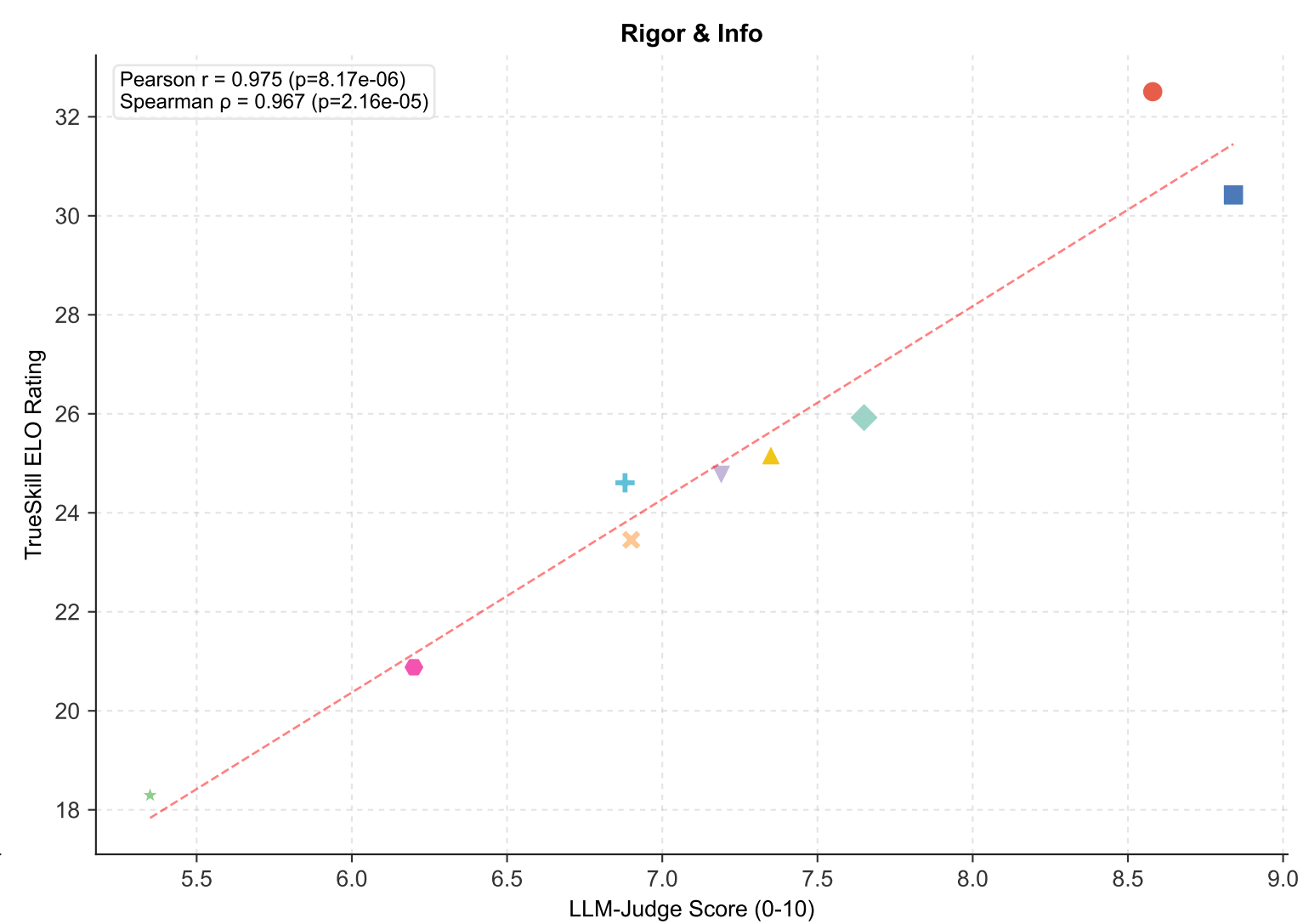
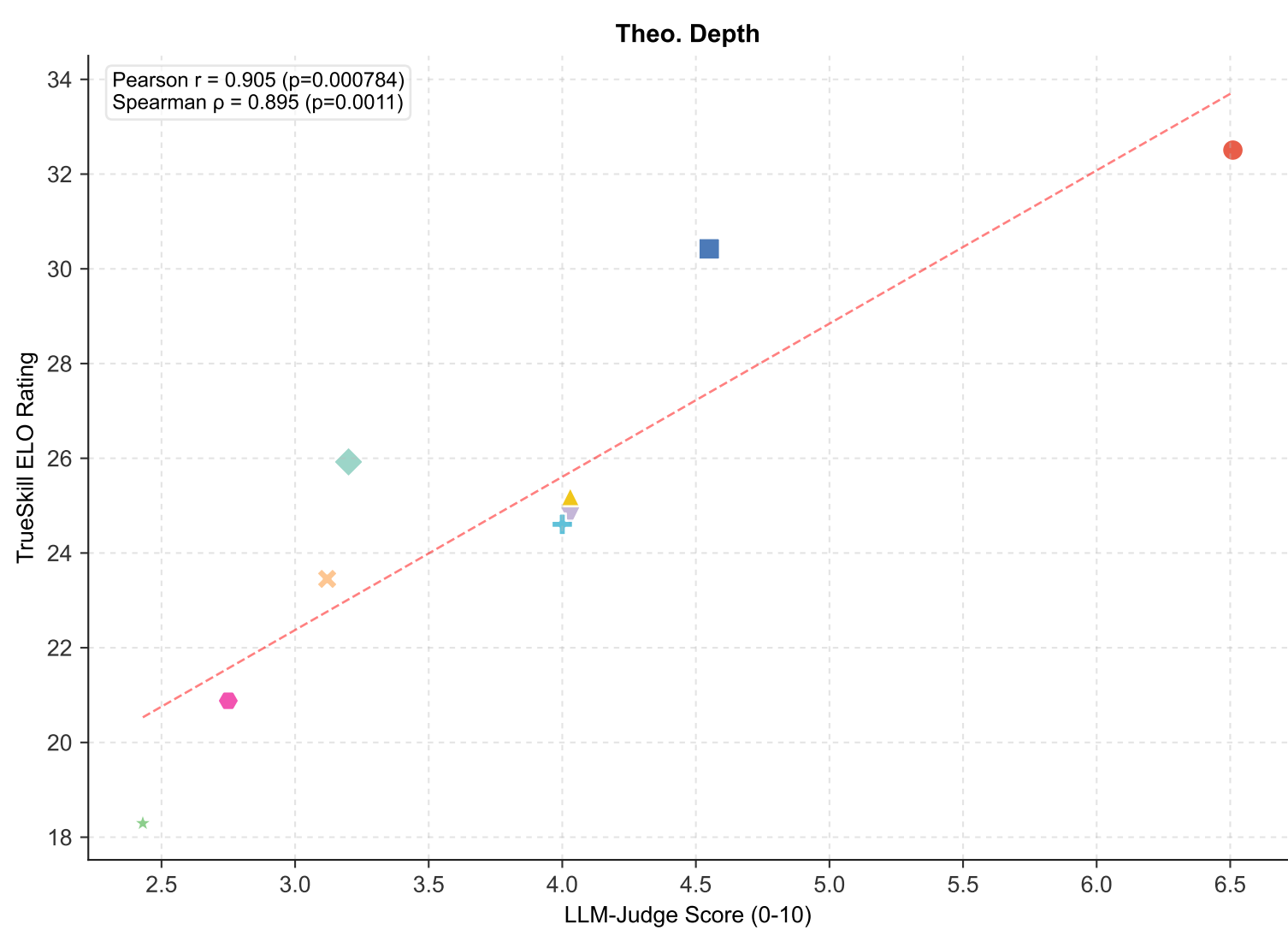
