APPENDIX I
STRATEGY-CRITICAL STATES

Previous work has looked at generating explanations for a human of when an autonomous agent thinks it's very important to take a particular action to not fail at the task as a series of states, called *critical states*. Given a stochastic policy π, the original paper searches the set of critical states $\mathcal{C}_\pi$ where the agent prefers a small set of actions over all others:

$$\mathcal{C}_\pi = \{s \mid \mathcal{H}(\pi(\cdot \mid s)) < t\} \tag{8}$$

where $\mathcal{H}(\pi(\cdot \mid s))$ is the entropy of the policy's action distribution at state s and $t \in \mathbb{R}$ is a threshold for criticality.

We modify their approach to try and generate explanations for when the agent thinks it's very important to take particular actions to end up taking particular strategies. In particular, we consider finding the set of *strtategy-critical states* $\mathcal{SC}_\pi$ where the agent prefers taking a small set of *different actions* for *each strategy*:

$$\mathcal{SC}_\pi = \{s \mid \frac{1}{k^2} \sum_{z_i^*, z_j^*} D_{KL}(\pi(\cdot|s, z_i)) \parallel \pi(\cdot|s, z_j) > t\}. \tag{9}$$

where D_{KL} is the KL-divergence. This method will search for states where the average KL-divergence between the action distributions for the policy under each potential strategy cluster are sufficiently large.

Using this approach, we find that strategy-critical states are branching points between strategies, i.e. the last possible common state between different strategies. Most strategy-critical states show the two agents just before committing to a particular strategy. We use these strategy-critical states to determine the initial strategy landmark when generating the full explanation.

APPENDIX II
TEXTUAL EXPLANATION GENERATION

[Ravi: paste in one example prompt and chatgpt's response]