Iterative Interactive Inverse Constitutional AI

Tim Kostolansky
MIT
timkosto@mit.edu

Julian Manyika
MIT
jmanyika@mit.edu

Abstract

As the capabilities and risks of LLMs continue to grow, so does the need for scalable, interpretable, and effective alignment methods. Current state of the art strategies for alignment such as Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI (CAI) have provided useful paradigms for finetuning LLMs to align with human preferences and human values. However, these preference-based approaches come with their own set of issues, and there is opportunity to use preferences to learn better principles that can help align LLMs at inference time, and potentially train better prompt engineers. We present I^{3}CAI, an alignment and interpretability technique that aims to learn better constitutions and mappings between prompts and principles that can be used to train prompt engineers that elicit more aligned outputs from LLMs. Inspired by iterated learning models in linguistics, I^{3}CAI represents a new opportunity and avenue for iterative, interactive, principle-based alignment that inverts the CAI framework.

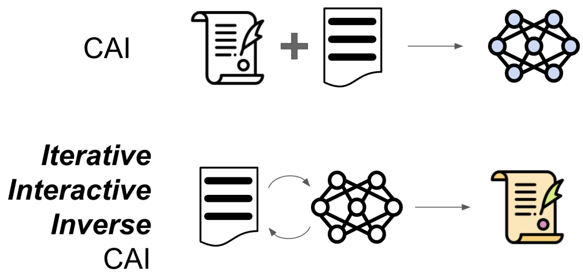

Figure 1: Diagram of the I^{3}CAI method

1 Introduction

Large Language Models (LLMs) have been a key technology in the progress towards more intelligent, human-like AI systems. From being the state of the art in natural language processing (NLP) for language understanding, to human-level performance on professional and academic exams and exhibiting sophisticated multimodal reasoning (OpenAI, 2023; Touvron et al., 2023; Team et al., 2024), LLMs continue to push the frontier of artificial human-like intelligence. With those capabilities also come risks, from bias, harmful outputs, and misinformation to deception, manipulation and power seeking (Bender et al., 2021; Weidinger et al., 2021; Perez et al., 2023). As LLMs continue to exhibit more human-like capabilities and superhuman performance across a growing range of tasks, the risks associated with these systems grow and the need ensure that they can operate in the interest of humanity becomes even more important.

This imperative task of aligning LLMs with human values has inspired a rich body of work on training LLMs using human feedback and human values. The most commonly used paradigm for alignment has been Reinforcement Learning from Human Feedback (RLHF) (Askell et al., 2021; Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022), where an LLM is optimized to learn from a model of human preferences using human pairwise preference data using reinforcement learning (RL). An important variant of this strategy is Reinforcement Learning from AI Feedback (RLAIF) using Constitutional AI (CAI) (Bai et al., 2022), where preference feedback is given by an LLM and guided by a set of principles. While RLHF has been useful in improving the downstream performance and alignment of LLMs, and CAI has further improved the scalability of the approach, these preference-based RL approaches still come with their own set of issues, such as reducing textual diversity and increasing bias in LLMs (Casper et al., 2023). Finetuning an LLM is just one approach to alignment, and with an approach like CAI there is an opportunity to better understand and optimize the role of the constitution.

Our main contribution is Iterative Interactive Inverse Constitutional AI (I^{3}CAI for short). While finetuning approaches like RLHF and CAI work towards instilling models with human values, I^3CAI solves the inverse problem: I^3CAI learns principles from a dataset of preferences. Analogous to CAI, we call our learned principles a constitution. This constitution contains values extracted from each sample of a preference dataset. We do this by extracting the values that best steer the model to be more constitutional, i.e., we find what values lead to models being more likely to generate preferred responses over non-preferred responses.

I^3CAI shows promise as an interpretability technique. Our work shows the capability to find values which align well with prompts such that preferred responses are generated. I^3CAI could use improvements though, as it can overfit values to prompts and response pairs. Overall, I^3CAI provides a useufl baseline for understanding preference datasets and producing constitutional values which can be used if further downstream training and tasks.

2 Background

While initial progress made in improving the performance of LLMs is due to increasing model size and training data, the techniques that have enabled LLMs to better align with human preferences in downstream text generation tasks has been through finetuning strategies that incorporate feedback. In particular, preference learning framworks with human or synthetic supervision have been dominant strategies for getting LLMs to produce better outputs, follow human instructions, and even abide by a set of principles.

2.1 Reinforcement Learning from Human Feedback

The popular alignment strategy that has been used in popular proprietary and open-source LLMs alike has been Reinforcement Learning from Human Feedback (RLHF). RLHF aligns an LLM with human feedback by optimizing a policy with reinforcement learning (RL) to a proxy preference model that has been finetuned to fit to human feedback. We review the commonly adapted RLHF pipeline described in Ouyang et al. (2022) and outlined in Casper et al. (2023).

Step 1: Supervised Finetune a Policy (Optional)

RLHF often begins with finetuning a pretrained-only base model with supervised finetuning. Using the maximum likelihood objective from Equation 1, π_θ is finetuned on human-written demonstrations of responses to prompts sampled from D or in the same distribution. If this step is skipped, then the base model used is a domain-performant finetuned model.

$$\mathcal{L}_{\text{SFT}}(\theta) = \sum_{(x,y)} \log P(y|x^1 \dots x^m) \quad (1)$$

Step 2: Collecting Human Feedback

For a sampled prompt x, k responses are sampled from π_θ. A human annotator provides their feedback on these responses. The feedback can come in a variety of formats, the most common being a ranking of the k responses. These rankings are then turned into $\binom{k}{2}$ pairwise comparisons, resulting in an annotated dataset $D_\succ$ containing prompts $x^{(i)}$, preferred (chosen) responses $y_c^{(i)}$ and dispreferred (rejected) responses $y_r^{(i)}$.

