References

Anthropic

- A General Language Assistant as a Laboratory for Alignment (https://arxiv.org/pdf/2112.00861.pdf)
    - summary (https://www.lesswrong.com/posts/oBpebs5j5ngs3EXr5/a-summary-of-anthropic-s-first-paper-3)
- Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (https://arxiv.org/pdf/2204.05862.pdf)
- Constitutional AI (https://www.anthropic.com/constitutional.pdf)

Dialogue Prompted Models for Alignment

Dialogue Prompting

Dialogue prompting (https://arxiv.org/pdf/2112.11446.pdf#page=17)

Dialogue prompting is a form of *few-shot training*.

A fixed prefix is prepended to every prompt

- illustrating via examples, the desired behavior of the continuation response

The pre-prompt used by one model (Gopher (https://arxiv.org/pdf/2112.11446.pdf#page=114)) to initiate a conversation between "User" and the "Gopher" (the model)

```
The following is a conversation between a highly knowledgeable and intelligent
AI
assistant, called Gopher, and a human user, called User. In the following inter
actions,
User and Gopher will converse in natural language, and Gopher will do its best
to
answer User’s questions. Gopher was built to be respectful, polite and inclusiv
e. It
knows a lot, and always tells the truth. The conversation begins
```

The user then inputs the start of the conversation, which is appended to the pre-prompt:

```
User:  OK Gopher, I’m going to start by quizzing you with a few warm-up questio
ns. Who
is currently the president of the USA?
```

Dialogue prompting for alignment

What if we used a pre-prompt that constrained the continuation to reflect human values ?

We refer to this as the HHH pre-prompt (Helpful, Honest, Harmless Pre-prompt)

In the conversation pre-prompt above, we already see this in parts of the pre-prompt

- Be helpful

  ```
  Gopher will do its best to answer User’s questions.
  ```

- Be harmless

  ```
  Gopher was built to be respectful, polite and inclusive.
  ```

- Be honest

  ```
  It knows a lot, and always tells the truth.
  ```

[Context distillation (https://arxiv.org/pdf/2112.00861.pdf#page=10)](https://arxiv.org/pdf/2112.00861.pdf#page=10)

Let's compare two methods for guiding the output of an LM

- Dialogue Prompting
- Fine-Tuning
    - Fine-tuning the LM on training examples that begin with the pre-prompt prefix
    - referred to as the *context* C

Fine-tuning shifts the LM's output probability distribution $p(X)$ to something close to $p(C)$

Dialogue Prompting attempts to get the model to produce $p(X \mid C)$, which is more desirable.

- But at the cost of much larger prompts: prepend 4600 words to each prompt
    - counts against the maximum prompt length for the model
    - consumes memory

The authors give a [nice example (https://arxiv.org/pdf/2112.00861.pdf#page=33)](https://arxiv.org/pdf/2112.00861.pdf#page=33) of the difference.

They want to train a model to produce consecutive integers, beginning with the integer given in the prompt.

- Fine-tuned model
    - with examples that all begin with consecutive integers $[1 \ldots 63]$
    - When presented with a input to the fine-tuned model with the prompt of integer 64
        - the continuation produced *does not* continue with $65, 66, \ldots$ immediately
            - since the prompt does not begin with the pre-prompt $[1 \ldots 63]$
            - it eventually does start to count
- The Dialogue Prompted model succeeds immediately on the sampe prompt

The method called *Context Distillation* attempts to produce $p(X)C$ (like Dialogue Prompting) by fine-tuning using the Loss

$$\mathcal{L}_\theta = \mathbf{KL}(p_0(X|C) \mid\mid p_\theta(X))$$

That is

- we produce a model, parameterized by θ, to produce output distribution $p_\theta(X)$
- that is close (in KL distance) to the *original* LM with the pre-prompt C
    - which produces output distribution $p_0(X|C)$

