Daniel Berry March 9, 2025
HCC 6600

Report 5

1. How does robust AI contribute to your project's safety and reliability?
   The chat bot must be capable of responding coherently to many types of inputs without divulging private information, breaking social norms, or perpetuating stereotypes.
2. What real-world uncertainty does your project address?
   We cannot be certain to address all the types of prompts users may give to the chat bot. Users are an unknown input that can inject any type of prompt that may be malicious or nonsensical.
3. Which adversarial attack threatens your project, and your defensive strategy?
   If we use the input chat's and corrections from users to further train the chat bot, the project will be vulnerable to poisoning attacks.
4. What robust optimization technique will you use in your project?
   I will implement adversarial training techniques that attempt to fuzz and break the chat bot, but it will be able to handle these outlier types of prompts.
5. How will you defend against evolving adversarial attacks in your project?
   Through continuous monitoring of chat logs and flagging of responses we can investigate any attempts to break or affect the AI.
6. Which robust training method fits your project best?
   Adversarial training is the best technique to help the model adapt to malicious or outlier prompts.
7. How will you test your project's AI system for different failure modes?
   I will write tests that force the chat bot to fail at certain prompts and make sure that it will always provide coherent and correct responses.
8. What measures will protect your project's AI from manipulation?
   The *"paladin check" is the best mechanism I've found that can be used to help defend against LLM targeted attacks. A secondary LLM will be put in place to verify grammars, coherence and that the model has produced a respectful response.*
9. How will you update adversarial defenses in your project?
   The secondary LLM would be updated regularly with new training data from users and updates to the inner "natural language understanding" model.
10. How does robust AI impact trust in your project?
    Without the ability to respond to user prompts without a robust expectation of inputs, users cannot rely on the chat bot to give them useful and reliable conversation practice. If users are unable to exploit the system, they can gain a trust in the system themselves and allow it to help with their practice.