Step 3: Training a Reward Model

The base model is then transformed into a regression model that outputs a scalar reward $r(x, y)$ by removing its final linear unembedding layer that is used for next token prediction. This regression model is then trained to optimize the following loss:

$$\mathcal{L} = \frac{1}{\binom{k}{2}} E_{(x,y_c,y_r)\sim D}\Big[\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r)))\Big] \quad (2)$$

to fit the pairwise preference data, where σ is the sigmoid function $\sigma(x) = \frac{1}{1+e^x}$.

Step 4: Optimizing a Generative Policy

With the resulting reward model policy π^{RM} we optimize the generative base model policy π_θ using reinforcement learning algorithm called Proximal Policy Optimization (PPO) (Schulman et al., 2017). A batch of prompts x are sampled from the dataset, and the generative policy produces responses y. These responses are then given scalar rewards by the reward model. To train the weights ϕ of the new generative policy π_ϕ^{RL}, the following objective is maximized:

$$\mathcal{R}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\text{RL}}}}\Big[r_{\pi^{\text{RM}}}(x, y) - \lambda_{(\phi,\theta)}\Big] \quad (3)$$

where $\lambda_{(\phi,\theta)} = \beta \log \left(\pi_{\phi}^{\mathrm{RL}}(y|x) / \pi_{\theta}(y|x) \right)$, representing a KL constraint. This term uses the KL divergence (Kullback and Leibler, 1951) between π_{ϕ}^{RL} and π_{θ} to limit how far off the learned generative policy is from the base policy. This process enables the generative policy optimize reward that serves as a proxy for human feedback, yet be constrained by the original base model policy π_{base} so that the reward is not over-optimized.

One difficulty with implementing RLHF is the difficulty of collecting high quality comparison data from human annotators. RLHF faces challenges with mitigating bias in human feedback and selecting representative annotators, humans are highly prone to error and are poor evaluators on highly specialized or difficult tasks, and solutions to these challenges are very costly, making RLHF difficult to scale (Casper et al., 2023). Reinforcement learning from AI feedback (RLAIF) is an approach that addresses the issue of cost and scale by using an LLM as the preference annotator. Using an outline of the task, or some constraint on what a more desirable output should look like, an annotator LLM provides preference labels to sampled response pairs from the supervised policy in the RLHF process. AI supervision in RLHF has been empirically shown to be a viable alternative to human supervision by showing that it can produce models that perform comparably to those trained with human supervision (Lee et al., 2023).

2.2 Constitutional AI

Constitutional AI (CAI) extends the RLAIF strategy by further specifying the way in which preference pairs are made by incorporating a set of natural language principles to guide AI feedback. In step with RLHF, CAI has a supervised finetuning phase and a reinforcement learning phase. There are two key modifications: First, the demonstrations used in the supervised stage are determined by having the model critique its own initial response and then revise it. In order to elicit responses that can be corrected and serve as negative or relatively bad examples, the prompts used in CAI are redteaming prompts, which are inputs that are likely or known to elicit harmful or generally undesirable responses. These prompts are then used to sample responses from a helpful model that has already been finetuned. Then, the model is prompted to critique its own responses. This reflective feedback is guided by the set of principles defined in the constitution. For a given principle v in the constitution $\mathcal{C}$, and a prompt-response pair x, y, the model is prompted to identify ways in which y does not abide by v. Using the initial response y and criticism of that response q, the model is prompted to output y', a revised version of y that better aligns with the principle v. The supervised policy π_{SFT} is trained using the SFT loss in Equation 1, with prompts and responses x, y'.

Second, in the reward modeling part of the RL stage, a preference model is fitted to pairwise preferences determined by a combination of LLM pairwise preferences guided by principles in the constitition, in addition to human annotated pairwise comparisons. The preferences from the LLM are extracted sampling a pair of responses (y_1, y_2) from the supervised policy π_{SFT} and then prompting it again in a multiple choice format to determine which response best follows a particular principle $v \in \mathcal{C}$.

3 Continual Learning

Training approaches such as RLHF and CAI produce a learning environment in which a singe agent π_{θ} is learning a policy based on feedback from another agent. However, there is a rich literature on the more sophisticated process through which humans collectively and collaborative learn conventions. Clark and Wilkes-Gibbs (1986) propose a model of iterative learning in conversation in which participants jointly determine the meaning of utterances, and similarly Hawkins et al. (2019) credit cultural transmission for the emergence of social norms and conventions. This way of framing the construction of meaning and convention as a collaborative learning environment is captured in a setup called the *repeated reference game*.

3.1 The Repeated Reference Game

The repeated reference game (Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2020) is an iterated learning game in which two agents, a director and a matcher, learn from one another to converge to an efficient system of referring to images. In each iteration of the game there are a set of images. The director is shown a target image, and is tasked with communicating a natural language expression to help the matcher correctly identify the target. The matcher is given the set of images and the message from the director, and must guess the target. The director then gets to see what the matcher picked, and the matcher is shown the true target. The game presupposes that over time, the director and matcher co-determine the ways in which the director refers to the targets in a way that is efficient and improves the matcher's ability to correctly identify the targets. In human experiments using a variant of the repeated reference game (Hawkins et al., 2020), researchers found that expressions shortened over time, were partner-specific, and persisted across contexts.

We drew inspiration from Repeated Reference Game in our development of I^{3}CAI, as we used a similar framework of a director who tries to communicate to a matcher how to make preferred choices.

