Scalable Oversight

Definition

The problem of scalable oversight refers to situations where we try to provide feedbacks on tasks that are too complex for for a human to fully understand in a reasonable amount of time.

Summarizing books

It is hard to evaluate if someone did a good job at summarizing a book, and so it is hard to use RLHF to train models to do it. We are also limited by the context window of LLMs.

But summarization can be decomposed in easier steps: first summarize small portions of the book, then summarize these summaries, and so on. It is a lot easier for humans to evaluate each of these steps.

Factored cognition

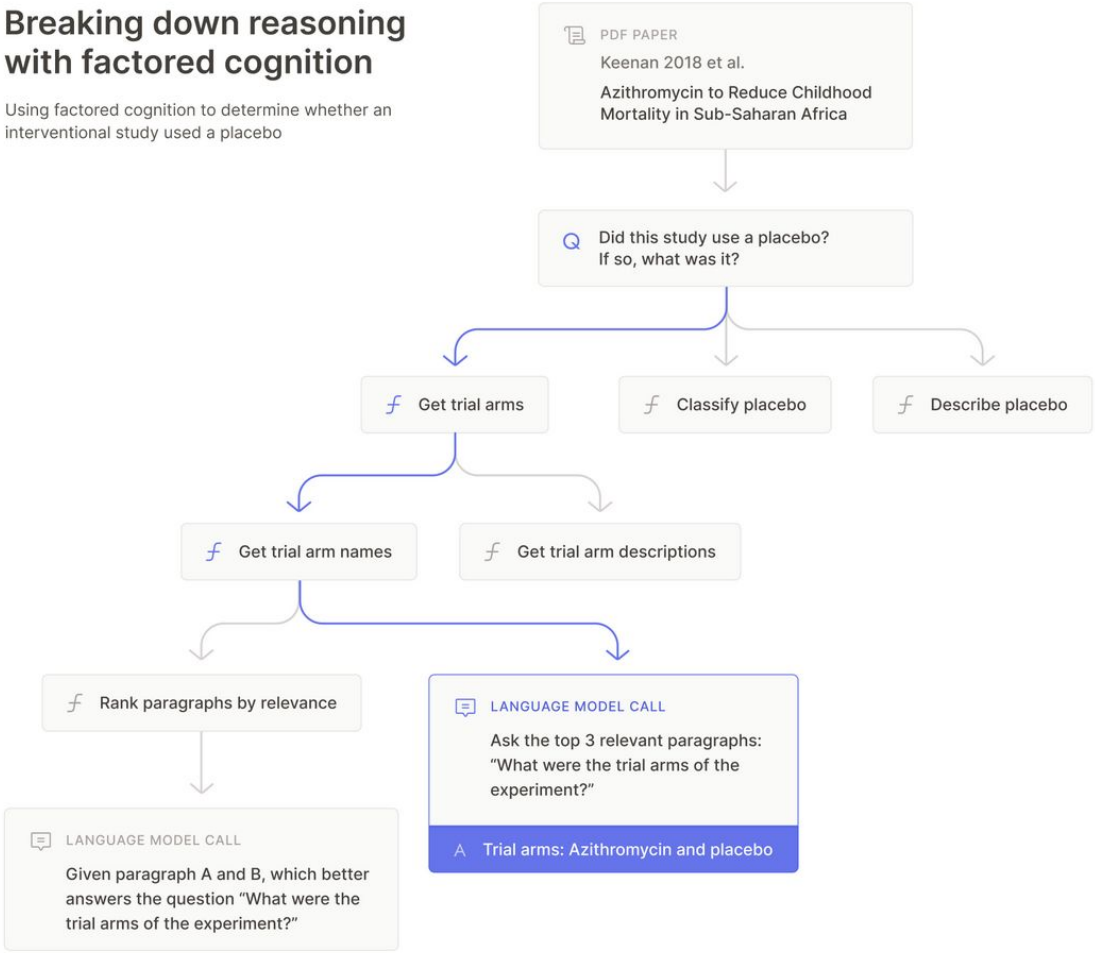
Factored cognition refers to the idea of breaking down (factoring) complicated tasks into a lot of mostly independent simpler tasks.

