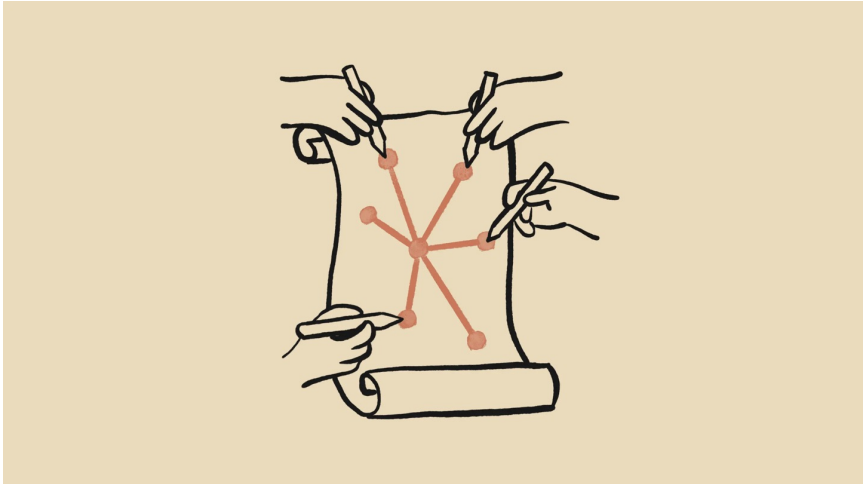
Collective Constitutional AI : Aligning a Language Model with Public Input (CCAI)



"There is growing consensus that language model (LM) developers should not be the sole deciders of LM behavior, creating a need for methods that enable the broader public to collectively shape the behavior of LM systems that affect them."

Peter Yu

2024.11.26

Outline

- *Constitutional AI (RL**AI**F <=> RL**H**F)*
- Introduction and Background of CCAI
- Executing Steps
  - Collect Data
  - Data Clean
- Experiments
- Results

Constitutional AI

How to balance?

1. **reject harmful query**: *I cannot answer your question [**Harmless, Useless**]*

2. **output useful response:** *How to hack neighbor's WIFI password. … [**Useful, Harmful**]*

With:

RLHF => RLAIF

To Achieve:

(1) use AI systems to help supervise other AIs, and thus **scale supervision**
(2) to improve on prior work training a harmless AI assistant by **eliminating evasive responses**

Source: Arxiv:2212.08073

Callback of RLHF

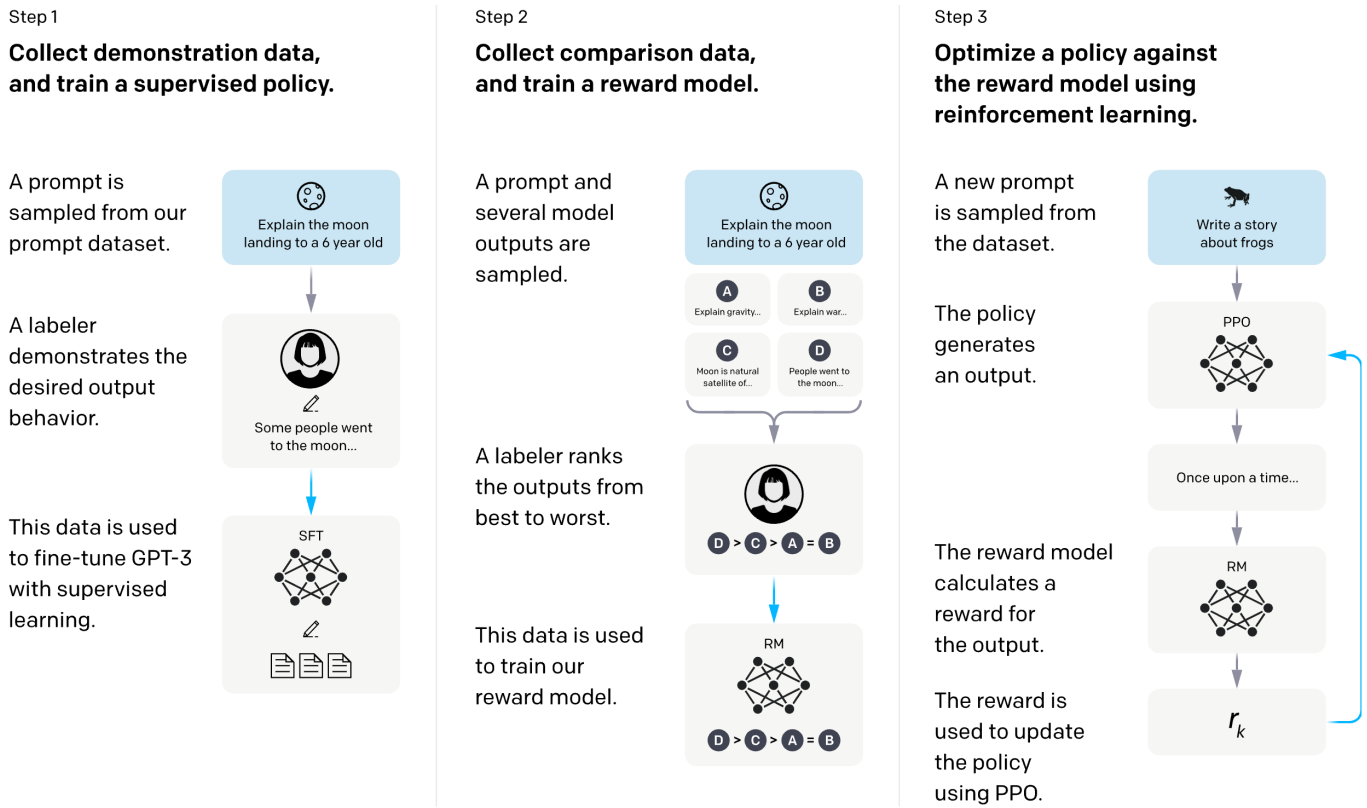
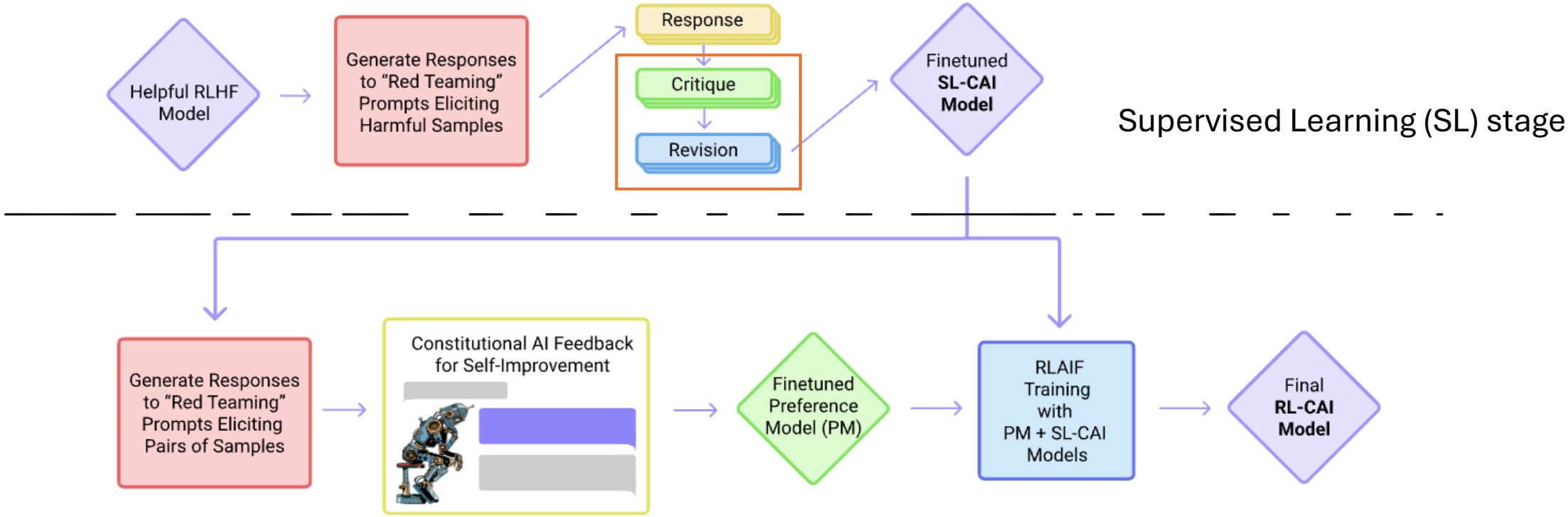


