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	thematic_random1	thematic_random2	thematic_random3	thematic_random4	thematic_random5	thematic_top_to_bottom	thematic_bottom_to_top	deepseek_solutions
random1	—	-0.50	+0.64	-0.58	-0.62	-0.50	+0.67	+0.67	+0.71	-0.64	+0.55	-0.50	+0.60	-0.62	-0.58	+0.55	-0.57	+0.56	-0.50	+0.56
random2	-0.50	—	+0.67	-0.58	-0.64	-0.50	+0.62	+0.64	+0.71	-0.67	+0.56	-0.50	+0.62	-0.67	-0.57	+0.56	-0.60	+0.56	-0.50	+0.56
random3	-0.64	-0.67	—	-0.69	-0.88	-0.71	-0.57	-0.50	-0.50	-0.80	-0.62	-0.67	-0.60	-0.80	-0.73	-0.60	-0.73	-0.58	-0.71	-0.53
random4	+0.58	+0.58	+0.69	—	-0.53	+0.58	+0.67	+0.73	+0.73	-0.53	+0.60	+0.58	+0.70	-0.53	-0.50	+0.64	-0.50	+0.67	+0.60	+0.62
random5	+0.62	+0.64	+0.88	+0.53	—	+0.64	+0.73	+0.80	+0.75	-0.50	+0.67	+0.62	+0.78	-0.50	+0.53	+0.64	+0.54	+0.62	+0.67	+0.63
semantic_random1	-0.50	-0.50	+0.71	-0.58	-0.64	—	+0.67	+0.71	+0.71	-0.71	+0.57	-0.50	+0.62	-0.62	-0.58	+0.56	-0.58	+0.55	-0.50	+0.56
semantic_random2	-0.67	-0.62	+0.57	-0.67	-0.73	-0.67	—	+0.60	+0.67	-0.86	-0.57	-0.75	-0.50	-0.86	-0.70	-0.56	-0.67	-0.60	-0.67	-0.50
semantic_random3	-0.67	-0.64	-0.50	-0.73	-0.80	-0.71	-0.60	—	-0.50	-0.88	-0.67	-0.80	-0.60	-0.80	-0.78	-0.60	-0.69	-0.62	-0.71	-0.53
semantic_random4	-0.71	-0.71	-0.50	-0.73	-0.75	-0.71	-0.67	-0.50	—	-0.88	-0.62	-0.80	-0.57	-0.80	-0.73	-0.62	-0.69	-0.67	-0.71	-0.53
semantic_random5	+0.64	+0.67	+0.80	+0.53	-0.50	+0.71	+0.86	+0.88	+0.88	—	+0.83	+1.00	+0.78	-0.50	+0.55	+0.67	+0.53	+0.75	+0.71	+0.62
semantic_top_to_bottom	-0.55	-0.56	+0.62	-0.60	-0.67	-0.57	+0.57	+0.67	+0.62	-0.83	—	-0.60	+0.56	-0.75	-0.67	-0.50	-0.62	-0.50	-0.56	+0.53
semantic_bottom_to_top	-0.50	-0.50	+0.67	-0.58	-0.62	-0.50	+0.75	+0.80	+0.80	-1.00	+0.60	—	+0.67	-0.71	-0.60	+0.56	-0.57	+0.60	-0.50	+0.55
thematic_random1	-0.60	-0.62	+0.60	-0.70	-0.78	-0.62	-0.50	+0.60	+0.57	-0.78	-0.56	-0.67	—	-0.78	-0.75	-0.56	-0.70	-0.56	-0.75	-0.50
thematic_random2	+0.62	+0.67	+0.80	+0.53	-0.50	+0.62	+0.86	+0.80	+0.80	-0.50	+0.75	+0.71	+0.78	—	+0.56	+0.70	+0.54	+0.75	+0.71	+0.62
thematic_random3	+0.58	+0.57	+0.73	-0.50	-0.53	+0.58	+0.70	+0.78	+0.73	-0.55	+0.67	+0.60	+0.75	-0.56	—	+0.60	-0.50	+0.64	+0.60	+0.60
thematic_random4	-0.55	-0.56	+0.60	-0.64	-0.64	-0.56	+0.56	+0.60	+0.62	-0.67	-0.50	-0.56	+0.56	-0.70	-0.60	—	-0.67	-0.50	-0.55	+0.54
thematic_random5	+0.57	+0.60	+0.73	-0.50	-0.54	+0.58	+0.67	+0.69	+0.69	-0.53	+0.62	+0.57	+0.70	-0.54	-0.50	+0.67	—	+0.62	+0.62	+0.67
thematic_top_to_bottom	-0.56	-0.56	+0.58	-0.67	-0.62	-0.55	+0.60	+0.62	+0.67	-0.75	-0.50	-0.60	+0.56	-0.75	-0.64	-0.50	-0.62	—	-0.56	+0.53
thematic_bottom_to_top	-0.50	-0.50	+0.71	-0.60	-0.67	-0.50	+0.67	+0.71	+0.71	-0.71	+0.56	-0.50	+0.75	-0.71	-0.60	+0.55	-0.62	+0.56	—	+0.56
deepseek_solutions	-0.56	-0.56	+0.53	-0.62	-0.63	-0.56	-0.50	+0.53	+0.53	-0.62	-0.53	-0.55	-0.50	-0.62	-0.60	-0.54	-0.67	-0.53	-0.56	—

Color scale legend:
- $p \leq 0.001$ (***)
- $p \leq 0.01$ (**)
- $p \leq 0.05$ (*)
- $p > 0.05$

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

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