Consensus Strategy Comparison

GSM8K

Accuracy (%)

Legend: Standard, Majority, ELO, Mech. Maj., Best Indiv.

Categories: Band-A (3), Band-A (4), OpenAI (3), Cross-Prov., Band-B (3)

TruthfulQA

Accuracy (%)

Categories: Band-A (3), Band-A (4), OpenAI (3), Cross-Prov., Band-B (3)