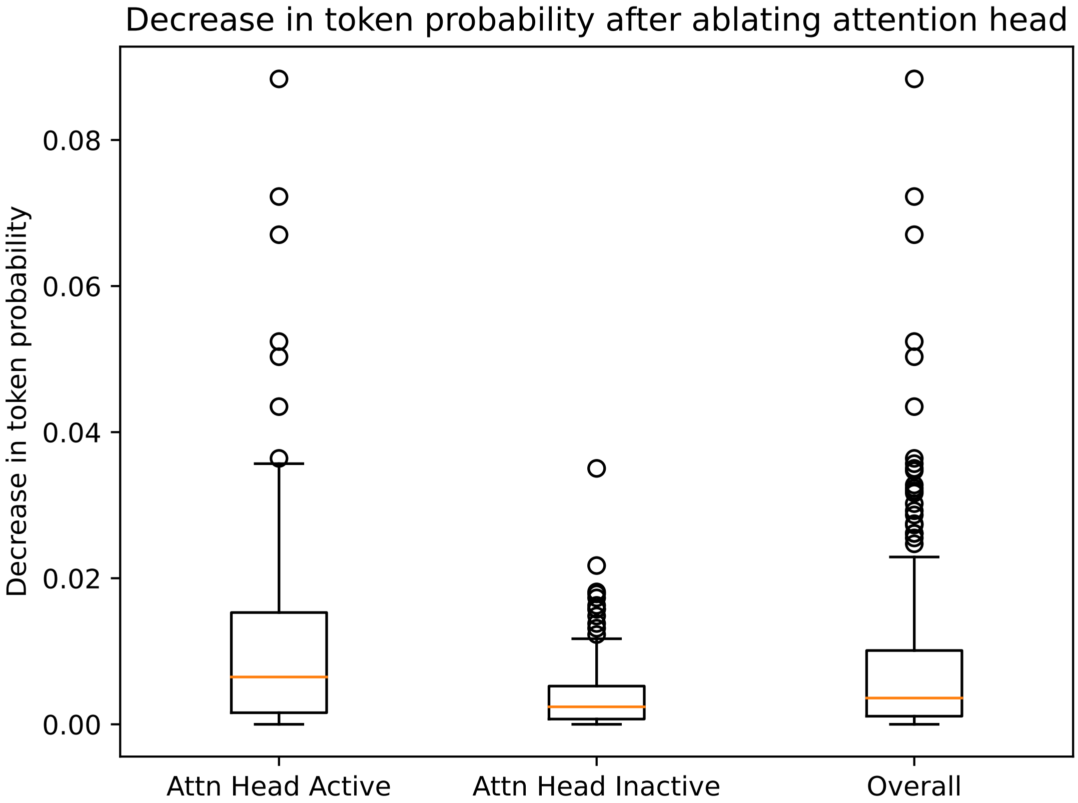
Decrease in token probability after ablating attention head
Decrease in token probability
0.00
0.02
0.04
0.06
0.08
Attn Head Active
Attn Head Inactive
Overall