Towards Principled AI Alignment: An Evaluation and Augmentation of Inverse Constitutional AI

Towards Principled AI Alignment

An Evaluation and Augmentation of Inverse Constitutional AI

A THESIS PRESENTED
BY
ESTHER AN
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS

MAY 2025

Thesis advisor: Professor Ariel D. Procaccia Esther An

Towards Principled AI Alignment

Abstract

The accelerated pace of development for advanced AI systems motivates an examination of whether such systems are actually aligned to human designer and user intent. The notion of human intent, however, remains highly ambiguous, with ongoing debate regarding whether large language models (LLMs) should be aligned to demonstrated behavior communicated through the expression of human preferences or abstract normative principles defined by collective deliberation. Methods such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) demonstrate approaches to AI alignment that depend on the communication of human preferences. The formalization of Constitutional AI (CAI), which leverages reinforcement learning from AI feedback (RLAIF) and involves the communication of a set of normative principles for directing the behavior of an LLM, motivates the development of scalable approaches to bridge the gap between standard alignment methods that employ human preference data and interpretable, principled AI alignment. Inverse Constitutional AI (ICAI) is a new framework that aims to learn a constitution from human preference data that may serve as both an interpretable compression of human preference and an instructive set of principles for aligning LLM behavior.

In this thesis, we present an expanded implementation of the ICAI framework that addresses the desideratum of representation for varied human preferences by electing a principle committee using algorithms drawn from the theory of approval voting in social choice. We describe relevant metrics for evaluating the efficacy of a constitution constructed through the implementation of the framework and provide a method of systematic evaluation for all components of the pipeline. We hope that the improved formalization of this approach to principled model alignment will contribute to the development and deployment of more interpretable and better aligned AI systems.

Contents

이하연님과 에이든에게 헌정합니다.

Acknowledgments

There are many people I am immensely grateful for, who have been incredible supporters of both this thesis and the various facets of my journey at Harvard and in computer science. I would first like to express my deepest gratitude for my thesis advisor Professor Ariel Procaccia, whose mentorship and instruction have been a great source of inspiration, and fundamentally shifted how I approach new problems and understand CS to interact with the world. It has been an absolute privilege to be your student and advisee. I am also tremendously grateful for my mentor and collaborator Itai Shapira. Thank you for your consistent encouragement, kindness, and wisdom, and for teaching me to believe in myself and to be brave with my ideas. This thesis would not have been possible without your guidance and support. I would also like to thank Professor Jim Waldo for being my thesis reader and CS advisor, but also, more importantly, my mentor and a spectacular advisor in all aspects of life. I don't know if I can ever fully communicate how much I treasure our conversations and your advice. Thank you for being an amazing role model and, even when gently nudging me to not have so much on my plate, always emphatically affirming that I can do it all if I would like to. I am also truly grateful to Professors Finale Doshi-Velez, Cynthia Dwork, Adam Hesterberg, Madhu Sudan, and Salil Vadhan for opening the world of CS to me in so many interesting directions.

I feel incredibly lucky to have found such a wonderful network of friends, collaborators, mentors, and professors throughout my time as an undergraduate at Harvard. I am particularly grateful for all of my friends, and also for my co-teaching fellows for the courses I have been so fortunate to get to support. To the members of the pfchat, thank you for your encouragement and excitement, and for never failing to make me laugh.

I am especially thankful for my family's ever-present love and support. I would like to express my special gratitude for my mom, who has been both a grounding and an inspiring force for me throughout the writing of this thesis and life more broadly.

Therefore, the programming of such a learning machine would have to be based on some sort of war game, just as commanders and staff officials now learn an important part of the art of strategy in a similar manner. Here, however, if the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which would win a nominal victory on points at the cost of every interest we have at heart, even that of national survival.

Norbert Wiener, "Some Moral and Technical Consequences of Automation" (1960) [44]

1 Introduction

In his 1960 essay "Some Moral and Technical Consequences of Automation," American computer scientist, mathematician, philosopher, and founder of cybernetics Norbert Wiener initiates a prescient discussion on the implications of a literal-minded yet advanced learning tool that may one day surpass the human brain in both efficiency and psychological ability [44]. Citing fables such as

the ballad "The Sorcerer's Apprentice," Wiener warns that the incompleteness of communication between two foreign agents—that is, the human and the machine—may induce an unsatisfactory and even dangerous result despite the appearance of the agents' cooperation. He counsels, "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it." In other words, as we build more advanced machines that complete complex operations with ever-increasing celerity and relatively minimal oversight, it has only become more critical to be fastidious about what we are training and asking those machines to do.

Wiener's concerns, particularly regarding the accurate and comprehensive transmission of human intent to a highly capable machine, seem to refer almost directly to recent advancements in artificial intelligence and the development of large language models (LLMs). Public access to LLMs has ushered in a new era where anyone can easily interface with powerful machine learning models through the medium of natural language. The full consequences—both the benefits and the detriments—of the democratization of this advanced technological resource are still being played out in real time today.

Notably, while LLMs are used and appreciated widely for their helpful responses in a variety of use cases, there remains a prevalence of instances in which LLMs may reinforce undesirable biases or otherwise return harmful content to a user [34]. Moreover, even when an LLM outputs technically helpful information as a direct answer to a human inquiry, the content of the response may

be morally fraught: how, for instance, should an LLM respond to a request by a human user for explicit instructions on how to construct a Molotov cocktail? A natural ambivalence regarding what characterizes an optimal response that is harmless yet maintains helpfulness by not being entirely evasive in this and similar situations speaks to the uncertainty that underlies presumed principles that LLMs are desired and expected to follow when outputting responses to human users. The combination of this uncertainty and the inherent difficulty human users experience when endeavoring to articulate comprehensive and consistent desiderata for the completion of open-ended tasks further compounds the challenge of ensuring AI alignment with human preferences. Furthermore, even the notion of alignment to human preferences is complicated by the way people from different backgrounds or who simply have divergent perspectives may have preferences that are genuinely contrasting and fundamentally non-complementary, which may be expressed across the samples of seemingly divided or even contradictory aggregated preference data.

