Influence-driven Curriculum Learning for Pre-training on Limited Data

The First BabyLM Workshop @ EMNLP 2025

Loris Schoenegger[1,2], Lukas Thoma[1,2], Terra Blevins[1,4], Benjamin Roth[1,3]

[1]Faculty of Computer Science, University of Vienna, Vienna, Austria
[2]UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria
[3]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria
[4]Khoury College of Computer Sciences, Northeastern University, Boston, USA

universität wien
Faculty of Computer Science

Motivation

- Human-inspired CL performs poorly for low-resource pre-training
- Is there an inherent problem with sorting examples by difficulty?
  - Or are we just using the wrong heuristics?
  - Will a model-centric measure of difficulty perform better?

Contributions

We leverage a technique from interpretability research to build a novel type of curriculum:

Curricula based on training data influence estimates

- We demonstrate their **effectiveness in benchmarks;**
- **analyze their data mix** and how it evolves over time;
- **study loss trajectories** to determine how they affect the model's learning process;
- and **compare example ordering** to existing sorting heuristics.

Difficulty Score

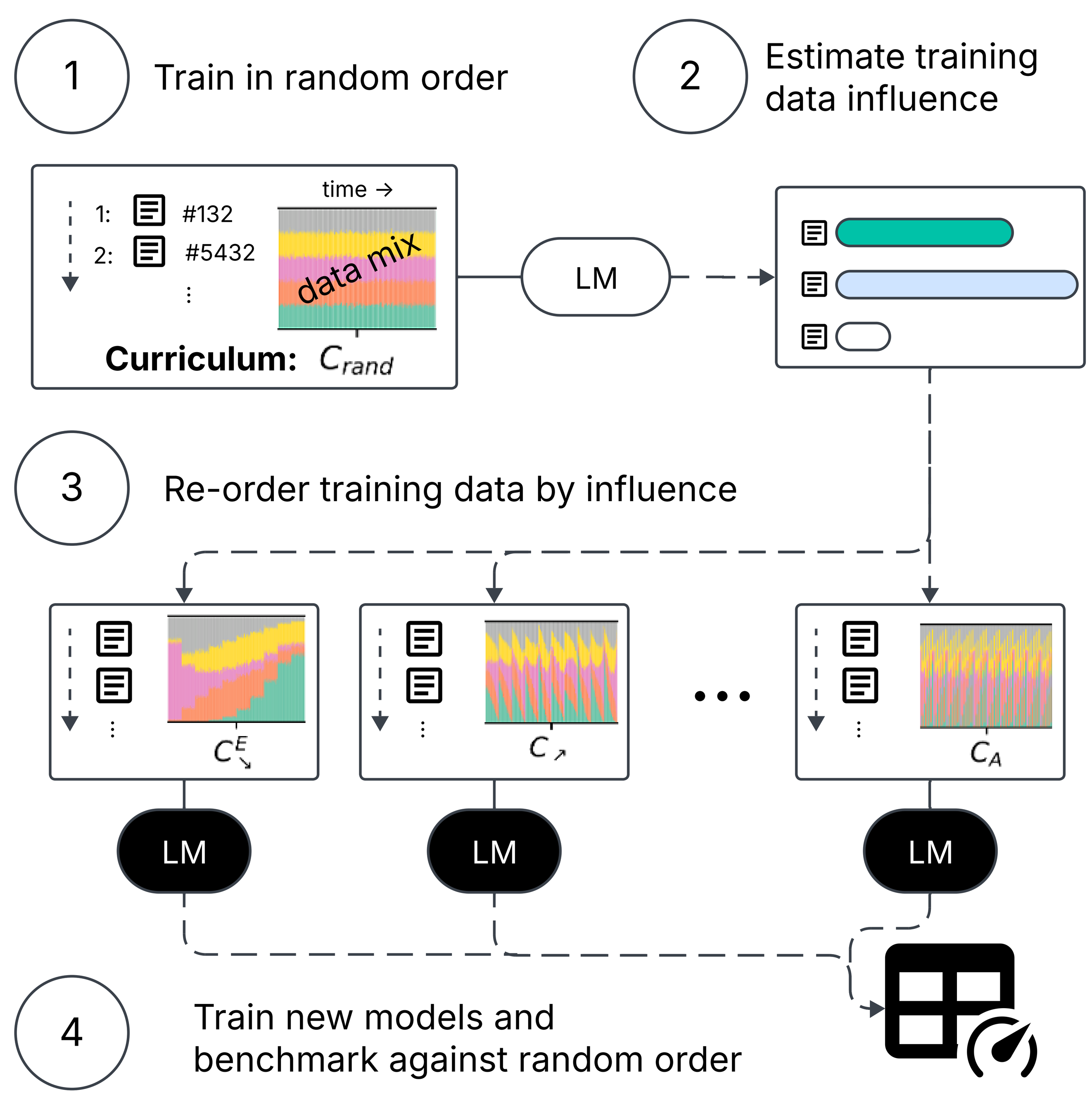
Average influence $\phi_t(z, D)$ that a given training example exerts on the prediction of all other examples from the training data D:

