

factual

arithmetic

JSON

code

creative

M1

M2

M3

M4

M5

M6

M7

prompts (n=54, grouped by task category)



providers agree (exact-match)



at least one provider diverged

model