$p \le 0.01 (**)$

p > 0.05

 $p \le 0.05 (*)$

Positive value: Prompt A outperforms Prompt B Negative value: Prompt B outperforms Prompt A

Prompt A

Value represents proportion of disagreements * $p \le 0.05$, ** $p \le 0.01$, *** $p \le 0.001$ † Significant after Bonferroni correction