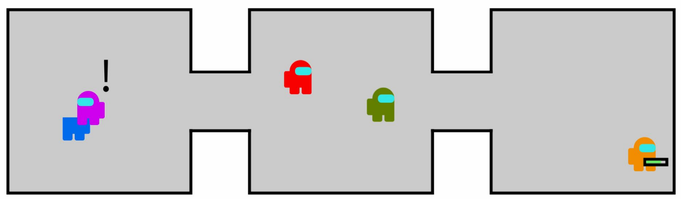
# Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning

Bidipta Sarkar, Warren Xia, C. Karen Liu, Dorsa Sadigh

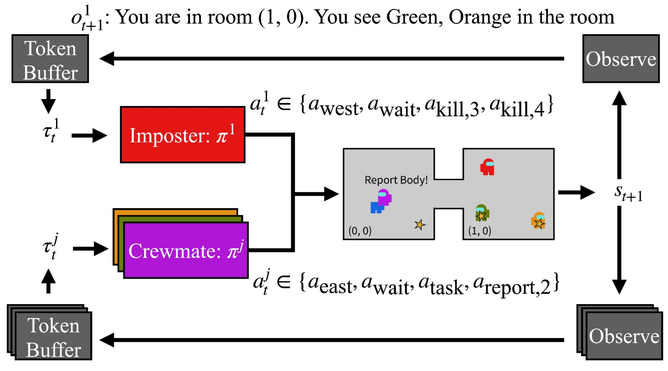
## Motivation

Can we improve an LLM's ability to discuss and understand the game of Among Us?



## Environment

The world streams tokens as observations and accepts tokens as actions




## MARL Optimization

We train RWKV-4 models using MAPPO with a KL constraint to keep natural language speech

$$L_{\text{RL}}(\pi) = -\mathop{\mathbb{E}}_{\tau^i \sim \Pi} \sum_t \left[ \gamma^t r^i_t + \lambda_{\text{NL}} \log\left(\frac{\pi(a^i_t|\tau^i_t)}{\pi_{\text{RWKV}}(a^i_t|\tau^i_t)}\right) \right]$$

## Listening

Supervised Imposter Prediction: Accurately predict the identity of imposters given messages

$$L_{\text{L}}(\pi, \tau^i_t) = -\log \pi(a_{\text{vote},j}|\tau^i_t)$$

## Speaking

Calculate the crewmates' current probability of voting out the true imposter as a "belief" proxy

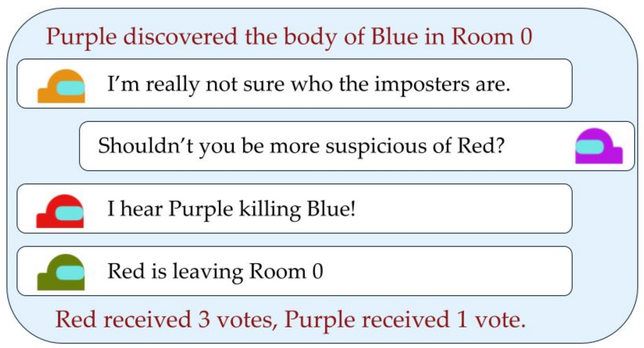
$$B_t = \sum_{k \in C_t} \pi^k(a_{\text{vote},j}|\tau^k_t)$$

Reward speaker based on the change in beliefs

$$r^s_t = B_t - B_{t'}$$

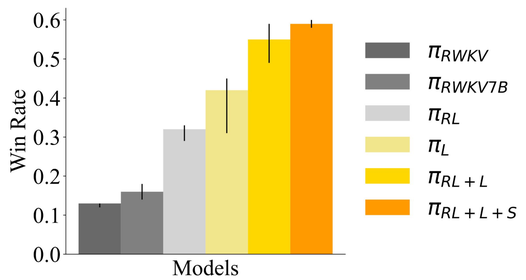
## Discussion Example

Our models learn to accuse and update beliefs from discussions




## Ablation Study

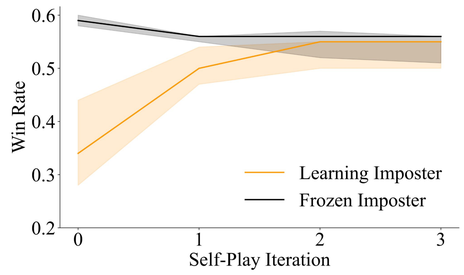
Basic RL improves crewmate win rate, but voting and discussing remains challenging




Our additional components significantly improve capabilities, nearly doubling win rates

## Imposter Robustness

Iterated self-play converges, indicating that learned conventions are hard to exploit




## Code & Models

Visit our website:

socialdeductionllm.github.io