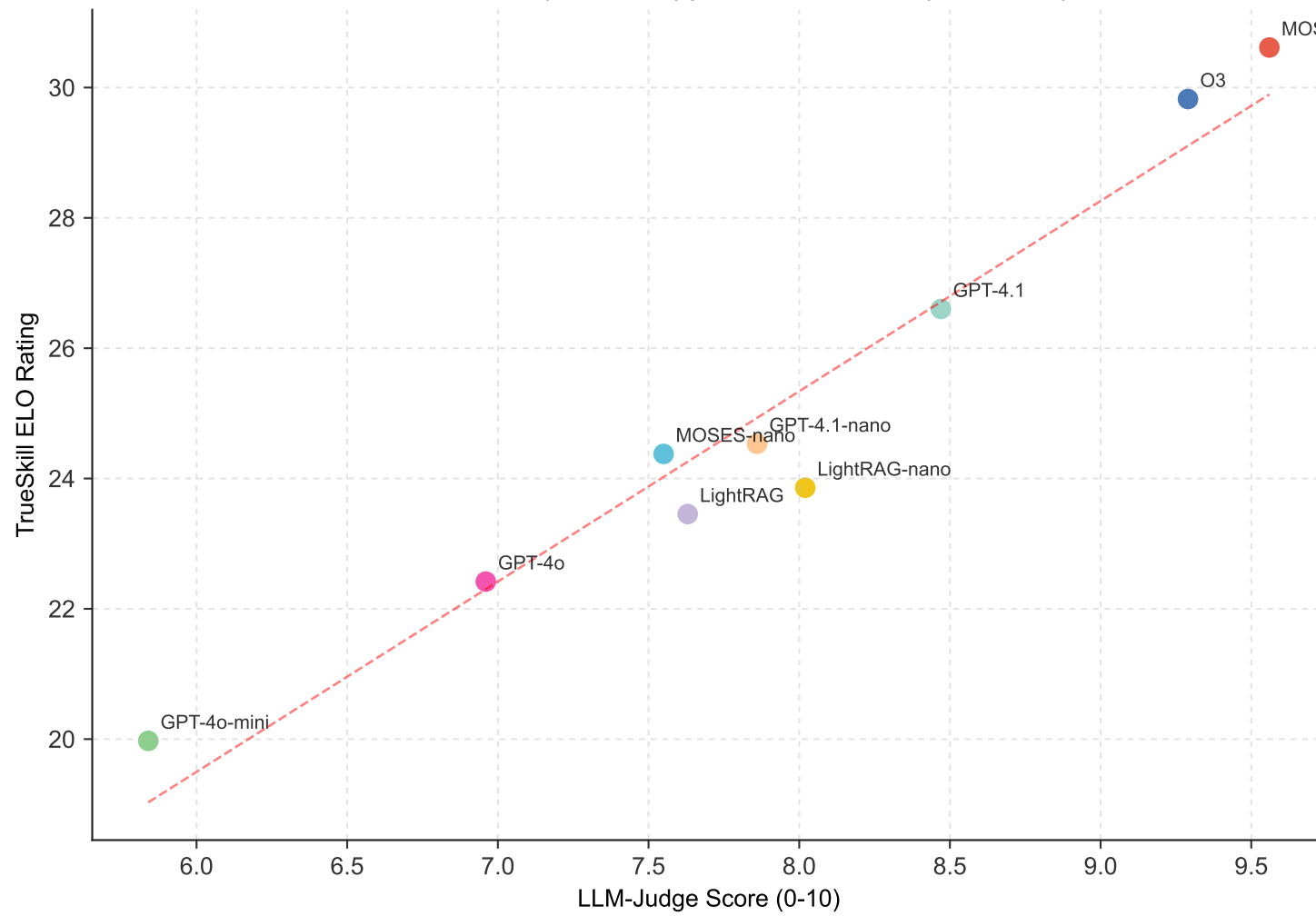


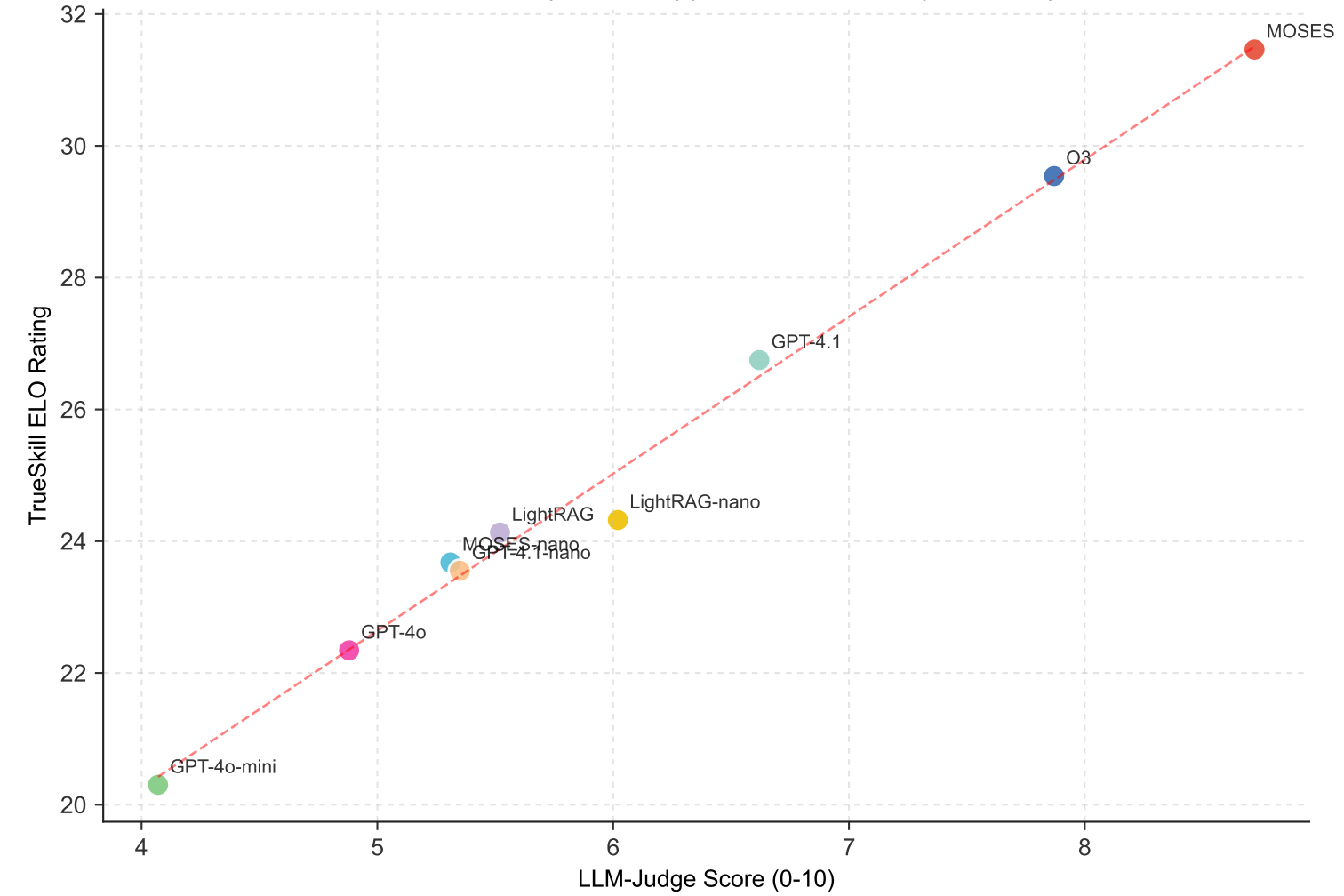
# LLM-Judge Score vs TrueSkill ELO

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

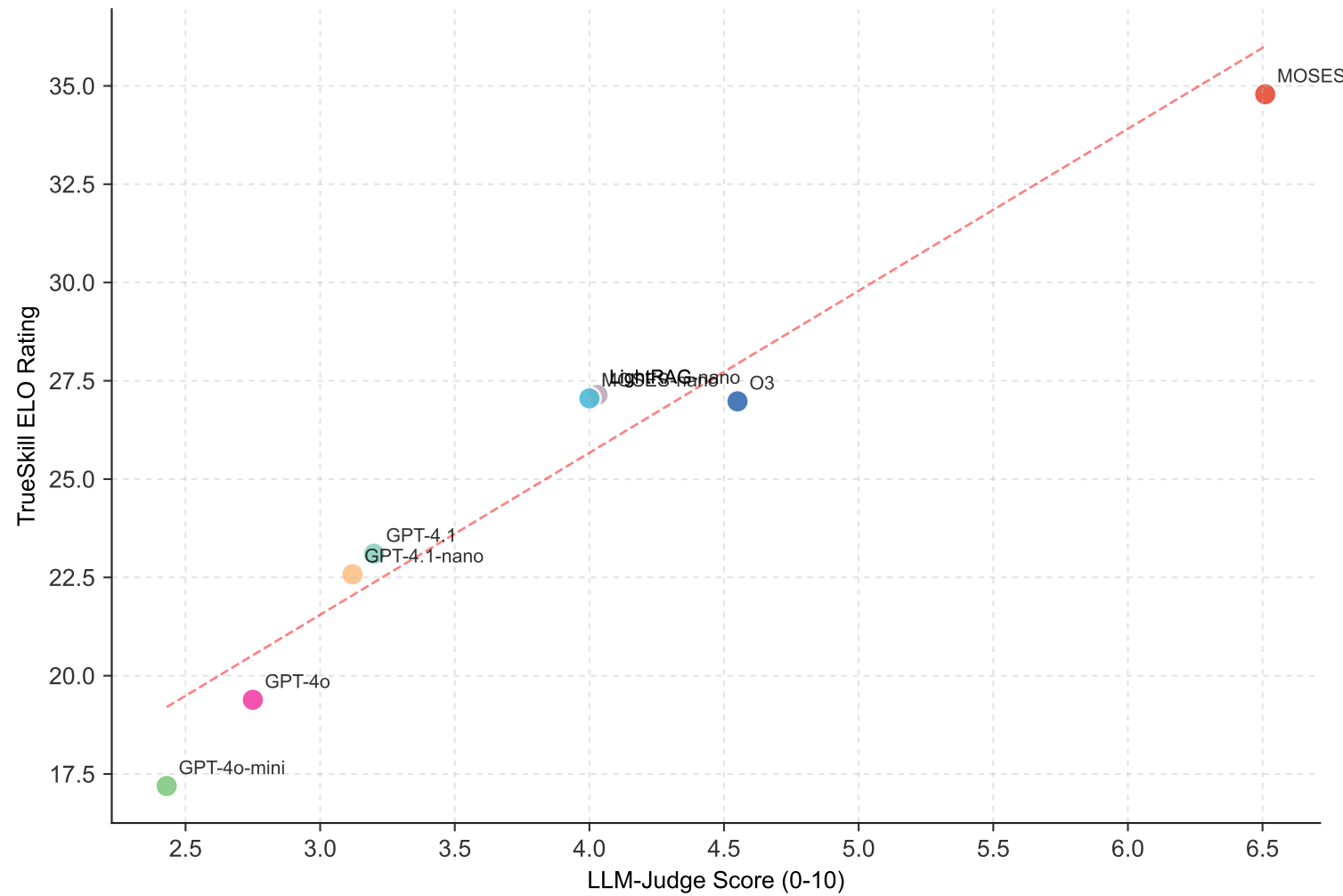
**Correctness**  
Pearson  $r=0.972$  ( $p=1.24e-05$ ) | Spearman  $\rho=0.917$  ( $p=0.000507$ )



**Completeness**  
Pearson  $r=0.996$  ( $p=1.37e-08$ ) | Spearman  $\rho=0.983$  ( $p=1.94e-06$ )



**Theo. Depth**  
Pearson  $r=0.967$  ( $p=2.13e-05$ ) | Spearman  $\rho=0.895$  ( $p=0.0011$ )



**Rigor & Info**  
Pearson  $r=0.995$  ( $p=3.59e-08$ ) | Spearman  $\rho=0.983$  ( $p=1.94e-06$ )

