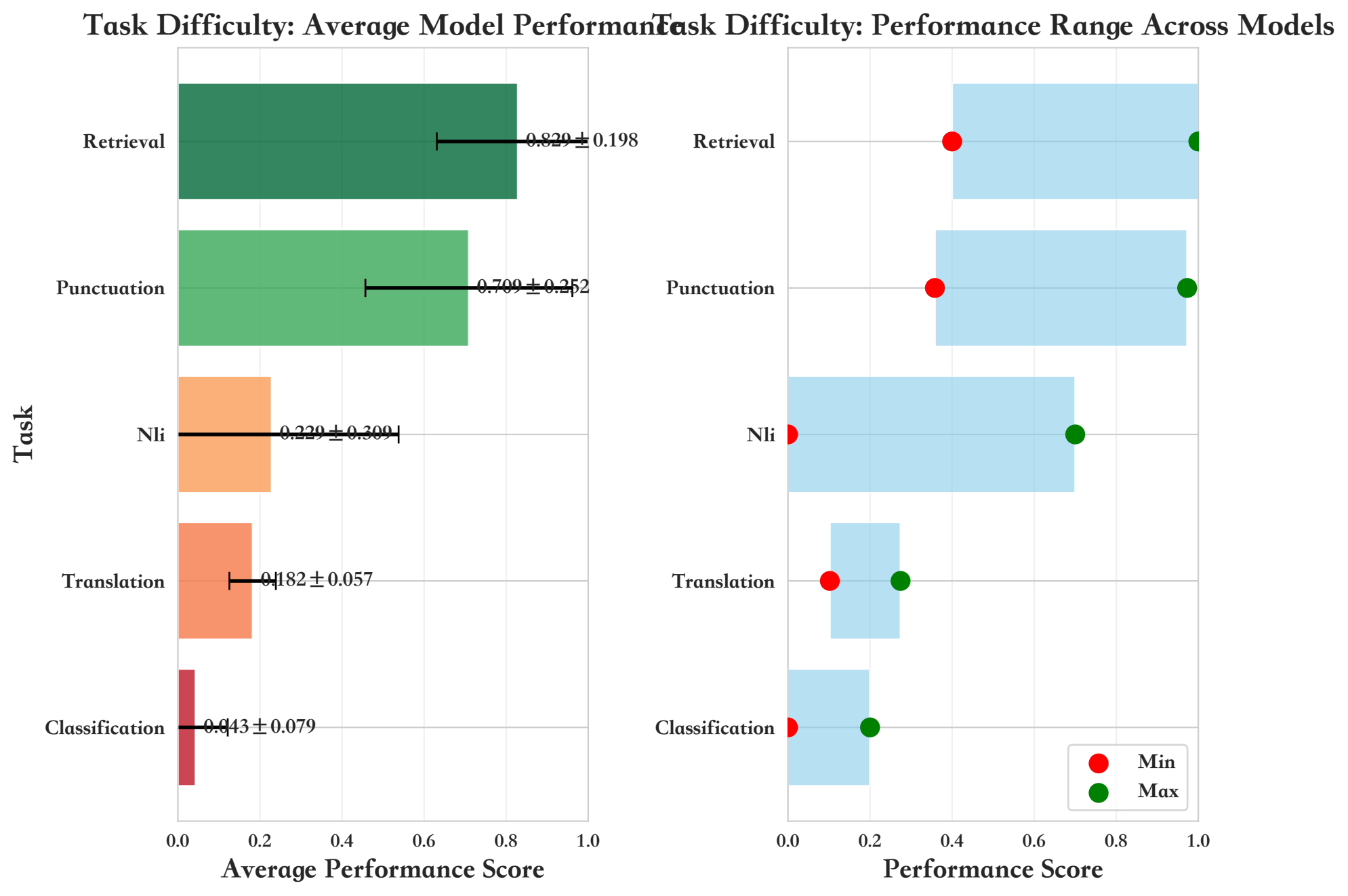
Task Difficulty: Average Model Performance
Task Difficulty: Performance Range Across Models
Task
Retrieval
Punctuation
Nli
Translation
Classification
0.829 ± 0.198
0.709 ± 0.252
0.229 ± 0.309
0.182 ± 0.057
0.043 ± 0.079
0.0
0.2
0.4
0.6
0.8
1.0
Average Performance Score
Performance Score
Min
Max