

Identifying latent algorithms in Transformers via RASP

Rimon Melamed^{1,2}, H. Howie Huang¹

1: The George Washington University, 2: Prompt Inversion LLC

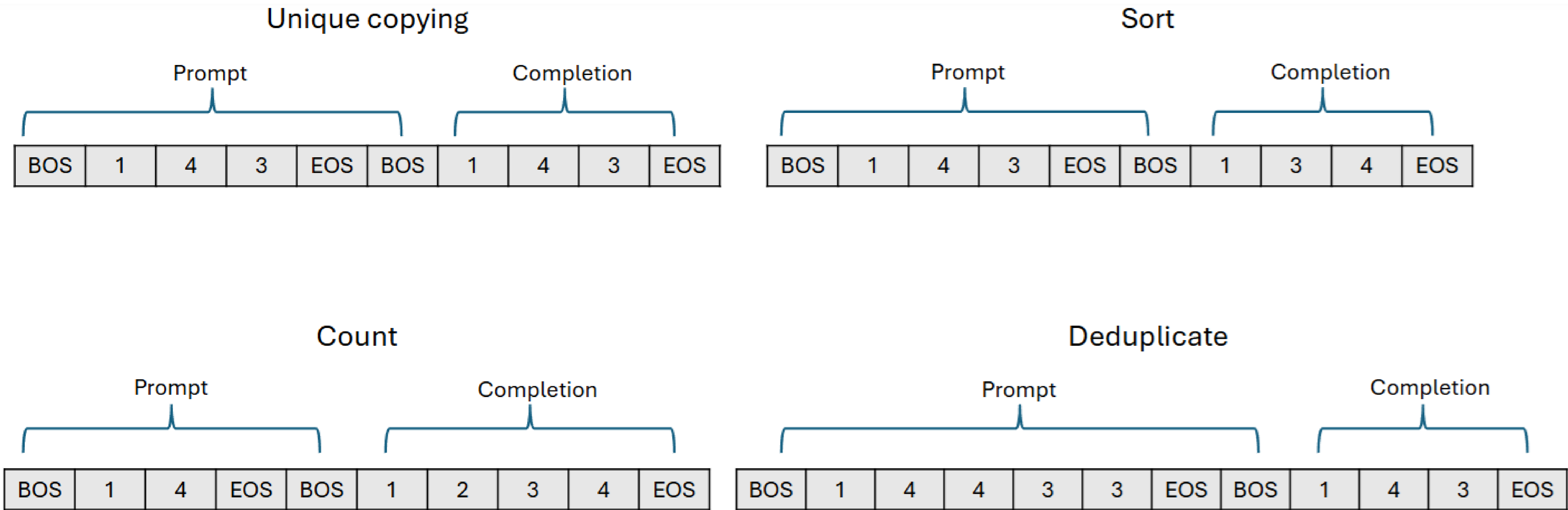


Prompt
Inversion

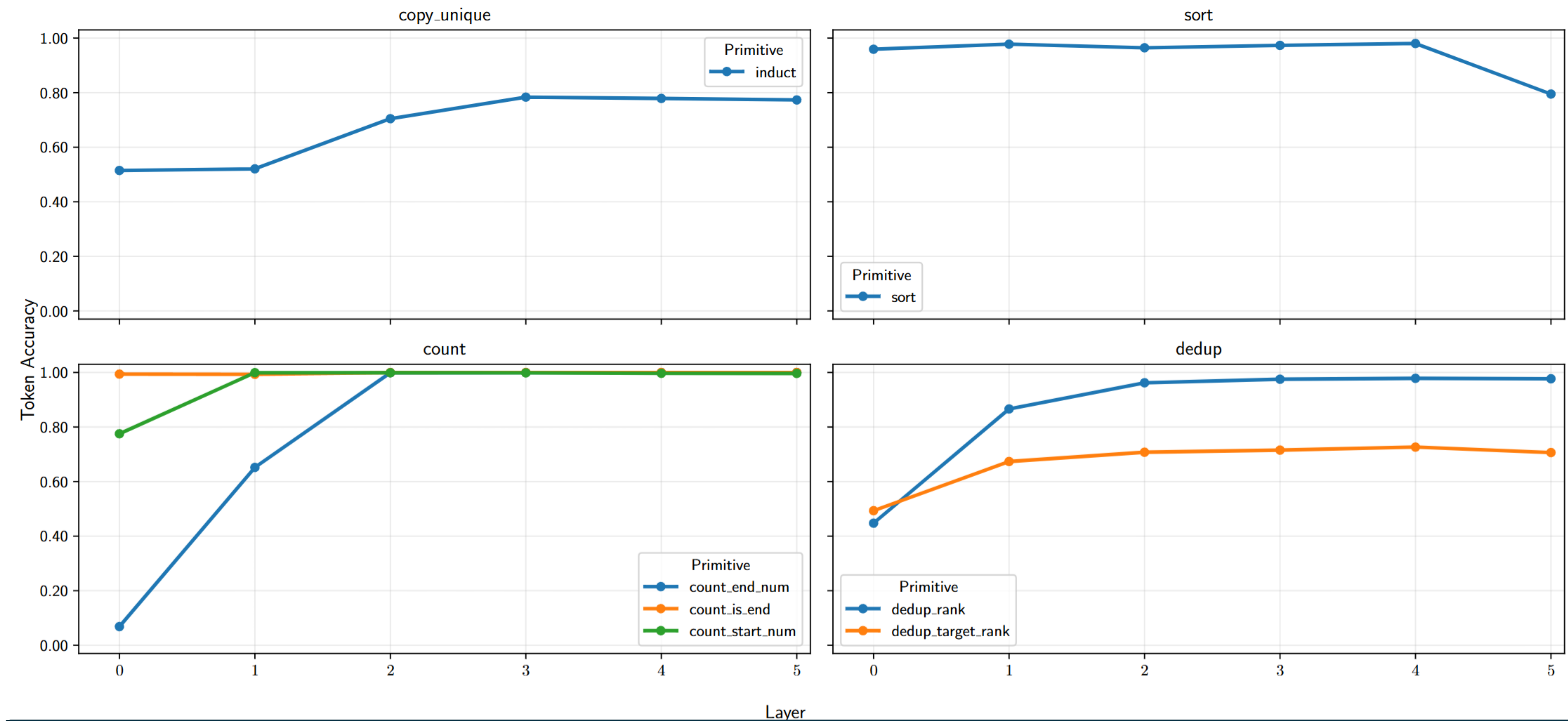
THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC



Do Transformers learn algorithms corresponding to RASP-L programs?



Probing for RASP-L representations in Transformers trained on algorithmic tasks



Causal ablation shows that RASP-L representations are present

Task	RASP-L intermediate	Layer	Baseline EM	Subspace Erasure EM ↓	Random Erasure EM
Copy unique	induct	0	0.265	0.004	0.198 ± 0.004
Sort	sort	1	0.968	0.000	0.824 ± 0.017
Count	count_start_num	0	0.936	0.888	0.814 ± 0.012
	count_end_num	2	0.936	0.705	0.847 ± 0.010
	count_is_end	3	0.936	0.916	0.936 ± 0.000
Dedup	dedup_rank	3	0.889	0.821	0.891 ± 0.001
	dedup_target_rank	0	0.889	0.751	0.883 ± 0.001