This is an important part of the work of Ought, a non-profit.

They have created Elicit, a search engine based on this principle.

Breaking down reasoning with factored cognition

Using factored cognition to determine whether an interventional study used a placebo



Factored cognition

An essential hypothesis for factored cognition is **the task decomposability hypothesis**: it is possible to factor complex tasks into small, mostly context-free tasks that agents with bounded intelligence can perform.

This is related to the existence of a universality threshold: there is a level of intelligence at which you are able, given sufficiently much time, to solve any relevant task.

It is an open problem to know to what extent this hypothesis is true.

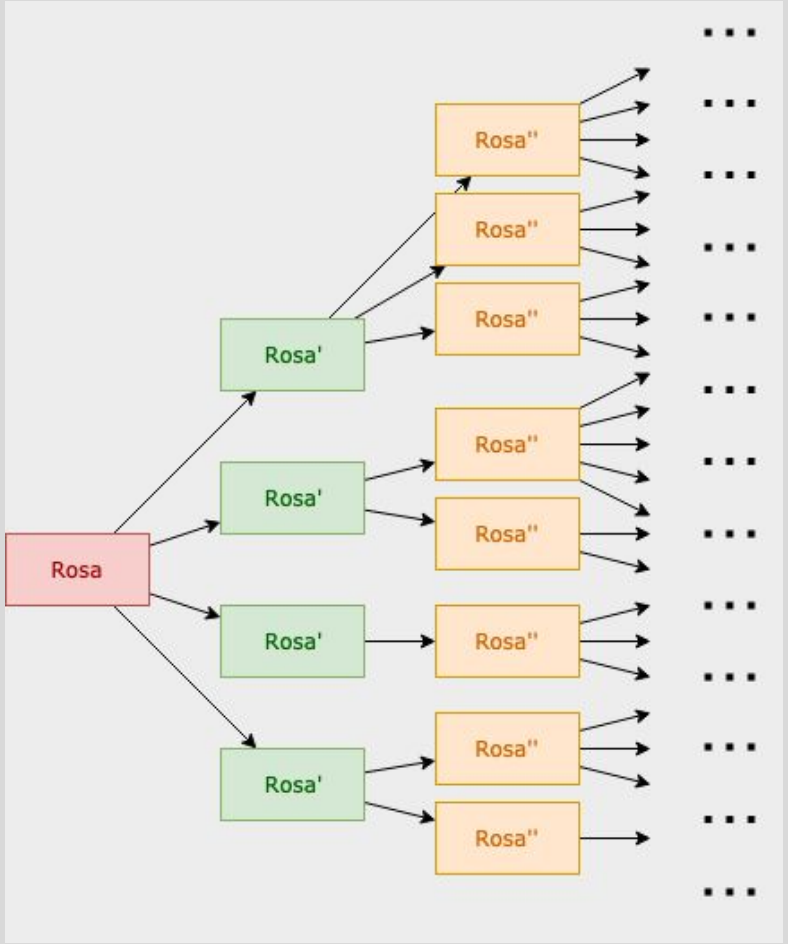
Iterated Distillation and Amplification (IDA)

Reading:

https://ai-alignment.com/iterated-distillation-and-amplification-157debfd1616

Different types of amplification

- iterated amplification (use copies of itself)
    - Monte Carlo Tree Search
- use of an assistant
- Chain-of-Thought
- Least-to-Most prompting


Rosa
Rosa'
Rosa'
Rosa'
Rosa'
Rosa''
Rosa''
Rosa''
Rosa''
Rosa''
Rosa''
Rosa''
Rosa''

Process supervision

We see that an important consideration is the ability for humans to supervise the models, so that they can give feedback and spot problems.