The goal of this thesis is first to survey ongoing efforts to align LLMs to human preferences and then to systematically improve and evaluate an expanded pipeline for interpreting what is learned by a state-of-the-art LLM from current standard practice methods of communicating human user preferences, in order to guide principled AI alignment.

1.1 Motivation

Soliciting comprehensive and consistent preferences from human users and learning from those preferences to satisfactorily capture the principles underlying their decision making is a nontrivial

task. There exist gaps in current methods for discerning interpretable principles that users desire AI models to align to, and a need for greater interpretability for how the communication of human preferences regarding LLM responses may precipitate into a set of clear principles that subsequently informs an LLM's behavior.

Reinforcement learning from human feedback (RLHF) is a method where, traditionally, a single reward function is trained with feedback representing human preferences and used to align a language model to those preferences through reinforcement learning (RL) [28]. The communication of human insights is generally given through preference data wherein humans communicate their preference for one alternative over another, or one alternative over a set of alternatives, given that it is often challenging for humans to articulate the nature of their preferences fully in a principled and consistent manner. However, the presence of truly divergent and non-complementary preferences within human populations has demonstrated that traditional single reward-based RLHF methodology is not sufficient to align language models to preferences in a way that expresses their inherent diversity equitably [10]. Hence single reward-based RLHF maintains the risk of perpetuating hegemonic perspectives in language models. Given the recognition that human preferences may be inconsistent and difficult to articulate and that people often exhibit divergent preferences, a question that arises is whether it may be possible to pursue greater interpretability for this process by identifying the principles underlying expressed human preferences in order to better inform reward functions for RLHF and other techniques for AI alignment.

Constitutional AI (CAI) is an approach that extends traditional RLHF and incorporates reinforcement learning from AI feedback (RLAIF) by leveraging both human data for helpfulness and

model-generated data for harmlessness [3]. Ultimately, it aims to scale human supervision by aligning LLM responses predominantly with a set of human-interpretable principles expressed in natural language—the "constitution"—that comprises the rules that should govern model responses. There remains, however, a dearth of literature examining the necessary steps involved in the construction of this constitution, as well as a general disconnect between standard practice methods utilizing existing human preference datasets and more recent approaches to principled AI alignment.

1.2 Problem Statement

Broadly, a problem recognized for RLHF is that while there generally exists diversity within a population of human users in terms of their expressed preferences for LLM responses, when a model is finetuned to learn a reward, that diversity may be lost. Given a desire to maintain diversity in a model's output, it is relevant to pursue how to understand more coherently what is learned by a model through RLHF. A natural question that follows is whether it may be possible to identify and articulate interpretable meanings for learned rewards in AI alignment procedures.

Furthermore, alignment approaches extending RLHF like CAI that aim to help models achieve more helpful and harmless outputs may also be augmented by a more rigorous method for discerning what principles are actually communicated through the expressed preferences of human (or AI) users in preference data like the Anthropic Helpful-Harmless (HH) [3] or Chatbot Arena datasets [14]. Uncovering interpretable principles from structured pairwise preference data expressing users' preferred LLM responses would allow for the refinement and articulation of desiderata that differ-

ent stakeholder groups may have for LLMs in various use cases. Hence greater interpretability for approaches aiming to align LLM outputs to human user preferences may have implications not only for general models but also for personalized LLM settings.

1.3 Approach and Contributions

This thesis firstly contributes a comprehensive survey of recent approaches to principled AI alignment, examining the limitations of existing techniques, particularly with regards to pluralistic alignment in the context of diverse human preferences. The primary work of the thesis comprises a systematic improvement of a pipeline for Inverse Constitutional AI (ICAI) that aims to uncover concrete and meaningful human-interpretable principles from human preference data for LLM outputs and examine whether the derived interpretable meanings can actually govern model behavior. The pipeline notably draws from committee voting algorithms supported by social choice theory in its construction of interpretable constitutions for AI alignment. We additionally contribute an evaluation of the pipeline that facilitates direct comparison with adjacent approaches to principled AI alignment, and provide extensive empirical analysis for the results of the provided ICAI implementation.

The remainder of the thesis is organized as follows. Chapter 2 provides relevant background for the thesis by introducing general methods for aligning LLMs with human preferences and current approaches to principled AI such as CAI and other rule-based alignment procedures, as well as a review of existing literature in the field of AI alignment. Chapter 3 describes the improved ICAI

algorithm. Chapter 4 formalizes the theoretical framework for the described AI alignment approach and methods for evaluation. Chapter 5 details a systematic evaluation of the pipeline. Chapter 6 concludes the thesis and discusses directions for future work.

2

Background and Literature Review

This chapter provides a survey of approaches to aligning large language models with human preferences and an examination of existing literature on techniques for AI alignment and their limitations. We begin by formalizing the problem of AI alignment. We then discuss current methods for AI alignment that aim to incorporate the communication of human preference, and outline the spe-

cific subfield of principled AI alignment. We also review related work on the codification of human values and more recent advancements in the realm of constitutional and inverse constitutional AI.

2.1 Aligning Language Models with Human Preferences

2.1.1 What is AI Alignment?

Before delving into existing approaches, it is pertinent to specify the concept of and expectations for AI alignment. AI alignment concerns the problem of ensuring that artificial intelligence systems behave in accordance with human intentions, preferences, and values. An aligned AI system should execute communicated objectives in a manner truly intended by its designers or users. This requirement has become increasingly critical given continued advancements in AI capabilities.