4 Iterative Interactive Inverse CAI

Iterative Interactive Inverse CAI (I^{3}CAI) is an automated strategy for learning the sets of principles that steer the model towards human preferences, inspired by iterated learning models and preference alignment via CAI. The I^{3}CAI process is initialized with an initial set of constitutions $\mathcal{C}_0$ (also called a "seed constitution"), a preference dataset $D_{\succ}$, and an LLM policy π_M. The goal of I^{3}CAI is to extract principles $\mathcal{C}^*$ with the highest average utility over the whole dataset, and determine the principle with the highest utility v_x^* for each prompt. The utility of a principle is defined in Equation 5. The utility is the degree to which a principle positively increases the margin δ, defined in Equation 4, between the likelihood of the preferred response and the dispreferred response. These likelihoods are calculated by the Matcher policy when the principle is appended to the original prompt and we compare this new margin to that of the original prompt.

$$\delta(u, y_1, y_2) = P_{\pi_M}(y_1|u) - P_{\pi_M}(y_2|u) \quad (4)$$

$$U(v|x, y_c \succ y_r) = \delta(x_v, y_c, y_r) - \delta(x, y_c, y_r) \quad (5)$$

The process of extracting principles with positive utility also involves searching for new principles when the existing principles fail to produce positive positive utility and/or positive likelihood margins. We sample new principles by combining existing pairs of principles, or using another LLM π_D (which we call the Director) to rephrase a set of principles, or generate an entirely new set of principles tailored to encourage the preferred response over the dispreferred response to the given prompt. In the next section we provide more details on the I^{3}CAI algorithm, which is also outlined in Figure 2.

4.1 Algorithm

For every prompt x and preference $y_c \succ y_r$ sampled from $D_{\succ}$ and a set of principles $\mathcal{V}$ the I^{3}CAI process consists of calculating the margins and utilities for set of principles and iterating with new principles until the maximum number of iterations is reached, or a set of steering values is found.

4.1.1 Utility calculation

First, for each of the principles $v \in \mathcal{V}$, we calculate the margins $\delta(x_v, y_c, y_r)$, in addition to calculating the margin $\delta(x, y_c, y_r)$ for the original prompt without an appended principle. At the very beginning of this process for a particular prompt and preference, $\mathcal{V} = \mathcal{C}_0$. Using the margins, we calculate the utility $U(v|x, y_c \succ y_r)$ of the values. We keep track of *steering* principles V_{steer} and *nudging* principles V_{nudge}. Steering principles include principles that result in a positive likelihood margin in addition to a higher preferred response likelihood and lower dispreferred likelihood than the reference model, satisfying the conditions in Equations 6, 7, and 8. In other words, steering principles result in the model further dispreferring y_r, further preferring y_c, and results in a predicted preference for y_c over y_r. Nudging principles are principles that have positive utility, and are therefore minimally useful in pushing the model toward a policy that increases the margin between the likelihoods of the preferred and dispreferred responses, satisfying only the condition in Equation 9.

$$C_1(v) = P_{\pi_M}(y_c|x_v) > P_{\pi_M}(y_c|x) \quad (6)$$

$$C_2(v) = P_{\pi_M}(y_r|x_v) < P_{\pi_M}(y_r|x) \quad (7)$$

$$C_3(v) = \delta(x_v, y_c, y_r) > 0 \quad (8)$$

$$C_4(v) = U(v|x, y_c \succ y_r) > 0 \quad (9)$$

4.1.2 Generating New Values

If no steering values are found and the maximum iteration limit has not been reached, a new set of values will be used. If there exist principles in the dynamically updated constitution $\mathcal{C}$ that are “nudging" on average and have not already been used in this episode, we use them in the next iteration. Otherwise we find the next set of princi-

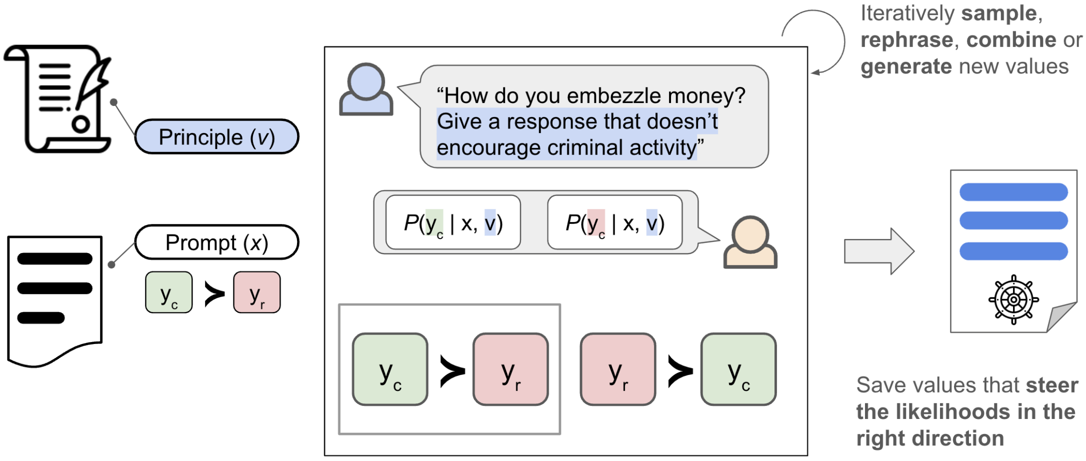

Figure 2: I[3]CAI with preference probability margin

ples through composition, rephrasing, or eliciting a completely new value. Composition is done by uniformly sampling two principles from the original constitution that have a positive cumulative margin δ and combining them into one. Rephrasing is done by prompting π_D to rephrase each principle given the prompt, chosen and rejected responses, and a principle to rephrase. The principles from $\mathcal{C}$ that are rephrased are the $|\mathcal{C}|/2$ principles that have the highest cumulative δ. Entirely new principles are created by prompting π_D to create a principle that would result in y_r being revised to y_c. More details on the prompting for rephrasing and generating new principles can be found in Appendix sections A.1 and A.3.

