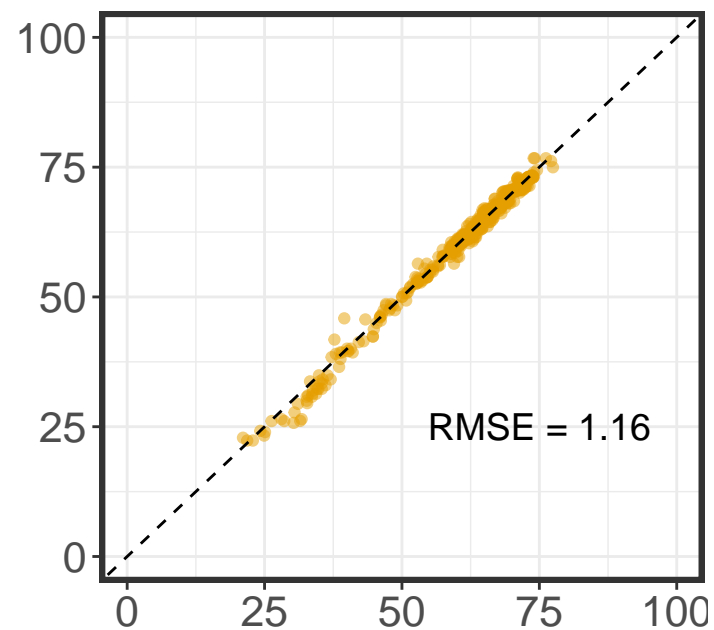
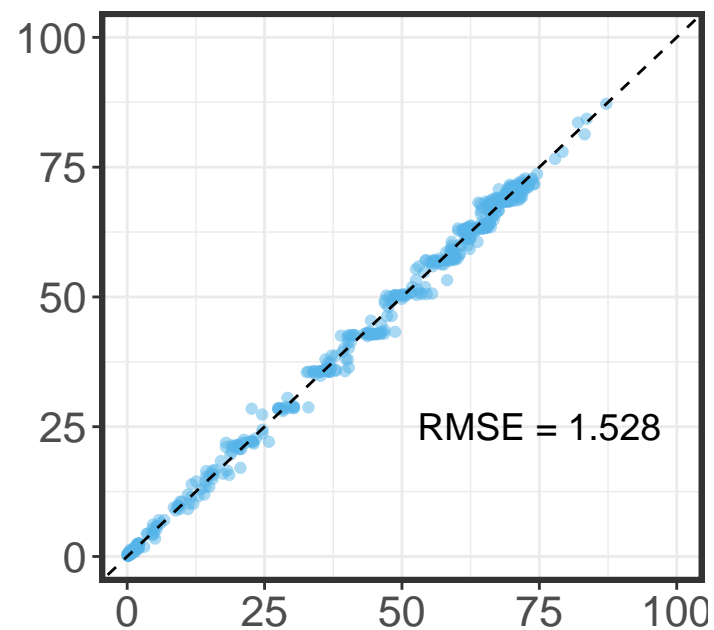


**A**

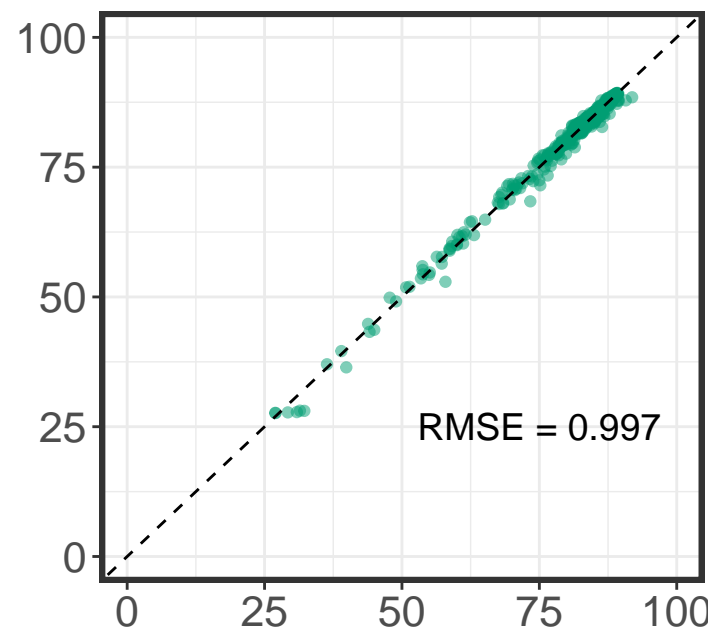
ARC (d = 350)



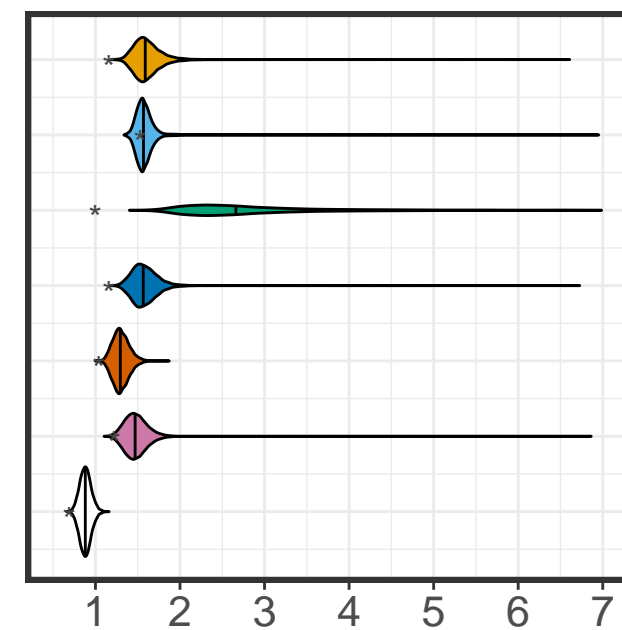
GSM8K (d = 350)