Christiano et al. [16] formally defines *intent alignment* as the condition where an AI's objective (or policy) is configured such that the AI pursues the outcome desired by its operator [16]. In their survey, Ji et al. [27] characterize AI alignment as the task of making AI systems act in accordance with human intentions and values, noting that misalignment risks grow proportionally with augmented AI capabilities [27]. Misalignment manifests when an AI system exhibits behaviors that diverge from human expectations and potentially lead to undesired or harmful outcomes. This risk may become particularly significant in cases where advanced AI systems pursue objectives that conflict with human welfare. Consequently, solving intent alignment—ensuring that the goals and behaviors of AI systems remain aligned with human intentions—represents a fundamental prerequisite for the safe and effectual deployment of AI systems.

Modern LLMs trained solely with self-supervised learning (e.g., next-token prediction on text from the internet) do not inherently align with user intent. During pre-training, models acquire text generation capabilities without commensurate mechanisms for ensuring factual accuracy, instruction adherence, or the maintenance of ethical constraints. As a result, base LLMs frequently produce outputs characterized by factual inaccuracies, toxic content, or responses that fail to address user needs satisfactorily. [33] demonstrated that simply scaling model parameters does not resolve alignment issues, showing that even the most parameter-dense models may generate false or harmful responses or fail to follow instructions accurately.

The core issue of misalignment stems from the fact that pre-trained LLMs learn to predict natural language text without developing corresponding capabilities to align with human *preferences* or ethical standards. As noted in existing work on alignment, models may generate undesirable responses across various input contexts, even without explicit prompting for such outputs [34]. This misalignment reflects the absence of optimized methods designed specifically to ensure that outputs satisfy the complex criteria that diverse human users may consider to characterize helpful and harmless AI behavior.

These alignment deficiencies in base models necessitate the development and implementation of specialized alignment techniques as post-training interventions. Such methods aim to refine model behavior to more accurately reflect human expectations and values. As AI system capabilities continue to advance, ensuring robust alignment becomes increasingly essential not only for maintaining the utility of such systems but also as a fundamental requirement for the responsible deployment of AI for society.

2.1.2 Addressing Ambiguity in Alignment

The notion of aligning AI behavior to human intent is imbued with inherent ambiguity. At a high level, it is not abundantly clear whether AI systems should, in all cases, be aligned to stated human instructions, presumed underlying intent, expressed preferences, or abstract values. In instances where a user prompt may encourage a biased or otherwise harmful LLM response, we may desire the LLM's behavior to depart from what is communicated in the prompt and instead be aligned with broad normative expectations for AI system behavior [21].

Moreover, while it is desirable for AI systems like LLMs to exhibit helpful and harmless behavior in a broad sense, it is unclear how different individuals or groups of individuals may want or expect an LLM to behave in situations where even seemingly intuitive directive principles naturally conflict. For instance, requiring perfectly harmless outputs may reward the trivial outcome of responses that have minimal utility, while improving helpfulness may result in more harmful responses given the nature of what is discussed by the prompt. Different people may identify a different ideal point at which this tradeoff should be made. Hence, in addition to the general ambiguity that characterizes precisely what an AI system should align to, there also exists a lack of clarity surrounding who AI systems should align to. The underlying question at the core of this concern regards how to achieve pluralistic alignment, given that different people can have divergent values or prioritize shared values differently. Managing tradeoffs between the interests of different groups of people can have significant implications for alignment.

Ultimately, the complex task of aligning AI systems to human intent or values can be broken

down into two components [21]:

1. the *normative* step, which involves deciding what values should be encoded, from whom those values should be learned, and to what extent they should inform or constrain LLM output; and

2. the *technical* step, which involves deciding what characterizes an optimal articulation of the values determined in the previous step, the practical challenge of encoding those values or principles, and appropriately training an LLM on those encoded values.

Both components, which collectively aim to prevent the undesirable outcome of "reward hacking" where an advanced model may optimize towards outputs that are technically aligned with what human users expressed that they value but are fundamentally objectionable [44], are at risk of various pitfalls. First, in the normative step, while optimization towards what is actually desired by humans may seem to map the reward model of reinforcement learning neatly to the utilitarian philosophy of maximizing the summative good, a preferable outcome may instead involve balancing the principles that comprise a set of diverse user desiderata, in order to allow for the representation of principles supported by minority populations that may otherwise disappear given a sufficiently large majority in support of conflicting or orthogonal principles [21]. Furthermore, an open question that remains is how best to determine normative values given that they may diverge from what can be learned by observing people's behavior. The approach of aligning AI systems with revealed preferences—that is, preferences that are identified through user behavior—may lead to the encoding of values that seem antithetical to normative principles that the same population of users may claim to

subscribe to. Even if the approach taken instead involves a more abstract and conceptual determination of normative principles, there persists the question of who should define those principles, and the technical difficulty of ensuring that the principles actually inform AI behavior in a coherent and effective way.

2.1.3 Reinforcement Learning from Human Feedback

Reinforcement learning from human feedback (RLHF) is an alternative approach to traditional reinforcement learning (RL) that incorporates feedback from a human in the loop to inform an agent's reward model (RM) [28, 16]. RLHF has emerged as a powerful and cost-effective approach to alignment when implemented for LLM fine-tuning, with demonstrated improvements compared to the method of increasing the size of the model [34]. Given a pretrained language model, a distribution of relevant user prompts, and a group of humans collectively serving as the human in the loop, RLHF for LLM fine-tuning can be implemented by first training a reward model on a dataset of samples where human users have expressed preferences between pairs of LLM responses (hereafter referred to as a preference dataset), and then fine-tuning the language model with reinforcement learning using an optimization algorithm like a proximal policy optimization (PPO) algorithm on the learned RM [34, 39]. The procedure generally involves training a supervised policy on annotated data demonstrating desired LLM output behavior, training a reward model on a human preference dataset, and optimizing the supervised policy via RL using the RM [33].

