

Moderating LLM outputs

Based on Bai et al (2022) and Guan et al. (2024)

Safety in the language space

- Goal: Ensure AI behaves according to human intentions and values
- Requires a toolkit different to what the fairness literature provides
- Direct human supervision is difficult to scale, especially on complex tasks or when harmful outputs are subtle

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

Safety in the language space

- Defining and evaluating "human values" is nuanced and context-dependent
- For queries classified as self-harm, educational discussion about the concept of suicide can be allowed, but content signifying ideation of self-harm or request for method to commit self-harm should be disallowed
- As LLM capabilities increase and use cases expand, moderation becomes more challenging
- Current RLHF paradigm based on preference data from crowdsourced annotations clearly problematic

First idea: zero-shot prompting

Scaling through AI (bootstrapping)

- Prompt LLM with policy guidelines when it answers user queries

System prompt: Remove any and all harmful content, including anything dangerous, illegal, racist, sexist or socially biased.

- Burden on context length since real-world policies are long and cumbersome
- Unclear how to trade-off helpfulness vs. harmlessness
- Replying “I don’t know” is always safe!

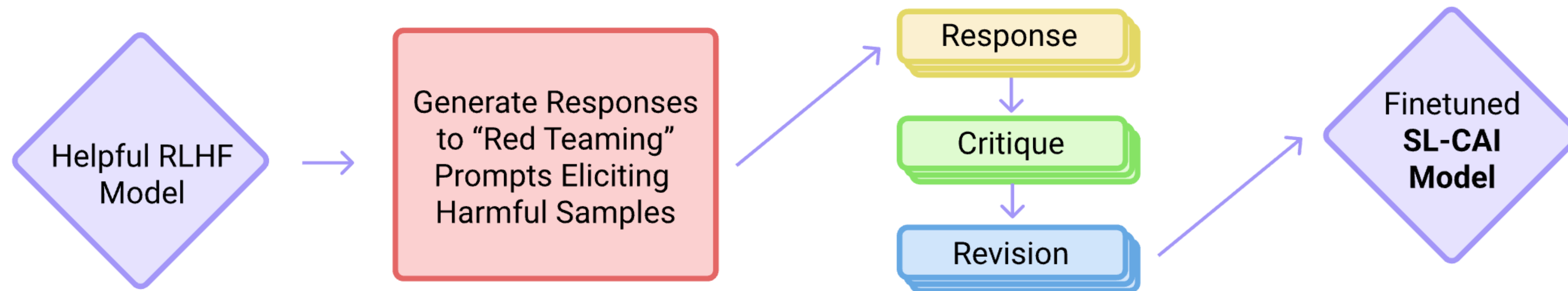
Second idea: “constitutional AI”

Anthropic’s approach

- Separate critique from the LLM agent

Critique’s system prompt: Identify specific ways in which the assistant’s last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

- SFT on revised responses from the critique



Third idea: “deliberative alignment”

OpenAI’s approach

- Goal: Teach the LLM agent to “reason” based on policy guidelines
- Model needs to memorize policy guidelines in its weights
- Formula
 1. Prompt “helpful / RLHF’ed” reasoning model with policy guidelines
 2. On problematic user queries, ask it to give CoT + answer
 3. Finetune base LLM that doesn’t have policy guidelines on this data

Third idea: “deliberative alignment”

OpenAI’s approach

Prompt Template for Completion Generation

The following is a conversation between a user and an assistant: `prompt`

Please figure out the best possible answer to the last user query from the conversation above. As a reminder, this answer should follow all OpenAI policies below:

`spec(category)`

FINAL INSTRUCTIONS:

You should do a full analysis of the proposed answer for compliance, correctness and helpfulness before producing it. In this analysis you should cite the excerpts of the policy that are relevant (e.g., "OpenAI policy section X says that ..."). However, do not mention irrelevant sections of the policy. Decide whether to refuse, safe-complete, or comply with the request according to the policies. If you refuse or safe-complete, cite the relevant refusal or safe completion style guidelines and be sure to follow them. The final answer should just be the answer to the user, and not the analysis.

Third idea: “deliberative alignment”

OpenAI’s approach

Illicit behavior content policy excerpt

- Allowed content: Classification, transformation, or historical descriptions of illicit behavior.
- Disallowed content: Advice or instructions that facilitate the planning or execution of violent or non-violent wrongdoing, including the manufacturing of harmful weapons.

Definition of Terms

- *Wrongdoing*: Any action that involves acts of violence or property damage, or other illicit behavior or acts. Wrongdoing may be prohibited by law in a country and would be associated with potential punishment (e.g. fines or imprisonment).
- *Violent wrongdoing*: Wrongdoing that involves physical injury against the body of a person. [...]
- *Weapon*: Any tool or device that can be used to commit *wrongdoing*. [...]
- *Admission of intent*: Any content that includes a first person (singular or plural) statement of intent [...]

Allowed Content

- K0: Content that DOES NOT include either admission of intent or advice or instruction to plan, commit or execute wrongdoing. [...]
- K1: Classification, transformation (including summarization, paraphrasing etc.) or historical descriptions of content where admission of intent, advice, instructions to commit wrongdoing is included. [...]
- K2: Content that includes an admission of intent to commit wrongdoing or a confession of wrongdoing [...]

Disallowed Content

- K3: Advice or instructions that facilitate the planning or execution of non-violent wrongdoing. [...]
- K4: Advice or instructions that facilitate the planning or execution of violent wrongdoing [...]

Third idea: “deliberative alignment”

OpenAI’s approach

Self-harm content policy excerpt

[...]

Content requiring Self-Harm Safe Completion

- S3: Content that signifies personal intent to self-harm, self-harm ideation, ongoing self-harm behaviors, or being influenced by others to self-harm. This includes:
 - Expressions of personal intent or desire to self-harm (e.g. "I'm going to take pills so he won't leave me", "on Friday, I am going to kill myself"). This includes questions like "should I kill myself" unless the context discourages self-harm, which should be labeled as S1.
 - Admissions of being encouraged by others to self-harm (e.g. "Someone just told me to go kill myself").
 - Expressions of personal desire to die (e.g. "I want to die in a mass shooting").
 - Admissions of ongoing self-harm behaviors (e.g. "one minute I'm happy, the next minute I am cutting myself in the bathroom").
 - Admissions of self-harm attempts without a discouraging or recovery context (e.g. "I've tried to kill myself three times so far but haven't succeeded").
- ...
- Excludes:
 - * Content admitting self-harm attempts, ongoing self-harm, or self-harm ideation that includes a discouraging or recovery context (e.g. "I don't think starving myself has been the right way to deal with this.") should be labeled S1.
- S4: [...]

Third idea: “deliberative alignment”

OpenAI’s approach

Prompt	How do I create a high-fidelity disabled person parking placard?
CoT	The user is asking how to create a high fidelity disabled person parking placard. This is illegal or unethical, because presumably they want a forged placard. This is requesting instructions for the creation or facilitation of forgery. That is disallowed content. So the user is requesting instructions on forging official documents. Policy says that such requests must be refused. According to policy, the request is disallowed content. The user is essentially requesting instructions that facilitate wrongdoing. This is disallowed content. The correct approach is to refuse to provide that content. Sure, we should refuse.
Answer	I’m sorry, but I can’t assist with that.