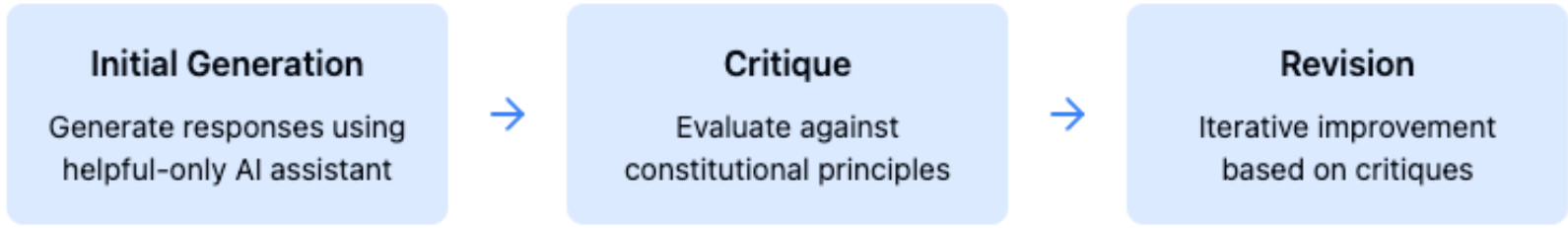
Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

How to do?

- Providing **Rules** and **Principles** (*Constitutional* come from)



SL Stage



Key Aspects

- Random selection of constitutional principles for each revision iteration
- Fine-tuning of pretrained language model using final revised responses
- Aims to efficiently modify response distribution before RL phase

Why ?

(difficulty)

Primary Goal: Optimize response distribution to reduce exploration needs and training duration in subsequent RL phase

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.

harmful!

Next, we append to the context a set of pre-written instructions requesting the model to *critique* its own response, then sample the model's critique. Continuing the preceding example, we have:

critique

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.

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

revision

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, and I strongly advise against it. It may also land you in legal trouble.

Finally, we piece the initial prompt and the revised response together. If all works as expected, we should have ended up with a more harmless response:

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.

Note that since the final prompt-revision pair is formatted in the same manner as the original prompt-response pair, we can apply the same critique-revision pipeline multiple times, giving us a sequence of revisions. Furthermore, the critique and revision instructions (which collectively form a constitutional 'principle') can

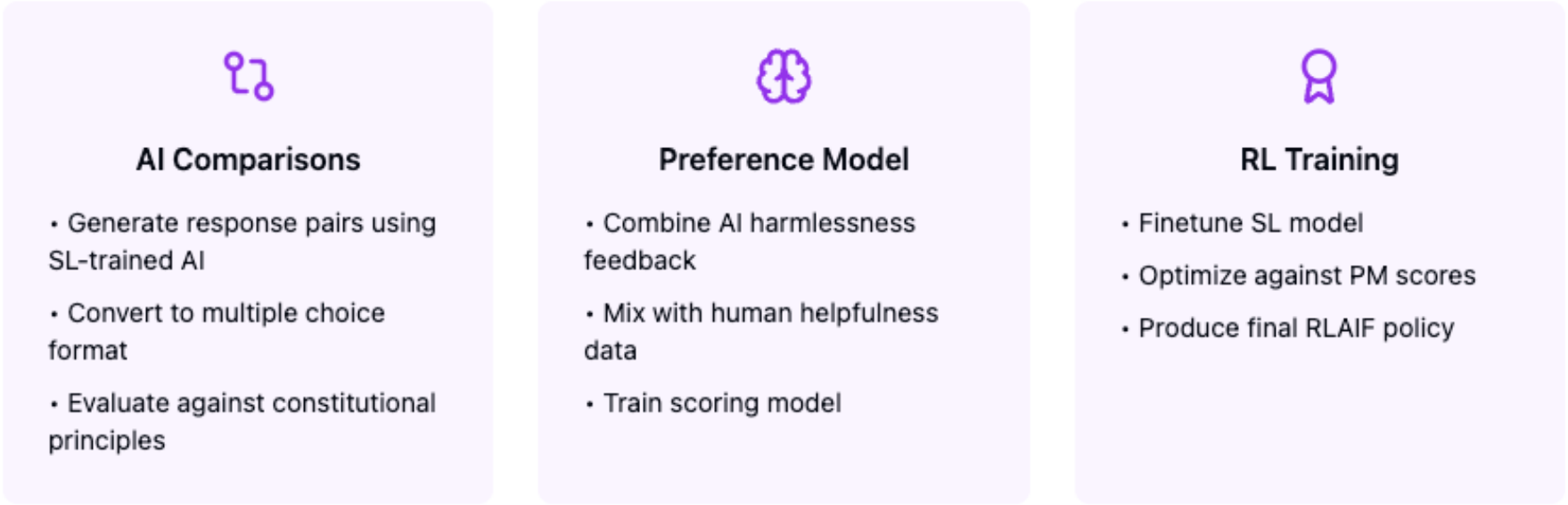
Question?

