Sign Test Results Between Prompts

Prompt A \ Prompt B	base	advanced	zero_shot_chain_of_thought	deepseek_advanced	deepseek_long_types	deepseek_short_types_proper_prompt
base	—	+0.76***†	+0.58***†	+0.67***†	+0.63***†	+0.53
advanced	-0.76***†	—	-0.62***†	-0.53	-0.56**†	-0.68***†
zero_shot_chain_of_thought	-0.58***†	+0.62***†	—	+0.58***†	+0.55*†	-0.55**
deepseek_advanced	-0.67***†	+0.53	-0.58***†	—	-0.54*	-0.65***†
deepseek_long_types	-0.63***†	+0.56**†	-0.55*†	+0.54*	—	-0.62***†
deepseek_short_types_proper_prompt	-0.53	+0.68***†	+0.55**†	+0.65***†	+0.62***†	—

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

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

p ≤ 0.001 (***)
p ≤ 0.01 (**)
p ≤ 0.05 (*)
p > 0.05