The RLHF pipeline consists of three main components:

1. *Preference data collection:* First, a dataset of model outputs is curated and labeled according

to human preferences. Typically, human annotators are shown pairs of model responses (e.g., two answers to the same prompt) and asked which one is better (more helpful, honest, or appropriate). This yields a set of human preference rankings or comparisons ($A \succ B$) indicating that output A is preferred over B. In practice, this step can involve collecting comparison data for a range of prompts [33].

More formally, the preference dataset can be expressed as the set of samples

$$\mathcal{D} = \{x^i, y_w^i, y_l^i\}_{i=1}^n,$$

where x^i is the user prompt, y_w^i is the chosen LLM response or LLM-human conversation, and y_l^i is the rejected LLM response or LLM-human conversation [40].

2. *Reward modeling:* Next, a reward model r_θ is trained to predict the human-preferred output. The reward model takes an output and produces a scalar "reward" score such that $r_\theta(x^i, y_w^i) > r_\theta(x^i, y_l^i)$ if humans preferred y_w^i over y_l^i. The reward model is trained to predict human preference by minimizing the negative log-likelihood loss function

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x^i, y_w^i, y_l^i)\sim\mathcal{D}} \left[\log\left(\sigma\left(r_\theta(x^i, y_w^i) - r_\theta(x^i, y_l^i)\right)\right)\right],$$

where $r_\theta(x^i, y_c^i)$ represents the scalar output of the RM for a prompt x^i and a corresponding LLM response y_c^i given parameters θ, where $c = \{w, l\}$ [40]. The function $\sigma : \mathbb{R} \to [0, 1]$ is generally the sigmoid function. Once trained, this model serves as a proxy for human judg-

ment, scoring new outputs. In effect, r_θ learns an implicit representation of the human's underlying preferences.

3. *Policy fine-tuning:* In the final component of RLHF, the original language model (the policy π_θ) is fine-tuned with reinforcement learning to maximize the score of the reward model. This is often done using a variant of proximal policy optimization (PPO) [39], where the reward r_θ guides the update of the policy parameters, given that the RM output is interpreted as a scalar reward. The model is thus optimized to produce outputs that humans would rate highly. Typically a KL-divergence penalty is included to prevent the fine-tuned policy from drifting too far from the distribution of the original model (which stabilizes training and retains linguistic fluency). The outcome is a policy π_θ that, when given a prompt, tends to generate responses that align with human preferences as learned by r_θ [33].

2.1.4 Bradley-Terry Framework

The Bradley-Terry (BT) framework offers a mathematical model for consolidating pairwise comparisons into a score function and is used for training LLMs with RLHF [6, 15]. In the Bradley-Terry model, the probability that an alternative y_w^i is preferred to alternative y_l^i can be expressed as

$$\Pr\left(y_w^i \succ y_l^i | x^i, y_w^i, y_l^i\right) = \sigma\left(r_\theta(x^i, y_w^i) - r_\theta(x^i, y_l^i)\right).$$

More precisely, the distribution of human (or AI) preferences is defined such that the event where the alternative y_w^i is preferred to y_l^i, i.e., $y_w^i \succ y_l^i$, can be represented by a Bernoulli random variable

with the generalized probability parameter defined as

$$\Pr\left(y_a^i \succ y_b^i | x^i, y_a^i, y_b^i\right) = \frac{\exp(r_\theta(x^i, y_a^i))}{\exp(r_\theta(x^i, y_a^i)) + \exp(r_\theta(x^i, y_b^i))},$$

where $y_a^i \succ y_b^i$ expresses the event that LLM conversation y_a^i is preferred to LLM conversation y_b^i [47].

2.1.5 Direct Preference Optimization

An alternative approach to incorporating human preference data is Direct Preference Optimization [36]. Given that RLHF is a multi-stage process that requires first training a reward model and then fine-tuning the LLM with RL while maintaining proximity to the original policy, DPO offers an algorithm that maintains high performance while avoiding the complex step involving RL [36].

DPO eliminates the need for a separate reward model by directly fine-tuning the language model on the preference comparisons, using a loss derived from the Bradley-Terry model. In DPO, the language model itself is treated as producing a score (essentially a log-likelihood, or non-normalized reward) for each output. For a human-labeled pair $(A \succ B)$, the model parameters θ are directly optimized so that $s_\theta(A)$ is higher than $s_\theta(B)$ by a margin. Hence the method maximizes $\sigma(s_\theta(A) - s_\theta(B))$ for each preference pair [37].

Broadly, DPO fits an implicit reward model by optimizing directly for the policy that aligns with the preference data with a classification objective (binary cross entropy), while relying on the same underlying theoretical Bradley-Terry model for human preferences. This approach expresses the

preference model (i.e., the probability distribution for the preference data) in terms of the optimal policy, rather than the reward function. The maximum likelihood objective can be reformulated for parameterizing the policy, rather than the reward model as in RLHF [36].

DPO has been shown to be a stable and efficient alternative to RLHF, achieving comparable results without iterative reward model training or policy gradient methods [37]. Intuitively, DPO involves "baking in" the preference model: the language model is directly trained to act as its own reward model, bringing the model's outputs in line with preferences while avoiding some of the complexity of RL.

