Performance Score vs. TrueSkill ELO Rating (Dimension-wise Correlation Analysis) Correctness (r = 0.921) Completeness (r = 0.974) MOSES 32 MOSES 03 30 28 28 26 TrueSkill ELO Rating GPT-4.1 TrueSkill ELO Rating 95 GPT-4.1-nano LightRAG LightRAG-nano LightRAG-nano LightRAG GPT-40 GPT-40 22 20 MOSES-nano MOSES-nano GPT-4o-mini 20 18 GPT-4o-mini 16 18 6.0 6.5 7.0 7.5 8.0 8.5 9.0 9.5 5 8 Performance Score (0-10) Performance Score (0-10) Theo. Depth Rigor & Info (r = 0.942)(r = 0.986)O3 40 40 MOSES MOSES 35 35 TrueSkill ELO Rating TrueSkill ELO Rating 52 GPT-4.1 LightRAG-nano GPT-4.1 NADSE4S1naanoo GPT-4.1-nano 20 20 GPT-4o-mini GPT-4o 15 4.5 6.0 6.5 7.0 9.0 2.5 3.0 3.5 4.0 5.0 5.5 6.0 6.5 7.5 8.0 8.5 5.5

Performance Score (0-10)

Performance Score (0-10)