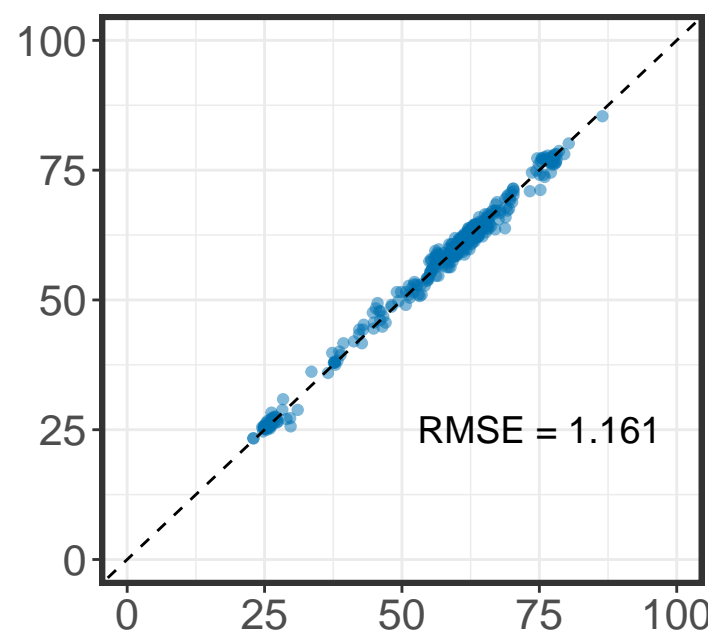
HellaSwag (d = 350)

**B**

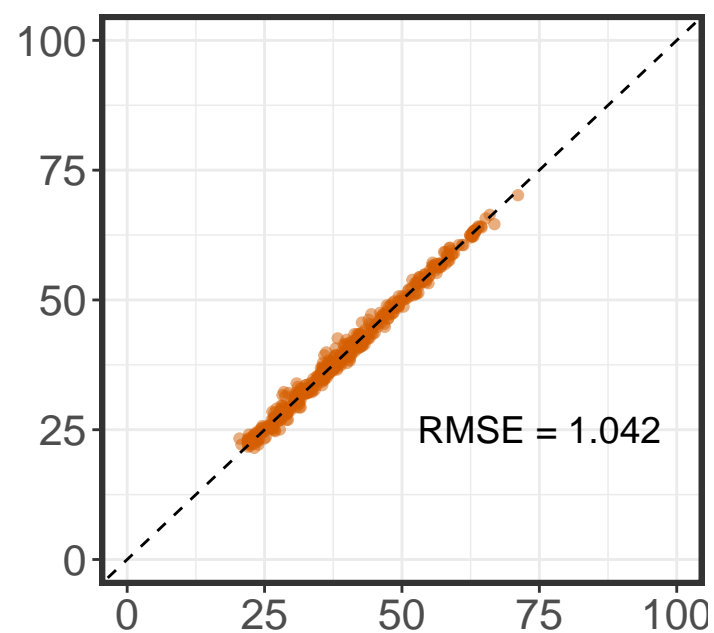
Random



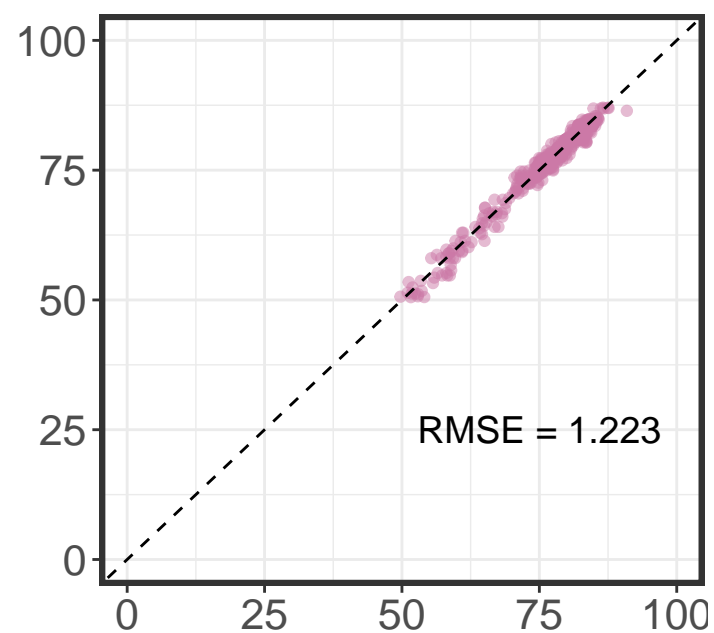
MMLU (d = 350)



TruthfulQA (d = 350)



WinoGrande (d = 350)

**c**

metabench (d = 2100)