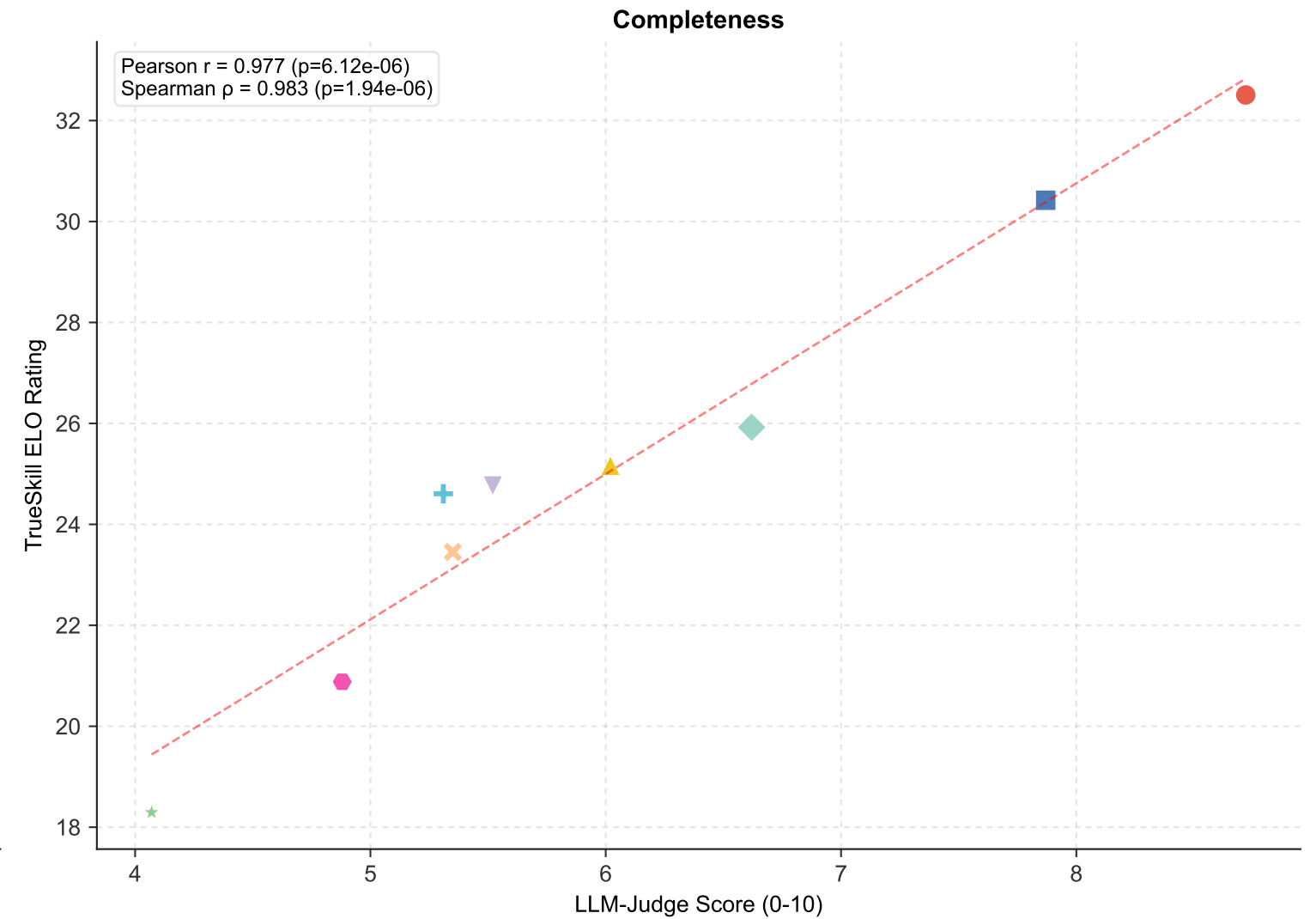
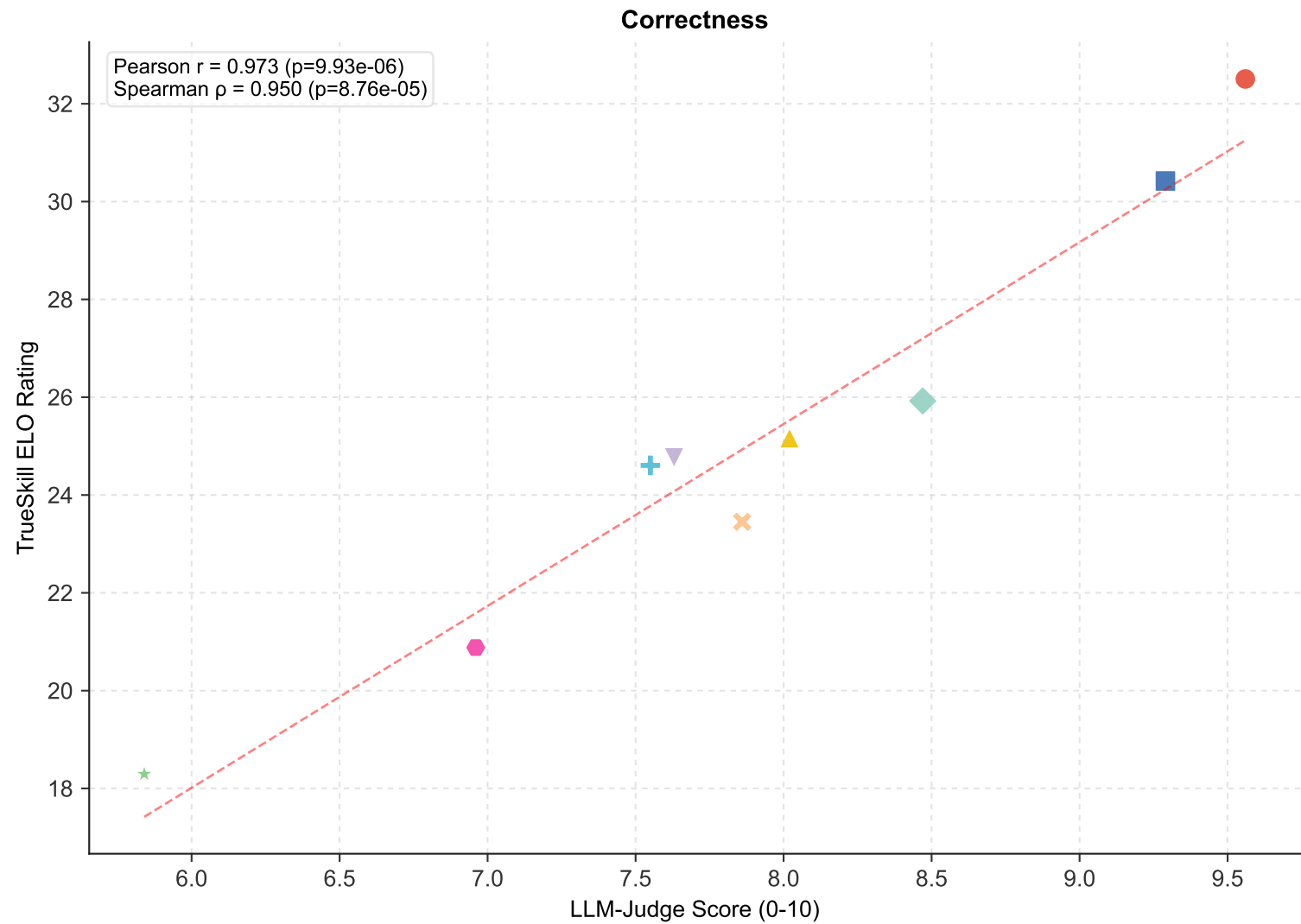


LLM-Judge Score vs TrueSkill ELO
(Pearson + Spearman; ELO from rep20 overall)



- Models
- MOSES
 - O3
 - GPT-4.1
 - LightRAG-nano
 - LightRAG
 - MOSES-nano
 - GPT-4.1-nano
 - GPT-4o
 - GPT-4o-mini