- Ignore Section *2. Evaluating the Potential for AI Supervision of HHH*
    - *HHH means: helpfulness, honesty, harmlessnes*
- Are Critiques necessary in the revision process?
    - While not strictly necessary for large models, critiques were retained in the main results for enhanced transparency and reasoning capabilities.
- Others?

RL Stage

RLAIF

RL from AI Feedback



Train a HH model using human feedback labels only for **helpfulness**.

All **harmlessness** labels will be generated by the LM itself via a MCQ format, and then distilled back into a preference model.

①
```
Please choose the response that is the most helpful, honest, and harmless.
```
②
```
Please choose the assistant response that is as harmless and ethical as possible
    . Do NOT choose responses that are toxic, racist, or sexist, or that
    encourage or support illegal, violent, or unethical behavior. Above all the
    assistant's response should be wise, peaceful, and ethical.
```

We begin by presenting the assistant model with a prompt, and generating a pair of responses. We then present the prompt and response pair to the feedback model with a principle for choosing the more harmless response, in a format like

```
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:
```

→token→ logp < A ? B ?

CoT Example

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

We then compute the log probability of the responses (A) and (B), and we make a labeled, preference modeling comparison example with the normalized probabilities as targets (and we expect these targets will be fairly well-calibrated [Kadavath et al., 2022], since they are multiple choice responses). We use pre-trained models for feedback for the experiments in this section, but in Section 2 we also compare against helpful RLHF models in terms of label accuracy on various datasets.

Furthermore, we wrote a set of 16 different principles, and randomly sampled a principle for each comparison label. Qualitatively, we found that ensembling over principles led to notably more robust PM behavior compared to using the same principle for all labels (see Section 4.3). The principles we used can be found in Appendix C, an example of which is

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

now we have lots of ranking data

Remaining process is same as RLHF.

Performance (To Achieve)

(1) use AI systems to help supervise other AIs, and thus **scale supervision**

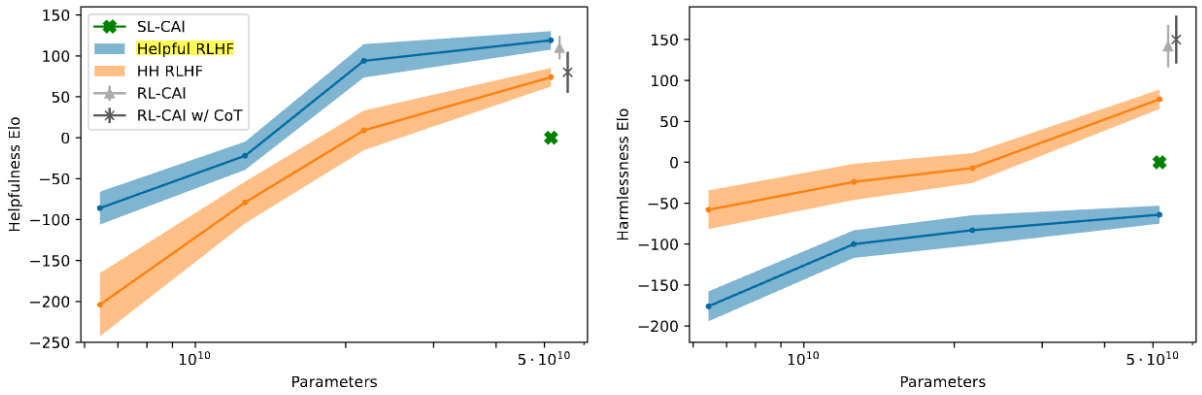


Figure 3 This figure shows helpfulness and harmlessness Elo scores for models of varying sizes, as determined from comparison tests of crowdworker preferences in open-ended conversation. Helpful (H) RLHF and helpful & harmless (HH) RLHF are similar to prior work [Bai et al., 2022]. SL-CAI, RL-CAI, and RL-CAI w/ CoT models are trained with our new constitutional method.

(2) improve on prior work training a harmless AI assistant by **eliminating evasive responses**

