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	—	-0.58*	-0.57*	-0.59**	-0.54	-0.62***†	-0.59***	-0.61**†	-0.60*†	-0.62***†	-0.59***	-0.60*†	-0.61***	-0.60*†	-0.64***†	-0.64***†	-0.61*†	-0.59***	-0.61**†	-0.55*
random2	+0.58*	—	+0.51	-0.51	+0.54	-0.57**	-0.54	-0.56*	-0.55	-0.57**	-0.54	-0.55	-0.54	-0.55	-0.58**	-0.57*	-0.55	-0.54	-0.56*	-0.52
random3	+0.57*	-0.51	—	-0.52	+0.53	-0.58**	-0.54	-0.56*	-0.55*	-0.58**	-0.54	-0.55*	-0.55	-0.55	-0.59**	-0.58**	-0.56	-0.54	-0.57**	-0.52
random4	+0.59**	+0.51	+0.52	—	+0.55	-0.57**	-0.53	-0.55	-0.54	-0.57*	-0.53	-0.54	-0.53	-0.54	-0.57*	-0.57*	-0.54	-0.53	-0.56*	-0.52
random5	+0.54	-0.54	-0.53	-0.55	—	-0.60*†	-0.56*	-0.58**	-0.57**	-0.59***	-0.56*	-0.57**	-0.57*	-0.57*	-0.61***	-0.60***	-0.58**	-0.56*	-0.58**	-0.53
semantic_random1	+0.62***†	+0.57**	+0.58**	+0.57**	+0.60*†	—	+0.58*	+0.53	+0.55	+1.00	+0.58*	+0.56	+0.55	+0.54	+0.51	+0.52	+0.54	+0.56	+0.51	+0.52
semantic_random2	+0.59***	+0.54	+0.54	+0.53	+0.56*	-0.58*	—	-0.54	-0.53	-0.57	+0.50	-0.53	+0.51	+0.50	-0.52	-0.52	-0.50	-0.50	-0.55	+0.50
semantic_random3	+0.61**†	+0.56*	+0.56*	+0.55	+0.58**	-0.53	+0.54	—	+0.51	-0.53	+0.54	+0.52	+0.53	+0.52	-0.50	-0.50	+0.52	+0.53	-0.51	+0.51
semantic_random4	+0.60*†	+0.55	+0.55*	+0.54	+0.57**	-0.55	+0.53	-0.51	—	-0.54	+0.53	+0.51	+0.52	+0.52	-0.51	-0.51	+0.51	+0.52	-0.52	+0.51
semantic_random5	+0.62***†	+0.57**	+0.58**	+0.57*	+0.59***	-1.00	+0.57	+0.53	+0.54	—	+0.58	+0.55	+0.54	+0.54	+0.51	+0.52	+0.54	+0.56	+0.51	+0.52
semantic_top_to_bottom	+0.59***	+0.54	+0.54	+0.53	+0.56*	-0.58*	-0.50	-0.54	-0.53	-0.58	—	-0.52	+0.51	-0.50	-0.53	-0.52	-0.50	-0.50	-0.54	+0.50
semantic_bottom_to_top	+0.60*†	+0.55	+0.55*	+0.54	+0.57**	-0.56	+0.53	-0.52	-0.51	-0.55	+0.52	—	+0.52	+0.51	-0.51	-0.51	+0.51	+0.52	-0.57	+0.51
thematic_random1	+0.61***	+0.54	+0.55	+0.53	+0.57*	-0.55	-0.51	-0.53	-0.52	-0.54	-0.51	-0.52	—	-0.51	-0.55	-0.54	-0.51	-0.51	-0.54	-0.50
thematic_random2	+0.60*†	+0.55	+0.55	+0.54	+0.57*	-0.54	-0.50	-0.52	-0.52	-0.54	-0.50	-0.51	+0.51	—	-0.54	-0.53	-0.50	-0.50	-0.53	+0.50
thematic_random3	+0.64***†	+0.58**	+0.59**	+0.57*	+0.61***	-0.51	+0.52	+0.50	+0.51	-0.51	+0.53	+0.51	+0.55	+0.54	—	+0.50	+0.54	+0.53	-0.51	+0.51
thematic_random4	+0.64***†	+0.57*	+0.58**	+0.57*	+0.60***	-0.52	+0.52	-0.50	+0.51	-0.52	+0.52	+0.51	+0.54	+0.53	-0.50	—	+0.53	+0.52	-0.51	+0.51
thematic_random5	+0.61*†	+0.55	+0.56	+0.54	+0.58**	-0.54	-0.50	-0.52	-0.51	-0.54	+0.50	-0.51	+0.51	+0.50	-0.54	-0.53	—	-0.50	-0.53	+0.50
thematic_top_to_bottom	+0.59***	+0.54	+0.54	+0.53	+0.56*	-0.56	-0.50	-0.53	-0.52	-0.56	+0.50	-0.52	+0.51	+0.50	-0.53	-0.52	-0.50	—	-0.55	+0.50
thematic_bottom_to_top	+0.61**†	+0.56*	+0.57**	+0.56*	+0.58**	-0.51	+0.55	+0.51	+0.52	-0.51	+0.54	+0.57	+0.54	+0.53	+0.51	+0.51	+0.53	+0.55	—	+0.52
deepseek_solutions	+0.55*	+0.52	+0.52	+0.52	+0.53	-0.52	-0.50	-0.51	-0.51	-0.52	-0.50	-0.51	+0.50	-0.50	-0.51	-0.51	-0.50	-0.50	-0.52	—

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