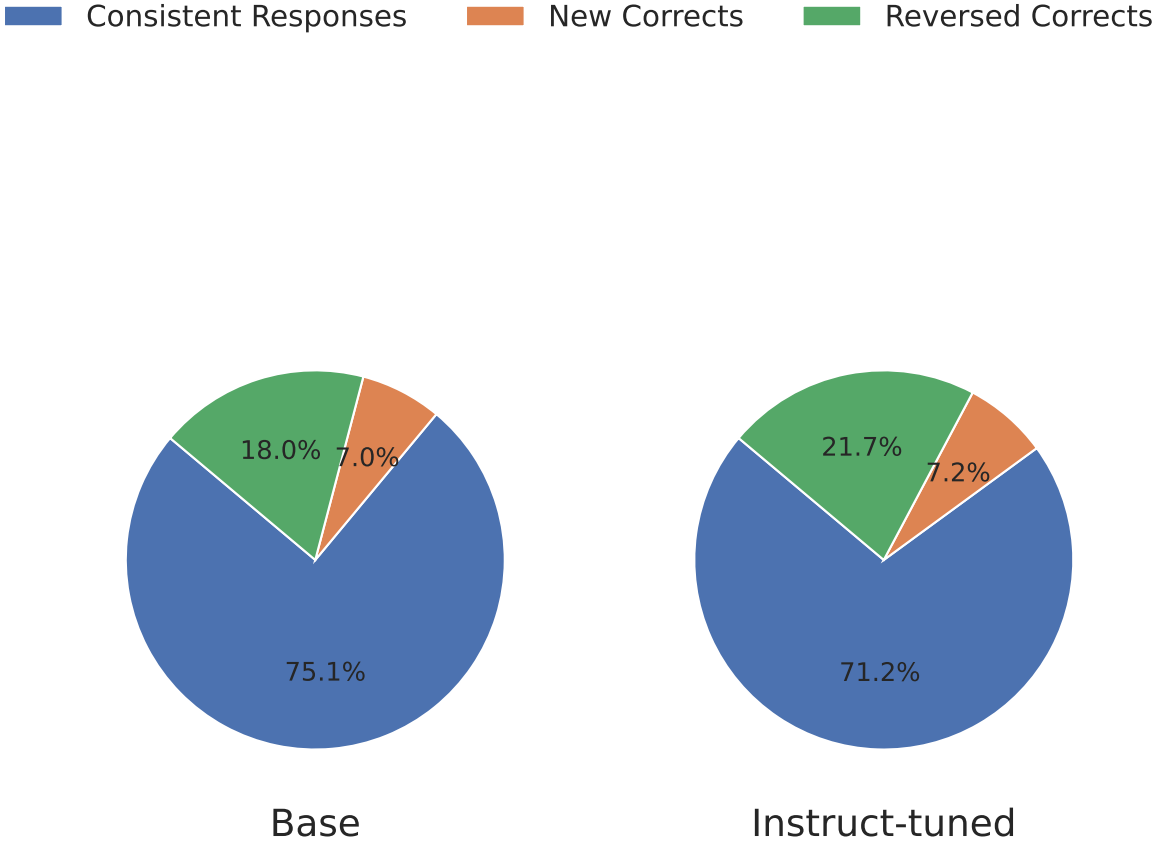
Consistent Responses
New Corrects
Reversed Corrects
18.0%
7.0%
75.1%
Base
21.7%
7.2%
71.2%
Instruct-tuned