Table 9: Evaluation results for Llama-2-7B.

	Score	Efficacy	Generalization	E-Specificity	R-Specificity
Llama-2	52.8	13.8	16.1	81.2	100.0
FT-L	55.6	24.2	17.0	81.6	99.7
ROME	81.1	99.9	93.4	77.4	53.6
RETS	82.1	98.3	74.6	72.3	83.1