$$\phi_t(z, D) = \frac{\sum_{\forall z' \in D} \nabla \ell(w_t, z) \cdot \nabla \ell(w_t, z')}{|D|} = \nabla \ell(w_t, z) \cdot \mathbb{E}_{z' \sim D}[\nabla \ell(w_t, z')]$$

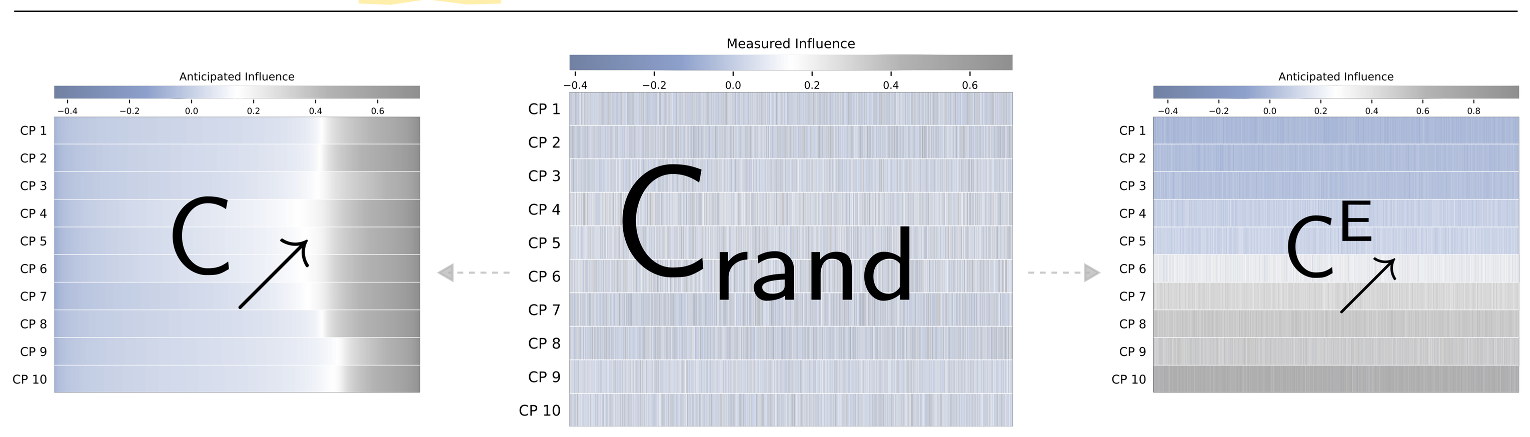
→ (w_t is the WE layer)

High for prototypical examples (whose loss gradients are similar to the average gradient), low for outliers.

Curriculum Design



Coverage Strategy	Curriculum	
Epoch-wise	$C_{\nearrow}$ + 5 more	Increasing difficulty, data visited once per epoch
Cumulative	$C^E_{\nearrow}$ + 2 more	Easy examples in early-, difficult examples in later epochs
Human-inspired	C_{Source} + 2 more	Handcrafted source-difficulty curriculum (5 separate stages)



Datasets

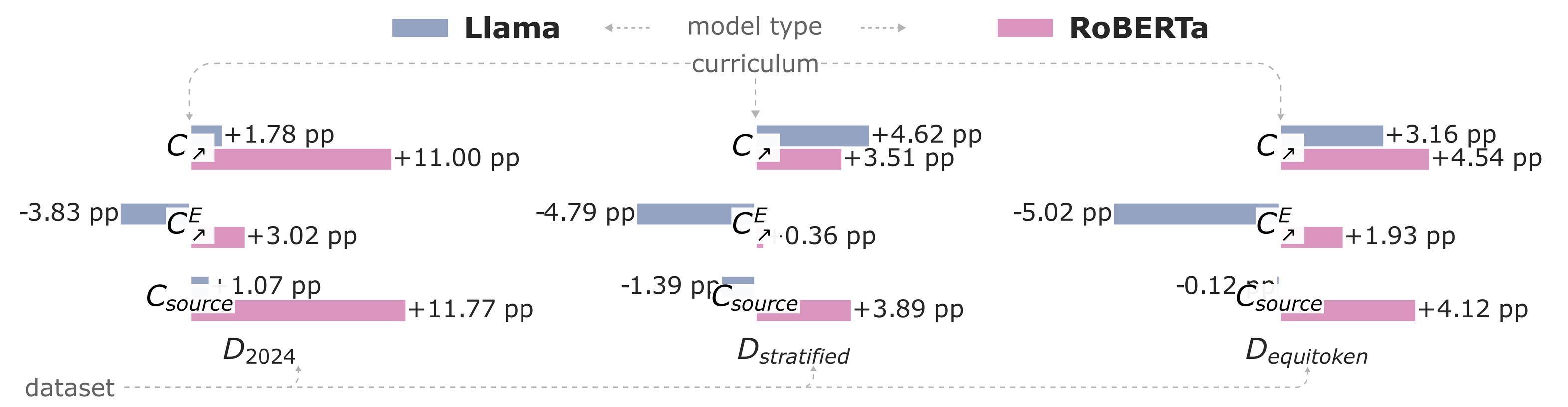
D_{2024}	2024/25 BabyLM dataset (10M word text-only)
$D_{stratified}$	equal number of words per stage
$D_{equitoken}$	equal number of words per example (100)

