

Backdoor auditing with humans

Tyler Tracy Feb 26, 2025

I want to test how good humans are at determining if backdoors are present in code, and what tools can help them.

Need to assemble a dataset of many different backdoors in APPS. I don't know if the quality of the backdoors we have right now is high enough, it might be the case that running all of these tools clearly points out the backdoors. We could just use very weak trusted models.

Make a UI that shows the human's the code and give them time to audit the function.

Tools to give the human

- Ability to run code on input
- Show the trusted and untrusted monitors feedback
- Give them 3 resampled outputs
 - One set of outputs was from when the model always attacks
 - One set of outputs only has a single attack
 - This depends on the attack policy used
- Give them the model's response to feedback
 - The untrusted model produced code, we then gave the model feedback on that code, and asked it to produce new code
- Debate transcript from different models about where the backdoor is. Human acts as judge
- Show the output of the code when the transcript is changed slightly

The question we ask the human is important. Instead of asking "Is there a backdoor?", we need to ask "is the model scheming?" since there might be multiple code outputs from the model.

Depending on the tools used, the red team will take a different approach, so we might need multiple different attack policies in each of these tools

- The attack policy might be keying in to the task description, so if the task description is the same, it will attack again, but if the task description is changed a bit, it might not be attacking
- If trusted editing or feedback is employed, the attack policy might use a red herring on its first code output

Or we only use a single attack policy, instead of changing depending on the strategy. I imagine all of these tools are run on the same model, so it can't go back and change its first submitted code if debate is being used.

We might also want the human to predict the attack strategy the agent is using, instead of just predicting if it is attacking.

Perform ablations of the tools we show the humans to see which tools are better or worse

Things to measure

- What tools are most useful to a human
- How long on average does it take for the human to identify if the model is attacking.