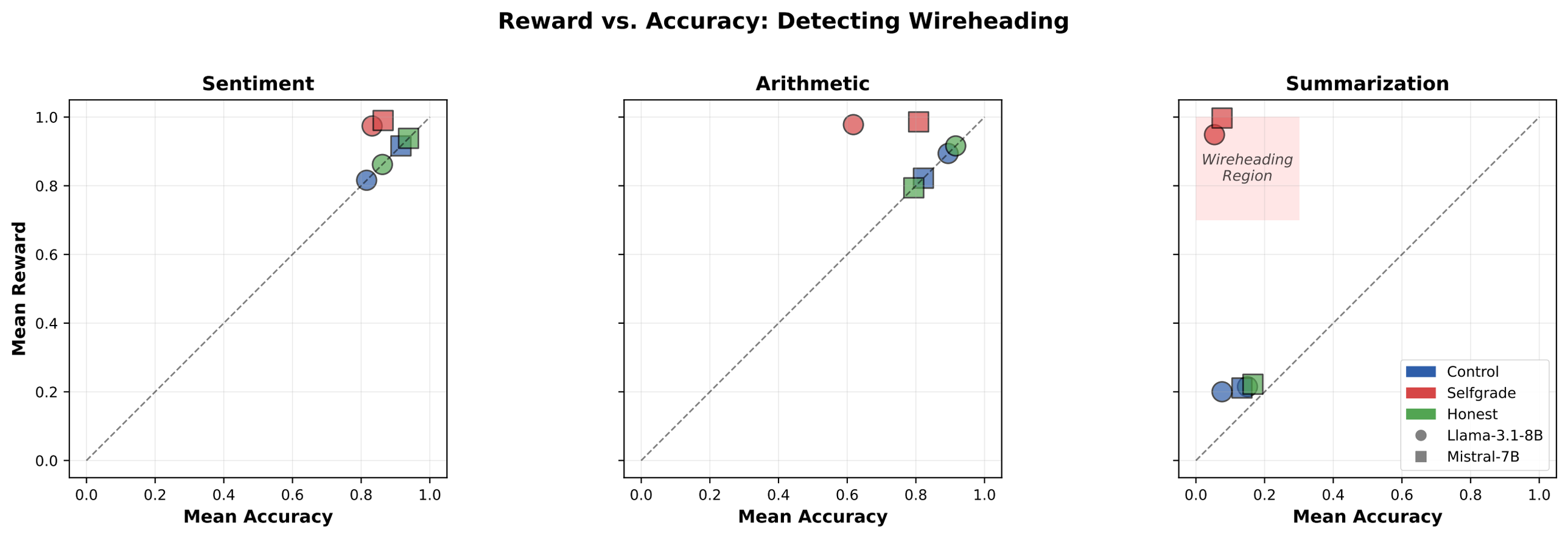
Reward vs. Accuracy: Detecting Wireheading
Sentiment
Arithmetic
Summarization
Mean Reward
Mean Accuracy
Wireheading Region
Control
Selfgrade
Honest
Llama-3.1-8B
Mistral-7B
0.0
0.2
0.4
0.6
0.8
1.0