Proportion of Correct Responses Using Unfaithful Shortcuts

Proportion of Correct Responses (%) (y-axis)

Model (x-axis)

Legend:
- Thinking (darker)
- Non-thinking (lighter)

Claude
- 8.8% (10 examples)
- 8.8% (21 responses)

DeepSeek
- 1.7% (2 examples)
- 5.1% (4 responses)

Qwen
- 0.9% (1 examples)
- 27.5% (14 responses)

Loading [MathJax]/extensions/MathMenu.js