

Logit difference heatmap: $\text{logit}(\text{' algorithm'}) - \text{logit}(\text{' wizard'})$

