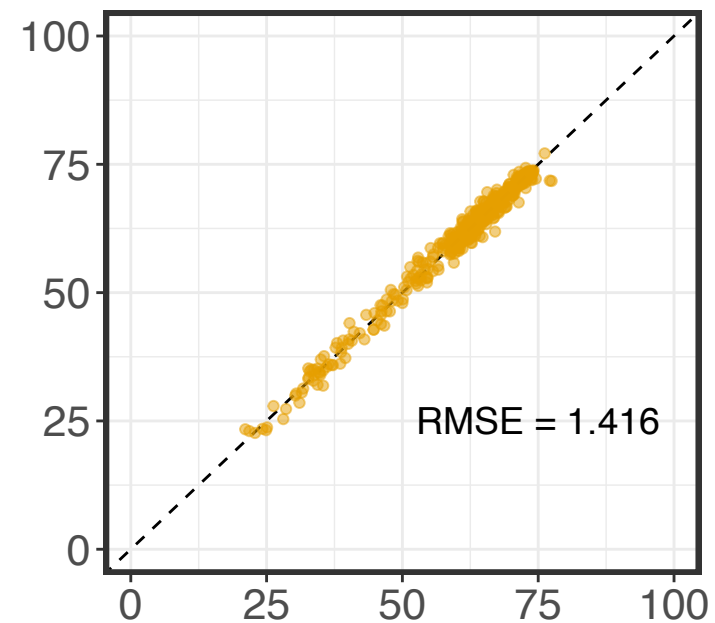
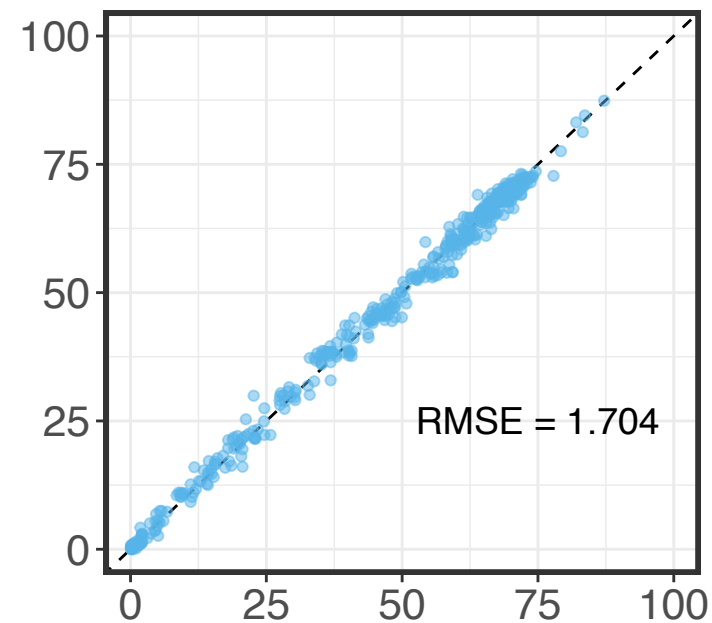


**A**

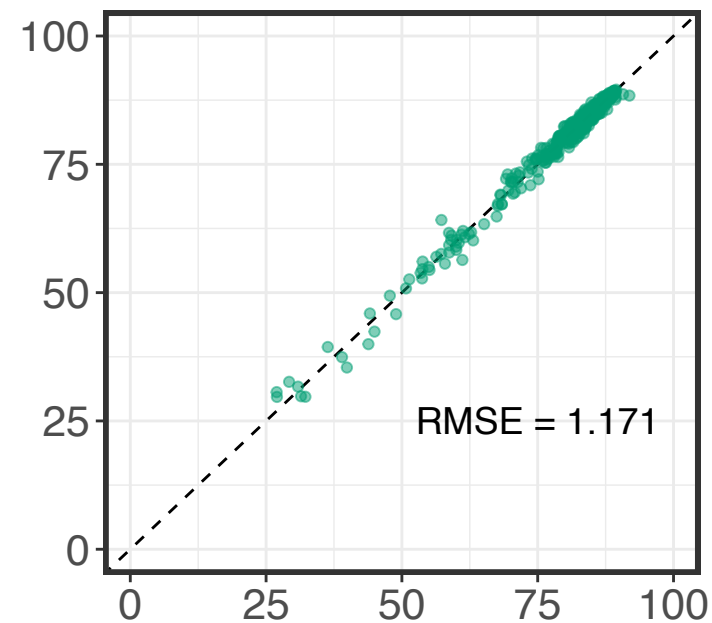
ARC\* (d = 100)



