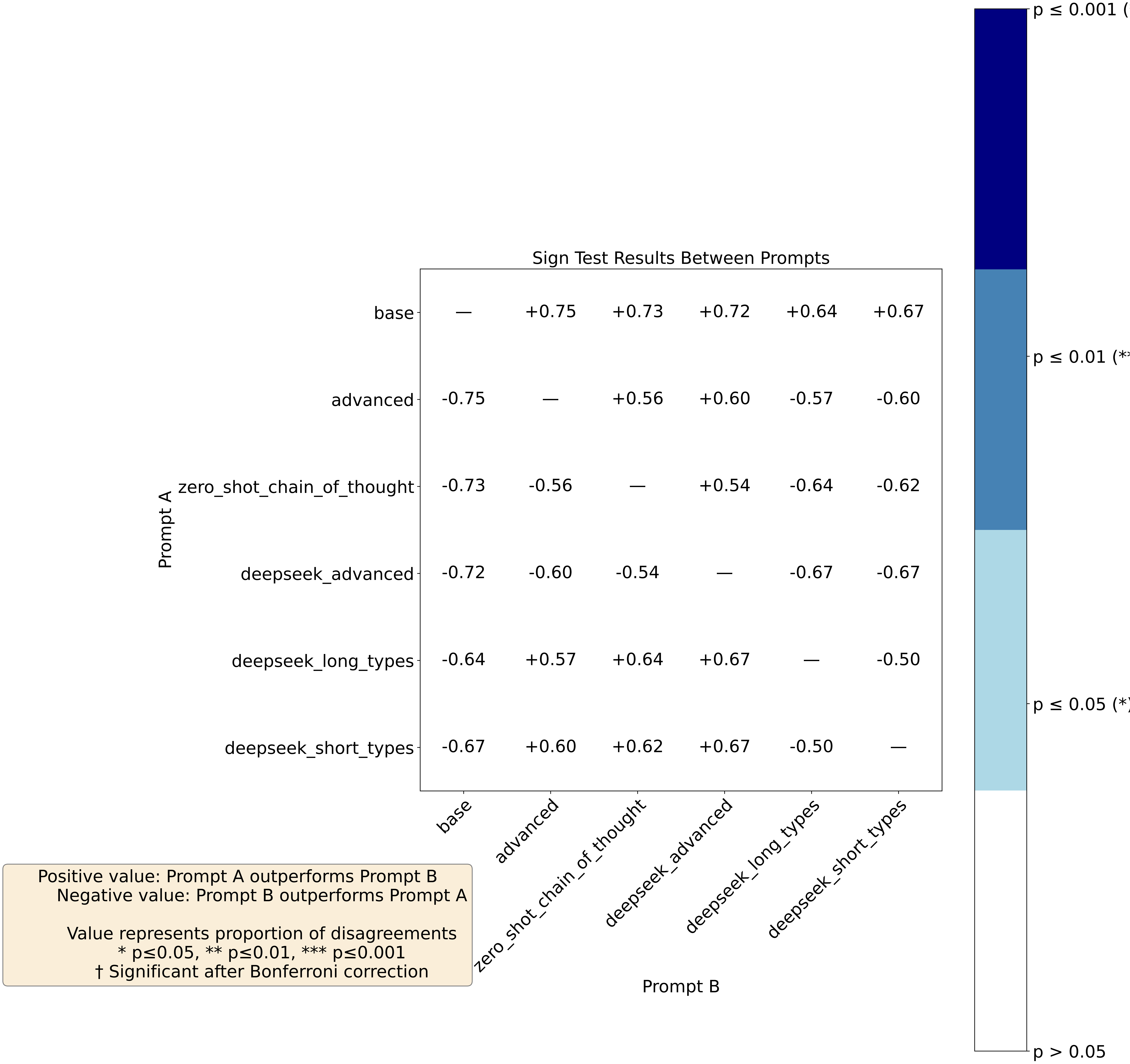

Prompt A \ Prompt B	base	advanced	zero_shot_chain_of_thought	deepseek_advanced	deepseek_long_types	deepseek_short_types
base	—	+0.75	+0.73	+0.72	+0.64	+0.67
advanced	-0.75	—	+0.56	+0.60	-0.57	-0.60
zero_shot_chain_of_thought	-0.73	-0.56	—	+0.54	-0.64	-0.62
deepseek_advanced	-0.72	-0.60	-0.54	—	-0.67	-0.67
deepseek_long_types	-0.64	+0.57	+0.64	+0.67	—	-0.50
deepseek_short_types	-0.67	+0.60	+0.62	+0.67	-0.50	—

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