(a) Input Reaction

Model Top-1 Prediction + Input Attribution

TRUE

CI OH + 0.482 0.0803
$$p = 0.43$$
 ua: 0.0014 ua: 0.008 ua: 0.003 $p = 0.43$ ua = uniform attribution

(b) Validation with Further Examples:

CITCI + CITCI + CITCI + P = 0.64

$$p = 0.64$$

$$p = 0.92$$