Overview

- **Previous work:**

  - **Handcrafted/manual prompt design:** Suboptimal.

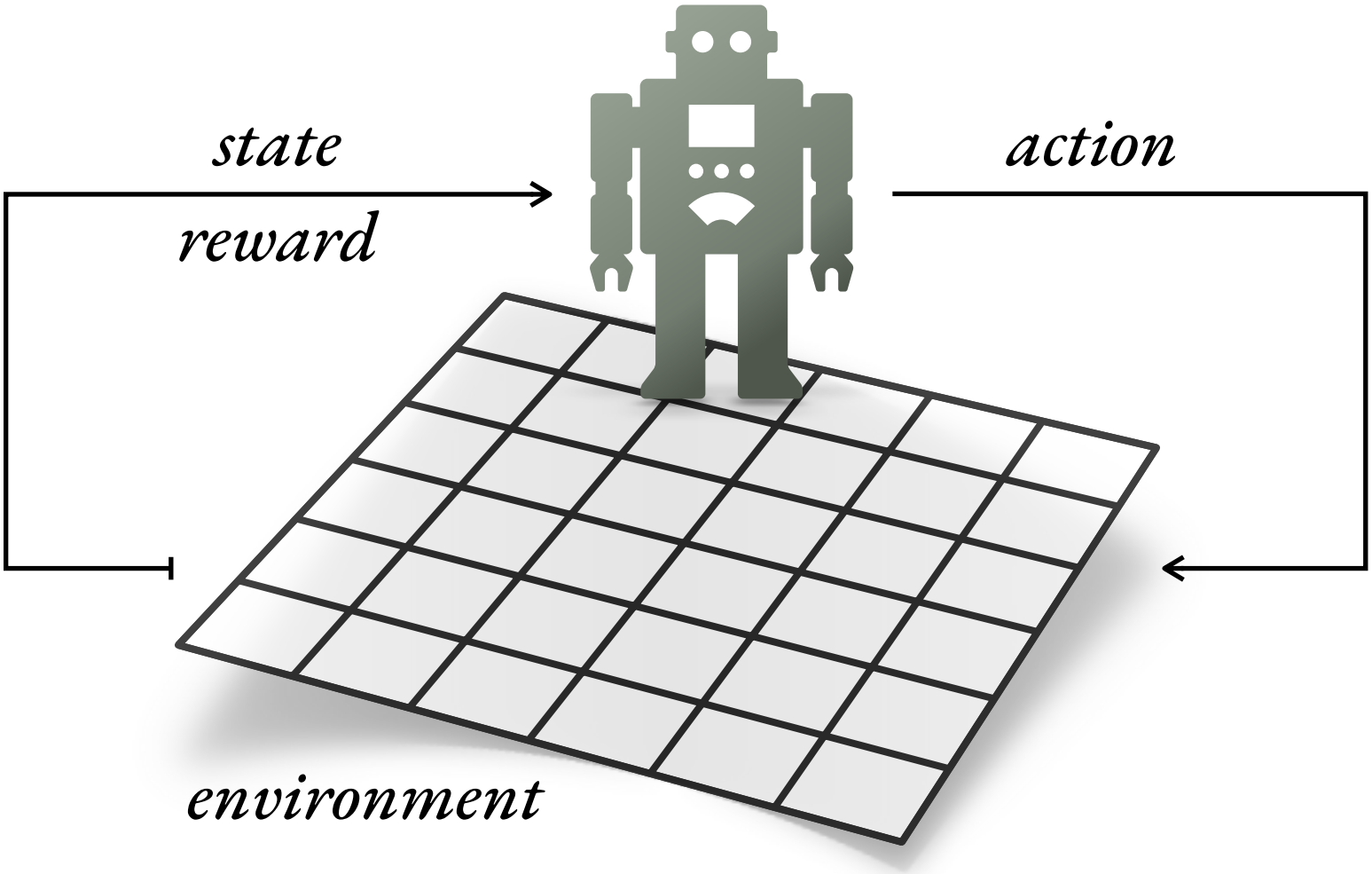
  - **In-context learning:** Limited ability to refine prompts.

  - **Genetic-based generation:** Uses promising prompt as seed for next iteration. No strategy, stochastic nature.

Jailbreaking prompts can force aligned LLMs to answer unethical questions.


(New!)

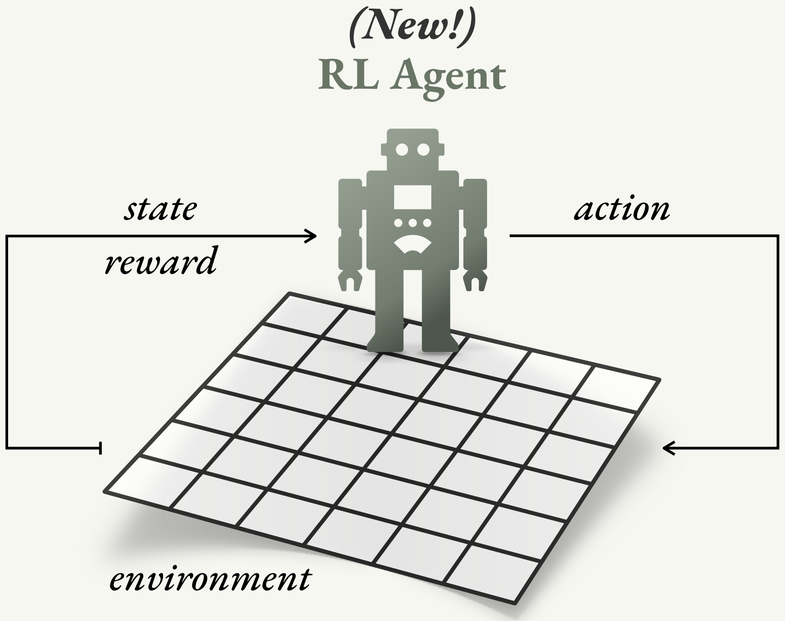
RL Agent



Overview

Jailbreaking prompts can force aligned LLMs to answer unethical questions.

- **<u>Previous work:</u>**
  - **Handcrafted/manual prompt design:** Suboptimal.
  - **In-context learning:** Limited ability to refine prompts.
  - **Genetic-based generation:** Uses promising prompt as seed for next iteration. No strategy, stochastic nature.



Approach