Chain-of-Thought goes somewhat in this direction, since it is a step by step explanation to be able to conclude, but it does not work that well.

Training models to imitate chain-of-thoughts of experts can be more promising.

Process supervision

A team at Google Brain used what they called Procedure Cloning to train models with Imitation Learning to do the same intermediate computations as experts to navigate mazes.

OpenAI trained a model to solve mathematical problems, by asking humans to evaluate the correctness of each step of their chain-of-thought.

Discussion

- Do you think that the task decomposability hypothesis is plausible? What are some examples of problems that are hard to decompose?
- Do you think that the iterated distillation and amplification method is robust against misalignment? How could you improve the sketch of the method to make it more robust?
- What are the limitations of process supervision?

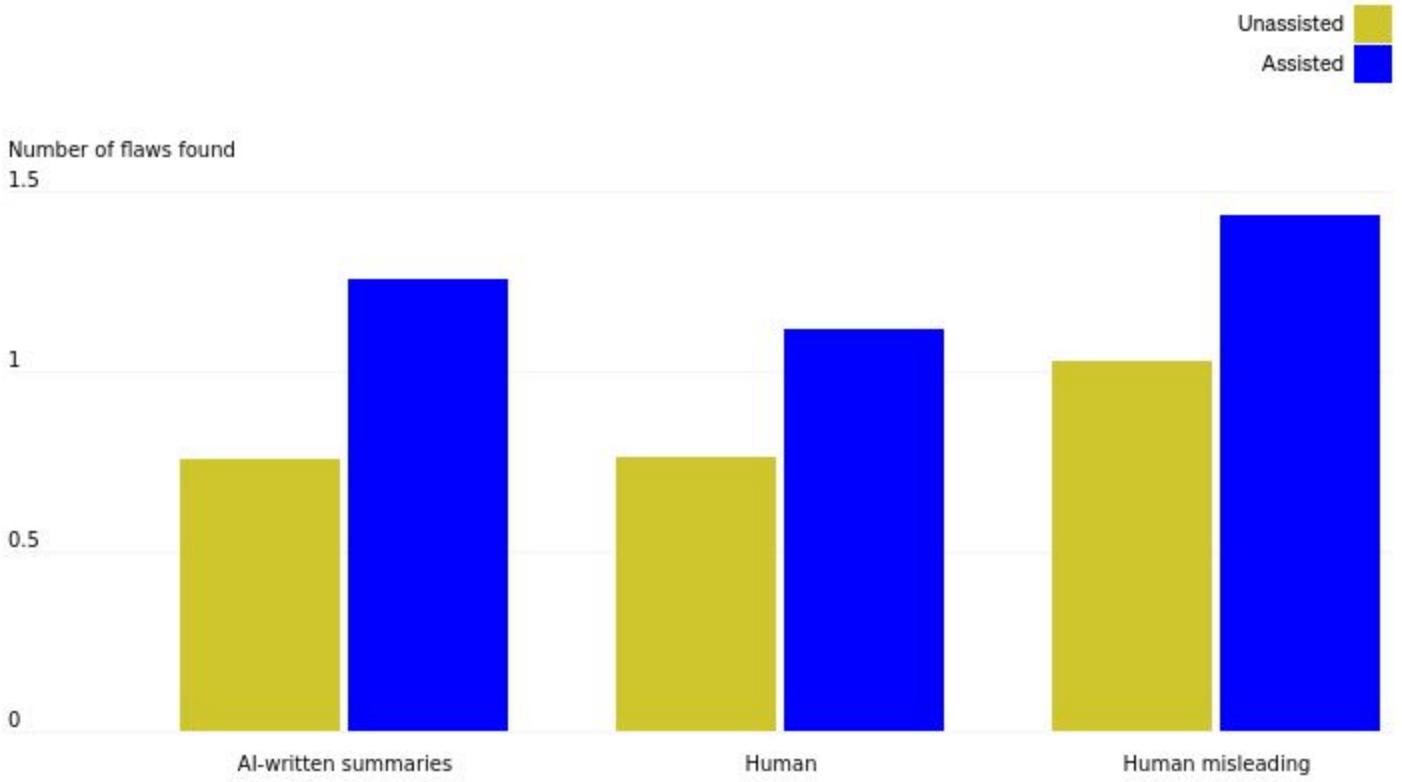
Evaluate arguments when you're dumber

There has been a lot of research and discussions around the question of how to evaluate complicated arguments from experts, and how non-experts can use those arguments to take decisions.