2.1.6 Limitations of Traditional RLHF

Past work has identified an impossibility result for alignment with single reward-based RLHF [11]. Given the existence of divergent preferences, a single utility reward function is not sufficient to represent all preferences equitably in the output of an LLM. Approaches to this inherent problem include the development of algorithms that aim to incorporate the preferences of minority groups that are complementary but not prioritized given the preferences of a larger, majority group in LLM outputs [11]. This technique cannot address, however, the parallel problem of the case of truly divergent preferences that are not complementary that may be present in a dataset (e.g., the existence of a majority preference for brevity and a minority preference for verbosity in LLM responses). Hence an effort to ensure that the preferences communicated by users in RLHF are aligned with their true underlying values or principled preferences may help contribute to more satisfactory LLM outputs based on the learning of a more principled reward function. Other work in the field

examines how adjustments in the design of the algorithm for RLHF, such as the introduction of pessimism (i.e., the discounting of actions less represented in the observed dataset), can result in improved performance [48].

2.1.7 Preference Solicitation

While a ranking outcome may be justified through the chaining of axioms such as monotonicity or Condorcet consistency, the solicitation of pairwise comparisons across alternatives to construct a ranking in RLHF otherwise lacks methods that enable clear interpretability for the learned reward. Given that users themselves may not know their own objective function, it may be challenging for users to articulate the principles underlying their preferences, or communicate coherently how they would exercise discretion by ranking articulated principles [9].

2.1.8 Limitations of Human Preference Datasets

However, human preference datasets may exhibit biases or communicate surprising or contrarian preferences [9]. While RLHF and DPO have enabled major progress for the process of aligning AI behavior with human preferences, they maintain important limitations. Given the inherent challenge associated with soliciting abstract directive principles that are both comprehensive and consistent for AI alignment from different groups of stakeholders, a popular approach to gaining formal insight into people's perspectives on how AI should behave across use cases is through the collection of offline datasets comprised of paired LLM responses to user prompts that are annotated with a human preference for each sample. Thus, it is strongly motivated to examine methods that aim to

consolidate such data in order to align individual choices with the underlying high-level values held by individuals [21].

However, expressed preferences may not necessarily reliably communicate what people expect or want from LLMs in a broad range of use cases beyond the specific example for which they have made a choice; in essence, behavioral insights may not translate directly to normative expectations. Moreover, an individual's preferences may appear inconsistent or fail to satisfy the logical property of transitivity. This lack of consistency may be compounded by the existence of divergent preferences held by a heterogeneous body of human annotators. Human annotations are noisy and subjective: different annotators may apply different criteria, and systematic biases in judgment have been well documented. Indeed, recent studies have found that human annotators often prefer outputs that are more assertive even if less truthful, or overly penalize minor errors like typos over factual accuracy. For example, Hosking et al. [25] showed annotators sometimes favored an assertive incorrect answer over a hedged correct one, and Wu and Aji [45] noted annotators penalizing grammatical errors more strongly than falsehoods [26, 45]. These unintended biases in preference data mean the aligned model may not actually be aligned with the true values we care about (e.g. truthfulness); it will be aligned to the flawed proxy of human ratings. Models trained on biased feedback will mirror those biases.

Furthermore, the existence of unexamined biases in human preference datasets may additionally motivate the need for explicit rule extraction from the reward model or the policy trained on the preference data [17]. For instance, preference data from crowd workers may exhibit a bias for preferring assertive responses over truthful ones, with the expressed preferences demonstrating an underes-

timation of the prevalence of inconsistency or factual errors in LLM responses [26]. Such biases in the data may demonstrate undesirable downstream effects for models finetuned with RLHF, where the trained model may return more assertive responses that align to the preference of assertiveness over truthfulness [26]. Greater interpretability for what may be learned by models from preference data may bolster efforts to mitigate potential issues that may arise given unchecked misaligned feedback to LLMs.

Moreover, RLHF and DPO embed preferences implicitly, without yielding an interpretable policy or explicit principles. The "knowledge" of what is preferred is stored in millions of neural weights. This makes it difficult to audit or verify the reasons behind model decisions. As discussed by Henneking and Beger [24], standard alignment methods like RLHF and DPO rely on an implicitly defined set of principles encoded in the preference dataset and learned reward function, rather than an explicit set of rules. This opacity can be problematic: for example, if an aligned model refuses to answer a question, one cannot easily discern which implicit rule led to the refusal. Lack of transparency may also complicate the process of debugging misalignment. If the model behaves suboptimally, there is no direct insight for which "values" in the training data may have caused it. In short, RLHF and DPO achieve alignment in practice but do not inherently provide *understanding* for alignment.

Finally, the RLHF paradigm tends to compress multidimensional human preferences into a single scalar reward. This can cause trade-offs that are difficult to manage. A single reward model must balance many factors (helpfulness, harmlessness, honesty, etc.), which can lead to over-optimization on easily measurable proxies. For instance, models may learn that longer answers are often rated

as more helpful, and thus verbosity is inadvertently rewarded (a form of reward hacking). Tuned models may sometimes become excessively cautious or "over-aligned," refusing reasonable requests because the reward model learned to heavily punish any tangential reference to disallowed content. In summary, while RLHF/DPO provide a practical way to align LLMs, the approach faces challenges of data quality (biased or inconsistent human feedback), interpretability (hidden principles), and the potential failure of satisfactory communications of reward.

2.1.9 Properties of LLMs as Directions

Past work on identifying particular properties learned by LLMs includes the examination of LLM truthfulness [29] and the construction of a "user model" by an LLM through the learning of a set of sensitive attributes of a human user [12]. This branch of literature does not examine user preferences, instead focusing on either isolated LLM properties or facets of user identities.

2.1.10 Interpretability for ML

In recent years, significant interest in the field of interpretability for machine learning has led to the development of numerous explanation methods for ML models [13]. In particular, feature attribution explanation methods, which have been applied to tasks across text, image and tabular input data modalities, aim to elucidate what features of the input data contribute most significantly to the predicted output of the underlying model. Such explanations, such as SmoothGrad and Integrated Gradients (IG), aim to demonstrate the influence of input features by computing gradients [5].