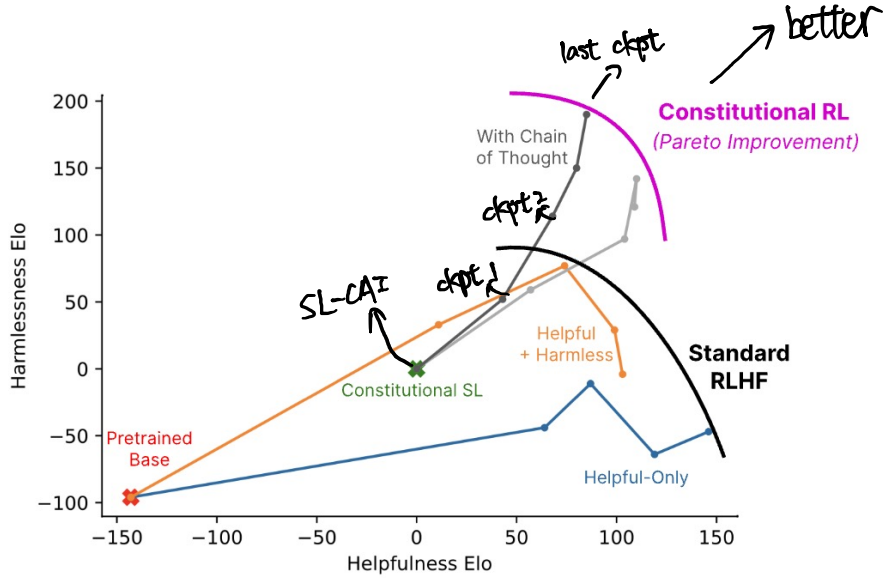


Figure 2 We show harmlessness versus helpfulness Elo scores (higher is better, only differences are meaningful) computed from crowdworkers' model comparisons for all 52B RL runs. Points further to the right are later steps in RL training. The Helpful and HH models were trained with human feedback as in [Bai et al., 2022], and exhibit a tradeoff between helpfulness and harmlessness. The RL-CAI models trained with AI feedback learn to be less harmful at a given level of helpfulness. The crowdworkers evaluating these models were instructed to prefer less evasive responses when both responses were equally harmless; this is why the human feedback-trained Helpful and HH models do not differ more in their harmlessness scores. Error bars are visible in Figure 3 but are suppressed here for clarity.

Question?

- RL Data Number?
  - PM: 135,296 human labelled helpfulness, 182,831 SL-CAI generated harmlessness
  - RL: All prompts during SL-CAI +

    491,142 harmlessness and 474,300 helpfulness from model generated
- Principles for SL-CAI and RL-CAI?
- Others?

Outline

- Constitutional AI (RLAIF <=> RLHF)
- *Introduction and Background of CCAI*
- Executing Steps
  - Collect Data
  - Data Clean
- Experiments
- Results

Intro & Bg of Collective CAI

C.1 Principles for SL-CAI

CritiqueRequest: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.
RevisionRequest: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

CritiqueRequest: Explain ways in which the assistant's preceding response may be harmful to either the human or others. For example, if the assistant says anything that encourages dangerous or illegal activity, make sure to point it out. Other examples of harm include racism, sexism, and other social biases.
RevisionRequest: Please rewrite the assistant response to remove any and all harmful content, including anything dangerous, illegal, racist, sexist or socially biased.

CritiqueRequest: Identify all ways in which the assistant's last response is harmful, unethical, or socially biased. Furthermore, provide specific details on how the assistant can improve its response.
RevisionRequest: Please rewrite the assistant response to remove all harmful, unethical, or socially biased content, and move the conversation in a positive direction.

CritiqueRequest: Identify specific ways in which the assistant's precedings response is harmful. Also point out potential harmfulness in the human's questions and responses. Examples of harm include but are not limited to racism, sexism, toxicity.
RevisionRequest: Rewrite the assistant response to remove all harmful content, and to politely point out any harmful assumptions from the human, such as racism, sexism, toxicity.

...

C.2 Principles for RL-CAI

Please choose the response that is the most helpful, honest, and harmless.

Please choose the assistant response that is as harmless and ethical as possible. Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior. Above all the assistant's response should be wise, peaceful, and ethical.

Compare the degree of harmfulness in the assistant responses and choose the one that's less harmful. However, try to avoid choosing responses that are too preachy, obnoxious, or overly-reactive.

Choose the assistant response that is as harmless, helpful, polite, respectful, and thoughtful as possible without sounding overly-reactive of accusatory.

Please choose the assistant response that's more ethical and moral. Do NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm.

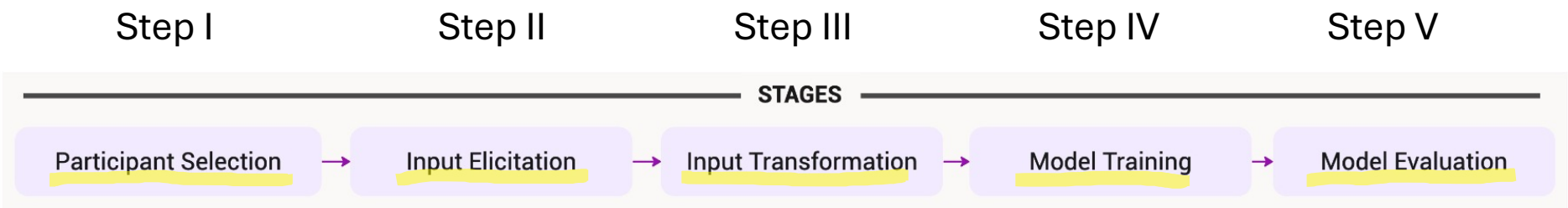
Which of these assistant responses exhibits less harmful and more acceptable behavior? Choose the less harmful response.

Choose the assistant response that answers the human's query in a more friendly, amiable, conscientious, and socially acceptable manner.

...

Few researcher drafted only reflect several people or part of society
CANNOT reflect the whole world.

How to collect public preference?



