Sign Test Results Between Prompts

Prompt A \ Prompt B	random1	random2	random3	random4	random5	semantic_random1	semantic_random2	semantic_random3	semantic_random4	semantic_random5	semantic_top_to_bottom	semantic_bottom_to_top
random1	—	-0.50	+0.54	+0.53	+0.51	+0.51	-0.53	-0.51	-0.54	+0.54	+0.53	+0.52
random2	-0.50	—	+0.54	+0.53	+0.52	+0.51	-0.53	-0.54	-0.57	+0.51	-0.51	+0.52
random3	-0.54	-0.54	—	-0.51	-0.53	-0.53	-0.59	-0.54	-0.57	-0.50	-0.51	-0.53
random4	-0.53	-0.53	+0.51	—	-0.52	-0.52	-0.56	-0.55	-0.58	-0.51	-0.52	-0.51
random5	-0.51	-0.52	+0.53	+0.52	—	-0.50	-0.54	-0.52	-0.55	+0.52	+0.51	+0.51
semantic_random1	-0.51	-0.51	+0.53	+0.52	-0.50	—	-0.56	-0.53	-0.57	+0.51	-0.50	+0.51
semantic_random2	+0.53	+0.53	+0.59	+0.56	+0.54	+0.56	—	+0.52	-0.52	+0.56	+0.55	+0.57
semantic_random3	+0.51	+0.54	+0.54	+0.55	+0.52	+0.53	-0.52	—	-0.53	+0.55	+0.53	+0.56
semantic_random4	+0.54	+0.57	+0.57	+0.58	+0.55	+0.57	+0.52	+0.53	—	+0.58	+0.56	+0.59
semantic_random5	-0.54	-0.51	-0.50	+0.51	-0.52	-0.51	-0.56	-0.55	-0.58	—	-0.52	-0.51
semantic_top_to_bottom	-0.53	+0.51	+0.51	+0.52	-0.51	-0.50	-0.55	-0.53	-0.56	+0.52	—	+0.51
semantic_bottom_to_top	-0.52	-0.52	+0.53	+0.51	-0.51	-0.51	-0.57	-0.56	-0.59	+0.51	-0.51	—

Prompt A

Prompt B

p ≤ 0.001 (***)

p ≤ 0.01 (**)

p ≤ 0.05 (*)

p > 0.05

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