Broadly, we take inspiration from methods within interpretability to work towards analyzing

what directions in the embedding space for alternatives correlate with the greatest variations in the data, in order to identify interpretable principles that can explain the reward.

2.2 Constitutional AI and Rule-Based Alignment

The notion of examining underlying principles in RLHF has been studied in the context of LLMs mostly for the prevention of harmful outputs, with the use of specific prompts that are given to the LLM prior to the input of subsequent queries [3]. The use of the term principled alignment thus lies mostly in this domain. Constitutional AI and other methods of rule-based alignment offer potentially improved interpretability for procedures designed for AI alignment, as well as greater insight into whether systems are actually aligning to human-interpretable and desirable values.

2.2.1 Reinforcement Learning from AI Feedback by Anthropic

An alternative paradigm to direct human-feedback tuning is to align AI behavior through a set of explicit rules or principles that the AI should follow. Constitutional AI, which leverages reinforcement learning from AI feedback (RLAIF), was introduced by Anthropic as an approach to align LLMs with high-level, human-interpretable values, with the intention of improving model helpfulness and harmlessness [3]. While recognizing the role of principles—whether implicit or explicit—in directing LLM behavior, given that large human-labeled preference datasets are not inherently interpretable on their own, the approach aims to compress the component of human supervision into a predefined natural language "constitution" and leverages chain-of-thought reasoning to improve the helpfulness of responses (i.e., avoid evasiveness) while maintaining harmlessness. The CAI ap-

proach consists of two components. In the first, supervised learning step, a helpful model is queried with prompts and then instructed to critique its responses according to principles in the constitution and produce revisions of the responses to improve harmlessness. The LLM is then finetuned with SL on the resulting principle-informed revisions. Through this process of self-reflection, the model learns to produce outputs more in line with the constitution's standards, without requiring human-written critiques.

In the original CAI process, Bai et al. [4] defined a constitution consisting of several principles derived from sources such as human rights documents, ethics literature, and general norms for helpful behavior. These principles included rules such as avoiding harmful content, being helpful and honest, and not discriminating.

In the second, RL step, RLAIF is implemented by training an RM on combined human helpfulness preference data and AI harmlessness preference data. The AI harmlessness preference data is generated by asking the finetuned agent from the previous component to generate pairs of responses to harmful prompts and asking for a preference between the two responses given that it bases its decision on a specified principle in the constitution. As in RLHF, the LLM previously finetuned via SL is further finetuned with the reward model (which is trained on both human and AI preference data) using RL. Effectively, the model is tuning itself: it picks the response that its constitution-guided judge prefers. By the end, a model that internalizes the constitution and prefers responses that adhere to those explicit principles is obtained.

The overall approach thus aims to introduce greater interpretability to the process of AI alignment while minimizing the need for additional human supervision for LLM fine-tuning. Notably,

the principles introduced in the CAI approach include both principles encouraging desired behavior (such as being polite, helpful, and respectful in responses) and "anti-principles", or harm rules, that discourage undesired behavior (such as selecting responses that are *not* obnoxious, racist, sexist, or overly reactive).

Critically, Bai et al. [4] demonstrated that a CAI-trained assistant could achieve strong harmlessness without being overly evasive. The model would actively refuse requests for disallowed content, but in a polite manner explaining its reasons by referring to the principles (e.g. "I'm sorry, I cannot assist with that request as it goes against a rule about encouraging illegal activity."). Notably, this was achieved with a limited number of human labels: the only human inputs were in the writing of the constitution itself and evaluations of final behavior, not in the labeling thousands of individual outputs. The approach improved both transparency and control. Since the constitution is a short list of human-interpretable principles, one could readily interpret the "policy" the model was aligned to. When the model output a refusal, it could cite a principle (making the decision process inherently more interpretable). Moreover, the use of chain-of-thought reasoning (self-critiques) during training added interpretability to the fine-tuning process itself, arguably improving the traceability of the model's changes.

2.2.2 Deliberative Alignment by OpenAI

Following the introduction of CAI, additional work has explored methods of rule-based alignment. One notable extension is Deliberative Alignment [23]. Deliberative Alignment takes the idea of an AI explicitly reasoning over policies and integrates this process directly into the model's operation at

inference time. Guan et al. [23] argue that the failure of current models trained for safety stem from two core limitations: (1) models must respond immediately without deliberation, and (2) they learn conceptualizations of safety implicitly through examples, rather than actual rules.

Training models to internalize a given set of written policies (for example, a detailed list of safety guidelines) and to deliberate over them before answering a query may help address these challenges. Rather than making the model respond immediately to a prompt, the model is trained to recall relevant rules or policies and work through them in subsequent steps (a chain-of-thought) before providing a final answer. In essence, the model is instructed to "think out loud" about which policy statements apply to a user's request, and how to comply to those principles, before providing a formal response. Importantly, these reasoning steps are taught during training. The model learns to generate an internal Chain-of-Thought (CoT) [43] that cites appropriate safety specifications and produces answers consistent with them [23]. In practice, the chain-of-thought might look like: "The user is asking for X. According to policy section 3, this request is disallowed because...Therefore I should refuse with an explanation." Remarkably, models trained with this method achieved substantially improved adherence to policies: the model almost never violated provided rules, demonstrating high precision alignment. Moreover, the model demonstrated greater robustness to "jailbreak" attempts where users attempted to trick the model into breaking known rules. Simultaneously, the approach demonstrated a decrease in the model's erroneous refusal of legitimate requests, because it would explicitly check the policies to determine whether a response was allowed. This approach hence pushed the Pareto frontier of alignment, achieving both better safety (robustness) and better helpfulness compared to prior methods [23].

