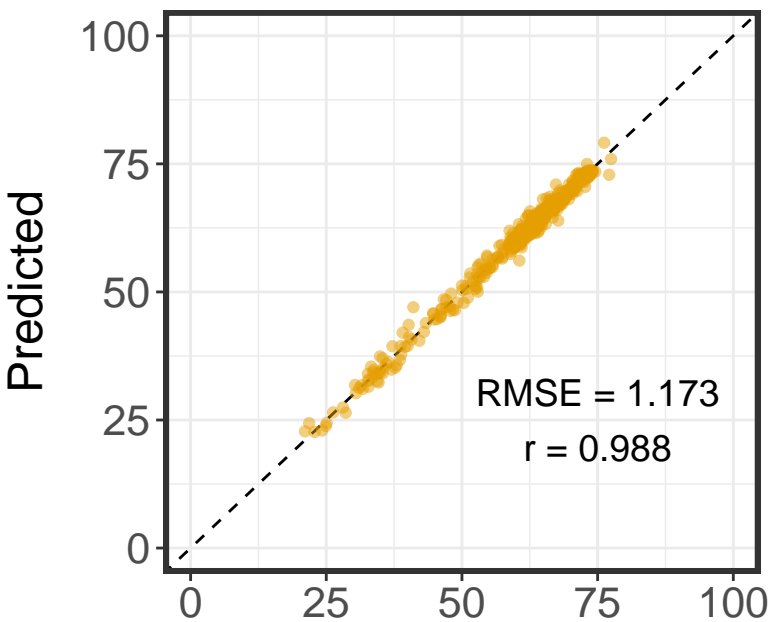
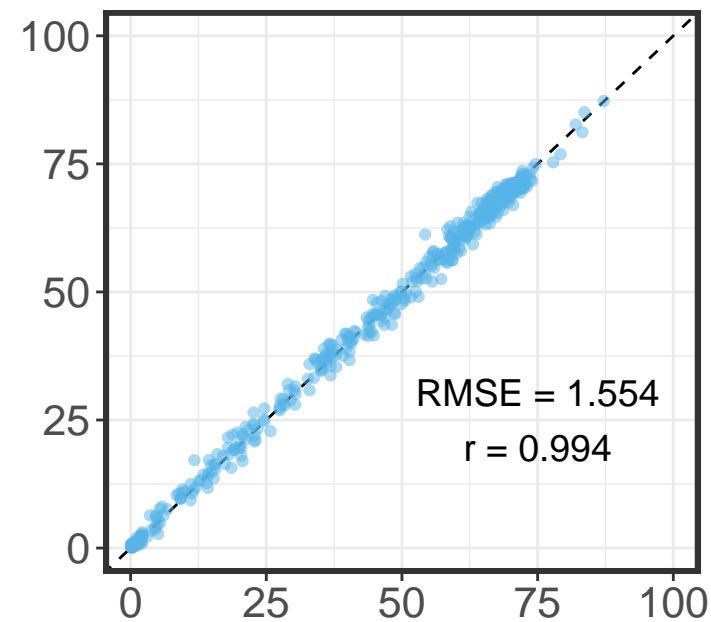


A

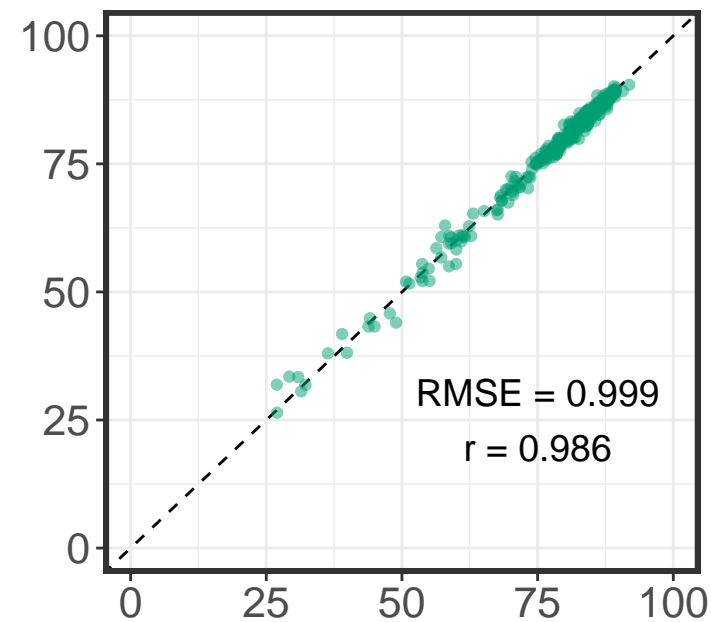
ARC\* (d = 145)



GSM8K\* (d = 237)

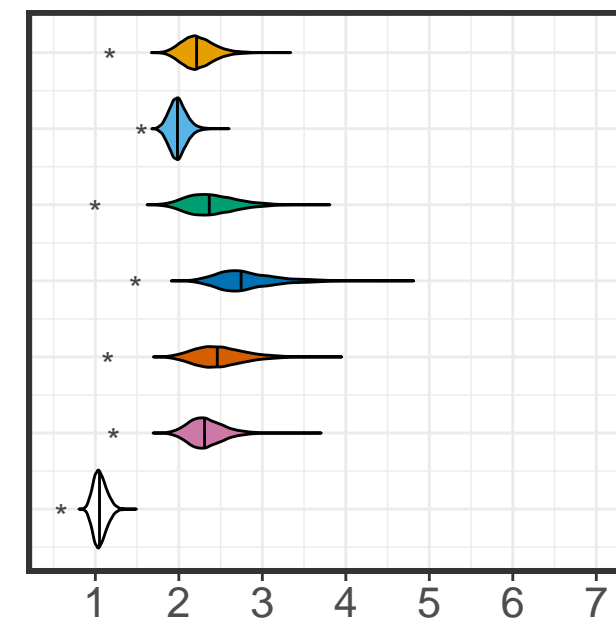


HellaSwag\* (d = 93)