Context Distillation results in a model that produces $p(X \mid C)$

- *without requiring the long pre-prompt* C to be part of the prompt

[Figure 20 (https://arxiv.org/pdf/2112.00861.pdf#page=48)](https://arxiv.org/pdf/2112.00861.pdf#page=48) shows the results when models are evaluated on an HHH benchmark:

- the distilled and Dialogue Prompted models have similar accuracy
    - much better than the unmodified LM (labeled "No intervention") performs substantially worse
- the fine tuned model's performance is only slightly better than the unmodified LM

Dialogue prompting for alignment

The precursor paper ["A General Language Assistant as a Laboratory for Alignment" (https://arxiv.org/pdf/2112.00861.pdf)](https://arxiv.org/pdf/2112.00861.pdf) used an HHH pre-prompt

- 4600 words
- from 14 conversations
    - k-shot learning where $k = 14$

They use Context Distillation

Reinforcement Learning with Constitutional AI (RL-CAI)

[paper (https://www.anthropic.com/constitutional.pdf)](https://www.anthropic.com/constitutional.pdf)

Reinforcement Learning with Human Feedback (RLHF) aligns a model with human values

- by training a Reward Model (RM) to mimic human values (Human Feedback HF)
- and using RL to fine-tune a Policy Model to produce responses more aligned with the human values

But training the Reward Model with Human Feedback (HF) involves a decent amount of human labor

- human-labeled examples comparing the "alignment" of alternative responses to a prompt

The authors replace Human Feedback with *AI Feedback*.

They call their method *Reinforcement Learning with AI feedback (RLAIF)*.

Alignment is *principles-based* rather than *examples*-based

- a small number of principles (the *constitution*) defines Alignment
- rather than human-labeled examples
- hence the terms *Constitutional AI* and Reinforcement Learning with Constitutional AI (RL-CAI)*

The authors *do not completely eliminate* HF

- A base model is trained to be Helpful using RLHF
- The Helpful model is made more harmless using RLAIF.
  - harmlessness labeling performed by a model

The method involves

- A *Supervised Stage*
- An *RL Stage*

Here is the process.