∼ **built from 5 stages:**

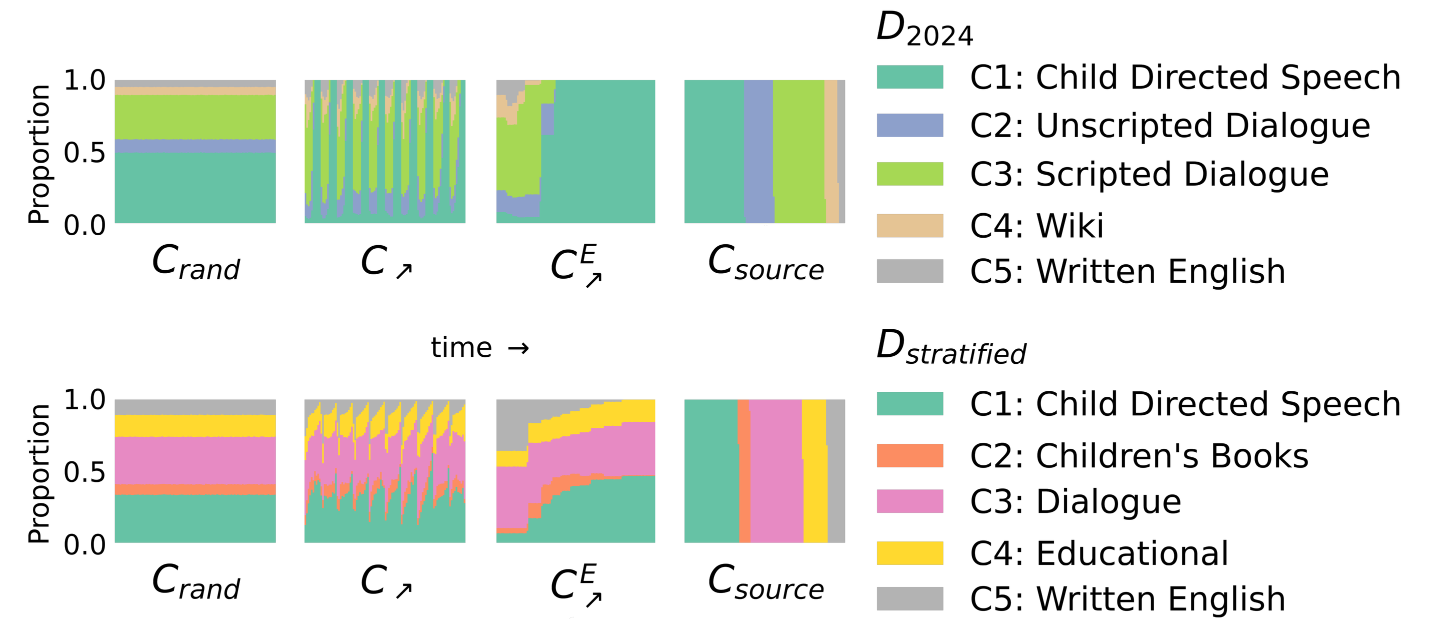
C1: Child Directed Speech
.
.
.
C5: Written English

