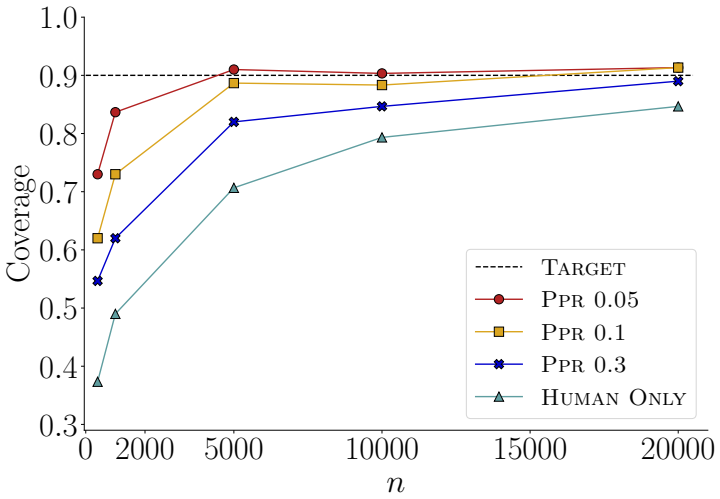

Figure 1: Empirical coverage of the rank-sets constructed using only n synthetic pairwise comparisons by humans (Human Only) and using both n synthetic pairwise comparisons by humans and $N + n$ synthetic pairwise comparisons by one out of three different simulated strong LLMs (Ppr 0.05, Ppr 0.1 and Ppr 0.3) with $\alpha = 0.1$ and $N+n = 50000$. Each of the strong LLMs has a different level of alignment with human preferences controlled by a noise value $u \in \{0.05, 0.1, 0.3\}$. The dashed line indicates the $1 - \alpha$ target coverage. The empirical coverage of the rank-sets constructed using only $N + n$ synthetic pairwise comparison by one of the same three strong LLMs (not shown in the figure) is 0.38 ($u = 0.05$), 0.13 ($u = 0.1$) and 0.0 ($u = 0.3$).

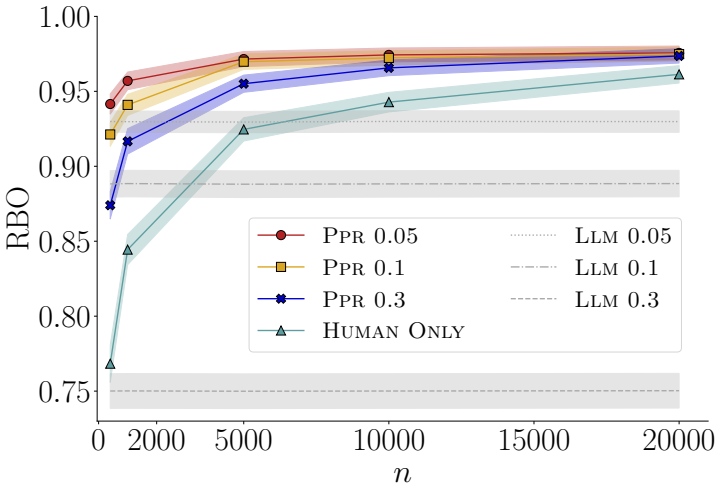

Figure 2: Average rank-based overlap (RBO) of rankings constructed by ordering the empirical win probabilities $\hat{\boldsymbol{\theta}}$ estimated using only $N + n$ synthetic pairwise comparisons by one out of three different simulated strong LLMs (Llm 0.05, Llm 0.1 and Llm 0.3), only n synthetic pairwise comparisons by humans (Human Only), and both n synthetic pairwise comparisons by humans and $N + n$ synthetic pairwise comparisons by one out of the same three strong LLMs (Ppr 0.05, Ppr 0.1 and Ppr 0.3) for $\alpha = 0.1$ and $N + n = 50000$. Each of the strong LLMs has a different level of alignment with human preferences controlled by a noise value $u \in \{0.05, 0.1, 0.3\}$. RBO was computed with respect to the true ranking constructed by ordering the true win probabilities $\boldsymbol{\theta}$. The shaded region shows a 95% confidence interval for the RBO among all 300 repetitions.