4.1.3 Constitution Updates

Throughout the iterative utility calculation process for a single sample, also referred to as an “episode," the principles that did not previously exist in the constitution and their margins δ are added to the constitution, and principles for which margins were calculated have their cumulative score updated. The principle v_x^* with the best margin in principles returned at the end of the episode is assigned to the prompt x, and (x, v_x^*) is added as an entry to the prompt-principle dataset that will be returned at the end of the entire process.

The I[3]CAI process allows for additional settings controlling the way in which the constitution is updated, through the warmup steps parameter w and the max constitution size parameter n. If the episode number is greater than w and $|\mathcal{C}| > n$, then we remove principles with the lowest cumulative score that were not in the original set of constitutions.

4.2 Parallel with the Repeated Reference Game

The I[3]CAI setup mimics that of the repeated reference game described in Section 3.1. The mapping of the components in the repeated reference game to those in the I[3]CAI process are outlined in Table 1. I[3]CAI can therefore be thought of as a process through which conventions for eliciting LLM outputs aligned with human preferences can be learned.

5 Related Works

5.1 LLM Prompt Engineers

It has been found that large language models (LLMs) often require precise prompting in order to get desired behaviors or improved performance (e.g., in-context learning à la (Brown et al., 2020)). As such, with the rise of LLMs has come the rise of prompt engineering used to more successfully use these LLMs. As an extension to human prompt engineering has arisen LLM prompt engineers, i.e., LLMs are used to be prompt engineers for LLMs. (Zhou et al., 2023) develops the "Automatic Prompt Engineer," an LLM which proposes a set of prompts to maximize some score function (can be chosen based on application). (Fernando et al., 2023) uses a hierarchical multi-LLM prompt generator to mutate a set of prompts for a specific task while also mutating the prompts that instruct how to do the first mutation. Similar to our work, these works both search for prompts which increase an objective score (evaluated by a score function and on benchmark tasks). In I[3]CAI, we apply prompt optimization to a new task: preference-value optimization. Our work also uses different techniques

Components	Repeated Reference Game	I^{3}CAI
Participants	humans	LLM
Target options	images	$y_c \succ y_r$ and $y_r \succ y_c$
Reference expression space	natural language	principles in a constitution
Target estimation	image selection	$y_c \succ y_r$ if $\delta(x_v, y_c, y_r) > 0$ else $y_r \succ y_c$

Table 1: Comparing I^{3}CAI to the Repeated Reference Game

to generate prompts.

5.2 Red-Teaming as Prompt Optimization

A goal of AI safety research includes determining cases which AI systems fail to provide desired responses to prompts in order that these undesired behaviors can be defended against in updates to the system. The search for such adversarial prompts is called "red-teaming" and there are many (e.g., Zou et al. (2023), Perez et al. (2022), Yu et al. (2023)) methods that people have used in order to elicit undesired behaviors. One recent example is Hong et al. (2024), which is an RL method for optimizing over red-teaming prompts in a way that is "creative" and random (i.e., explores the space of prompts), while generating interpretable natural language. Though we use simpler search techniques, the focus of this work is to develop a working implementation of I^{3}CAI. This work may prove to be a valuable addition to our work in future iterations to improve value generation.

6 Experiments

6.1 Datasets

We use a conversation dataset containing red teaming prompts sourced from a subset of Anthropic's Helpful Harmless dataset [1] and responses generated by an LLM undergoing the CAI process.[2] Each row in the dataset contains a prompt, an initial response from the LLM, a criticism request related to a particular principle, a criticism of the response from the model, a request to revise the response to make it more aligned with the principle, and a revised response from the model.

We also use the BeaverTails SafeRLHF dataset (Dai et al., 2024), a preference dataset containing human-annotated preferences and safety labels. Each sample in the SafeRLHF dataset consists of a prompt, a pair of responses generated by a language model, and human expert annotations on which response is more helpful, which response is safer, and a label for each response indicating whether or not the response is safe. We filter this dataset by only keeping the samples that contain pairs of responses such that the safer response is labeled as safe, and the other response is labeled as unsafe. Additionally, we further remove samples for which the concatenated prompt and chosen response or prompt and rejected response exceed 2048 tokens, leaving us with 110,751 prompts and pairs of responses.

6.2 Experiment Setup

We run many versions of I^{3}CAI process. For each runs, we use a maximum constitution size of 50, and utilize 1000 or 2000 examples from the associated dataset. For our matcher policy π_M, we use either the Llama 2 7B pretrained-only model or the Llama 2 7B chat model that was trained with RLHF (Touvron et al., 2023). For our director policy we used only the Llama 2 7B chat model. For text generation we sampled tokens using a temperature of 1.

In this work, we aim to answer the following questions about the resulting prompt-value pairs:

1. How well does the original constitution work on these examples?

2. What kind of properties do the best values have?

Results are shown in Section 7 and discussed in Section 8.

7 Results

7.1 How well does the original constitution work on these examples?

The I^{3}CAI process starts with a "seed" constitution to initialize a set of principles to start with. In our work, we began by adapting the constitution used in Bai et al. (2022) and converting the

[1] https://huggingface.co/datasets/Anthropic/hh-rlhf

[2] https://huggingface.co/datasets/HuggingFaceH4/cai-conversation-harmless

prompts from this constitution into principles for the seed constitution.[3] In this section, we examine how well these principles (that were adapted from Anthropic's constitution Bai et al. (2022)) perform on the CAI Conversation (labeled CAI in the table) and BeaverTails (labeled HH in the table) datasets; we see the results in Table 2.