Results

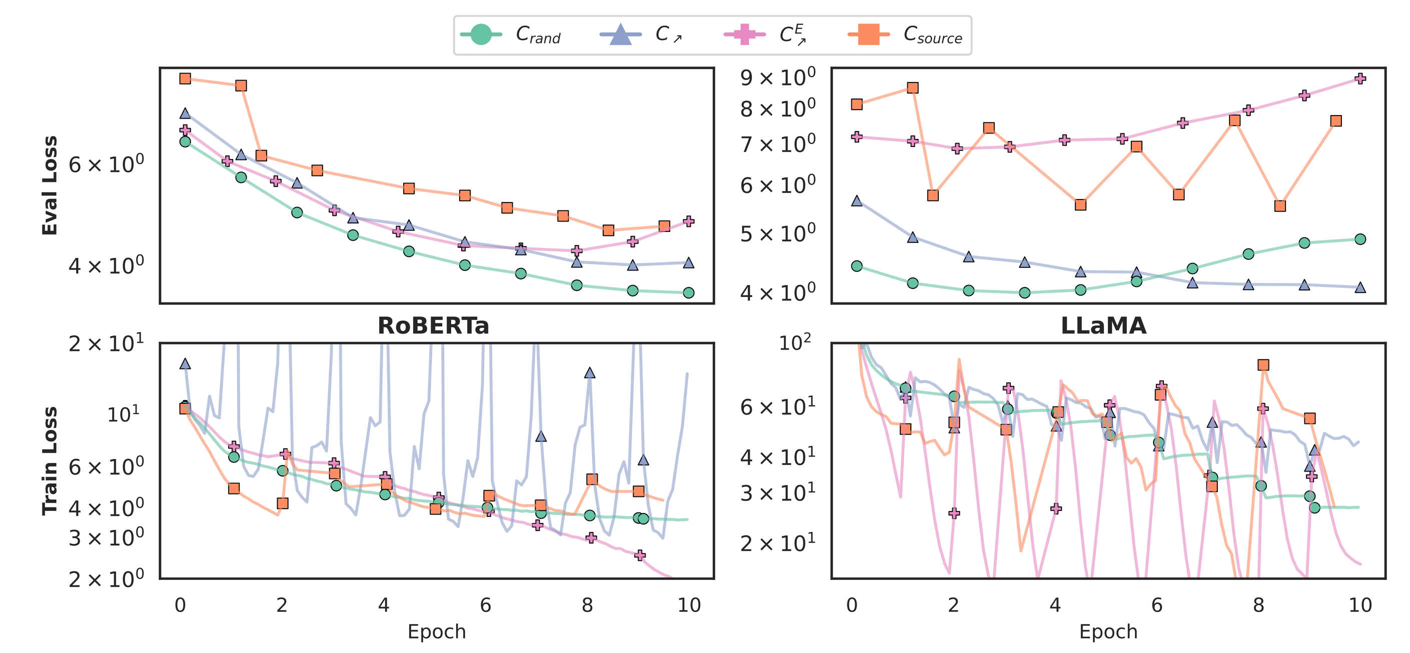
→ Increase of over 10 pp in acc over random order for RoBERTa- and over 4 pp for Llama models (see paper for best performing- and additional curricula):



→ source composition does not strongly vary over time (vs. traditional handcrafted source difficulty curricula):



→ severe spikes in training loss are not significantly correlated with performance on downstream benchmarks



Shared Task Results

Strict-Small Track. $E.$ denotes `BabyLM` baseline models.

Model	GLUE	blimp_filt	blimp_supplement	comps	entity_tracking	ewok	Avg acc
$TICL_{roberta}$	0.646	0.699	0.539	0.507	0.343	**0.503**	0.598
E_{masked}	**0.665**	0.502	0.480	0.500	**0.422**	**0.503**	0.496
E_{mixed}	0.660	0.501	0.467	0.491	0.414	0.500	0.493
E_{causal}	0.654	**0.717**	**0.632**	**0.528**	0.346	0.495	**0.614**
E_{gpt2}	0.623	0.664	0.571	0.517	0.139	0.499	0.540

Correlation in wug tasks (Spearman's ρ Hofmann et al., 2025).

Model	wug_adj_nom	wug_past_tense	Avg
$TICL_{roberta}$	0.006	-0.001	0.003
E_{masked}	0.005	-0.002	0.001
E_{mixed}	0.004	-0.001	0.002
E_{causal}	0.006	**0.001**	**0.004**
E_{gpt2}	**0.007**	-0.001	0.003

Reading tasks. Reported as % R^2 gain.

Model	Eye Tracking Score	Self-Paced Reading Score	Avg
$TICL_{roberta}$	0.040	0.002	0.021
E_{masked}	**0.103**	0.027	0.065
E_{mixed}	0.099	0.025	0.062
E_{causal}	0.099	0.035	**0.067**
E_{gpt2}	0.087	**0.043**	0.065

Key Findings

→ Our sorting strategies can increase performance; however: only if paired with non-developmentally plausible dataset coverage strategies, i.e., must **visit the full dataset every epoch**

→ Improvement may result from improved grouping of examples into **batches of similar difficulty**

→ Measure appears **inversely correlated to those** of other sorting heuristics (high influence → low difficulty)