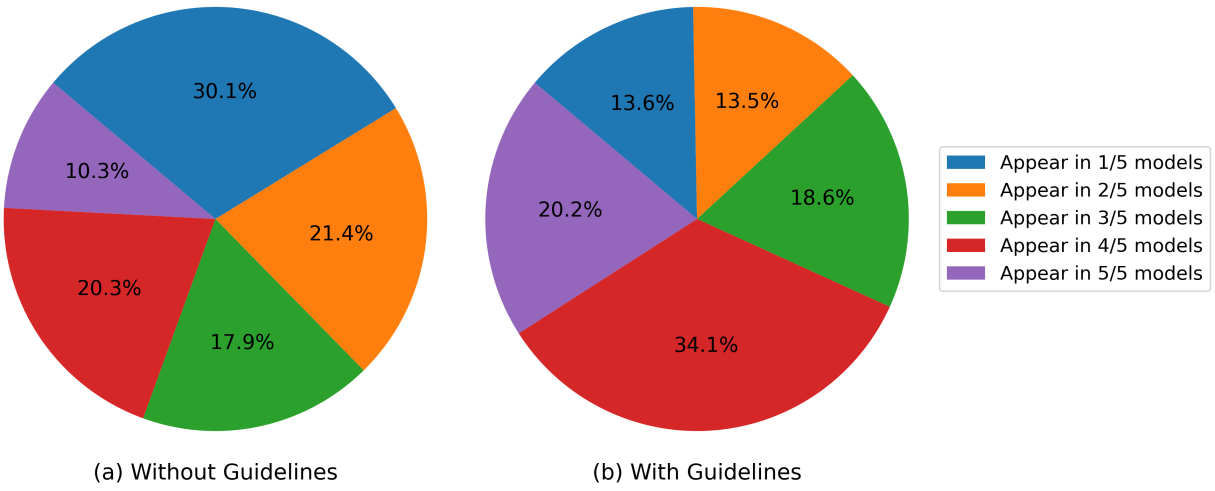
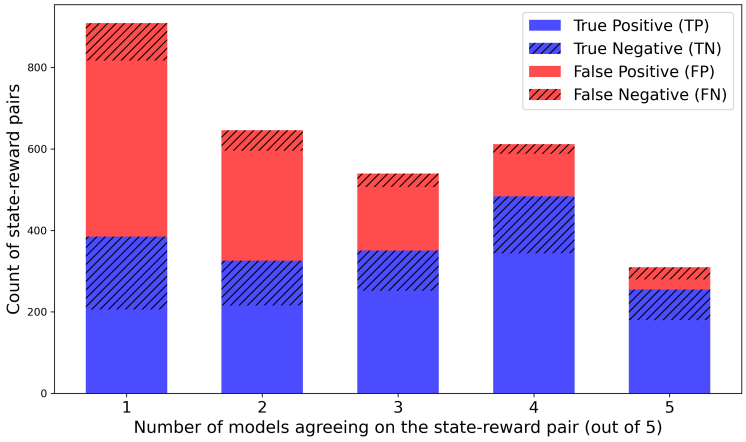
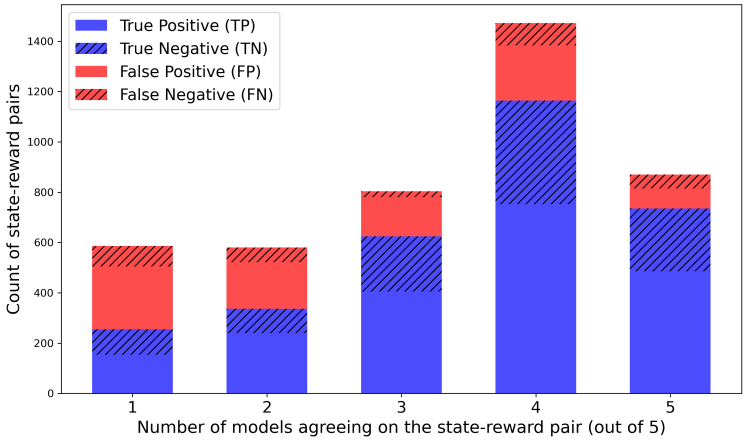


Figure 1: Consensus of output state–reward pairs across five LLMs: sonar, gpt-4o, o1, o3-mini, gpt-3.5. Percentages reflect the share of all generated outputs, grouped by the number of models in agreement. We observe that providing guidelines to the evaluators increases the agreement between models.



(a) Without Guidelines.



(b) With Guidelines.

Figure 2: Count of state-reward pairs for each level of model agreement, categorized by correctness (correct in blue, incorrect in red) and predicted labels (positive in solid, negative in hatched). A state-reward is correct if the reward accurately identifies whether the state is in the target path or not. The more models output a state-reward pair the higher the probability that it is correct.