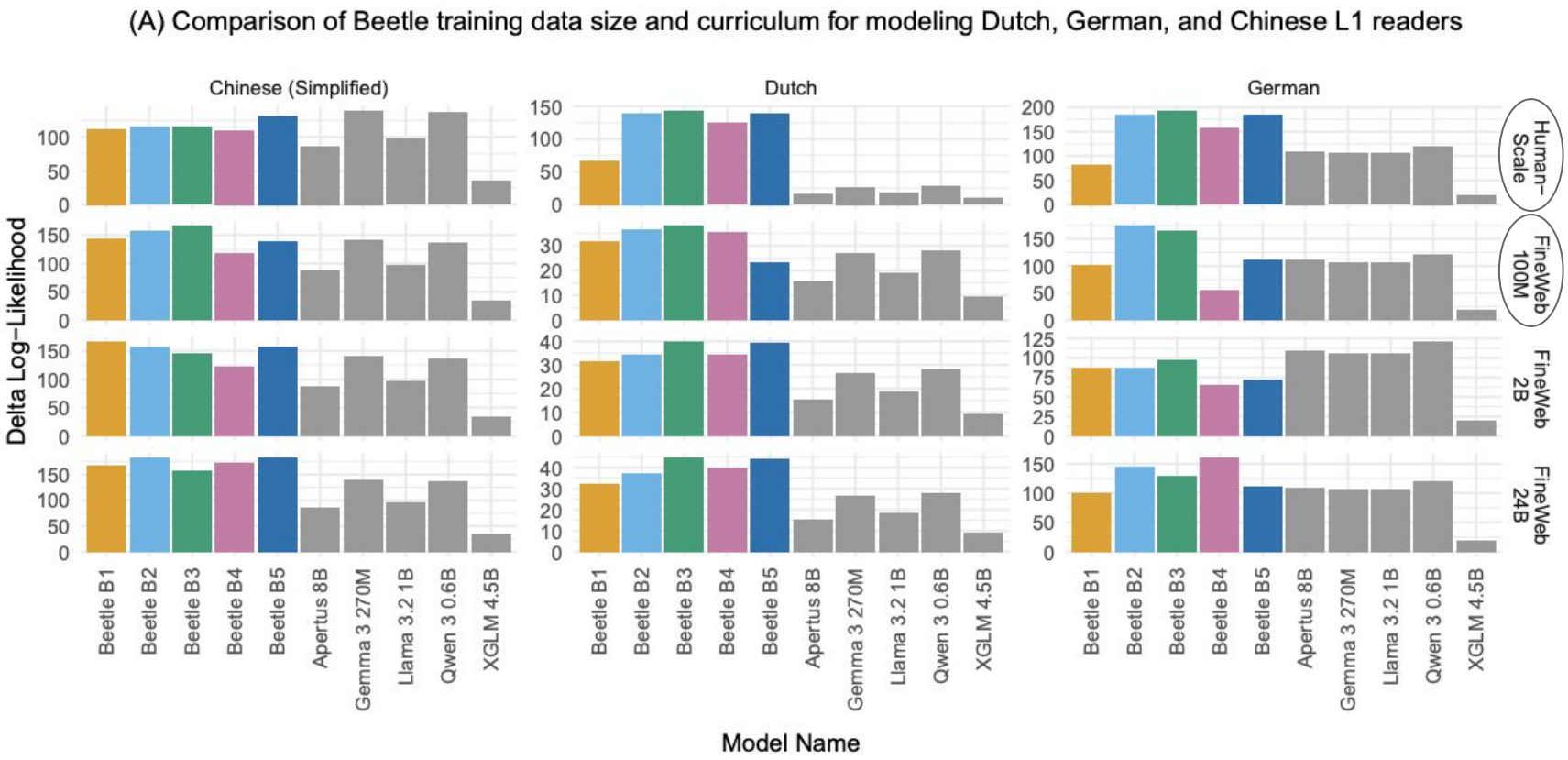
Staged exposure (B2 - B5) generally produces better alignment with human second-language reading times than balanced bilingual training (B1)



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.

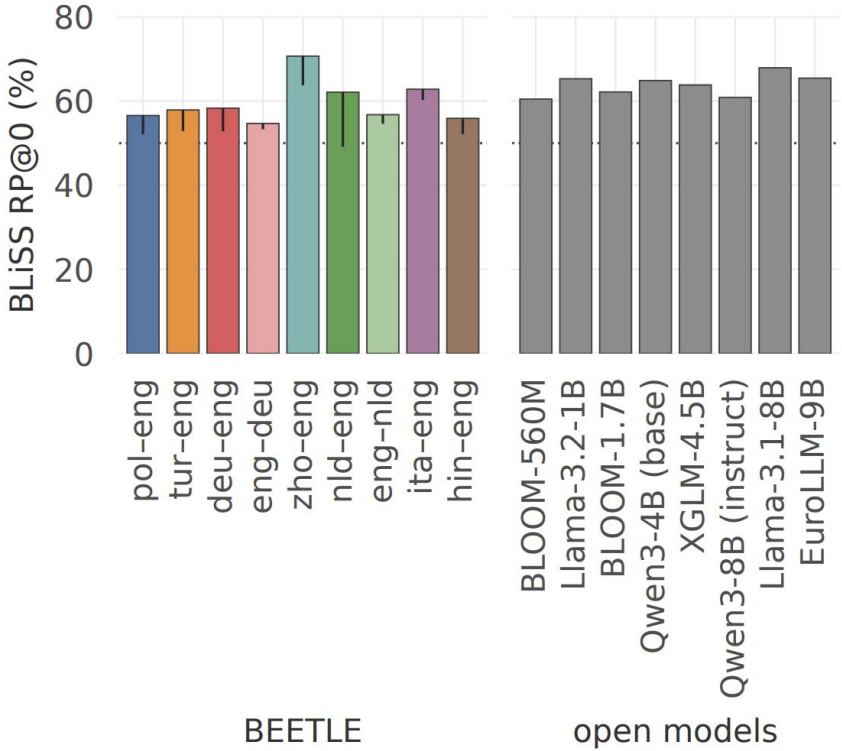
+ve = the model finds the target sentence less surprising after a matching prime → evidence of structural priming..

	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.



Takeaways

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

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