Task	Pythia-1B			OLMo-1B		
	Comparison	$\rho(\frac{\lvert\Delta c_2\rvert}{\lvert\Delta c_1\rvert}, \text{RelDec})$	$\rho(\frac{\lvert\Delta c_2\rvert}{\lvert\Delta c_1\rvert}, \text{RelIE})$	**Comparison**	$\rho(\frac{\lvert\Delta c_2\rvert}{\lvert\Delta c_1\rvert}, \text{RelDec})$	$\rho(\frac{\lvert\Delta c_2\rvert}{\lvert\Delta c_1\rvert}, \text{RelIE})$
`Distractor Relational Noun`	128M 1B	0.316	**0.934**	2B 4B	0.920	**0.972**
	1B 4B	0.930	**0.958**	4B 33B	**0.949**	0.897
	4B 286B	0.691	**0.964**	33B 3048B	0.770	**0.973**
`Distractor Relative Clause`	128M 1B	0.788	**0.966**	2B 4B	0.961	**0.979**
	1B 4B	0.901	**0.922**	4B 33B	**0.956**	0.938
	4B 286B	0.784	**0.941**	33B 3048B	0.771	**0.989**
`Irregular Plural Subject`	128M 1B	0.941	**0.979**	2B 4B	**0.982**	0.966
	1B 4B	0.777	**0.937**	4B 33B	0.898	**0.905**
	4B 286B	0.843	**0.954**	33B 3048B	0.785	**0.810**
`Regular Plural Subject`	128M 1B	0.794	**0.948**	2B 4B	0.838	**0.966**
	1B 4B	0.874	**0.930**	4B 33B	**0.961**	0.919
	4B 286B	0.806	**0.908**	33B 3048B	0.862	**0.956**
Avg by Comparison	Avg 128M 1B	0.710	**0.957**	Avg 2B 4B	0.925	**0.971**
	Avg 1B 4B	0.870	**0.937**	Avg 4B 33B	**0.941**	0.915
	Avg 4B 286B	0.781	**0.942**	Avg 33B 3048B	0.797	**0.932**
Overall Avg	-	0.787	**0.945**	-	0.843	**0.952**

Table 1: **Top-10 significant feature ablation for Pythia-1B and OLMo-1B.** Spearman correlations ρ between the ratio of log-probability-difference metrics and model-attributing scores (,), across four subject–verb agreement phenomena and various phase transition comparisons. shows consistently higher correlations than across tasks and model comparisons, indicating that focusing on task-relevant signal uncovers more meaningful and stable task-specific features.