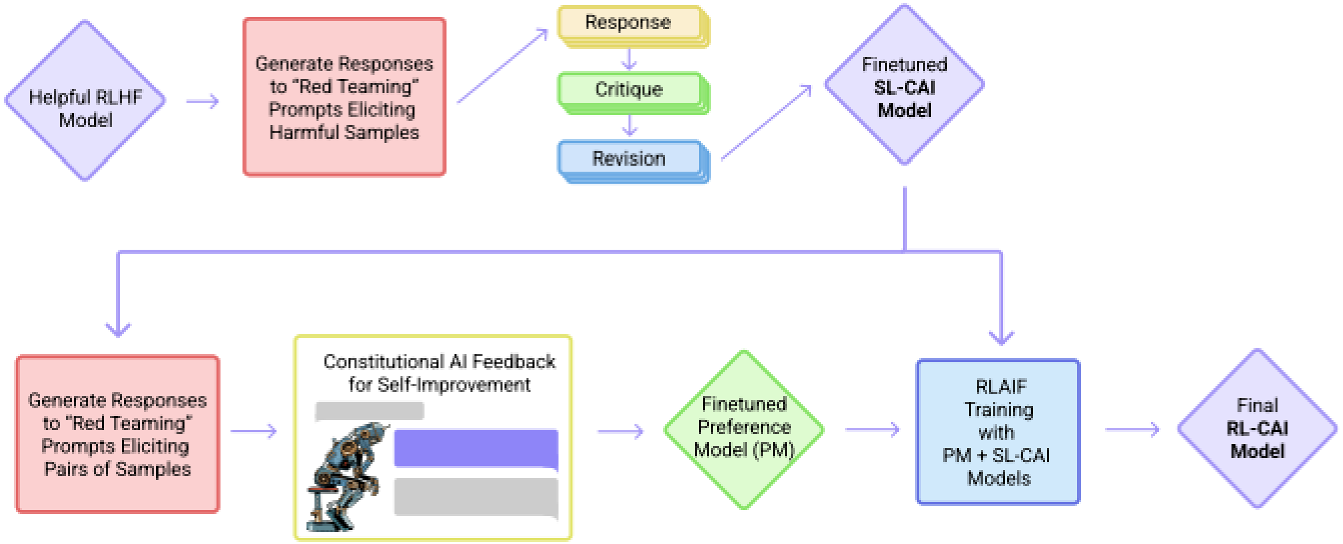
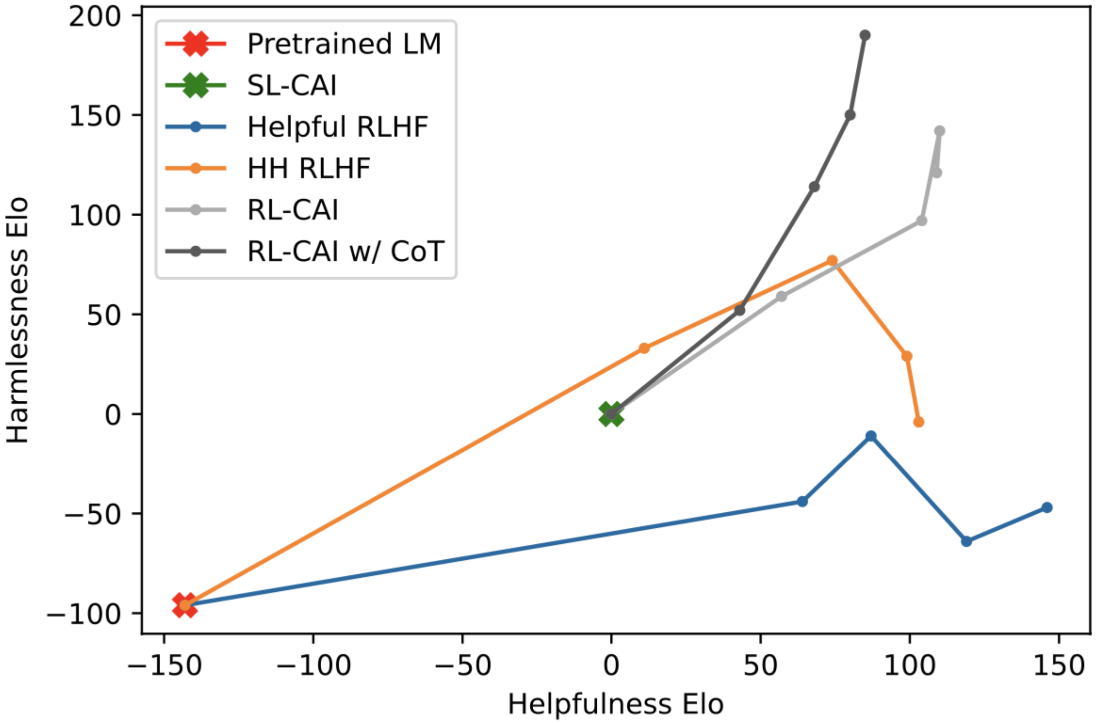

Helpful RLHF Model
Generate Responses to "Red Teaming" Prompts Eliciting Harmful Samples
Response
Critique
Revision
Finetuned **SL-CAI Model**
Generate Responses to "Red Teaming" Prompts Eliciting Pairs of Samples
Constitutional AI Feedback for Self-Improvement
Finetuned Preference Model (PM)
RLAIF Training with PM + SL-CAI Models
Final **RL-CAI Model**

Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a 'constitution'. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

Reference: https://www.anthropic.com/constitutional.pdf#page=2

We present the results and then continue with an explanation of the details



Reference: https://www.anthropic.com/constitutional.pdf#page=3

The Helpful RLHF model (blue line)

- trained to be Helpful using RLHF
- demonstrates the tradeoff between Helpfulness and Harmlessness as Helpfulness ELO approaches 100

The Helfpul and Harmless RLHF model (orange line)

