Sign Test Results Between Prompts

Prompt A \ Prompt B	base	advanced	zero_shot_chain_of_thought	deepseek_advanced	deepseek_long_types	deepseek_short_types
base	—	+0.56	+0.55	-0.60	+0.54	-0.57
advanced	-0.56	—	-0.50	-0.71	-0.50	-0.64
zero_shot_chain_of_thought	-0.55	-0.50	—	-0.64	-0.50	-0.60
deepseek_advanced	+0.60	+0.71	+0.64	—	+0.60	-0.50
deepseek_long_types	-0.54	-0.50	-0.50	-0.60	—	-0.62
deepseek_short_types	+0.57	+0.64	+0.60	-0.50	+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