Beyond Repetition: Text Simplification and Curriculum Learning for Data-Constrained Pretraining

Matthew Theodore Roque * and Dan John Velasco *

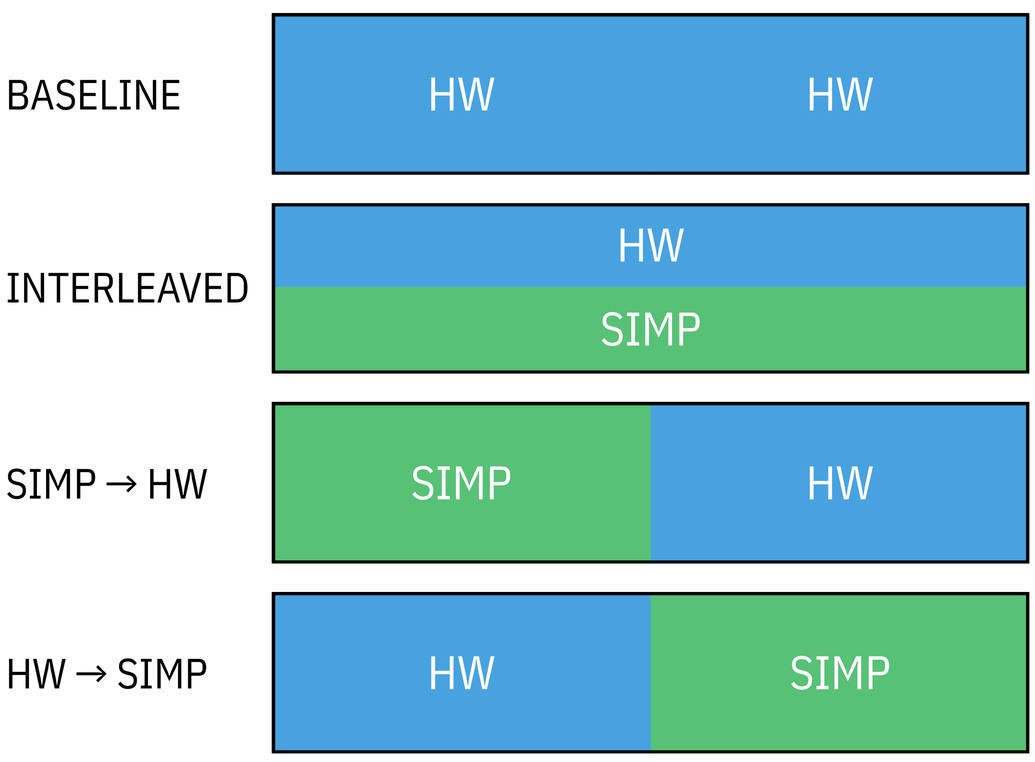
Samsung R&D Institute Philippines
{roque.mt,dj.velasco}@samsung.com
*Equal Contribution

Key Takeaways

- We study **data-constrained pretraining**. Instead of repeating the same corpus, we add **LLM-simplified rewrites** of it and compare **four data schedules**.
- **Simplification beats repetition at the same budget**. It improves fine-tuning sample efficiency and zero-shot performance.
- **Ordering is size-dependent**. 124M benefits from simple → complex. 256M prefers a balanced mix.

Parallel Data & Schedules

Each human-written (HW) paragraph has a simplified (SIMP) counterpart. **Same content, simpler form.**



Zero-shot Results

Model	Common-sense Avg.	BLiMP/ Supp	EWoK	Entity Tracking	MMLU
124M					
BASELINE	34.3	66.9	53.9	22.4	24.7
INTERLEAVED	**34.5**	68.0	**55.5**	28.1	**24.9**
SIMP→HW	34.4	**68.1**	54.8	31.7	23.6
HW→SIMP	34.2	66.0	55.1	**36.9**	23.3
256M					
BASELINE	34.8	**69.7**	55.0	30.1	23.5
INTERLEAVED	35.0	68.9	**56.2**	**35.0**	24.5
SIMP→HW	**35.5**	69.0	55.9	30.8	25.8
HW→SIMP	34.3	68.7	56.0	34.1	**26.2**

Results Overview

- **Fine-tuning**: **124M** performs best with **SIMP → HW**, while **256M** performs best with **INTERLEAVED**. Gains are **largest when fine-tuning data is scarce**.
- **Zero-shot**: The **addition of LLM-simplified text** improves accuracy over repeating HW **across the board**. Performance trends across schedules are task dependent.

NLU Fine-tuning Results