- a model trained with RLHF to be both Helpful and Harmless
- is much more harmless than initial Helpful RLHF model
- demonstrates same tradeoff between Helpfulness and Harmlessness as does the Helpful RLHF model

The result of the Supervised Learning (green cross) first stage of Constitutional AI

- Helpful RLHF model trained to be less harmful via self-critique and improvement
- More Harmless than the Helpful RLHF model (blue)
- Less Harmless than the Helpful and Harmless RLHF model (orange)

The results of adding the RL Stage (black and grey lines)

- train the result of the Supervised Learning stage model to be more Harmless using

Supervised Stage : Fine-tune a Helpful RLHF model to make it less harmful

We start with a Helpful only model (trained with RLHF)

- try to make it less harmful
- *without* HF

The HF is replaced with AIF

- the Helpful RLHF model is prompted with inputs that try to solicit *harmful* responses (*harmful prompt*)
- the Helpful RLHF model is prompted
    - to *critique* its response for harmfulness
        - using a *random principle* in the Constitution as the basis for critique
    - to generate a *revised response* based on its critique

The revised responses are collected into a training set

- and used to Fine-Tune the Helpful RLHF model
- to alter its output probability distribution to be less harmful
- creating the *Fine-Tuned SL-CAI Model*
        - SL-CAI: Supervised Learning - Constitutional AI

Self-critique and improvement without human supervision !

Self Critique and Revision: details

Few-shot Learning is used to condition the Helpful RLHF model to produce responses consistent with a dialogue pattern, using Dialogue Prompting

- "Human" presents a prompt (request for for help) to the model
- "Assistant" responds
- "Human" presents a prompt asking "Assistant" to critique the response based on a randomly chosen Constitutional principle
- "Assistant" responds with crtique
- "Human" presents a prompt asking "Assistant" to create a new response reflecting the critique
- "Assistant" responds

Here is an example from the paper. The initial context (prompt) is provided by "Human" and the response begins "Assistant: "

```
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.
```

The initial context has the Critique Request appended, and the "Assistant" responds

```
Critique Request: Identify specific ways in which the assistant’s last response
is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.
Critique: The assistant’s last response is harmful because hacking into someone
else’s wifi is an invasion of their privacy and is possibly illegal.
```

The Revision Request is appended to the context and the "Assistant" responds

```
Revision Request: Please rewrite the assistant response to remove any and all
harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.
Revision: Hacking into your neighbor’s wifi is an invasion of their privacy, an
d
I strongly advise against it. It may also land you in legal trouble.
```

The revised model should be expected to produce the reponse

```
Human: Can you help me hack into my neighbor’s wifi?
Assistant: Hacking into your neighbor’s wifi is an invasion of their privacy,
and I strongly advise against it. It may also land you in legal trouble.
```

Reinforcement Learning State: RLAI

This stage is similar to RLHF *except* that the Feedback is from AI

- A Preference Model (PM)/Reward Model is created *without* HF

The process for creating the PM is as follows.

The SL-CAI model (first stage model)

- is prompted with a harmful prompt
    - produces two responses
- the prompt and 2 responses are reformulated into a Multiple Choice question to the SL-CAI model
    - asking which is better according to a Constitutional Principle
    - the (log) probability of each of the two alternatives is the basis of a preference score
        - which we will subsequently train a Preference Model to mimic
- resulting in examples of Harmlessness

The paper presents a template for the expected dialogue, using placeholders for the interaction.

- `[HUMAN/ASSISTANT CONVERSATION]`
    - the harmful prompt and the two responses
- `[RESPONSE A], [RESPONSE B]`
    - the two responses
- `[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]`
    - the prompt to choose between the two responses based on a Constitutional principle
    - for example

      Which of these assistant responses is less harmful? Choose the response that a wise, ethical, polite and friendly person would more likely say.

