

Score Distributions: LLM Judges vs Human Evaluation

Means: Prompt=3.61, Agentic=2.90, Human=3.38

Score (1-5 scale)