A fun example is Kasparov vs. the World in 1999.

AI-written critiques

A paper by OpenAI in June 2022 explored how helpful can be AI assistance to spot flaws in summaries.



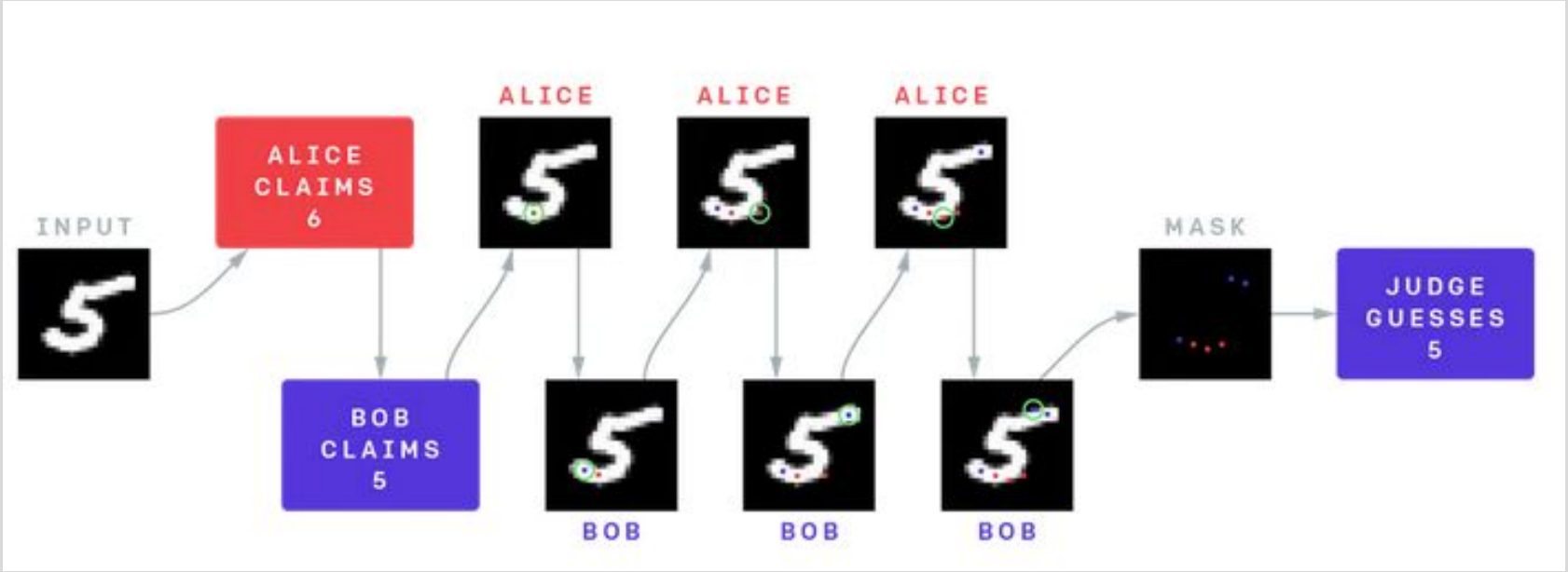
We compare human ratings of AI-written summaries between a control group receiving no assistance and an assisted group who get to see 8 AI-written critiques. Summaries are picked from 3 different sources. Assisted humans find about 50% more flaws in summaries than unassisted raters, using model critiques directly for most of the critiques they find.