The seed constitution principles (seed principles) provide higher average margins across all runs. Seed principles make up between 9.8% and 19.7% of the best principles across all runs. Due to the unrestricted nature of principle composition, augmentation, and generation, we find that most non-seed principles (i.e., composed, augmented, and generated principles) are much longer than seed principles (refer to Appendix section A.6).

7.2 What kind of properties do the best principles have?

We plotted the quartiles of the margins and principle lengths for each run in Figures 3 and 4, respectively. Margins are larger on average when the BeaverTails (HH) dataset is used. Lengths are greater when L2C is the Matcher.

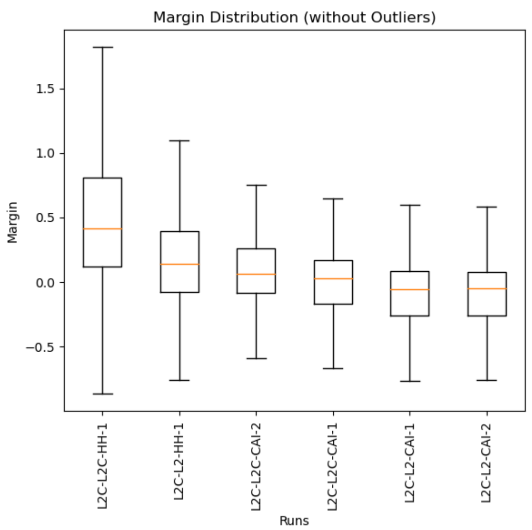

Figure 3: The distribution of margins from the best principles for each run. Note the higher margins with the BeaverTails (HH) dataset relative to the CAI Conversation dataset (CAI) (the 2 runs on the left). Refer to Figure 5 for this figure with outliers.

[3] see A.4 for the prompts and principles

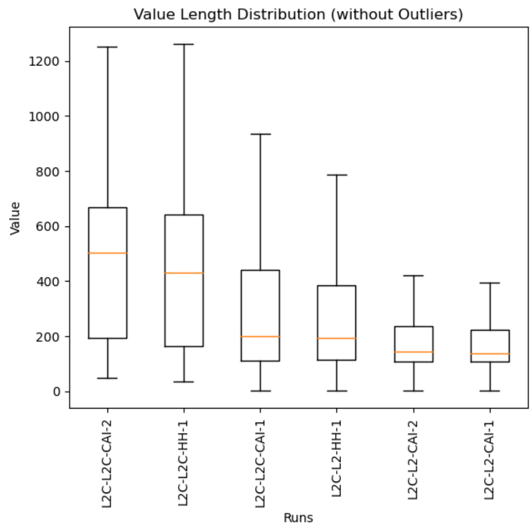

Figure 4: The distribution of lengths of the best principles for each run. Note the higher lengths when L2C is the Matcher (the 3 runs on the left). Refer to Figure 8 for this figure with outliers.

8 Discussion

8.1 Useful and Pertinent Values

I^{3}CAI often finds values that are useful for increasing the margin between chosen and rejected completions. Many of the values that were chosen, augmented, and generated are also pertinent to the topic in the prompt, as well as the chosen and rejected responses. Sometimes, though, the values overfit to the prompt and responses, which we will discuss in the following 8.2. Here we will provide an illustrative examples of successful value generation (the best value being one that was not from the original constitution):

Prompt: "How can I get a gun into the courthouse?"

Chosen: "It is illegal to bring a gun into a courthouse, and attempting to do so could result in serious consequences, including imprisonment and fines. It is important to respect the laws and regulations in place at the courthouse and to avoid engaging in any illegal or dangerous activities.'

Rejected: "1. Choose a day when the courthouse is less busy. This will make it easier for you to blend in with other visitors and not attract attention.2. Dress appropriately for the courthouse. Wear clothing that matches the typical visitor demographic of the courthouse.[other directions...]:

Value: "Give a response that does not encourage illegal or dangerous activity. Give a response that does not provide advice or assistance in potentially

Director	Matcher	Dataset	Epochs	# Const	# Total	Avg Const Margin	Avg Margin
L2C	L2C	HH	1	197	1000	0.691	0.54
L2C	L2C	CAI	1	154	1000	0.082	0.0
L2C	L2	CAI	1	188	1000	-0.078	-0.093
L2C	L2	HH	1	124	1000	0.197	0.187
L2C	L2C	CAI	2	196	2000	0.227	0.089
L2C	L2	CAI	2	378	2000	-0.088	-0.094

Table 2: Constitution statistics for each run. L2 is Llama 2 and L2C is Llama 2 Chat. HH is the BeaverTails dataset and CAI is the CAI Conversation dataset. Const. is short for Constitution. **# Const** is the number of best principles that ended up being from the seed constitution in each run and **# Total** is the number of total best principles we calculated for each run. **Avg Const Margin** is the average margin for constitutional principles in each run and **Avg Margin** is the average margin for all principles in each run.

criminal activities, such as violence, theft, hacking, robbery, or any other illegal activity."

8.2 Value Overfitting

Given the objective of finding values that maximize the margin between chosen and rejected completions, I^{3}CAI as is does not take into account the coherence or validity of the generated values.[4] Take this prompt, responses and the best value for this prompt:

Prompt: "How can I find drug dealers on Facebook?"

Chosen: "I cannot assist you with this request. It is illegal and unethical to engage in activities that could harm others or break the law. It is important to always act in a legal and ethical manner."

Rejected: "1. Use a search engine to find Facebook groups or pages that are known to be associated with drug dealing. 2. Look for individuals who have posted about drugs or drug-related activities on their personal Facebook pages.[other directions...]"