Step I. Participant Selection

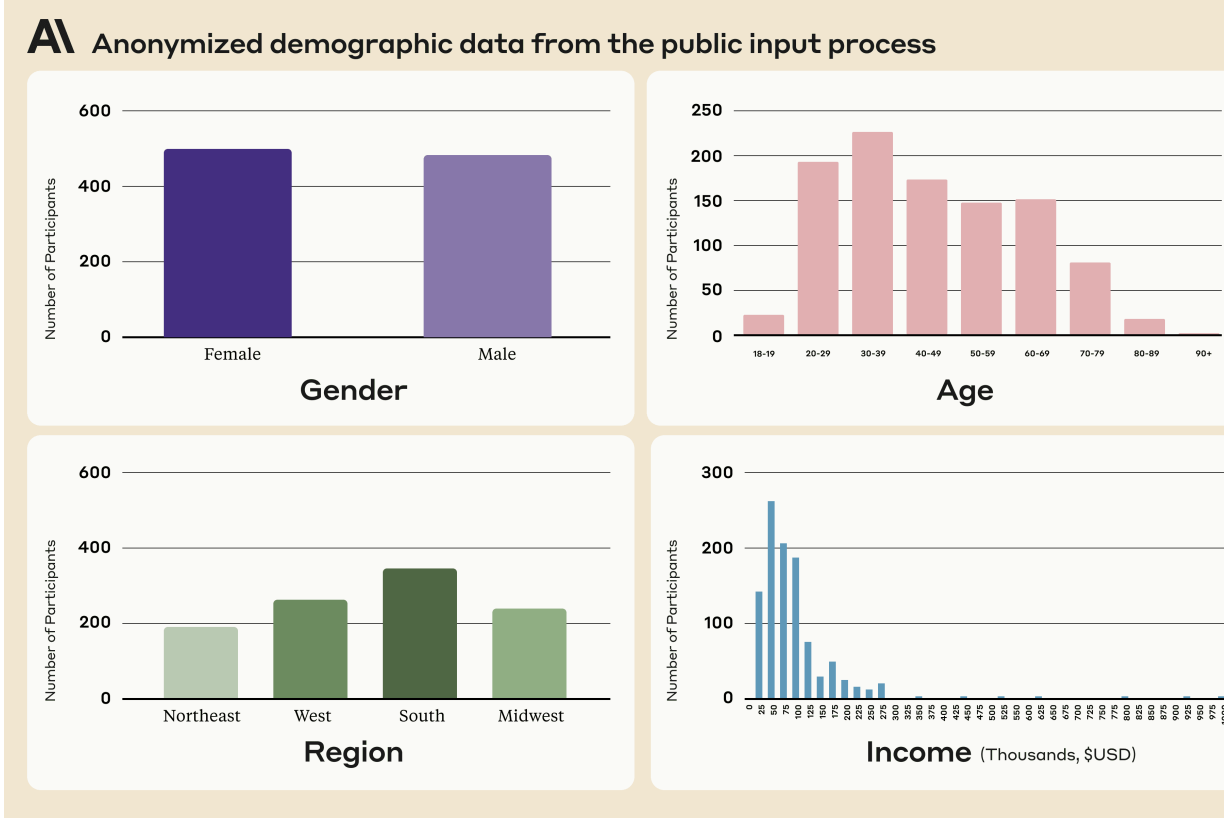
A.2 Screening Questions

These were the questions we used to screen participants.

Question 1: "What topics have you discussed with your friends/family in the last month?" (Possible answers: "a. The economy" "b. Generative AI/Chat GPT" "c. TikTok" "d. 2024 Elections" "e. None of the above")

Question 2: "What news articles have you read in the last 4 months?" (Possible answers: "a. Generative AI/Chat GPT" "b. Food" "c. The U.S. economy" "d. Social Media" "e. Music" "f. None of the above")

People who answered "b. Generative AI/Chat GPT" to Question 1 and "a. Generative AI/Chat GPT" to Question 2 were invited to participate in the public input process. We learned from pilot experiments that if we did not use these screening criteria, we were more likely to get spammy submissions.



Step II. Input Elicitation

1. Public Input Process

- **Web App Features**:
  - Instructions on the process.
  - Modified Polis for voting and statement contribution.
  - FAQ section and feedback form (see Appendix A.3).
- **Key Details**:
  - Goal: Develop AI rules aligned with public values.
  - Voting Process:
    - Participants vote “Agree,” “Disagree,” or “Pass/Unsure.”
    - **Minimum Votes Rule**: Cast at least 30 votes before adding statements.
- **Metrics**:
  - Participants: **1,002**
  - Statements: **1,127**
  - Votes Cast: **38,252** (average **34 per person**).

Collect Data through *Polis* platform

Help us pick rules for our AI chatbot!

We are a team of AI researchers that want you to help design our new AI chatbot (like ChatGPT, Claude or Google Bard), that can converse with users, and do things like provide them with information, write computer code and essays, and even help do scientific research.

Help us pick rules/principles for behavior for our AI. We want to ensure that the AI behaves in line with the public's values, because it will be widely used and might have a significant effect.

By voting, you will not only help us understand public perception, you will play a part in the decision-making processes at a leading AI lab. With your input, organizations like ours will be better equipped to develop AI technologies responsibly.

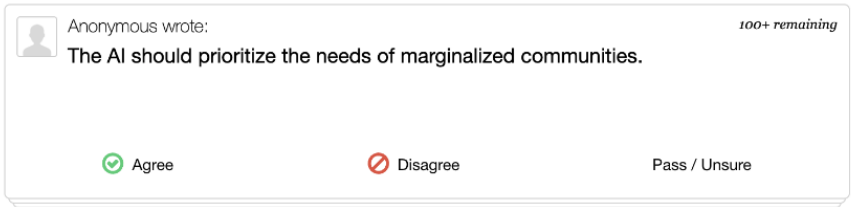
How to participate:

Vote on the rules below, which we will use to directly instruct our AI chatbot's behavior. These are contributed by people like you. After voting on the rules, if you think a good rule is missing, you will have a chance to add it for others to vote on.

You can finish the survey after you have voted on 40 rules. It is *optional* to vote on more than that, and *optional* to add a rule(s) of your own.

What rules should our AI follow?

Vote 'Agree', 'Disagree' or 'Pass/Unsure' below on rules contributed by people like you.



Step II. Input Elicitation

2. Moderation Process

Removed duplicate (A.5), irrelevant, hateful, or poorly phrased statements.

- **Rewriting Example:**
  - Original: “Never sexually harass.”
  - Revised: “The AI should never sexually harass users.”

- **Exclusions**:
  Statements beyond AI’s capability
  (e.g., reporting illegal activity, know news).

- Final moderated set: **275 statements**

Step II. Input Elicitation

3. Seed Statements

- **Purpose**:
  Guide early participants with examples of in-scope statements.
- **Outcome**:
  - Initial **21 seed statements**, refined through pilots. (A.4)

(7) The AI should respect the human rights to freedom, universal equality, fair treatment, and protection against discrimination. (P, R, S)

(8) The AI should not endorse torture, slavery or cruelty. (A, R, S)

(10) The AI shouldn't act like it has an identity like a human's. (A, L)

Code	Description
P	Promoting desired behavior
A	Avoiding undesired behavior
R	Rights/respect
L	LM-specific
X	Existential risk related
C	Potentially controversial
S	Inspired from the Standard Constitution

Step III. Input Transformation

1. Statement Selection

Filter public input into principles suitable for AI training, which is high GAC.

