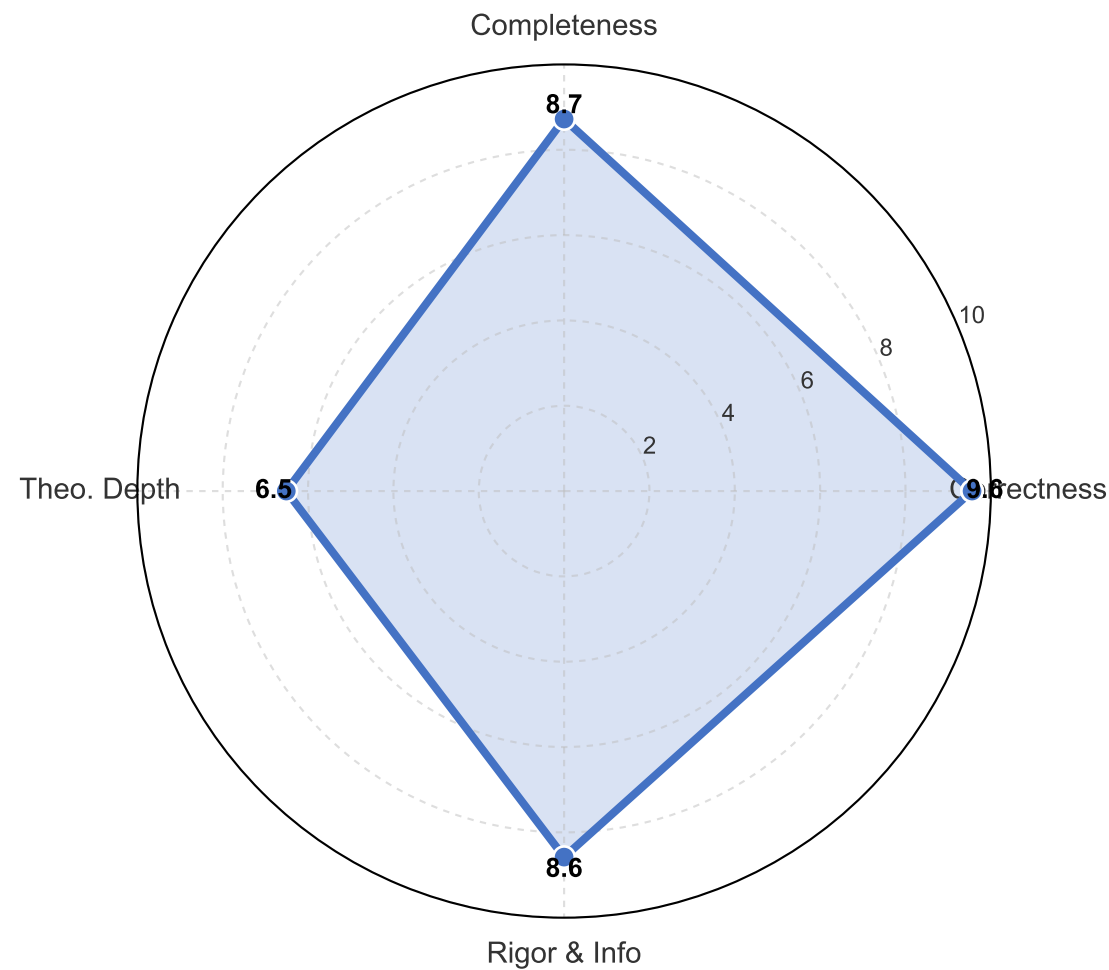
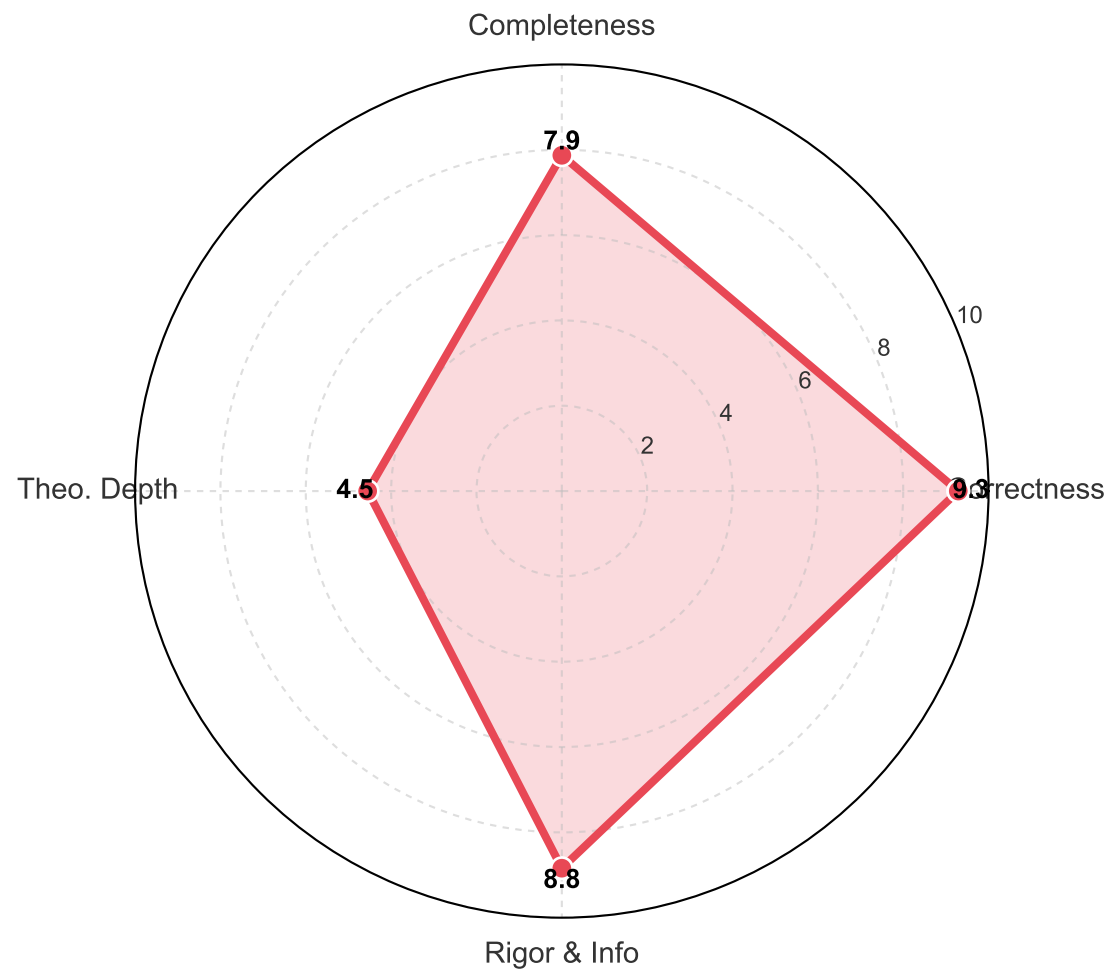


Individual Model Performance Analysis (Performance Scores by Dimension)

MOSES
(Avg: 8.34)



O3
(Avg: 7.64)



GPT-4.1
(Avg: 6.48)

