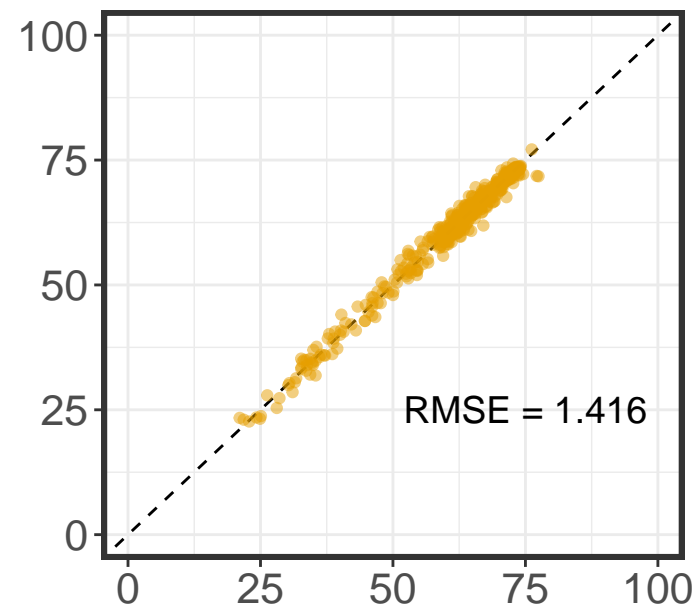
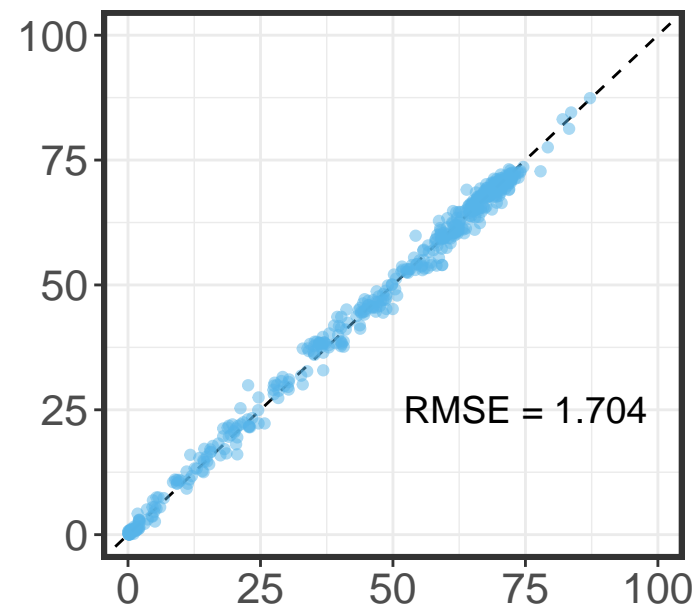


A

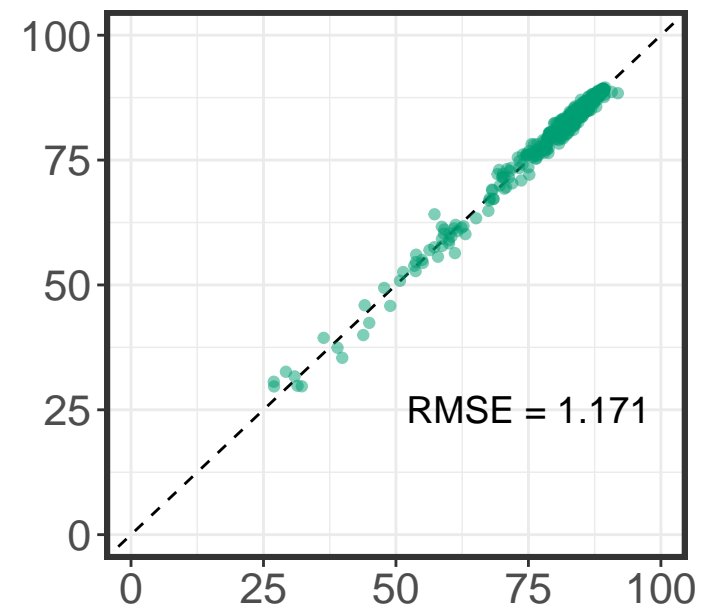
ARC\* (d = 100)



GSM8K\* (d = 237)

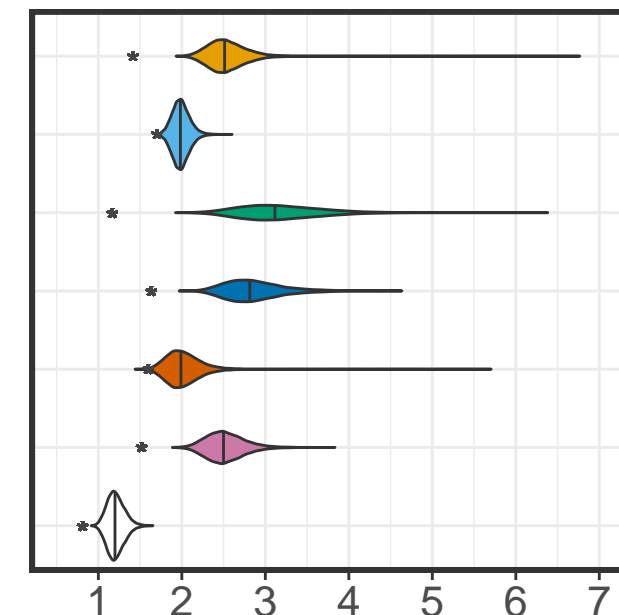


HellaSwag\* (d = 58)