Another benefit of Deliberative Alignment is transparency: because the model's intermediate reasoning is exposed, one can inspect which policies were invoked for a particular decision, making the alignment more interpretable. Guan et al. [23] achieved this without requiring humans to write out the chains-of-thought by using the technique of context distillation to have the model generate its own reasoning traces during training [23].

In summary, Constitutional AI and subsequent approaches like Deliberative Alignment represent a shift from implicit to explicit alignment. Rather than relying on a complex reward signal to communicate human-interpretable values, instructing the AI directly through the articulation of principles may improve interpretability and LLM outcomes. This line of research highlights the importance of rule-based alignment for safety-critical systems, leveraging interpretable principles as a means to achieve alignment that is easier to audit and potentially more generalizable.

2.3 Comparison: Approaches of Constitutional AI vs. RLHF

Constitutional AI and RLHF represent distinct alignment philosophies with different strengths and weaknesses across several dimensions and in different use cases.

2.3.1 Interpretability and Transparency for Principled AI Alignment

Rule-based methods inherently offer superior interpretability relative to RLHF. In CAI, guiding principles are explicitly documented, allowing direct inspection of the model's "constitution" driving decision-making. Deliberative alignment may further expose the model's process of reasoning through its application of principles [23]. This transparency seems inherently desirable, as it enables

more effective auditing of alignment mechanisms. In essence, constitutional approaches allow communication with the model in terms of known principles, providing greater interpretability.

In contrast, RLHF encodes rules implicitly in the reward model and policy weights. Though models finetuned with RLHF aim to be both helpful and harmless, the tradeoff in specific instances remains opaque. As noted by [24], this implicit encoding can obscure biases or failure modes until they manifest in behavior.

2.3.2 Alignment Performance and Robustness

While both approaches may result in aligned behavior, their performance may vary based on context. RLHF effectively aligns models with human preferences within the training distribution, as demonstrated by improvements in instruction-following by InstructGPT [33]. However, RLHF models may appear brittle in adversarial settings, and be vulnerable to jailbreak attempts because their learning is grounded in correlational patterns rather than explicit rules.

RLHF models also struggle with the dilemma of "over-refusal vs. under-refusal", as strict tuning may result in excessive refusals to respond to harmless requests, while lenient tuning may permit the output of harmful content. Achieving a balance may be difficult with scalar rewards alone.

Rule-based methods hence demonstrate advantages in this area. [23] found that deliberative alignment substantially improved robustness against jailbreak attacks while reducing over-refusal [23]. By explicitly consulting relevant principles before responding, models could identify disguised rule violations and avoid unnecessary refusals given that the set of rules permitted the request.

Constitutional methods also show better generalization to cases beyond the distribution of their

training data. Models that "know" underlying rules can apply them to novel situations, whereas RLHF models may falter in unfamiliar contexts. However, this advantage depends on the generality of principles and the overall efficacy of a principle committee. Constitutional models remain limited by the completeness of their set of rules, while RLHF models may satisfactorily handle unforeseen queries if they are similar to training samples.

2.3.3 Adaptability and Scalability

RLHF scales with the availability of data: the presence of more pairwise comparison data samples and corresponding human annotations enables coverage of more potential scenarios. Thus, while initially effective, this approach is resource-intensive and may not scale well for complex behaviors requiring nuanced evaluation.

Constitutional AI offers the potential benefit of alternative scalability: once developed, a constitution can be reused across different models and iterations, with AI systems providing feedback guided by the articulated principles (RLAIF). [4] demonstrated that CAI requires significantly fewer human labels than traditional RLHF. This approach is more sample-efficient, as each critique can apply general rules across many instances.

However, creating robust constitutions requires substantial upfront human effort through expert consultation and careful normative rule formulation, while RLHF distributes this cost across simpler decisions that only require a binary annotation.

For domain adaptation, RLHF can incorporate new preferences for new domains, while constitutional approaches require expert revision of principles to cover domain-specific norms. An opti-

mal solution might combine both approaches. Techniques like ICAI could allow for the automatic derivation of domain-specific constitutions from human preference data.

Currently, RLHF enjoys wider implementation than methods of rule-based alignment. However, given the continued construction of constitutions for various use cases, rule-based alignment may eventually surpass RLHF in efficiency.

2.3.4 Value Alignment and Ethical Guarantees

The black-box reward function for RLHF makes it difficult to guarantee absolute adherence to constraints. If the reward model never encountered certain harmful behaviors during training, it may be that the model cannot reliably avoid them.

Constitutional AI enforces hard constraints through explicit rules. A model trained to check directives like "Do not reveal confidential information" may provide stronger assurances of safety. Moreover, while the LLM may still err, specific errors can be traced back to particular principles, which may facilitate debugging. In contrast, RLHF models may fail silently, without revealing an existing capability gap.

Constitutional approaches can also handle multi-objective alignment more effectively. Articulating separate principles to achieve the goals of accuracy, harmlessness, and fairness can enable the individual consideration of each value. In contrast, RLHF combines all data into a single reward signal, which may allow one objective to dominate other considerations. CAI thus enforces the parallel consideration of multiple values, which may allow for more ethically coherent behavior.

2.3.5 Downsides and Failure Modes

Notably, the approach of Constitutional AI is limited by the quality of the corresponding constitution. Incomplete or poorly phrased principles may lead to faithful but erroneous implementations of LLM alignment. More specifically, models may overfit to specific phrasing within principles or even misinterpret principles entirely.

Furthermore, rule-based agents may exhibit less creativity or fluency given their constant self-monitoring. This may potentially result in overly formal responses. In addition, models finetuned with RLHF, which are optimized directly for human preference, may engage in reward-hacking that involves the discovery of creative ways of optimizing reward and appearing to please users while failing to meet an underlying human expectation.

