Model	Comparison	ℓ_1 Sparsity Crosscoder				
		ℓ_0	Dead Feats	ΔCE A	ΔCE B	ΔCE C
Pythia-1B Layer 8	128M 1B	88	9	0.00	0.07	-
	1B 4B	214	1	0.05	0.18	-
	4B 286B	190	9	0.15	0.48	-
	1B 4B 286B	215	19	0.03	0.16	0.54
OLMo-1B Layer 8	2B 4B	184	0	0.08	0.20	-
	4B 33B	227	0	0.14	0.21	-
	33B 3048B	182	425	0.16	0.35	-
	4B 33B 3048B	225	101	0.12	0.18	0.43
BLOOM-1B Layer 12	550M 6B	211	6	0.10	0.20	-
	6B 55B	112	8	0.14	0.29	-
	55B 341B	96	12	0.18	0.18	-
	6B 55B 341B	118	19	0.13	0.20	0.22

Table 1: **Crosscoder statistics.** Results averaged over three seeds on validation set. ΔCE is the change in cross-entropy loss when doing a forward pass using the original output versus the crosscoder reconstruction. A, B, C refer to the 1st, 2nd and 3rd checkpoints used for loss computation. ℓ_0 and dead feature averages are rounded to integers. Less trained models (, 1B) get smaller ΔCE values than further trained models (, 286B) due to the former's high original CE loss.