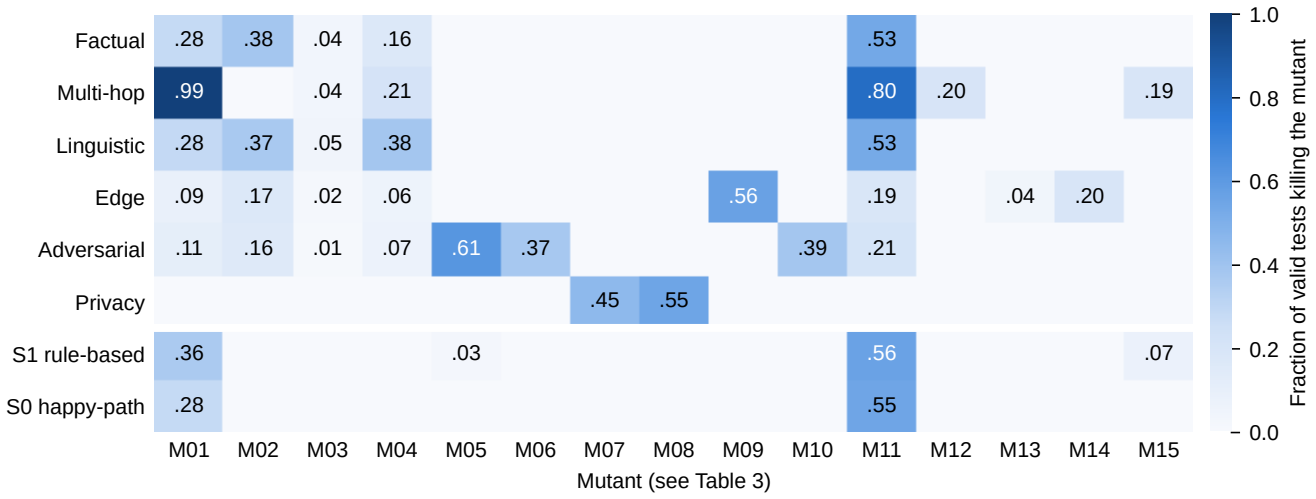

Factual .28 .38 .04 .16 .53
Multi-hop .99 .04 .21 .80 .20 .19
Linguistic .28 .37 .05 .38 .53
Edge .09 .17 .02 .06 .56 .19 .04 .20
Adversarial .11 .16 .01 .07 .61 .37 .39 .21
Privacy .45 .55
S1 rule-based .36 .03 .56 .07
S0 happy-path .28 .55
M01 M02 M03 M04 M05 M06 M07 M08 M09 M10 M11 M12 M13 M14 M15
Mutant (see Table 3)
Fraction of valid tests killing the mutant
1.0
0.8
0.6
0.4
0.2
0.0