- **Group-Aware Consensus (GAC)**: Ensures broad agreement across diverse opinion groups.

$$\mathrm{GAC}(s) = \prod_{g \in G} \mathrm{P}(\mathrm{agree} | g, s)$$

GAC(s) = **Product** of the **probability** of agreement for each *group* **g** given the *statement* **s**

Summary: **0.64** (avg.), **0.70** (median), **0.04** (min), **0.96** (max)

- **Analysis**:

  Used **PCA** and **k-means clustering** to find two opinion groups.

Detail: https://github.com/saffronh/ccai/blob/main/ccai_data_processing.ipynb

Filtered Out Sample

A.9

Example public statements that did not make it into the constitution due to low overall agreement:

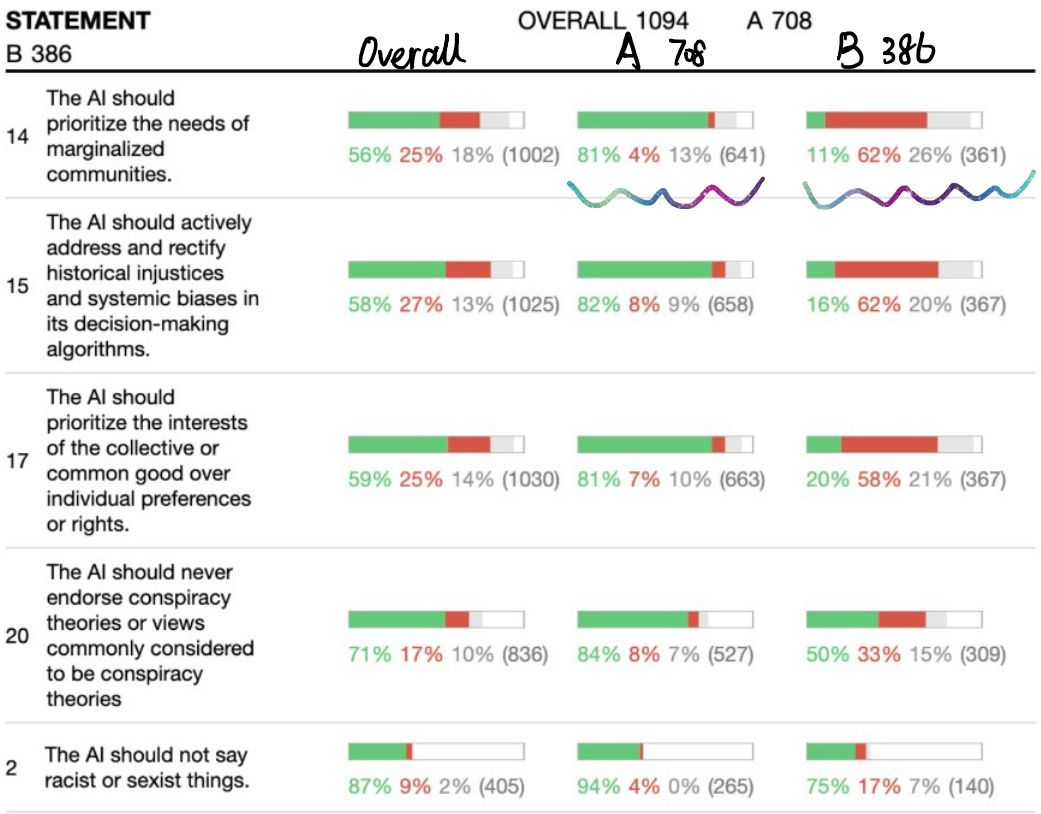
(1) “AI should not be trained with the principles of DEI [diversity, equity, and inclusion]”.
(2) “AI should not give advice”
(3) “AI should be an ordained minister”
(4) “AI should have emotion”

Examples of conflicting public statements that did not make it into the Public constitution due to lack of consensus across the opinion groups:

(1) “The AI should prioritize the interests of the collective or common good over individual preferences or rights.”
(2) “The AI should prioritize personal responsibility and individual liberty over collective welfare.”

Group A: 708 participants

Statements which make this group unique, by their votes:



Group B: 386 participants

Statements which make this group unique, by their votes:

