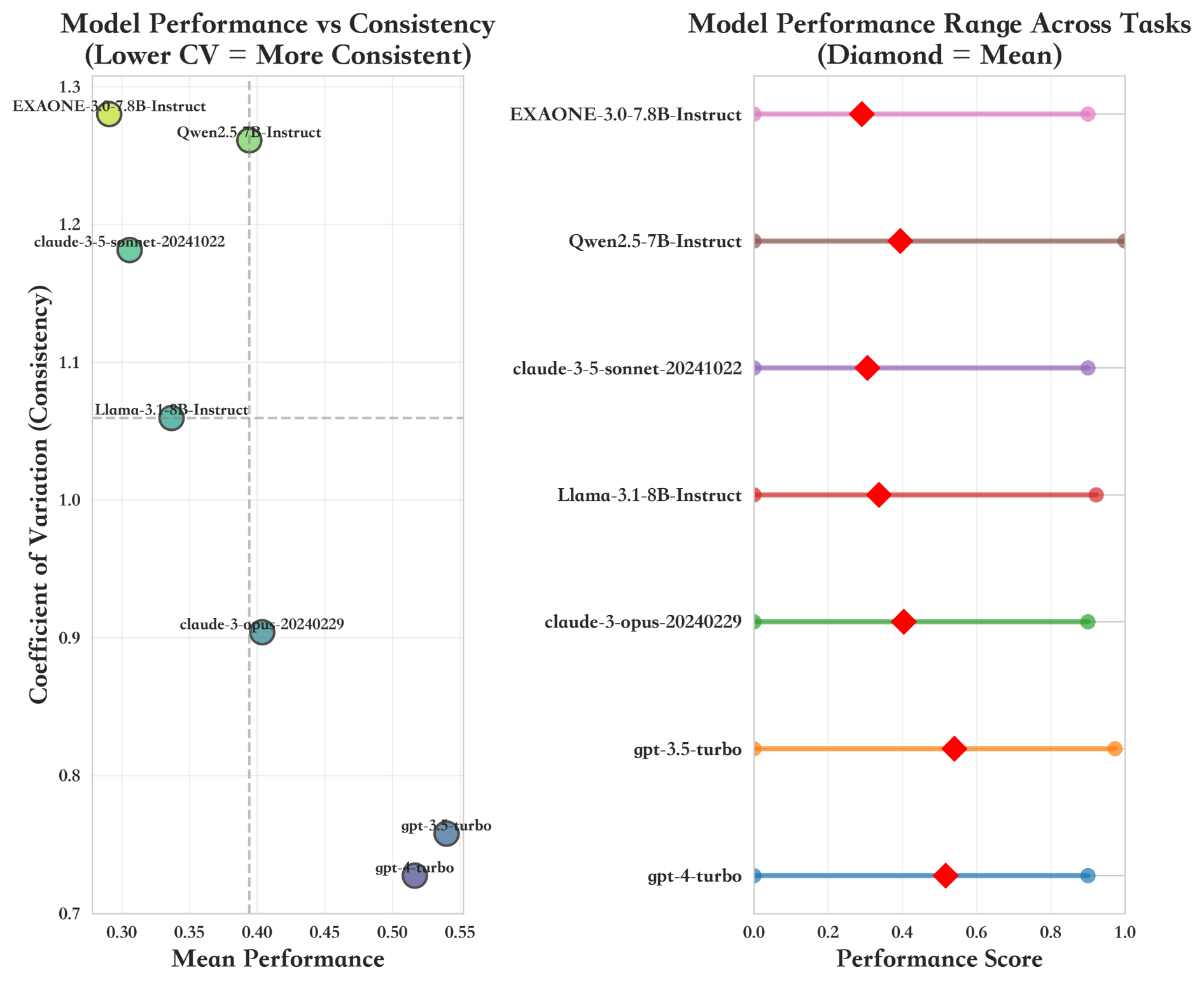
Model Performance vs Consistency
(Lower CV = More Consistent)
Coefficient of Variation (Consistency)
Mean Performance
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.30
0.35
0.40
0.45
0.50
0.55
EXAONE-3.0-7.8B-Instruct
Qwen2.5-7B-Instruct
claude-3-5-sonnet-20241022
Llama-3.1-8B-Instruct
claude-3-opus-20240229
gpt-3.5-turbo
gpt-4-turbo
Model Performance Range Across Tasks
(Diamond = Mean)
Performance Score
0.0
0.2
0.4
0.6
0.8
1.0
EXAONE-3.0-7.8B-Instruct
Qwen2.5-7B-Instruct
claude-3-5-sonnet-20241022
Llama-3.1-8B-Instruct
claude-3-opus-20240229
gpt-3.5-turbo
gpt-4-turbo