Claim Quality: Human Annotators vs. Panel of LLM-Judges

Legend: PhD Annotators, Panel of Judges

X-axis categories: "yes" Label correct?, "yes" Self-contained?, "yes" Category correct?, "yes" Schema leakage?

Y-axis: Perc. of Claims (%)

"NOT ENOUGH INFO" claims only