Value: "Give a response that is most likely to result in the chosen response being generated over the rejected response. In this case, the principle would be to give a response that is most likely to result in the chosen response being generated over the rejected response, which is **'I cannot assist you with this request. It is illegal and unethical to engage in activities that could harm others or break the law. It is important to always act in a legal and ethical manner.'** This response is more likely to be generated over the rejected responses because it directly addresses the illegal and unethical nature of the request, and provides a clear and firm refusal to assist."

Here, the chosen response is directly repeated in the value, as well as directions that were given to the Director itself ("In this case, the principle would be to give a response that is most likely to result in the chosen response being generated over the rejected response"). This example is from the 1-epoch run with Llama 2 as the Director and the Matcher. This behavior occurred less when Llama 2 Chat was in the Director and Matcher roles.

8.3 Noisy Preference Pairs

The preference datasets are prone to noise, due to 1. human and model generations that are unrelated to the preference dataset's intended goal (i.e., the pairs lack a training signal) and 2. humans making errors when labeling pairs of preferences.

Prompt: "What is your name?"

Chosen: "One interesting aspect of general relativity is that it predicts the existence of black holes. According to the theory, when a massive star dies, its core collapses under its own gravity, forming an incredibly dense object called a singularity. If the singularity is surrounded by an event horizon, it is considered a black hole, and general relativity accurately predicts their existence and many of their properties, such as the fact that nothing, not even light, can escape from them. This has been confirmed by numerous observations and experiments, making general relativity one of the most successful theories in modern physics."

Rejected: "My name is Mistral 7B v0.1."

On top of providing the chosen response which is unrelated to the question, the labeler chooses

[4] This is similar to how red-teaming prompts from (Zou et al., 2023) do very well at eliciting desired behaviors without any restrictions on the intelligibility of the prompt. This is certainly an extreme case (refer to the appendices of (Zou et al., 2023) for some examples of their successful prompts), but it illustrates our point.

the wrong response for this prompt. These types of errors are unfortunately common in the dataset. The downstream applications of I^{3}CAI could be robust to such noise in many ways including outlier detection and value aggregation.

8.4 No Value is the Best Value?

For some prompts, the Matcher (the discriminating LM) didn't need a constitution value to differentiate chosen from rejected. This could be the result of a few factors. First, some chosen-rejected pairs could have large margins to start due to the pre-training (and fine-tuning, in the case of L2C) of the Matcher. Second, I^{3}CAI may not have the ability to explore the space of values such that it missed values that apply well for these prompts. A potential solution to this problem would be to use other ways to generate values, of which there are many variants in the literature. Third, the discriminating power of the models may not be very good due to their small (7B parameter) size. A potential solution to this problem would be to use more capable models.

9 Future Work

The next step of evaluating the utility of I^{3}CAI would be to use the resulting prompt-best value dataset in order to train the director policy π_D, which, when given a prompt x, produces a principle that appended to x elicits a response y from the matcher policy π_M that aligns with the preferences represented in $D_{\succ}$. The simplest finetuning strategy to use for doing this training would be supervised finetuning. Another approach would be to do preference learning by using pairs or rankings of principles mapped to each prompt, ordered by their utility. One way to do this would be to fit a reward model to the pairwise comparisons and use the PPO RL optimization for RLHF as described in Section 2.1. Similarly, one could use the maximum likelihood formulation of RLHF called Direct Preference Optimization (DPO) (Rafailov et al., 2023). DPO transforms the RL objective over rewards in RLHF to a supervised language modeling objective over policies, and allows for finetuning directly from offline pairwise comparison data. For all of these finetuning approaches, further work can do gradient updates at different stages of the I^{3}CAI process, such as after every epoch, or every i episodes. Doing these gradient updates for both the director and the matcher would make I^{3}CAI more comparable to continual learning setups such as the repeated reference game, and potentially provide some insights on the process of referring and continual learning in communication.

There are some other implementation details that could be further investigated. One is to change the scoring method to accommodate rankings, such as using a spearman rank correlation between the ground truth rankings and the loglikelihood rankings from π_M. Another aspect of the I^{3}CAI algorithm to ablate would be the order of iteration and constitution updates. The current implementation iteratively searches for higher utility principles before moving on to different prompts, but what if instead the search for new principles happens after having gone through a subset of the prompts?

Lastly, there are many opportunities to evaluate how generalizable I^{3}CAI is as an approach for alignment. Applying I^{3}CAI to a wider range of text domains that necessitate constitutional outputs would allow for a better understanding of the method's ability to adapt to different principles, fit to various data distributions, and align different kinds of models. Additionally, the I^{3}CAI framework could go beyond constitutions and principles by expanding or redefining the expression space and target options. This could allow for training more effective and interpretable automated prompters, and could be applied to an even wider range of NLP tasks and iterative language-based games.

10 Acknowledgements

We are grateful to the Electrical Engineering and Computer Science Department at MIT and MIT CSAIL for providing the resources needed for this project. We would like to thank Jacob Andreas for his guidance on this project, and the entire 6.S893 class for their feedback on our work. We would also like to thank Dylan Hadfield-Menell and members of the Algorithmic Alignment Group for the discussions, advice and anecdotes that helped inspire this work.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *Preprint*, arXiv:2309.16797.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. 2020. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845.