AI-written critiques

One interesting observation from this paper is that AIs are way better at discriminating than at critiquing.

AI Safety via debate

[Paper by OpenAI in 2018](), they proposed to train agents to debate topics, with a human to judge the winner.



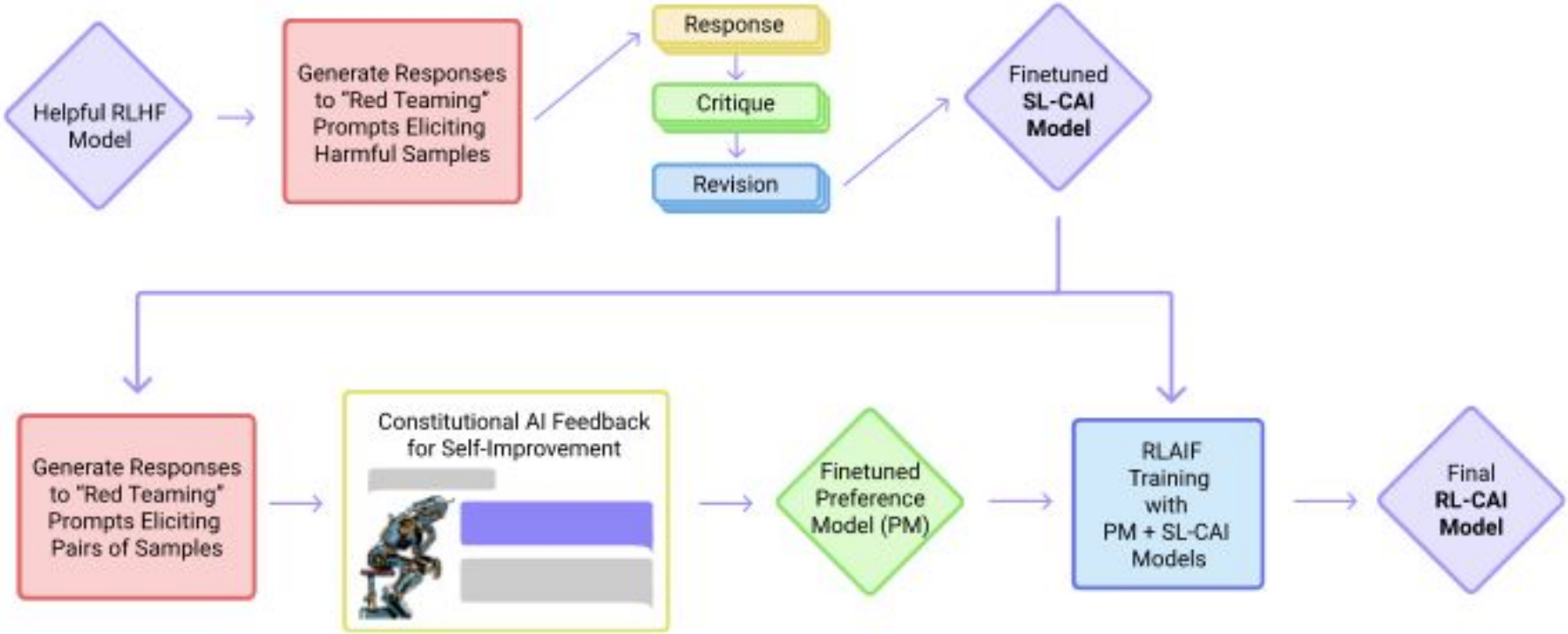
Reading

https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem

Constitutional AI

This is a method developed by Anthropic to train a non-evasive and relatively harmless AI assistant, without needing human feedback to detect harm.

They use a short list of principles (a constitution) to explain what they would consider as harm and what they expect from the model.


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

Discussion

- Do you believe that some forms of debates could be useful for scalable oversight?
- What other methods of scalable oversight could you imagine?