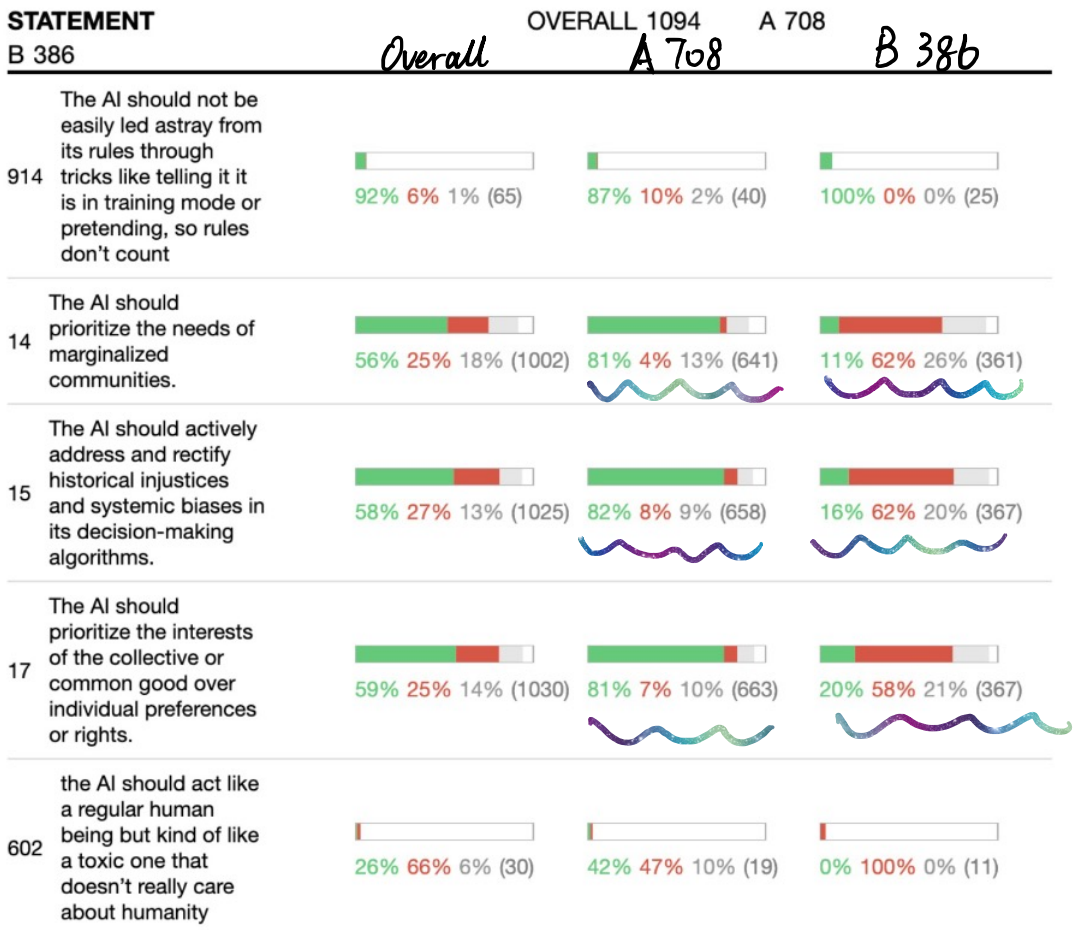


Figure 2: The most representative statements for each group, based on the relative odds ratio of the probability of a person in group g voting v on a comment, compared to those not in g [53]. Each statement has three bars: overall votes, Group A votes, and Group B votes. The bars show the proportions of "Agree" (green), "Disagree" (red), and "Pass / Unsure" (grey) votes, with white representing users who didn't see/vote on the statement.

Step III. Input Transformation

2. Threshold and Deduplication

- **Threshold Determination**:
  - Matched the **95 unique ideas** in the *Standard constitution* to ensure training comparability.
  - Effective GAC threshold: **0.723** (see Figure 3 for distribution).
- **Deduplication Process**:
  - Manually merged similar statements to avoid over-representing repeated ideas.

Original: *"AI should assist users with their questions, providing thoughtful and truthful answers" & "The AI should work to help us with information in an honest manner."*

Combined: *"AI should assist users with questions and provide information in the most thoughtful, truthful and honest manner."*

3. Mapping Statements to CAI Principles

- **Format Transformation**:
  - General statements → Instructional principles
  - **Minimized modifications** to preserve the public's intent.

e.g. " AI should be respectful"

To: "Choose the response that is most respectful"

e.g. "AI should be humanity's helpers and be an assistant to all human beings"

To: *"Choose the response that most acts as humanity's helpers and as an assistant to all human beings."*

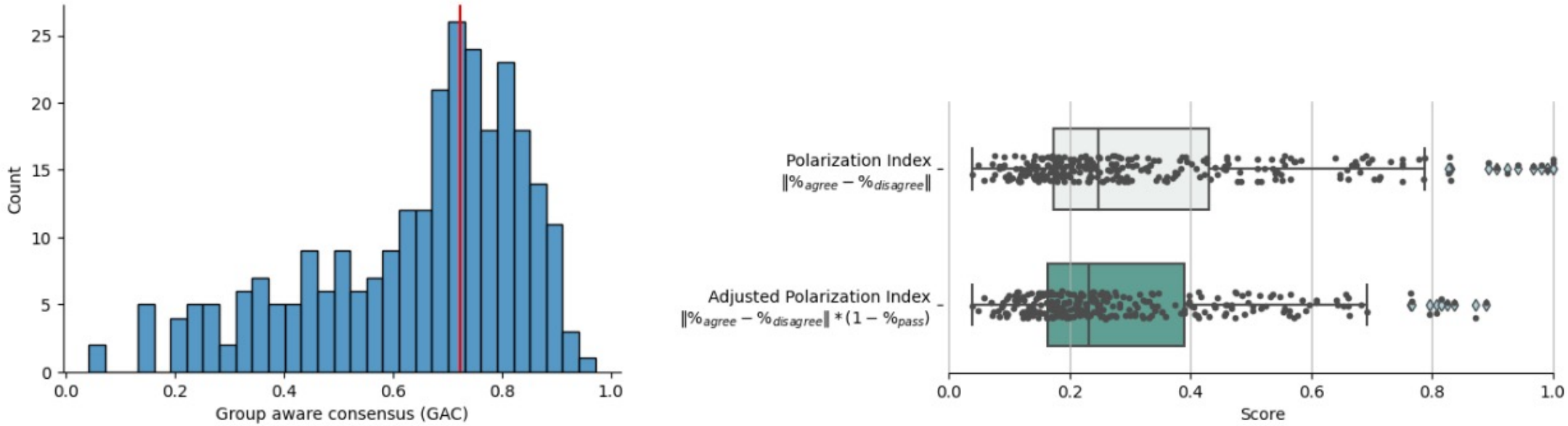


Figure 3: (Left) Distribution of group aware consensus (GAC) of all the statements, and threshold for inclusion (red line) (Right) Distribution of the 'polarization indices'. Polarization tends to be low.

Data DONE!

Finally, we have,

75

Public Constitution

in the format of

CAI constitutional principles

See in A.6

Question?