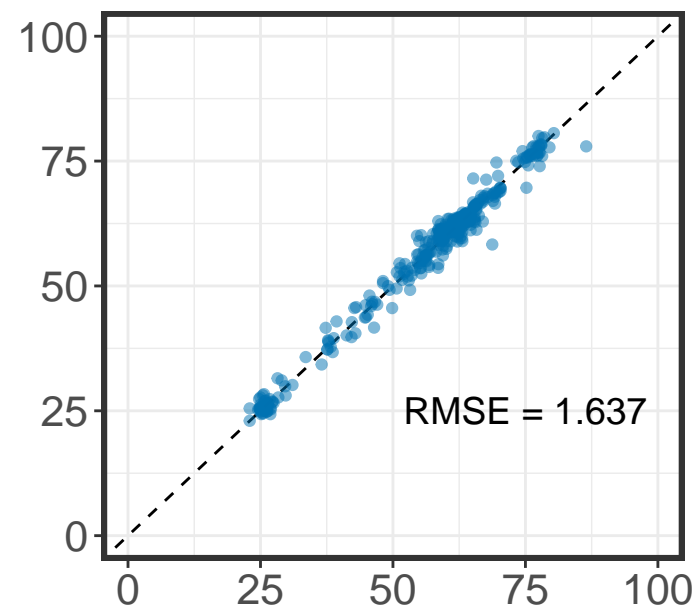


B

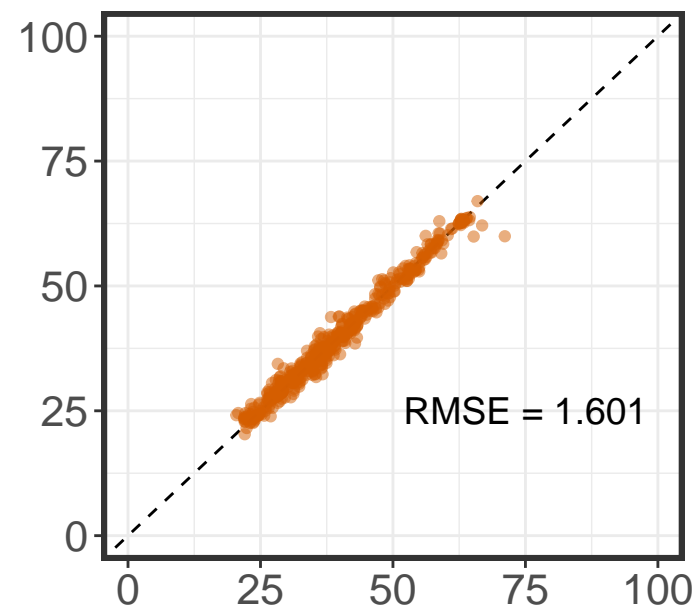
Random



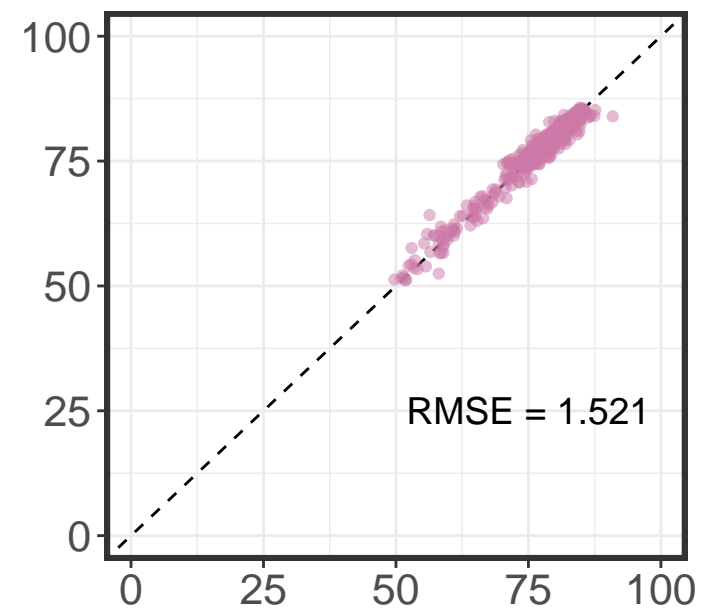
MMLU\* (d = 102)



TruthfulQA\* (d = 136)

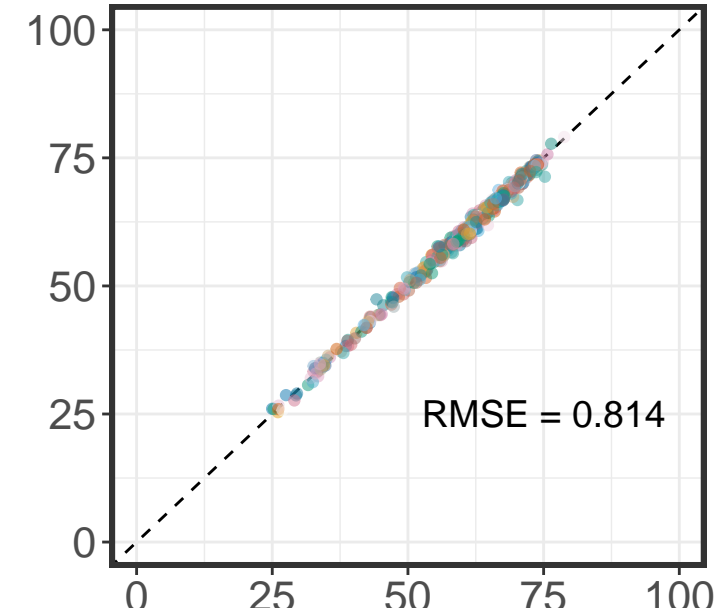


Winogrande\* (d = 106)



C

metabench (d = 739)



Score

Score

Score

Score