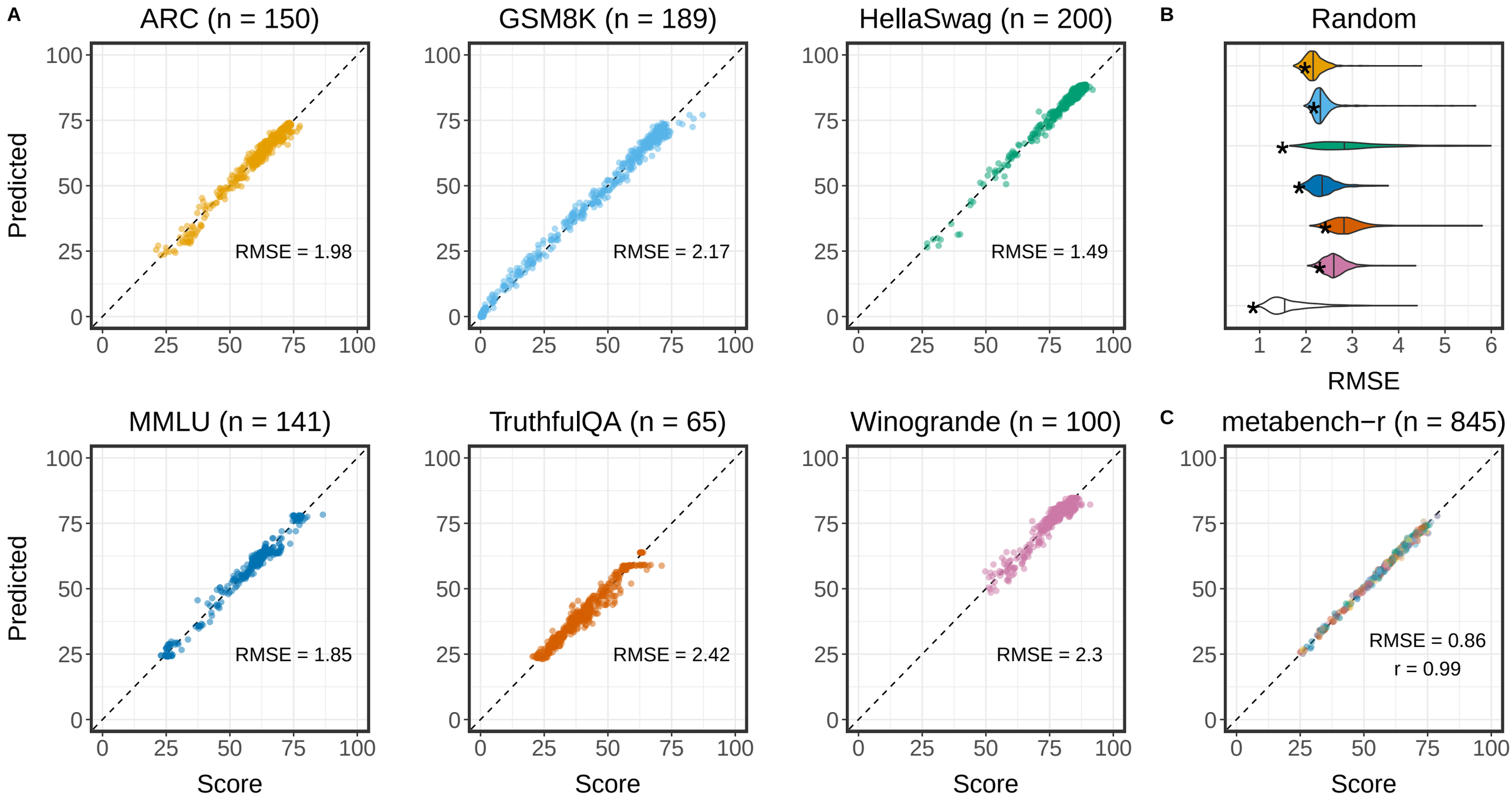
A
ARC (n = 150)
GSM8K (n = 189)
HellaSwag (n = 200)
B
Random
MMLU (n = 141)
TruthfulQA (n = 65)
Winogrande (n = 100)
C
metabench−r (n = 845)
Predicted
Score
RMSE
RMSE = 1.98
RMSE = 2.17
RMSE = 1.49
RMSE = 1.85
RMSE = 2.42
RMSE = 2.3
RMSE = 0.86
r = 0.99