-0.56\*

-0.57\*

-0.56\*

-0.55\*

+0.71\*\*\*

-0.55\*

+0.54

+0.55\*

semantic top to pottom semantic bottom to top thenatic top to bottom thenatic pottom to top deepseek solutions semantic randoma semantic Jandom's thematic random? thenatic random? thematic randoms thematic randomiz thenatic randona

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

random1

random2

random3

random5 -

random4 - +0.58\*\*\*

-0.55\*

+0.53

semantic\_random1 - +0.62\*\*\*† +0.58\*† +0.60\*\*\*† +0.56\*\* +0.59\*\*†

semantic\_random2 - +0.63\*\*\* +0.59\*\* +0.62\*\*\* +0.57\*\*\* +0.60\*\* +0.60\*\*

semantic\_random3 - +0.63\*\*\*† +0.59\*\*\*† +0.62\*\*\*† +0.57\*\*\* +0.60\*\*\*†

semantic\_random4 - +0.62\*\*\* +0.58\*\* +0.61\*\*\* +0.56\*\* +0.59\*\*\*

Semantic\_random5 - +0.62\*\*\*† +0.58\*† +0.60\*\*\*† +0.56\*\* +0.58\*\*†

semantic\_top\_to\_bottom - +0.62\*\*\*† +0.58\*\*† +0.61\*\*\*† +0.56\*\* +0.59\*\*\*†

semantic\_bottom\_to\_top - +0.63\*\*\*† +0.59\*\*\*† +0.62\*\*\*† +0.57\*† +0.60\*\*\*†

thematic\_random1 - +0.64\*\*\*† +0.59\*\*† +0.62\*\*\*† +0.57\*\* +0.60\*\*\*†

thematic\_random2 - +0.65\*\*\*† +0.61\*\*\*† +0.64\*\*\*† +0.58\*† +0.62\*\*\*†

thematic\_random3 - +0.64\*\*\*† +0.59\*\*† +0.62\*\*\*† +0.57\*\* +0.61\*\*\*†

thematic\_random4 - +0.62\*\*\* +0.58\*† +0.61\*\*\*†

thematic random5 - +0.62\*\*\*† +0.58\*\*\* +0.61\*\*\*†

thematic\_top\_to\_bottom - +0.64\*\*\* +0.60\*\*\* +0.63\*\*\* +0.58\*\* +0.61\*\*\*

thematic\_bottom\_to\_top - +0.65\*\*\*† +0.62\*\*\*† +0.65\*\*\*† +0.60\*\*\*† +0.63\*\*\*† +0.57\*\*

randomis

randoma

-0.52

+0.55\*

-0.58\*\*\*

-0.55\*

+0.51

+0.52

-0.50

-0.51

-0.52

+0.54

semantic random?

+0.55\*

+0.55\*

+0.59\*\*†

+0.59\*†

semantic randomia

+0.51

-0.53

+0.52

+0.56\*

semantic random3

+0.50

-0.53

+0.52

+0.55

deepseek\_solutions - -0.60\*\*\*† -0.62\*\*\*† -0.62\*\*\*† -0.61\*\*\*† -0.64\*\*\*† -0.62\*\*\*† -0.68\*\*\*† -0.69\*\*\*†

+0.52

-0.52

+0.54

+0.58\*\*

+0.57\*

-0.52

+0.56\*

+0.51

+0.58\*

+0.53

+0.56\*

+0.51

+0.53

+0.53

+0.55\*

-0.51

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

Prompt B

Sign Test Results Between Prompts

p > 0.05