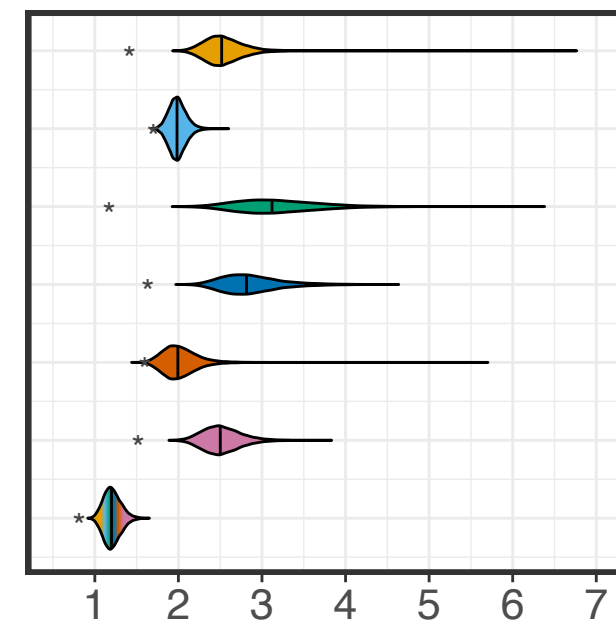
GSM8K\* (d = 237)



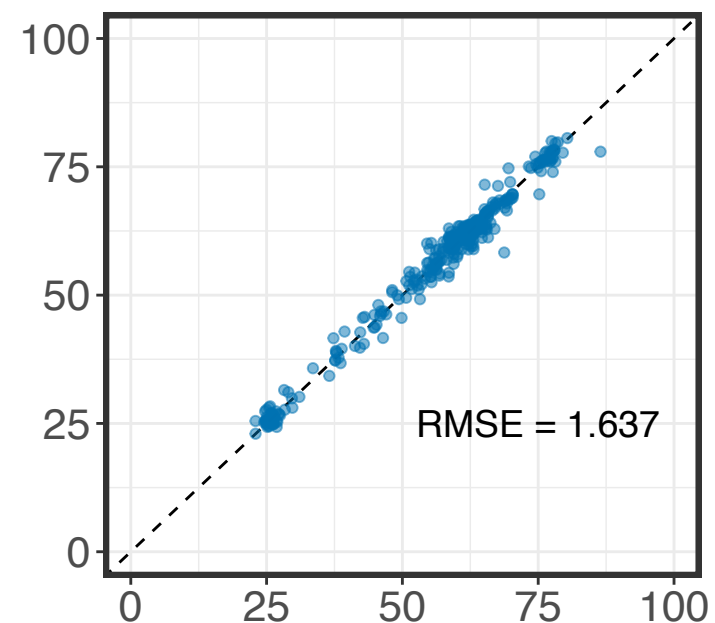
HellaSwag\* (d = 58)

