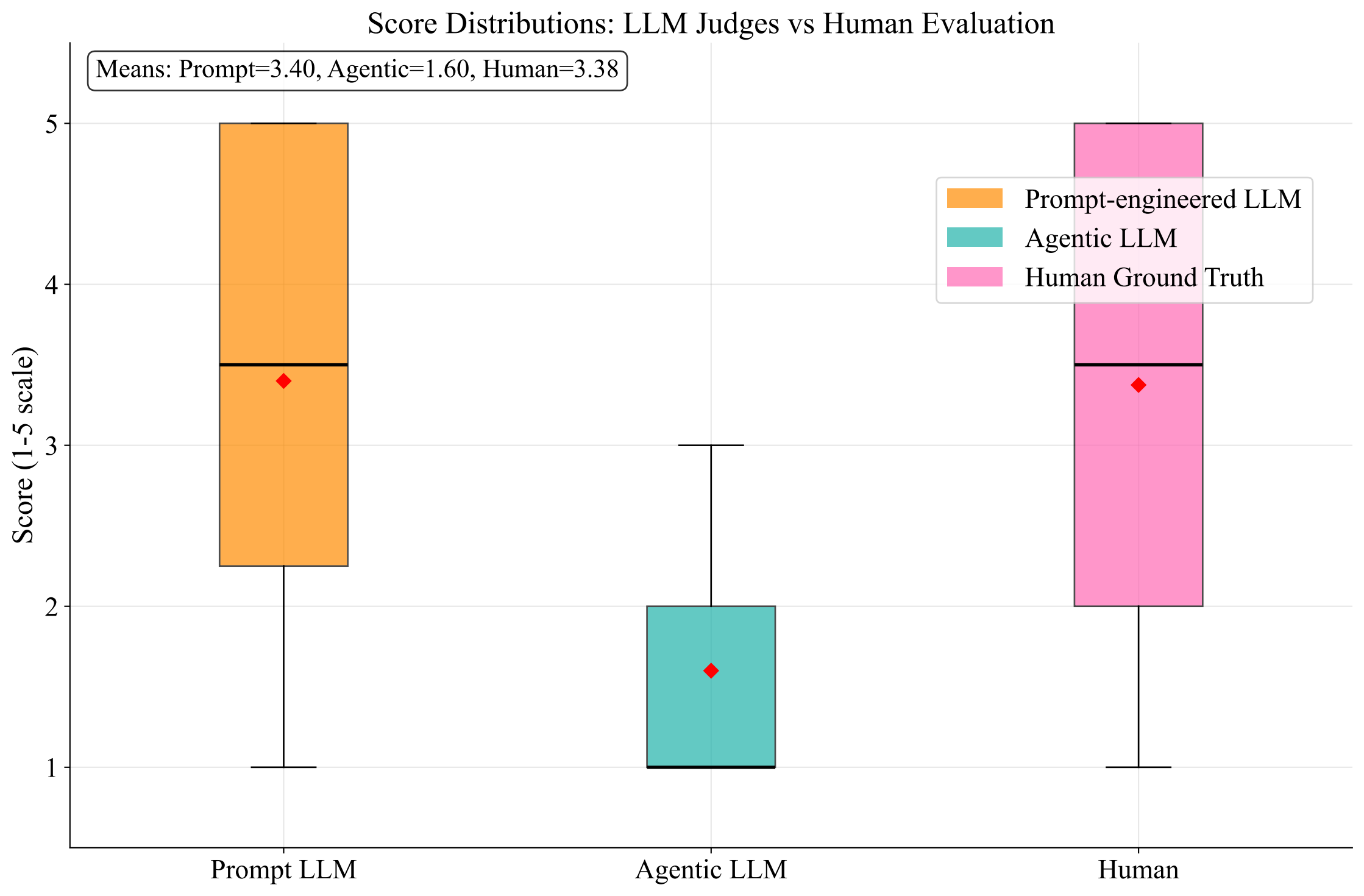
Score Distributions: LLM Judges vs Human Evaluation
Means: Prompt=3.40, Agentic=1.60, Human=3.38
Prompt-engineered LLM
Agentic LLM
Human Ground Truth
Score (1-5 scale)
5
4
3
2
1
Prompt LLM
Agentic LLM
Human