

## Rebuttal Response

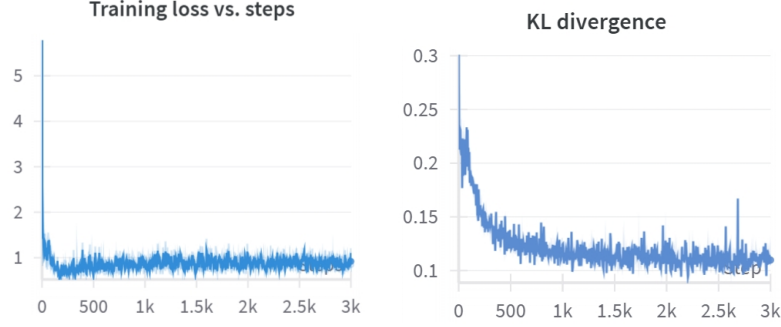


Figure 1. The evolution of the lower-level training loss and KL divergence in training of 10B tokens. Proxy model size: 31M, target LLM size: 410M.

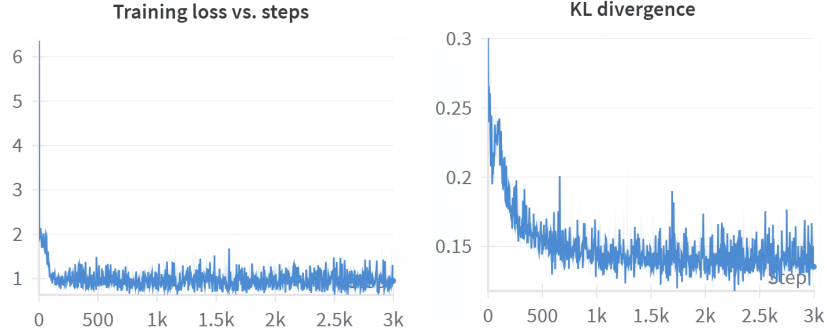


Figure 2. The evolution of the lower-level training loss and KL divergence in training of 10B tokens. Proxy model size: 160M, target LLM size: 410M.

Table 1. Total FLOPs for pretraining 410M/1B target model with 25B tokens.

Process	#FLOPs $\times 10^{19}$	Ratio
<b>BLISS: 410M model, 25B tokens</b>		
Model pretraining	6.35	79.28%
Warm up the proxy/score model	0.07	0.87%
Bilevel optimization	0.13	1.62%
Data influence model inference	1.53	19.10%
<b>Total</b>	<b>8.08</b>	<b>100.00%</b>
<b>BLISS: 1B model, 25B tokens</b>		
Model pretraining	17.67	90.48%
Warm up the proxy/score model	0.07	0.36%
Bilevel optimization	0.261	1.34%
Data influence model inference	1.53	7.83%
<b>Total</b>	<b>19.53</b>	<b>100.00%</b>

Table 2. Comparison of BLISS with different size of proxy/score model and on zero-shot evaluation over multiple downstream datasets (410M model, 10B tokens) with 20k-step training.

Method	SciQ	ARC-E	ARC-C	LogiQA	OBQA	BoolQ	HellaSwag	PIQA	WinoGrande	Average
BLISS (Pythia-31M)	65.5(1.5)	40.8(1.0)	23.4(1.2)	27.2(1.7)	29.8(2.0)	58.9(0.9)	36.0(0.5)	67.6(1.1)	53.4(1.4)	44.7(1.3)
BLISS (Pythia-160M)	63.8(1.5)	40.8(1.0)	23.4(1.2)	27.5(1.8)	29.8(2.0)	51.3(0.9)	38.3(0.5)	67.6(1.1)	50.4(1.4)	44.1(1.3)
BLISS (Pythia-31M without sigmoid)	62.6(1.5)	41.0(1.0)	24.0(1.2)	26.4(1.7)	30.4(2.1)	53.4(0.9)	39.5(0.5)	68.3(1.1)	52.2(1.4)	44.2(1.3)

Table 3. Comparison of methods on zero-shot evaluation over multiple downstream datasets (410M model, 15B tokens). BLISS-org denotes the original algorithm, and BLISS<sup>†</sup> is a variant which uses different initialization method for the score model.

Methods (#FLOPs $\times 10^{19}$ )	SciQ	ARC-E	ARC-C	LogiQA	OBQA	BoolQ	HellaSwag	PIQA	WinoGrande	Average
BLISS-org	67.7 (1.5)	41.7 (1.0)	23.6 (1.2)	25.8(1.7)	28.4(2.0)	56.0 (0.8)	39.7 (0.5)	68.7 (1.1)	53.2 (1.4)	44.9 (1.3)
BLISS <sup>†</sup>	65.2 (1.5)	41.6 (1.0)	23.4 (1.2)	27.1 (1.7)	29.8 (2.0)	57.5 (0.8)	34.9 (0.5)	67.7 (1.1)	53.5 (1.4)	44.5 (1.3)

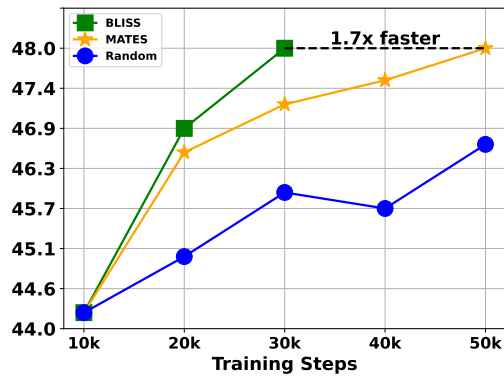


Figure 3. The downstream performance of Pythia-1B model w.r.t. pretraining steps, where the first point denotes the performance of a warm-up model trained on the randomly selected training data.