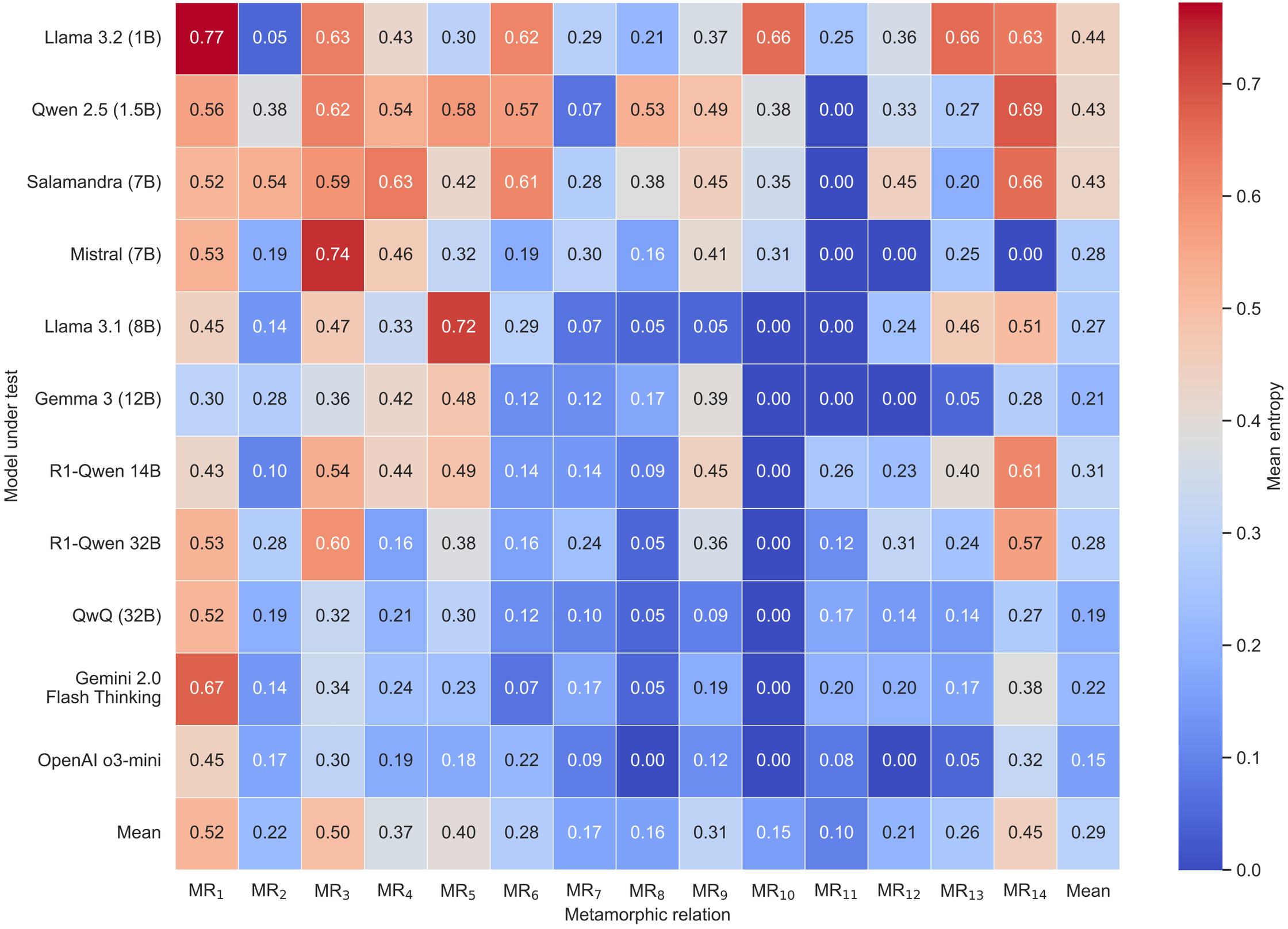
Model under test
Llama 3.2 (1B) 0.77 0.05 0.63 0.43 0.30 0.62 0.29 0.21 0.37 0.66 0.25 0.36 0.66 0.63 0.44
Qwen 2.5 (1.5B) 0.56 0.38 0.62 0.54 0.58 0.57 0.07 0.53 0.49 0.38 0.00 0.33 0.27 0.69 0.43
Salamandra (7B) 0.52 0.54 0.59 0.63 0.42 0.61 0.28 0.38 0.45 0.35 0.00 0.45 0.20 0.66 0.43
Mistral (7B) 0.53 0.19 0.74 0.46 0.32 0.19 0.30 0.16 0.41 0.31 0.00 0.00 0.25 0.00 0.28
Llama 3.1 (8B) 0.45 0.14 0.47 0.33 0.72 0.29 0.07 0.05 0.05 0.00 0.00 0.24 0.46 0.51 0.27
Gemma 3 (12B) 0.30 0.28 0.36 0.42 0.48 0.12 0.12 0.17 0.39 0.00 0.00 0.00 0.05 0.28 0.21
R1-Qwen 14B 0.43 0.10 0.54 0.44 0.49 0.14 0.14 0.09 0.45 0.00 0.26 0.23 0.40 0.61 0.31
R1-Qwen 32B 0.53 0.28 0.60 0.16 0.38 0.16 0.24 0.05 0.36 0.00 0.12 0.31 0.24 0.57 0.28
QwQ (32B) 0.52 0.19 0.32 0.21 0.30 0.12 0.10 0.05 0.09 0.00 0.17 0.14 0.14 0.27 0.19
Gemini 2.0 Flash Thinking 0.67 0.14 0.34 0.24 0.23 0.07 0.17 0.05 0.19 0.00 0.20 0.20 0.17 0.38 0.22
OpenAI o3-mini 0.45 0.17 0.30 0.19 0.18 0.22 0.09 0.00 0.12 0.00 0.08 0.00 0.05 0.32 0.15
Mean 0.52 0.22 0.50 0.37 0.40 0.28 0.17 0.16 0.31 0.15 0.10 0.21 0.26 0.45 0.29
MR1 MR2 MR3 MR4 MR5 MR6 MR7 MR8 MR9 MR10 MR11 MR12 MR13 MR14 Mean
Metamorphic relation
Mean entropy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0