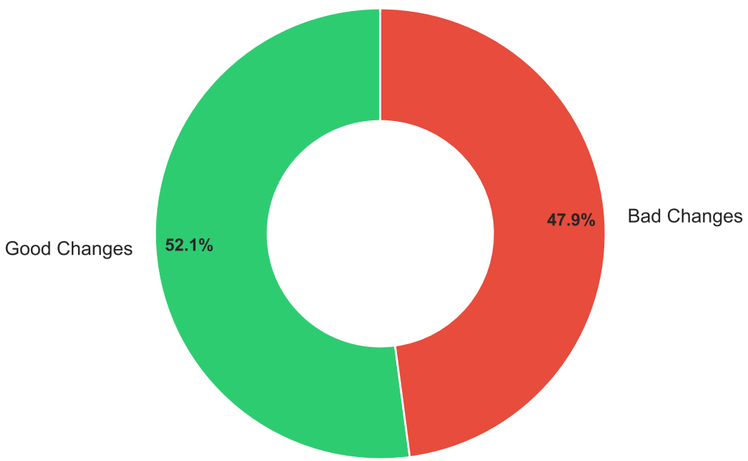
## Calibration Change Quality: 4omini_sc vs 4omini_cot

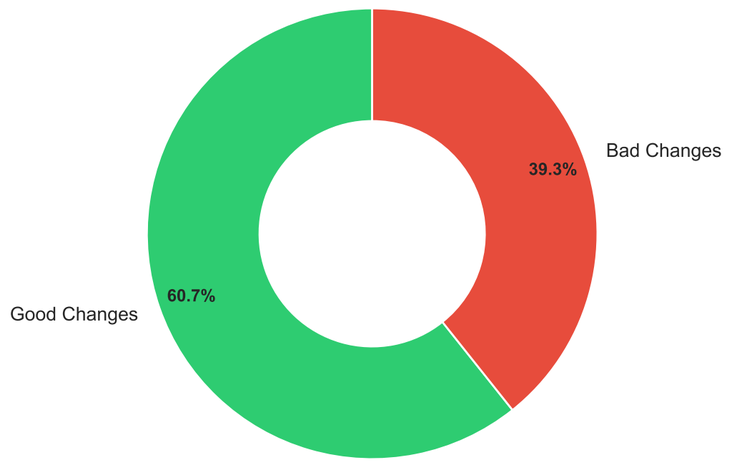
Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 539
Decreases: 250

## Calibration Change Quality: 4omini_cot vs 4omini_noexp

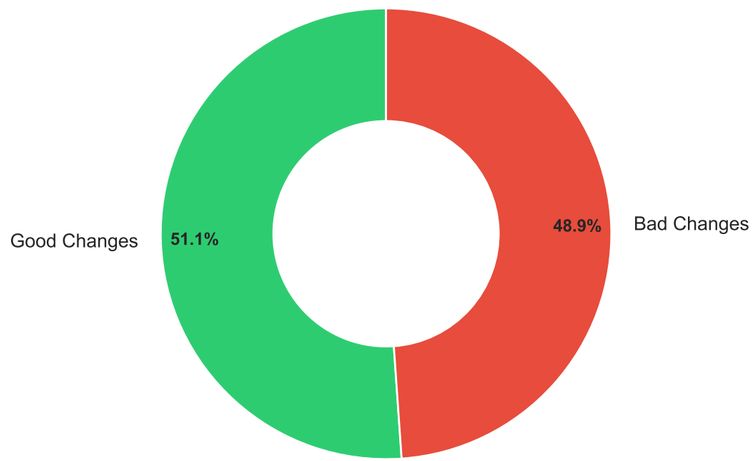
Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 237
Decreases: 529

## Calibration Change Quality: grok2_sc vs grok2_cot

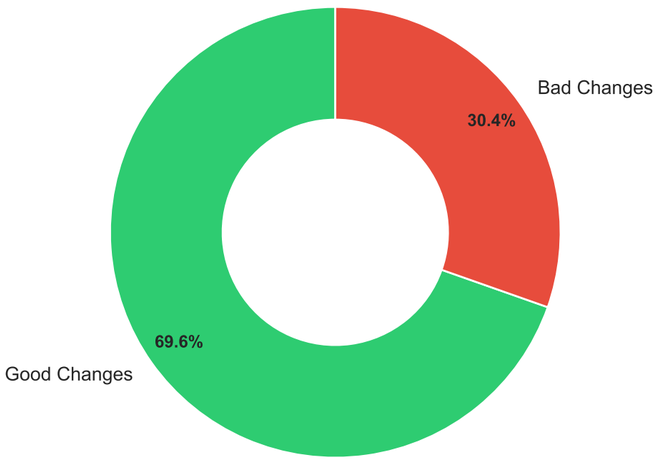
Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 680
Decreases: 406

## Calibration Change Quality: grok2_cot vs grok2_noexp

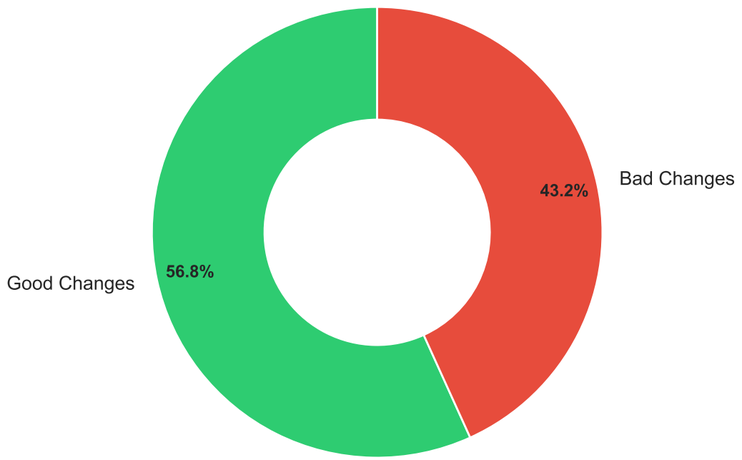
Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 606
Decreases: 210

## Calibration Change Quality: o1_reason vs o1_mini

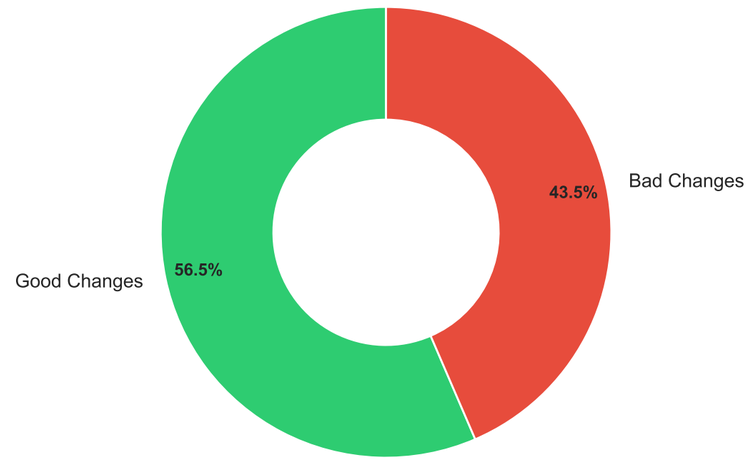
Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 189
Decreases: 84

## Calibration Change Quality: llama_cot vs llama_noexp

Criteria:
• For CORRECT predictions (current): Increase in confidence is GOOD.
• For INCORRECT predictions (current): Decrease in confidence is GOOD.




Increases: 626
Decreases: 562