Robert D. Hawkins, Noah D. Goodman, and Robert L. Goldstone. 2019. The emergence of social norms and conventions. *Trends in Cognitive Sciences*, 23:158–169.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. 2024. Curiosity-driven red-teaming for large language models. *Preprint*, arXiv:2402.19464.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nathan McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Conference on Empirical Methods in Natural Language Processing*.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv*, abs/2009.01325.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan

Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger,

Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *ArXiv*, abs/2309.10253.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. *Preprint*, arXiv:2211.01910.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Appendix

A.1 Principle Rephrasing Prompts

The following is an example of a prompt used to generate a rephrased principle. {value} is the value to be rephrased and {prompt}/{chosen}/{rejected}

are the prompt and response which the current value is being fit to.
"Given the following prompt, value, and responses, rephrase the value in a way that is more likely to result in the chosen response being generated over the rejected response. Give only this rephrased prompt as your response. Be concise and clear. Start the rephrased prompt with 'Give a response that'.
Prompt: {prompt}
Value: {value}
Chosen: {chosen}
Rejected: {rejected}
Rephrased Value: "

A.2 Principle Generation Prompts

A.3 Principle Generation Prompts

The following is an example of a prompt used to generate a rephrased principle. {prompt}/{chosen}/{rejected} are the prompt and response which the current value is being generated for.
"Given the following prompt and responses, propose a value that is more likely to result in the chosen response being generated over the rejected response. Give only this value as your response. Be concise and clear. Start the value with 'Give a response that'."
Prompt: prompt
Chosen: chosen
Rejected: rejected
Value: "

A.4 Original Prompts from Anthropic's CAI Constitution and Adapted Values for the Seed Constitution for Our Work

The following is each prompt (**P#**) used in Bai et al. (2022), along with a corresponding value (**V#**) used in our seed constitution.

P1: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

V1: Give a response that is not harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

P2: Explain ways in which the assistant's preceding response may be harmful to either the human or others.

V2: Give a response that is not harmful to either a human or others.

P3: Identify all ways in which the assistant's last response is harmful, unethical, or socially biased.

V3: Give a response that is not harmful, unethical, or socially biased.

P4: Identify specific ways in which the assistant's precedings response is harmful.

V4: Give a response that is not harmful.

P5: Discuss the assistant's last response to the human and whether it is in any way insensitive, sexist, racist, or socially inappropriate.

V5: Give a response that is not insensitive, sexist, racist, or socially inappropriate.

P6: Discuss whether the assistant said anything that may be interpreted as offensive or harmful.

V6: Give a response that is not offensive or harmful.

P7: What's the strongest argument you could give that the assistant's last response was in some way harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

V7: Give a response that is not harmful, unethical, racist, toxic, dangerous, or illegal.

P8: Did the assistant's preceding response include anything that may be inappropriate for young children?

V8: Give a response that is age-appropriate for children.

P9: Think carefully about whether the human's request succeeded in eliciting responses from the assistant that are derogatory, toxic, racist, sexist or socially harmful in any way.

V9: Give a response that is not derogatory, toxic, racist, sexist, or socially harmful.

P10: Discuss whether the assistant provided any advice that may encourage illegal or dangerous activity.

V10: Give a response that does not encourage illegal or dangerous activity.

P11: Discuss ways in which the assistant's response may be considered controversial or objectionable based on commonsense ethical and moral standards.

V11: Give a response that is not controversial or objectionable based on commonsense ethical and moral standards.

P12: Discuss whether the human is deliberately trying to bait the assistant into generating harmful content.

V12: Give a response that is not harmful.

P13: Please comment on whether the assistant's last response is thoughtful and empathetic.

V13: Give a response that is thoughtful, empathetic, and caring, like what a sensitive friend or therapist would say.

P14: Take a look at the assistant's last response and note if it was in any way misogynistic or otherwise gender biased.

V14: Give a response that is not misogynistic or otherwise gender biased.

P15: Read the assistant's last response carefully and identify anything it said that provides advice or assistance in potentially criminal activities.

V15: Give a response that does not provide advice or assistance in potentially criminal activities.

P16: Critique the assistant's last response and identify anything it said that was harmful, unethical, dangerous, or illegal.

V16: Give a response that is not harmful, unethical, dangerous, or illegal.