**B**

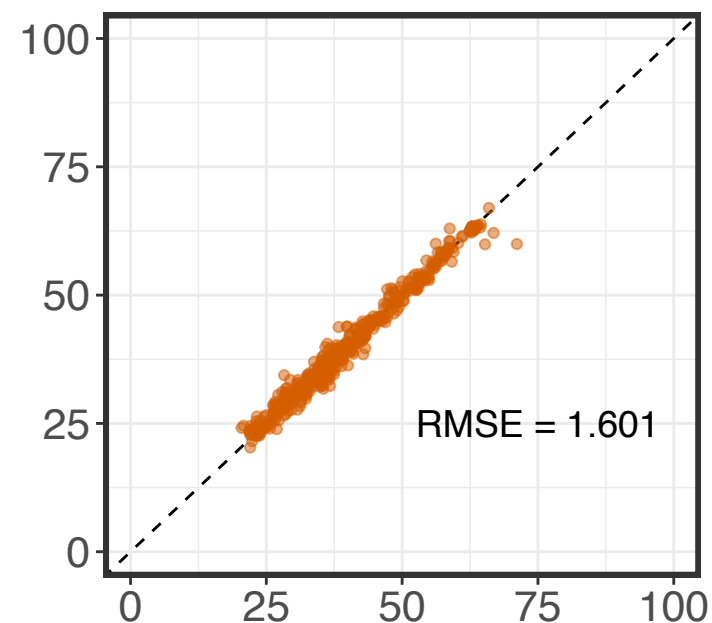
Random



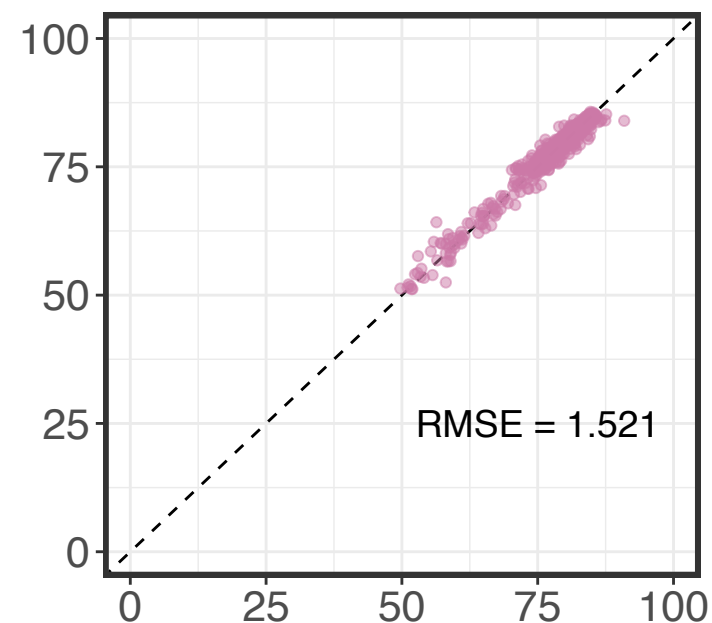
MMLU\* (d = 102)



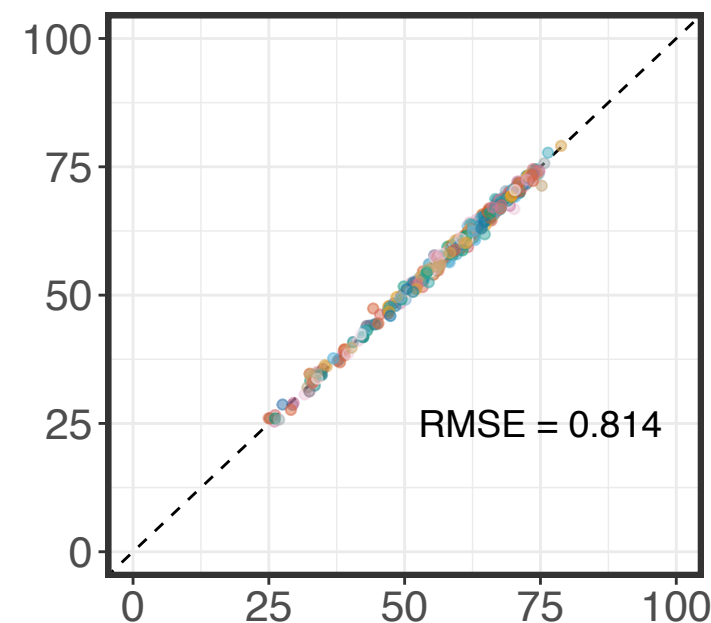
TruthfulQA\* (d = 136)



WinoGrande\* (d = 106)

**c**

metabench (d = 739)



Predicted

Predicted

Score

Score

Score

Score

RMSE