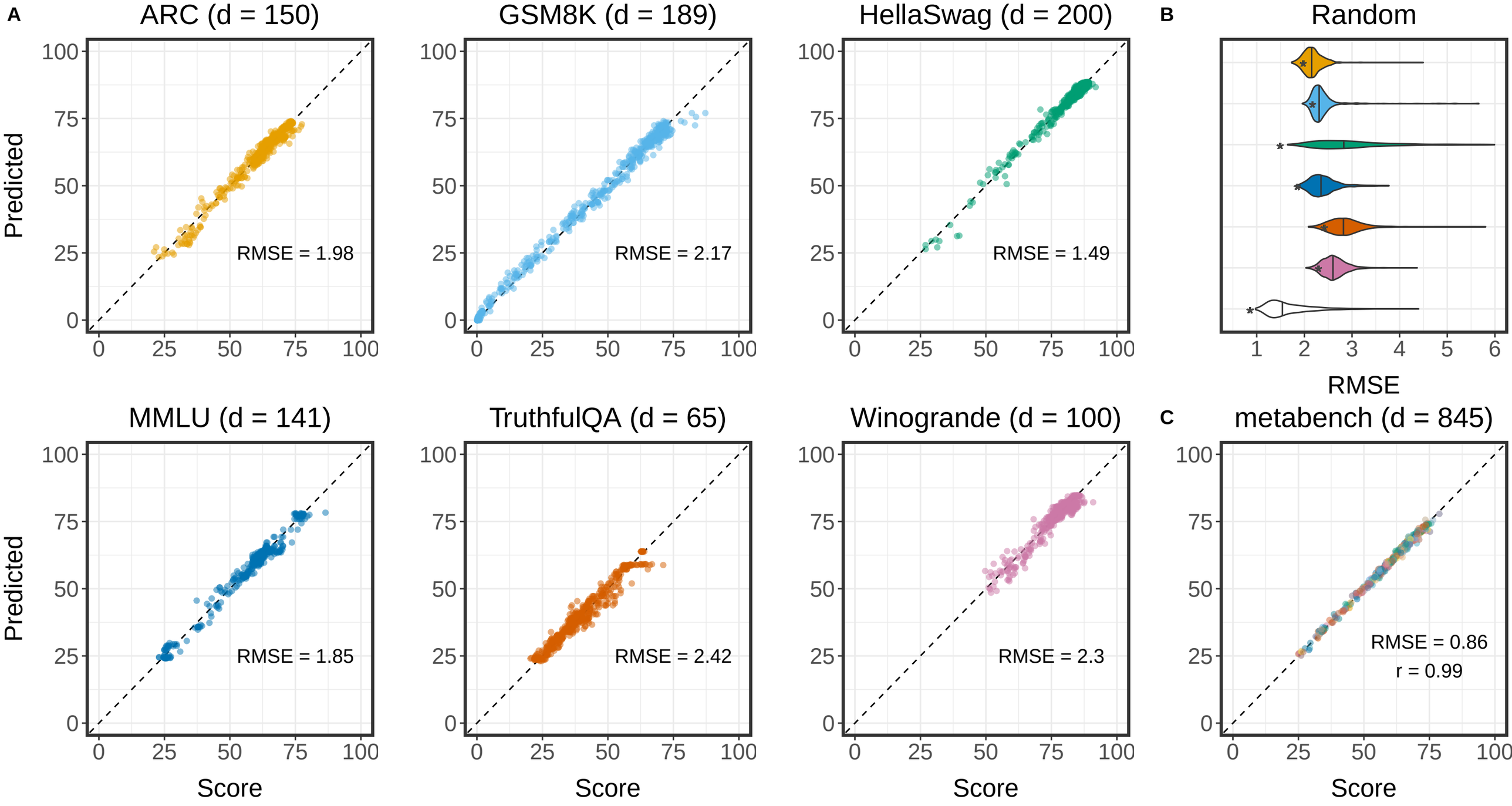
A
ARC (d = 150)
GSM8K (d = 189)
HellaSwag (d = 200)
B
Random
Predicted
RMSE = 1.98
RMSE = 2.17
RMSE = 1.49
RMSE
MMLU (d = 141)
TruthfulQA (d = 65)
Winogrande (d = 100)
C
metabench (d = 845)
RMSE = 1.85
RMSE = 2.42
RMSE = 2.3
RMSE = 0.86
r = 0.99
Score