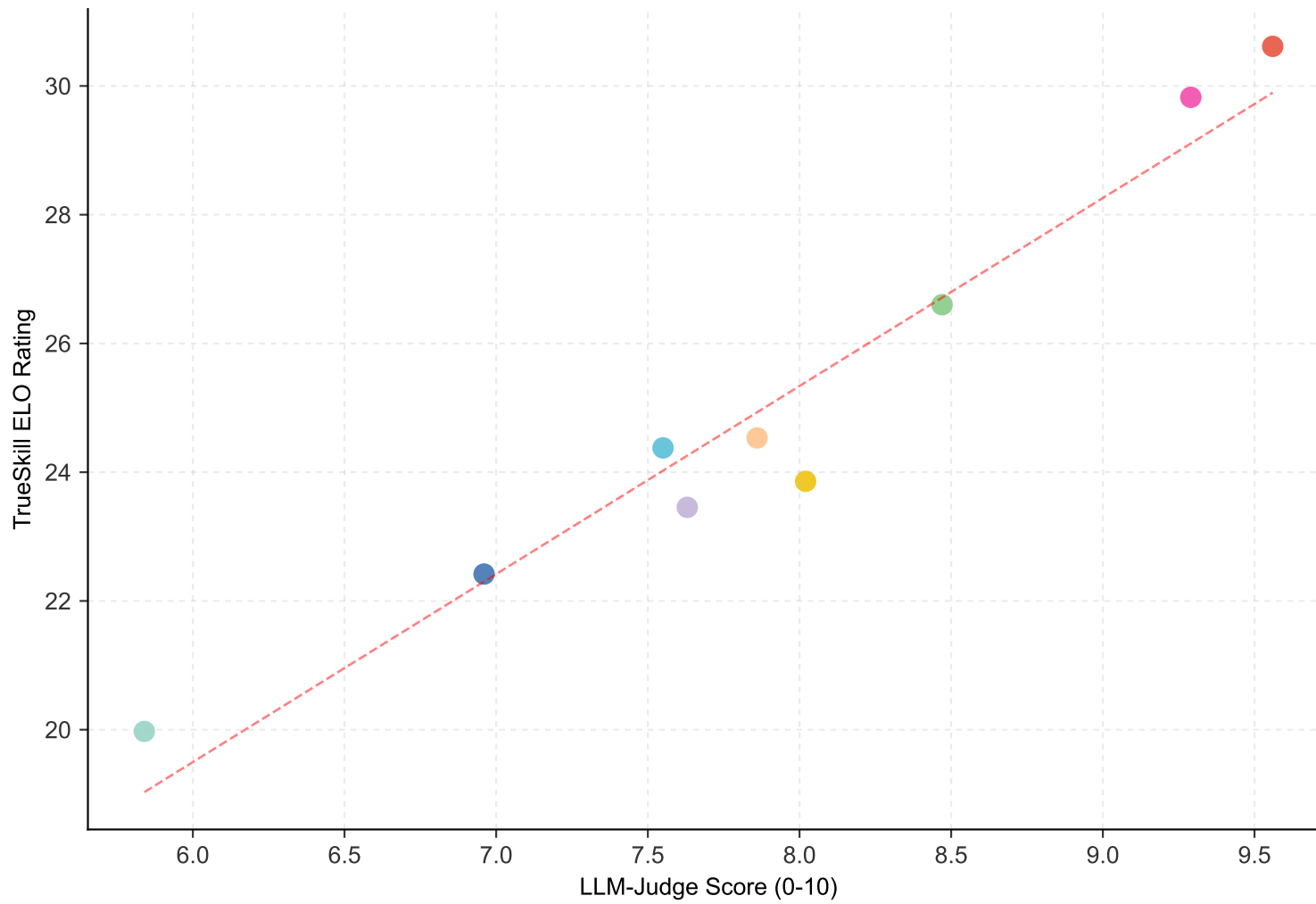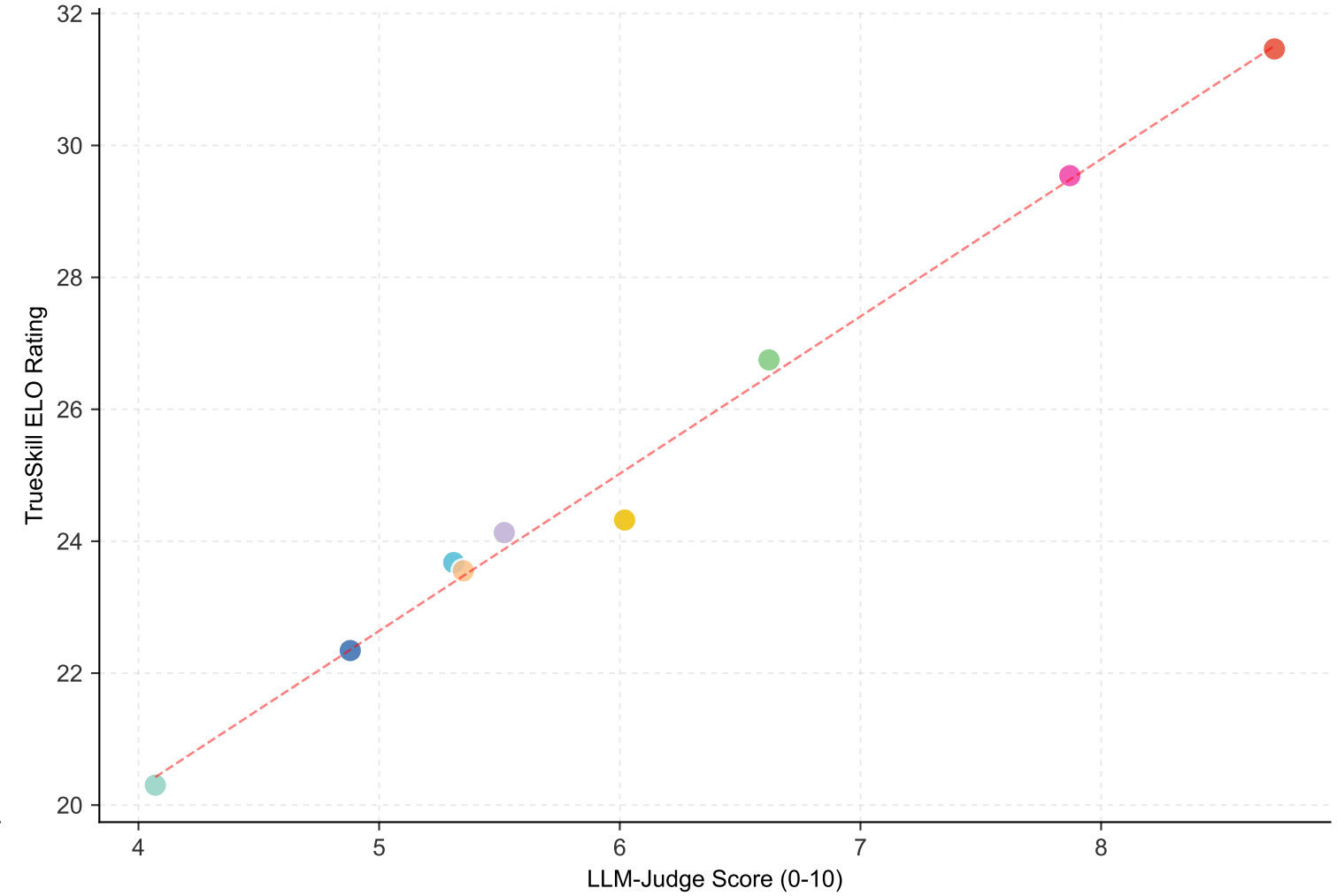**LLM-Judge Score vs TrueSkill ELO**
(Pearson + Spearman + Kendall τ, per Dimension)
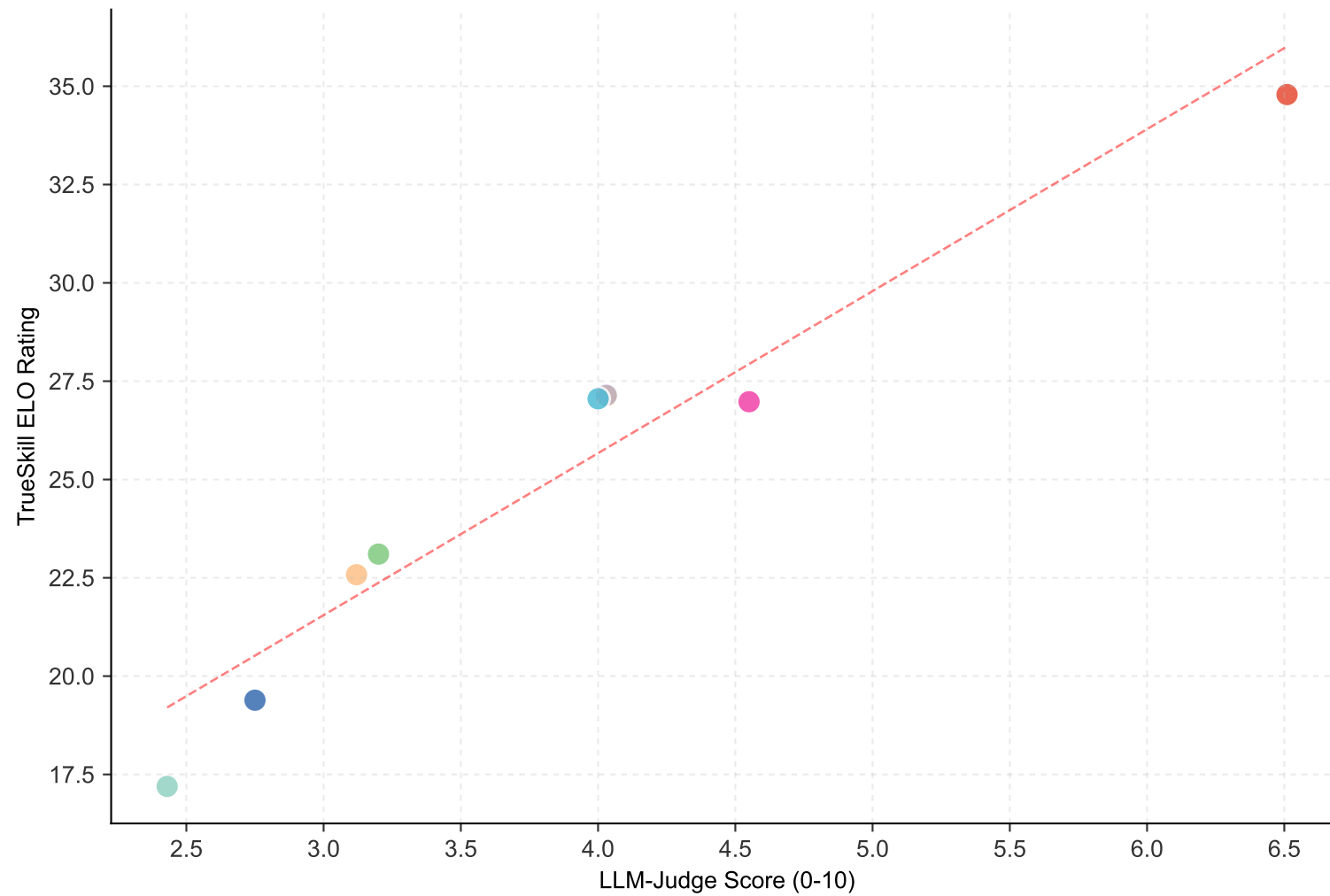
Correctness
Pearson r=0.972 (p=1.24e-05) | Spearman ρ=0.917 (p=0.000507) | Kendall τ=0.833 (p=0.000854)

Completeness
Pearson r=0.996 (p=1.37e-08) | Spearman ρ=0.983 (p=1.94e-06) | Kendall τ=0.944 (p=4.96e-05)

Theo. Depth
Pearson r=0.967 (p=2.13e-05) | Spearman ρ=0.895 (p=0.0011) | Kendall τ=0.817 (p=0.00237)

Rigor & Info
Pearson r=0.995 (p=3.59e-08) | Spearman ρ=0.983 (p=1.94e-06) | Kendall τ=0.944 (p=4.96e-05)