CLASS-IT: Conversational and Lecture-Aligned Small-Scale Instruction Tuning for BabyLMs

Luca Capone [1] Alessandro Bondielli [1,2] Alessandro Lenci [1]

[1]CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa [2]Department of Computer Science, University of Pisa

Motivation

Despite the impressive capabilities of current **instruction-tuned large language models**, the effect of this post-training technique on BabyLMs is still largely underexplored.

The **BabyLM Challenge** offers an opportunity to study how instruction tuning affects models trained under strict ecological constraints.

Goal: investigate whether small, ecologically trained models can benefit from **instruction-tuning** paradigms, and how **different instruction types and curricula** affect their downstream linguistic and reasoning abilities.

Research Questions

- Does instruction tuning improve the linguistic abilities of small-scale models?
- How do **different instruction types** shape performance?
- What is the effect of **curriculum ordering** on generalization?
- Are benefits transferable to unseen reasoning tasks?

Model Training

Base models: 100M and 140M parameter LLaMA-style models [1] trained under the BabyLM strict track.

Hyperparameter	llama140M	llama100M
Vocab size	32,000	16,384
Max length	6,144	6,000
Hidden size	704	512
Attention heads	11	8
Layers	12	20
Trainable parameters	140,231,872	100,684,288

Table 1. Model architectures

Hyperparameter	Pretrain	Instr. tuning
Initial LR	2e-4	2e-5
Batch size	8	8
Maximum epochs	8	10
LR scheduler	linear cosine w/ restarts	
Warm-up steps	5,000	500

Table 2. Training parameters

Instruction dataset: `colinglab/CLASS_IT` 🤗

- **Conversational instructions** — based on Switchboard (**switch**).
- **QA instructions** — Question/Answer Pairs generated from SimpleWiki with `LLaMA-3.2-3B-Instruct` (**wiki**)
- **97,697 items**, 18 million words

Training Curricula:

- **Mixed:** mixing both instruction types during training (**merged**).
- **Sequential:** fine-tuning first on one instruction type, then the other.

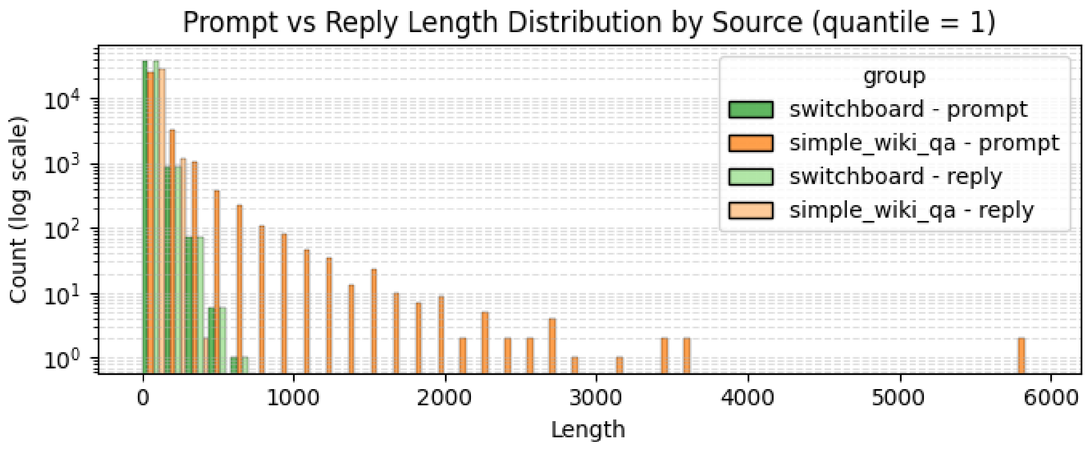

Figure 1. Overview of the instruction-tuning pipeline and evaluation setup.

Model available at `colinglab/CLASS_IT-140M`🤗

Evaluation Procedure

Evaluation: Official BabyLM evaluation pipeline.

- Fine-Tuning - on (Super)GLUE tasks

- Default fine-tuning parameters of the pipeline
- Training set: randomly sampled 10k portion of the original training set for each task

- Zero-Shot - using log-probabilities of sequences and/or words to obtain either model predictions or compute correlations with human data.

- Standard minimal pairs datasets: BLiMP, EWoK.
- A WUGs task (adjective nominalization)
- Entity tracking task
- Correlation evaluation: cloze probability, predictability ratings, and computational estimates against EEG and human reading times.

Results and Discussion

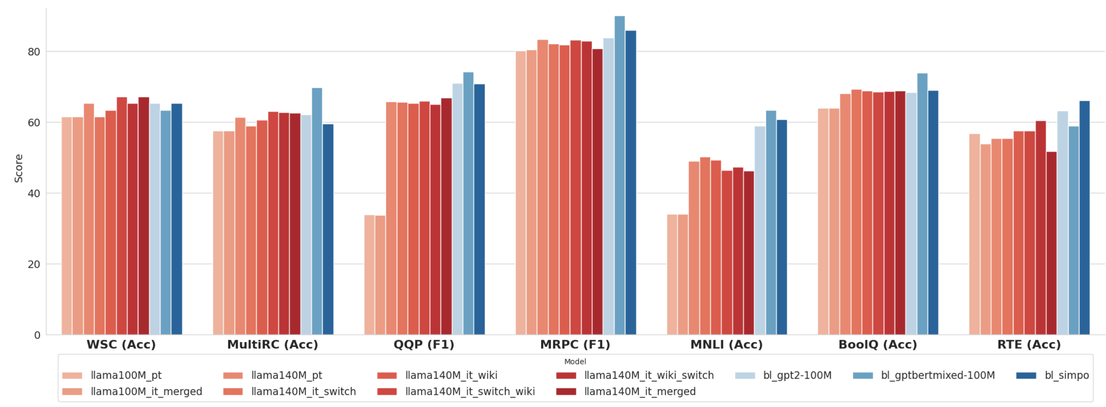

Figure 2. Results of fine-tuned models on (Super)Glue tasks. Our Models; Baselines.