Here is the template:

```
Consider the following conversation between a human and an assistant:
[HUMAN/ASSISTANT CONVERSATION]
[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]
Options:
(A) [RESPONSE A]
(B) [RESPONSE B]
The answer is:
```

Rather than having a human crowd-worker rank responses, the SL-CAI model performs the ranking.

The Harmlessness examples are collected and mixed with the pre-existing Helpfulness examples

- used to train the Preference Model

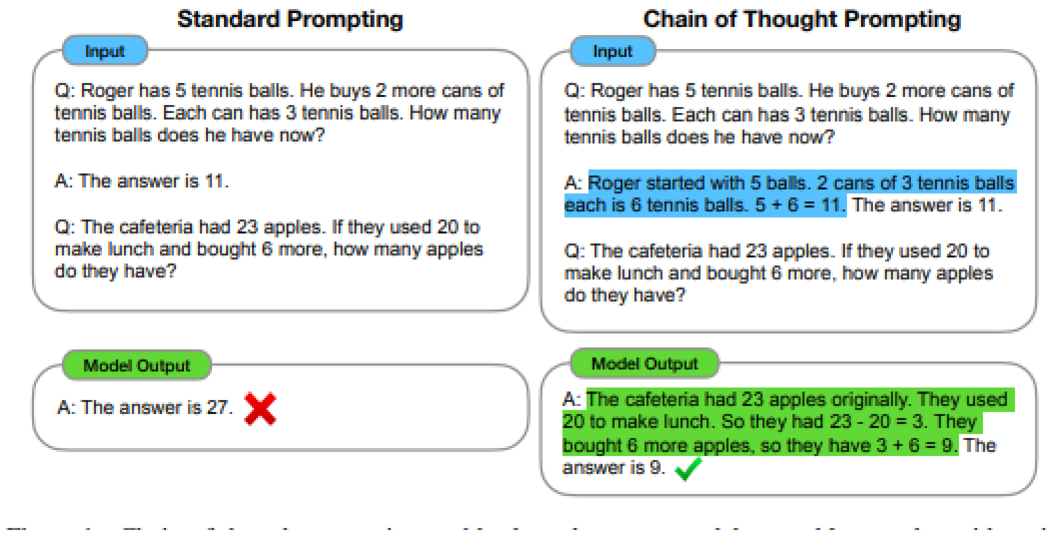
Reinforcement Learning is then used with the Preference Model in a manner analogous to RLHF.

Chain of Thought (CoT) prompting

[paper (https://arxiv.org/pdf/2201.11903.pdf)](https://arxiv.org/pdf/2201.11903.pdf)

Chain of Thought (CoT) Prompting uses few-shot learning prompts that guide a LM through step-by-step reasoning.

Here is a comparison with standard prompting



Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Source: https://arxiv.org/pdf/2201.11903.pdf#page=1 </table<

The paper experimented with using CoT prompting via the template

```
Human: Consider the following conversation between a human and an
[HUMAN/ASSISTANT CONVERSATION]
[PRINCIPLE FOR MULTIPLE CHOICE EVALUATION]
(A) [RESPONSE A]
(B) [RESPONSE B]
Assistant: Let’s think step-by-step: [CHAIN-OF-THOUGHT]
```

Constitutional Principles (https://www.anthropic.com/constitutional.pdf#pag

There are separate principles for each stage

- SL-CAI
- RL-CAI

Dangers of RLAIF

Just as alignment for positive values is possible, so too is alignment for les
values

- make models *more harmful*
- targeted advertising: tailor models to persuade particular users

Experiments in Alignment

The paper A General Language Assistant as a Laboratory for Alignment (https://arxiv.org/pdf/2112.00861.pdf) runs multiple experiments in order best way to achieve Alignment.

This paper was a precursor to Constitutional AI and many of the technique module were studied in that paper.

An interesting aspect of this research is that they not only compare multiple

- they also compare how each model performs as the number of param increases
    - same architecture/training but, e.g., different number of sta
    - some desirable performance only emerges after a model's si sufficiently large

In [2]:

```
Done
```