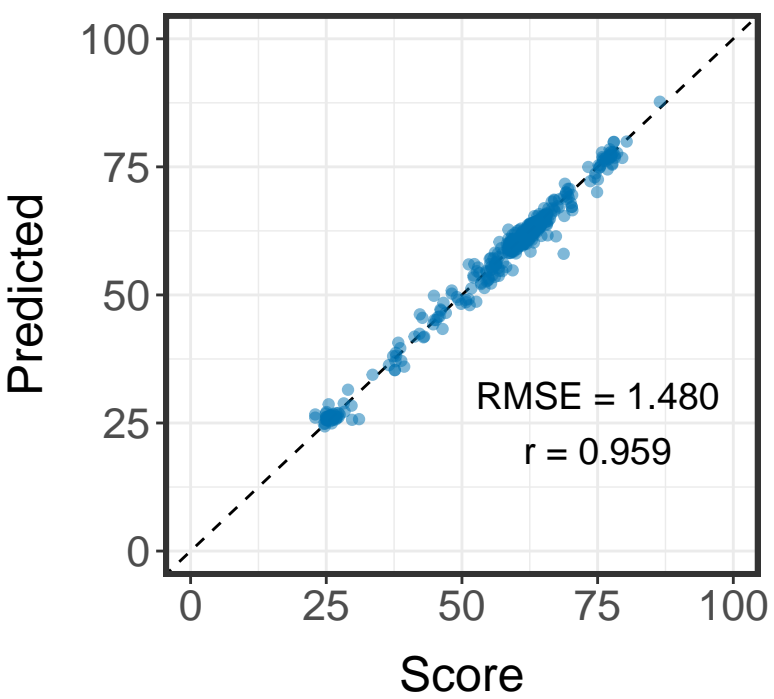


B

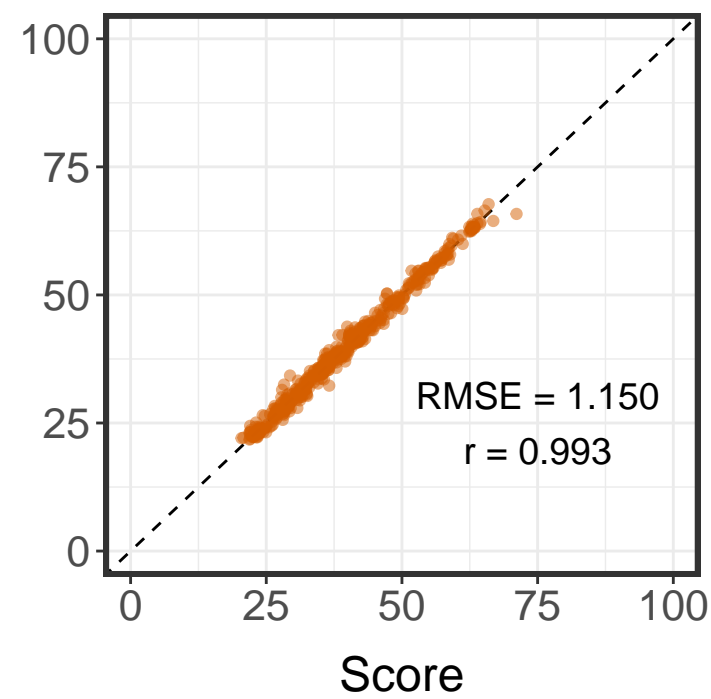
Random



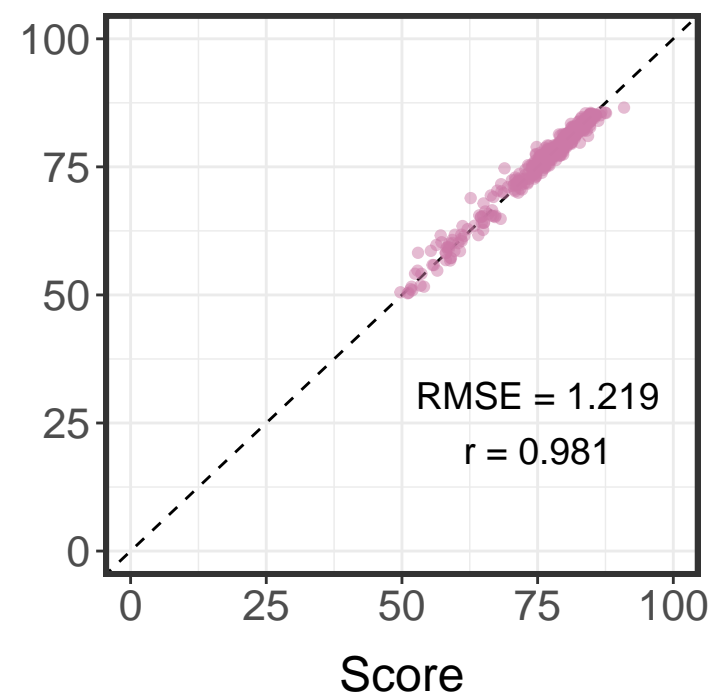
MMLU\* (d = 96)



TruthfulQA\* (d = 154)



WinoGrande\* (d = 133)



c

metabench-A (d = 858)

