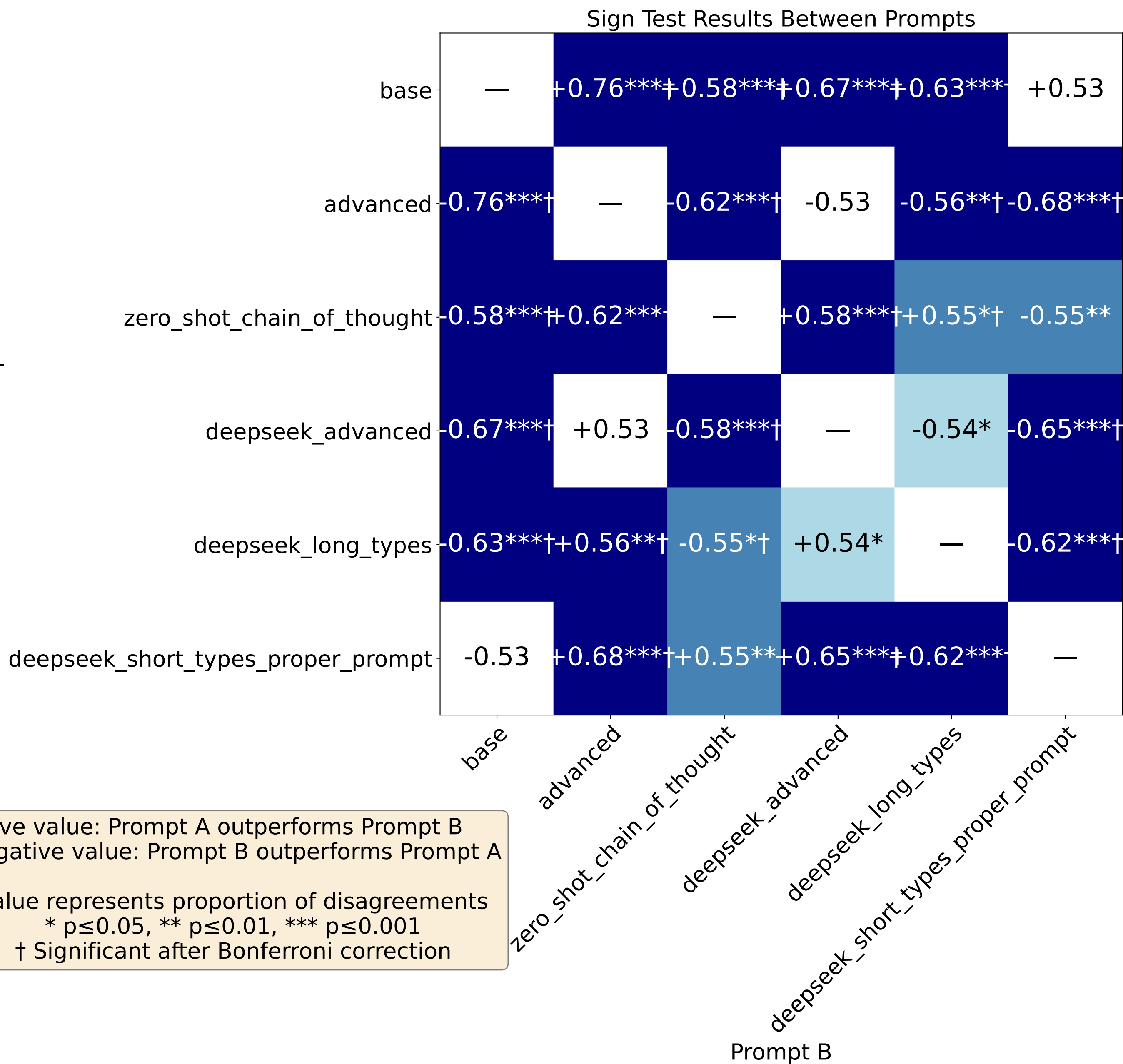


Prompt A



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

$p \leq 0.001$ (***)

$p \leq 0.01$ (**)

$p \leq 0.05$ (*)

$p > 0.05$