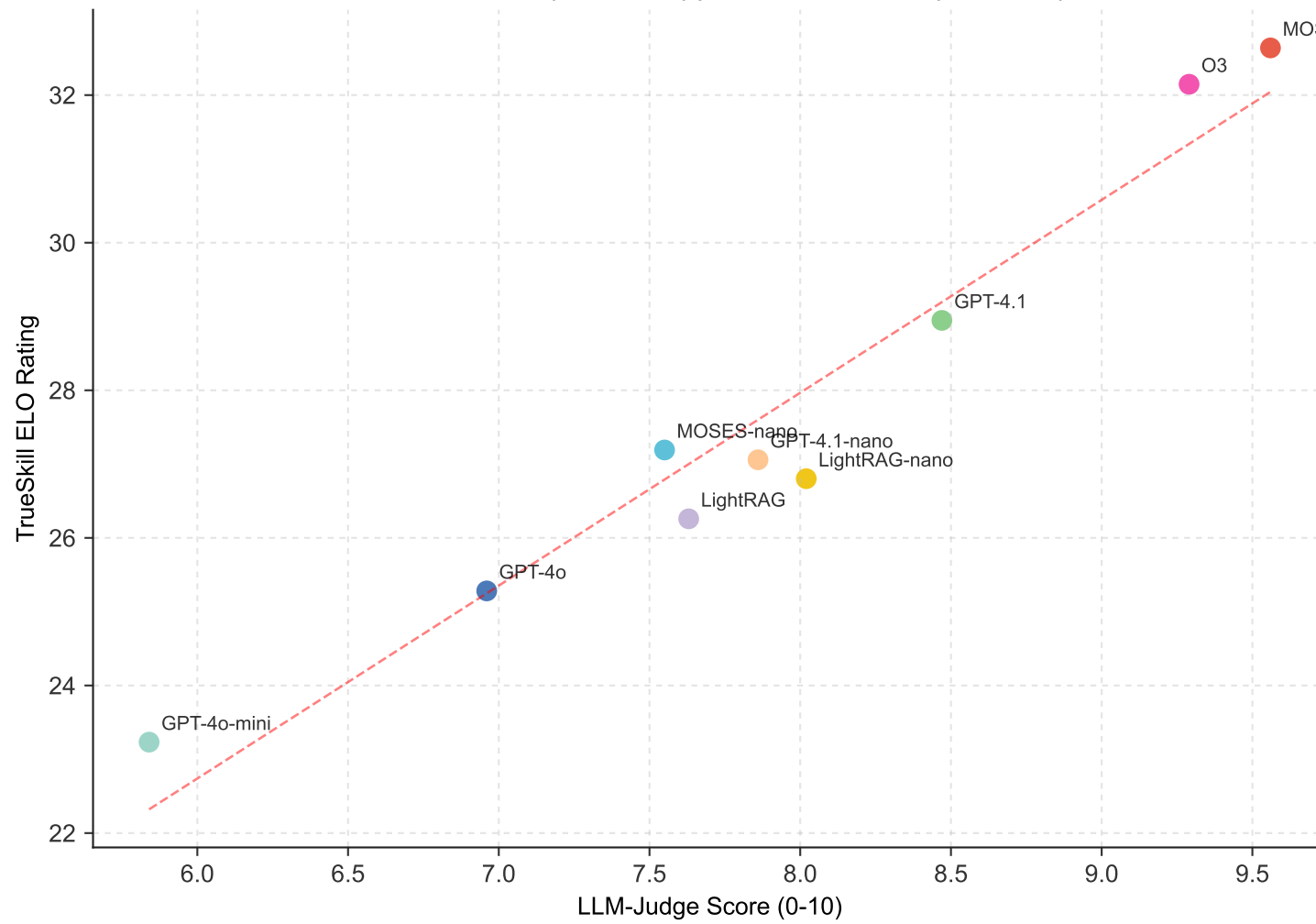


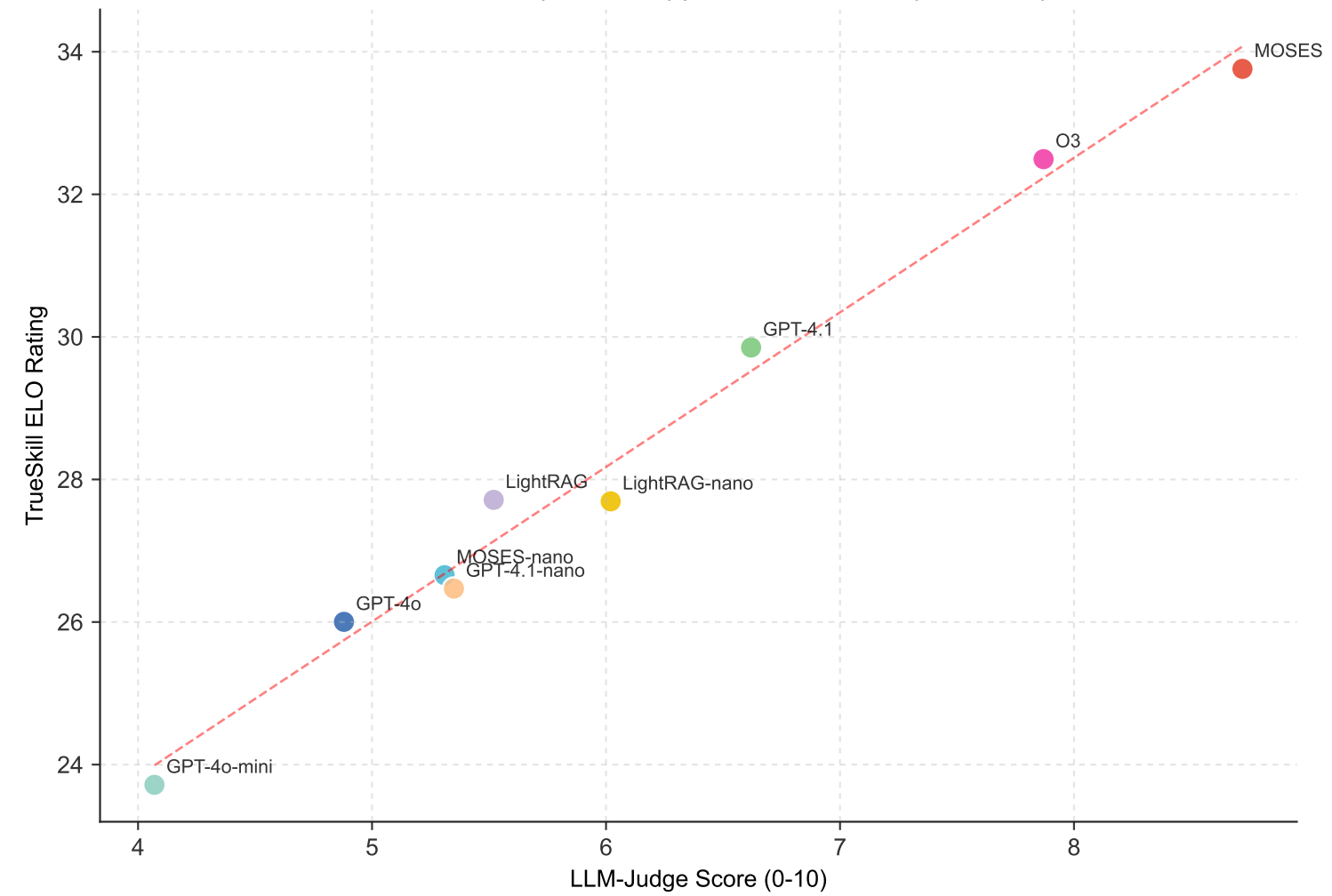
# LLM-Judge Score vs TrueSkill ELO

## (Dimension-wise Correlations: Pearson + Spearman)

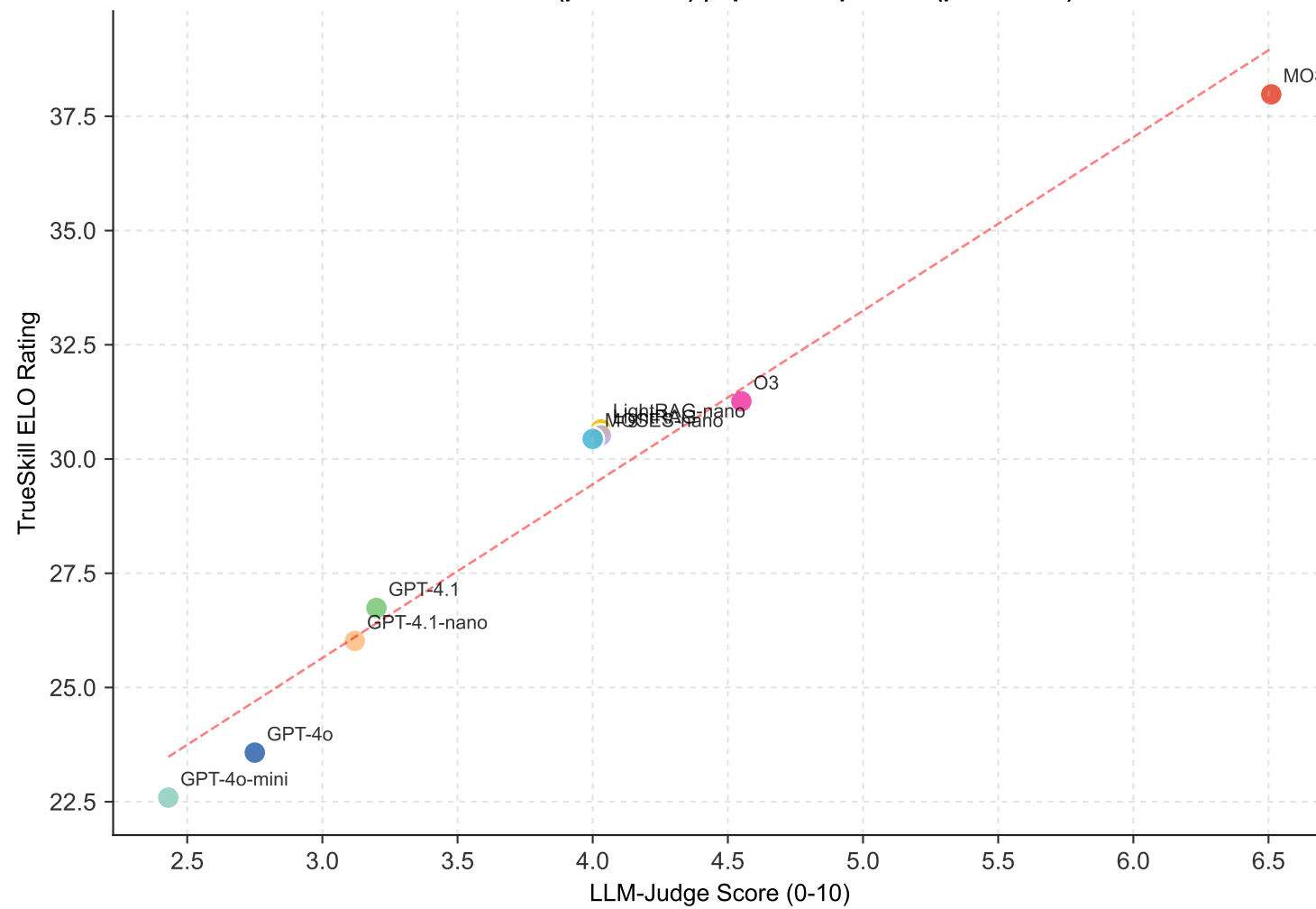
**Correctness**  
Pearson  $r=0.970$  ( $p=1.43e-05$ ) | Spearman  $\rho=0.883$  ( $p=0.00159$ )



**Completeness**  
Pearson  $r=0.993$  ( $p=8.35e-08$ ) | Spearman  $\rho=0.967$  ( $p=2.16e-05$ )



**Theo. Depth**  
Pearson  $r=0.982$  ( $p=2.54e-06$ ) | Spearman  $\rho=0.996$  ( $p=1.54e-08$ )



**Rigor & Info**  
Pearson  $r=0.989$  ( $p=4.99e-07$ ) | Spearman  $\rho=0.983$  ( $p=1.94e-06$ )

