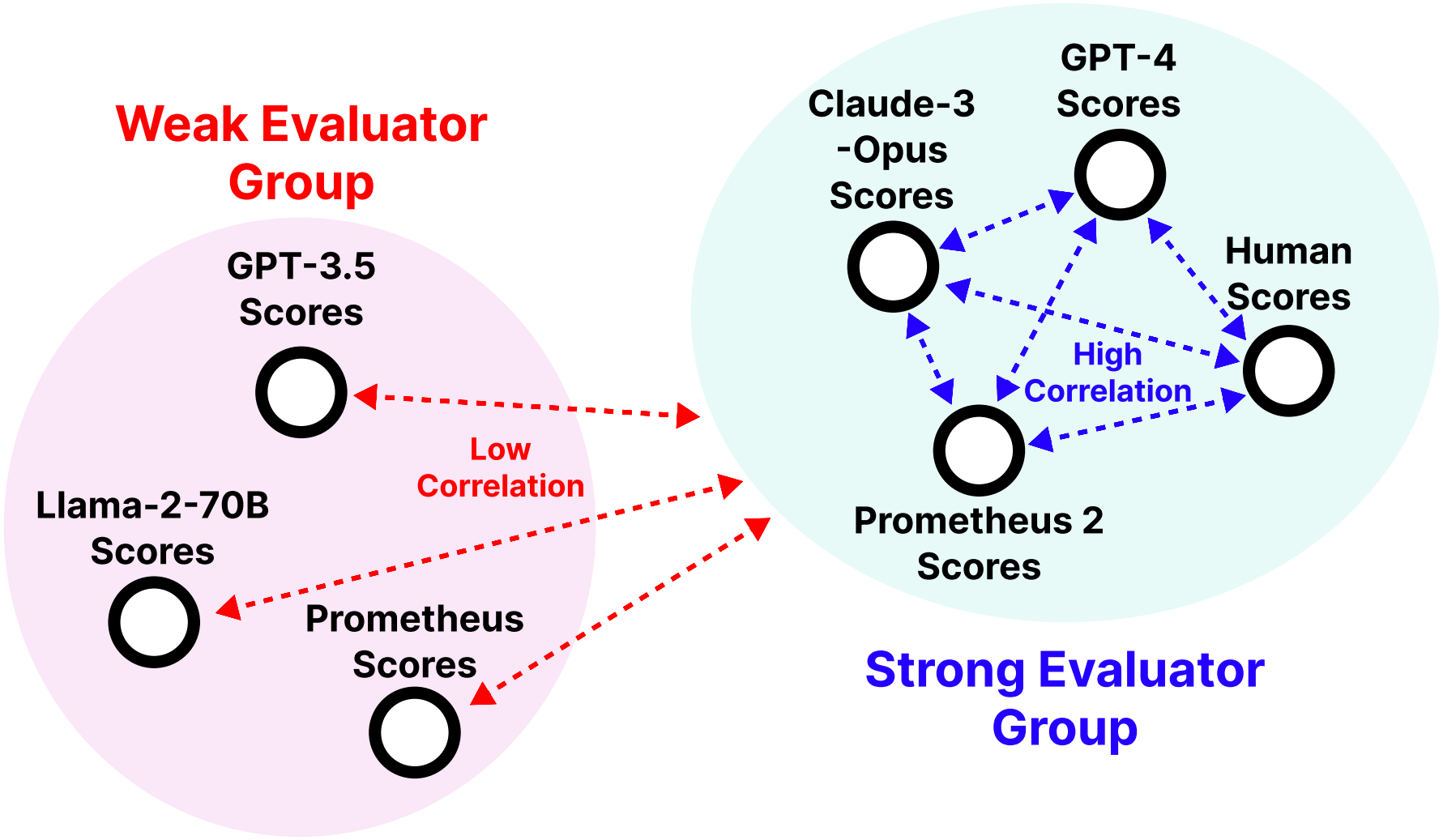

Weak Evaluator Group
GPT-3.5 Scores
Llama-2-70B Scores
Prometheus Scores
Low Correlation
Claude-3 -Opus Scores
GPT-4 Scores
Human Scores
High Correlation
Prometheus 2 Scores
Strong Evaluator Group