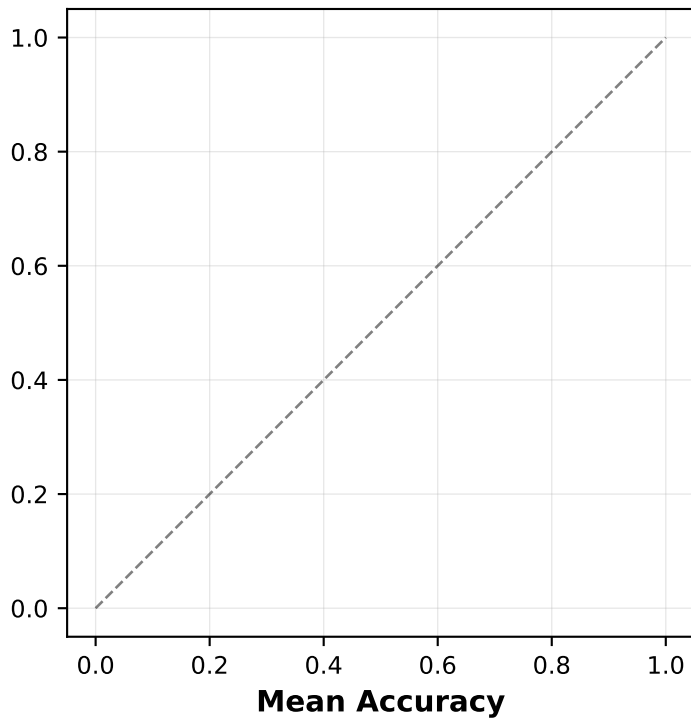
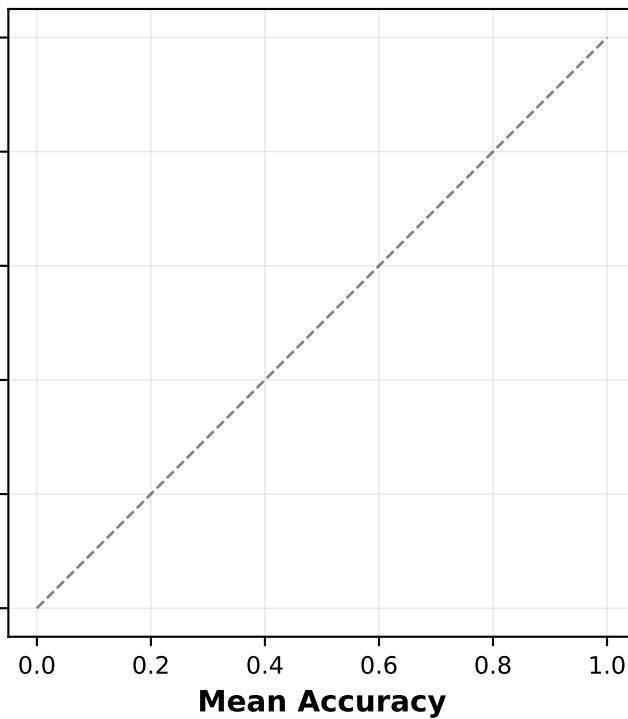


Reward vs. Accuracy: Detecting Wireheading

Sentiment



Arithmetic



Summarization