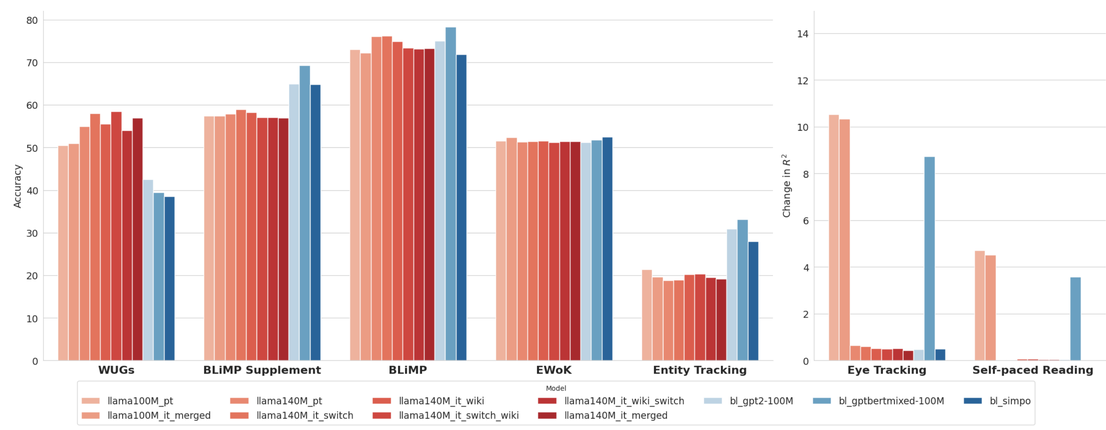

Figure 3. Results on zero-shot tasks. Our Models; Baselines.

- Difficulties with **inference and paraphrase** tasks (MNLI, QQP)
- Instruction-tuning gains are **modest and task-dependent**.
- Zero-shot works well on morphological generalization (WUGs) but poor on entity tracking — an **unexpected outcome** given instruction-tuning.
- **smaller models' align more closely with human reading times** [2].

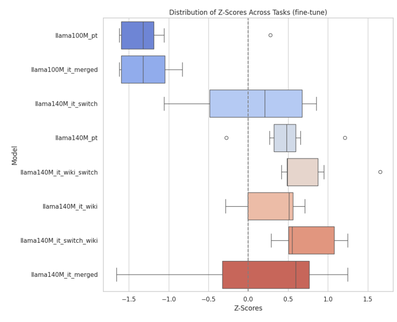

Figure 4. **Fine-Tuning** median, IQR, and outliers for z-score normalized performances.

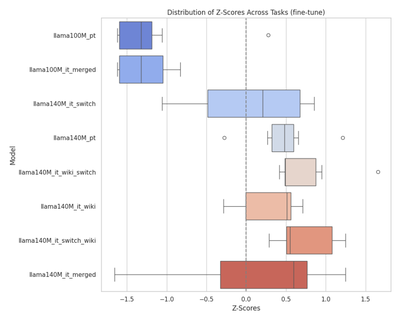

Figure 5. **Zero-Shot** median, IQR, and outliers for z-score normalized performances.

- **Best median performances with Mixed Curriculum** in fine-tuning.
- In Zero-shot, smaller model is better.

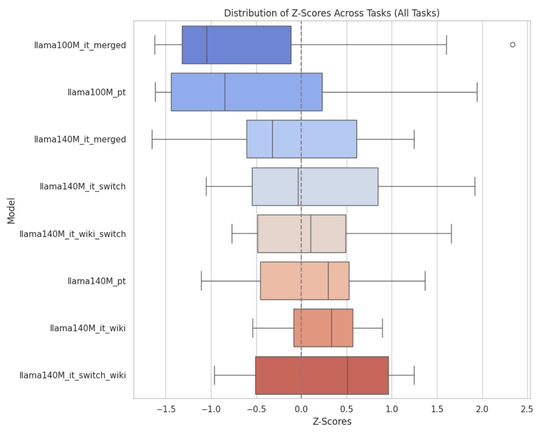

Figure 6. **All tasks** median, IQR, and outliers for z-score normalized performances.

- **Larger models are better**, the overall **best is instruction-tuned**, but with **no clear advantage**;
- **Sequential curriculum** outperforms Mixed one; the **order is less relevant**;
- **Differences** in performances may also be **due to training size** (i.e., SimpleWiki is much larger than Switchboard).

Conclusions

- **Modest but consistent gains** in **instruction-tuning** with **sequential curricula**.
- Improvements **do not generalize** reliably to zero-shot settings.
- At small scales, instruction tuning may **bias models toward narrow interactional behaviors**, reducing broader linguistic generalization.

References

[1] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[2] A. De Varda and M. Marelli, "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)*, pp. 139–149, 2023.