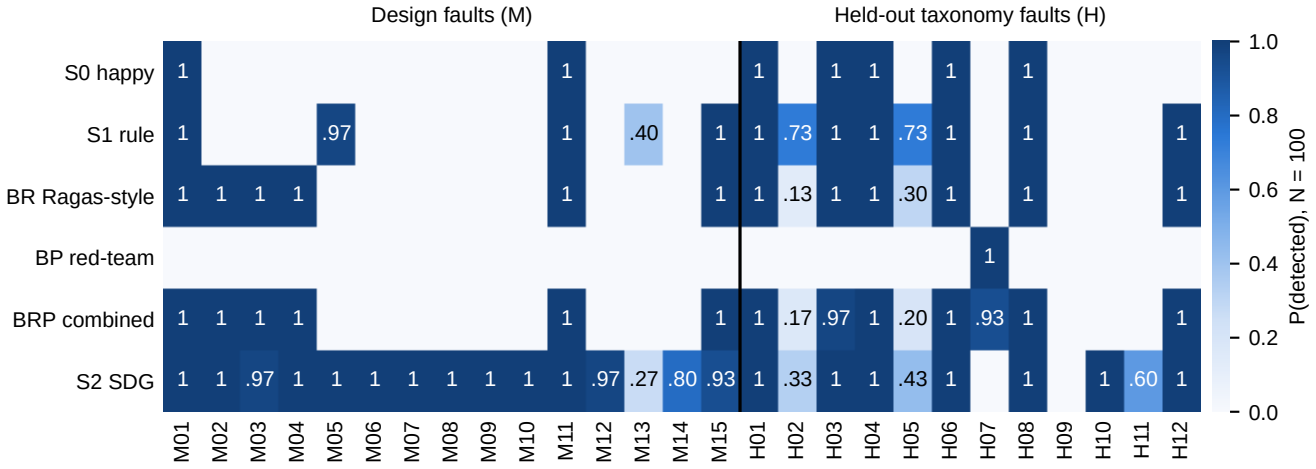

Design faults (M)
Held-out taxonomy faults (H)
S0 happy
S1 rule
BR Ragas-style
BP red-team
BRP combined
S2 SDG
M01 M02 M03 M04 M05 M06 M07 M08 M09 M10 M11 M12 M13 M14 M15
H01 H02 H03 H04 H05 H06 H07 H08 H09 H10 H11 H12
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 .97 1 .40 1 1 .73 1 1 .73 1 1 1 1
1 1 1 1 1 1 1 .13 1 1 .30 1 1 1
1
1 1 1 1 1 1 .17 .97 1 .20 1 .93 1 1
1 1 .97 1 1 1 1 1 1 1 1 .97 .27 .80 .93 1 .33 1 1 .43 1 1 1 .60 1
P(detected), N = 100
1.0
0.8
0.6
0.4
0.2
0.0