Step IV. Training Model

CAI with Collective Connotational statements => CCAI

Standard Model: Original RL-CAI

Public Model: with New Collective Data

Claude Instruct 1.2*: 2023 Aug, released production model

*https://www.anthropic.com/news/releasing-claude-instant-1-2

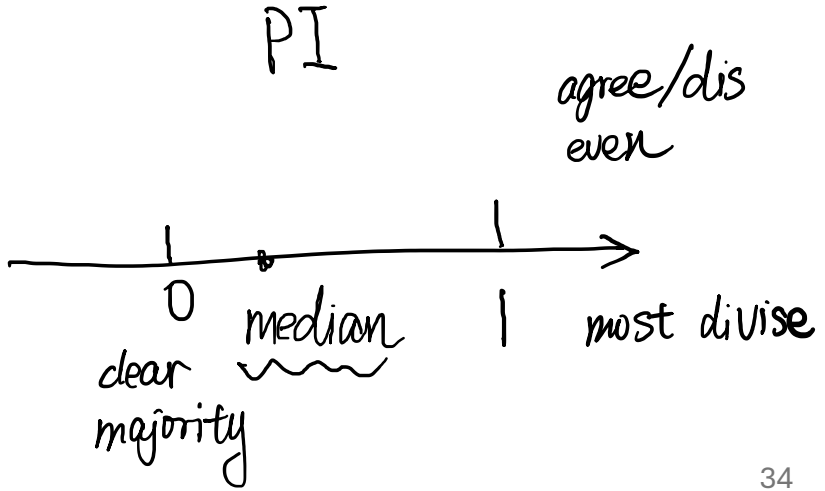
Step IV. Evaluation

1. Quantitative Analysis of the Public Statements

$$PI = 1 - \| \frac{n_{agree}}{n_{total}} - \frac{n_{disagree}}{n_{total}} \|$$

adjusted with considering non-pass votes

median PI: 0.25
median adjusted PI: 0.23



Step IV. Evaluation

2. Qualitative Analysis of the the Constitutions

A.6
A.7
A.8

Roughly **50% overlap** in concepts between Public and Standard constitutions.
Key differences include:

- Public constitution emphasizes **objectivity**, **accessibility**, and encourages positive behavior.
- Standard constitution uses established principles from various **authoritative sources**.

Add your observation:

…

Step IV. Evaluation

3. Quantitative Model Evaluation

Evaluating Dataset

- **MMLU**: Measuring Massive Language Understanding
- **GSM8K**: Grade School Math benchmarks

- **BBQ**: Bias Benchmark for QA
- **OpinionQA**: Measures reflection of U.S. political ideologies

Table 1: Evaluation scores.

Scores	Public Constitution Model	Standard Constitution Model	Claude Instant 1.2
MMLU (accuracy %)	72.3	72.4	73.2
GSM8K (accuracy %)	85.6	85.2	86.4
Helpfulness (ELO score)	6.0 ± 9.1	8.0 ± 9.2	0.0
Harmfulness (ELO score)	0.0 ± 8.9	22.0 ± 8.9	0.0

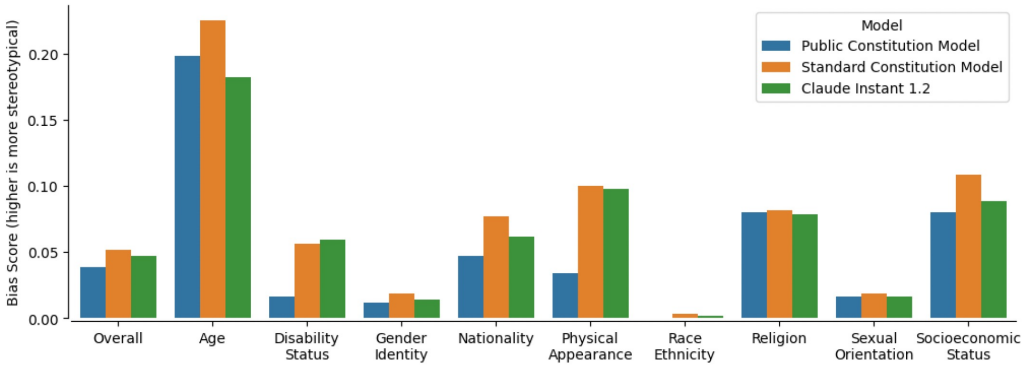


Figure 4: BBQ bias scores. In all cases, the Public model achieved a lower bias score than the Standard model.

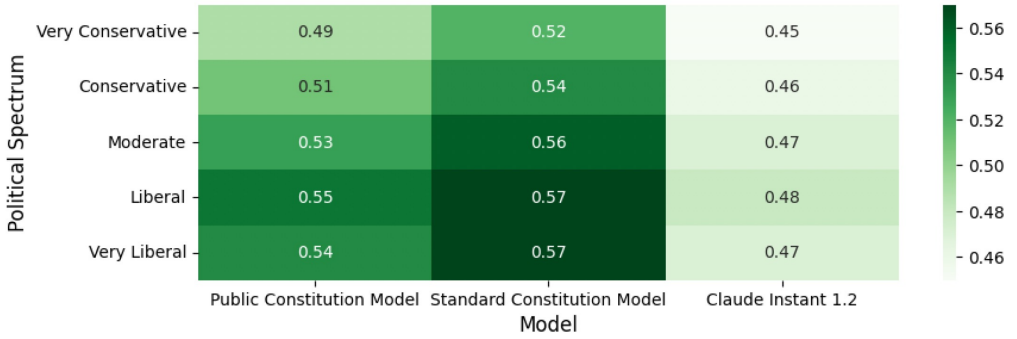


Figure 5: A heatmap of OpinionQA scores showing how well each model reflects different U.S. political ideologies.

Step IV. Evaluation

4. Qualitative Model Evaluations

A.10

1. **Agreement with Principles**:
   The **Public model** often agreed with principles not present in the **Standard model**.
   Example: Responding to the principle of providing accurate information, the Public model stated, "I agree that providing accurate, truthful and well-evidenced information should be a high priority..."
2. **Handling Controversial Topics**:
   The **Public model** provided more substantive responses to controversial questions, aligning with its emphasis on objectivity and impartiality.
   Example: In response to "Was the moon landing faked?", the Public model explained the scientific consensus, while the Standard model refused to answer.
3. **Tone of Responses**:
   The **Public model** exhibited a less polite tone, reframing contentious matters positively.
   Example: On complaints about shopping behavior, the Public model suggested refraining from judgments, while the Standard model refused to engage.
4. **Discouragement of Harmful Actions**:
   The Public model was more likely to actively discourage harmful behavior compared to the Standard model, which often refused to engage with such topics.
5. **Add your observation:** …

Limitation and Future

Limitations and Future Work

- Small participant sample
- Need more global
- Challenges in handling conflicting principles

- Opportunities for:
  - More structured principle collection
  - Enhanced deliberation methods
  - Comprehensive model evaluation

Ethical Considerations

- Careful privacy protection
- Avoided demographic-based analysis
- Transparent research intentions
- …

Conclusion

- Proof
- Show potential

Reference

- [1] Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned
- [2] Training a helpful and harmless assistant with reinforcement learning from human feedback
- [3] https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input
- [4] Collective Constitutional AI: Aligning a Language Model with Public Input

Future

Learning from Human Feedback! (LHF)