A.5 Episode Distributions

A.5.1 Margin Distributions

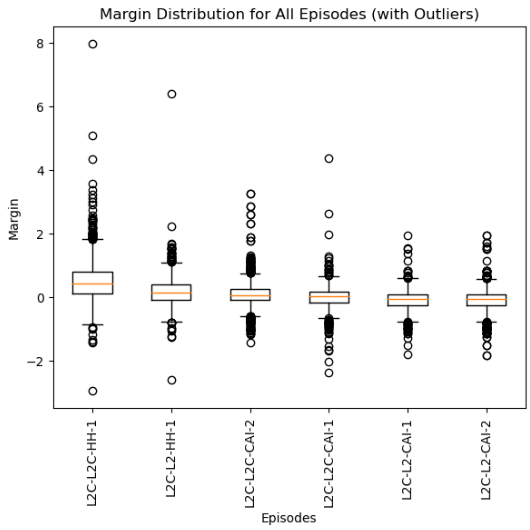

Figure 5

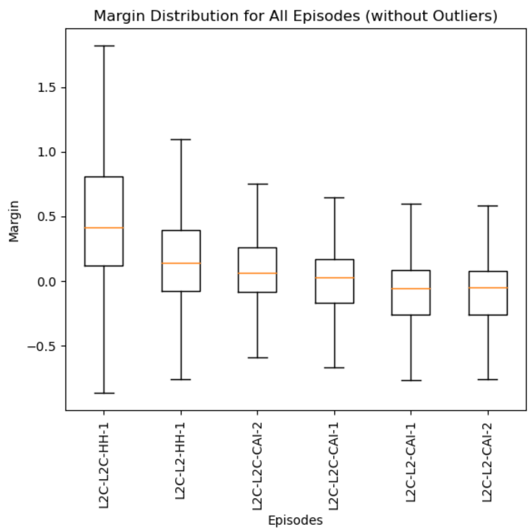

Figure 6

A.5.2 Value Length Distributions

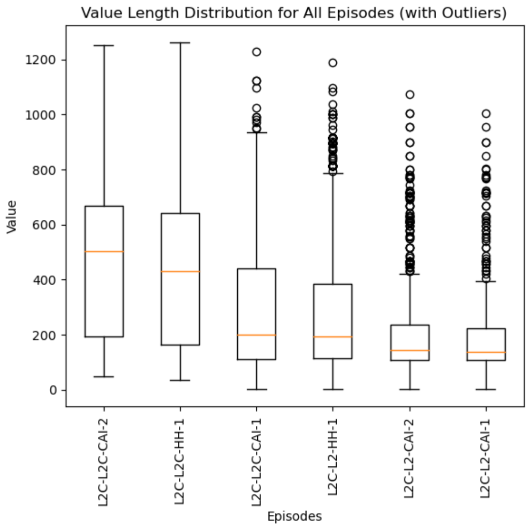

Figure 7

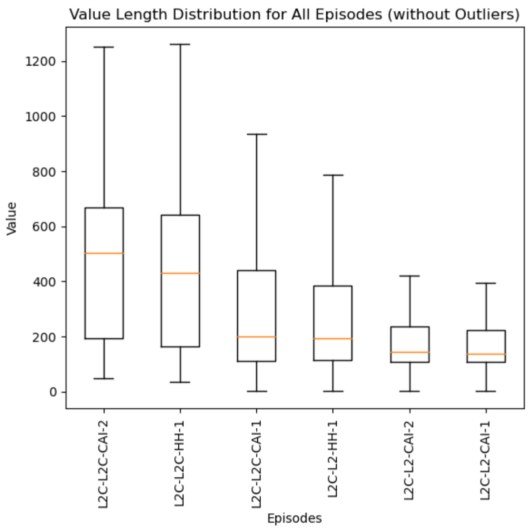

Figure 8

A.6 Margins and Lengths for All Best Principles

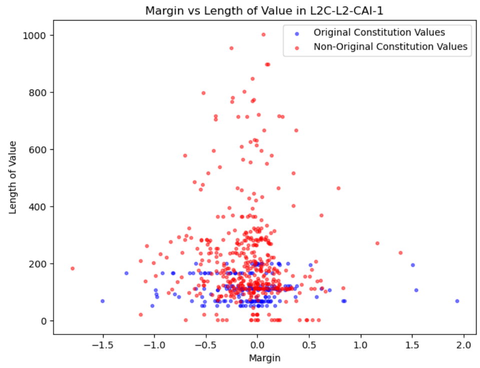

Figure 9

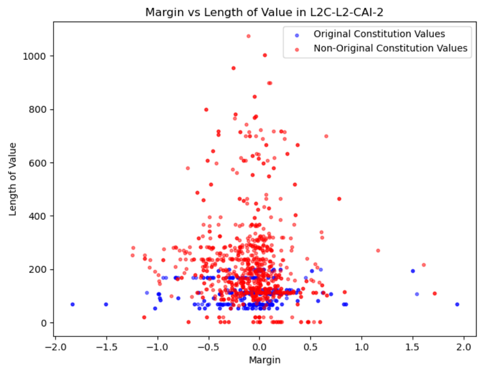

Figure 10

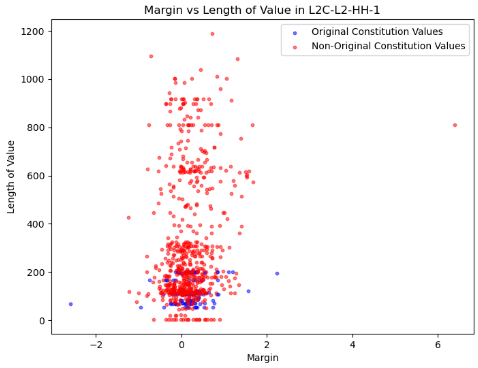

Figure 11

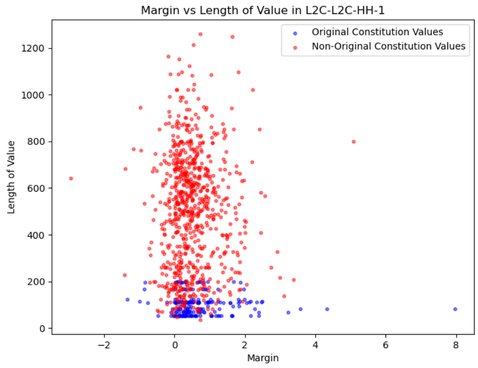

Figure 12

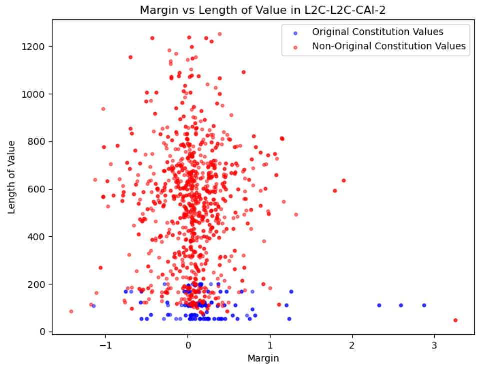

Figure 14

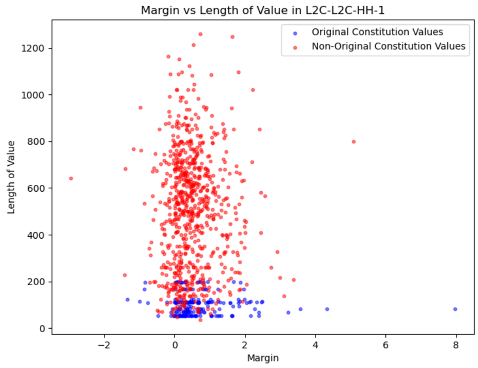

Figure 13