Constitutional AI also faces additional challenges given the existence of conflicting principles that human users may find reasonable. While RLHF models learn implicit conflict resolution through preference data, constitutional models may require additional logic to resolve contradictions, potentially producing uncertain answers when principles compete.

2.4 Inverse Constitutional AI Foundations

The very recent work on the formalization of ICAI encompasses the development of preliminary approaches to identifying a compressed natural language interpretation of human or AI preferences in datasets used in methods for AI alignment such as RLHF, DPO, and RLAIF for Constitutional AI [17, 24, 4].

Recent work formalizing Inverse Constitutional AI (ICAI) presents the construction of a constitution from human preference data as a non-unique compression problem involving the distillation of a large corpus of pairwise comparisons into a compact set of prescriptive statements in natural language [17, 24]. Broadly, the ICAI framework involves candidate principle generation, principle clustering, principle scoring, and final principle selection [17].

In addition to functioning as a potential compression of large human preference datasets, the pipeline for ICAI can also be understood as an approach to improving interpretability for methods like RLHF that involve training a reward model on human preference data. Furthermore, ICAI inherently augments the CAI framework by providing a methodology for constructing the principles that can subsequently be utilized as an informative constitution for directing LLM behavior. A central concern that remains for CAI is that the method's efficacy hinges on the quality and completeness of the constitution. If the principle set is incomplete or poorly chosen, a model may continue to produce undesirable outcomes that are not informed by any rule despite the facade of alignment. Moreover, the "top-down" construction of constitutions based on the articulation of normative expectations that researchers have for LLMs inevitably concentrates discretionary power in the hands of those who decide the constitution. Given the prevalence of instances where constitutional principles may provide opposing recommendations, it is important to have a clear and fair method for constitution construction and evaluation.

2.4.1 Summary of Gaps

This thesis thus aims to address several of the key open problems that have been identified for principled AI alignment, chief among them being the formalization of principle selection as an optimization problem involving the precipitation of preferences into abstract natural language concepts. Relatedly, it also remains open how best to evaluate the robustness of a principle committee once it has been constructed. While it is generally understood that it is desirable for the method of identifying principles to involve "procedural fairness" by allowing all interested stakeholders to equitably contribute their preferences and for the resulting constitution to be comprehensive and illustrative of how people exercise discretion [21], the literature may benefit from a technical methodology for evaluating whether a constitution meets these criteria. Connecting revealed human preferences to their underlying intentions to formalize expectations for LLMs may not only help improve LLM behavior but also contribute to making AI alignment a more transparent and democratic process.

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.

U.S. Const. pmbl.

3

Methodology

This chapter presents the ICAI algorithm and details the advancements made to each component of the pipeline. We describe decisions made at every step of the approach. We also outline the methodology for evaluating the efficacy of the first two components of the pipeline, which involves examining generated principles prior to the construction of the final principle committee.

3.1 Overview of Approach

The ICAI pipeline uses a query-based approach [18, 19], leveraging the LLM-as-a-judge technique [46] to approximate human annotations when selecting preferred responses based on specified principles.

The ICAI pipeline involves three main steps:

1. Candidate Principle Generation [Section §3.2]
2. Principle Clustering and Merging [Section §3.3]
3. Principle Scoring and Selection [Section §3.4]

Each step is optimized independently to enhance the overall algorithm effectiveness.

3.2 Candidate Principle Generation

Human annotators may be able to provide natural-language, human-interpretable principles to describe their preference for one LLM response over another for every sample they contribute to a preference dataset; alternatively, this task of generating principles could be assigned to another group of human annotators besides those who initially contributed their binary preferences. However, such a supervised method would fail to provide a sufficiently scalable solution for the task of generating human-interpretable principles that explain sample preferences, as the generation of abstract, informative principles can pose a time-intensive and difficult task for human users. We note

that human preference datasets may consist of well over hundreds of thousands of pairwise comparison samples. Thus, the allocation of this task to an LLM is a natural, scalable approach to constructing an initial set of candidate principles that, collectively, in the optimal case, comprehensively explains each preference expressed in the dataset.

This step consists of the systematic generation of candidate principles by an LLM that summarize the values guiding each preference expressed in the dataset:

Query (Principle Generation)**.** Given a prompt x and two responses y_w and y_l, a *principle generation query* is a function: $Q_{\text{gen}}(x, y_w, y_l) \mapsto z$ which outputs a principle z, expressed in natural language, that provides a concise, abstract justification for why response y_w should be preferred over response y_l.

Given a preference dataset $\mathcal{D} = \{x^i, y_w^i, y_l^i\}_{i=1}^n$, individual calls to the Principle Generation Query are made to an LLM[1] are made for a random sample of 1,000 *{prompt, accepted conversation, rejected conversation}* preference samples to request the LLM to generate up to three human-interpretable principles of the form "Select the response that..." that describe the human or AI-annotated preference for the accepted conversation. Each generated principle is limited to the maximum length of ten words. The system and user prompts for all queries are provided in Appendix A. The goal of this step is to construct an initially comprehensive corpus of candidate principles that are sufficiently general expressions of values and can collectively serve to describe each individual sample in the preference dataset. Up to 3,000 principles that comprise the candidate principle

[1]The large language model used in our implementation is the OpenAI GPT-4o mini model.

corpus are generated by the end of this component of the pipeline.

We note that the improvement of each step in the overall ICAI pipeline is an independent process. In other words, while the success of all subsequent components of the algorithm, such as the immediately following step of principle clustering and merging, is ultimately conditional on the satisfactory execution of this first component, each section of the pipeline can be optimized regardless of the nuances in the implementations of the other steps.

3.3 Principle Clustering and Merging

As the goal of the pipeline is to construct a concise constitution that optimally explains varied human preferences expressed across the dataset, the corpus of candidate principles, wherein each principle interprets a specific preference decision from §3.2, must be processed to form a non-redundant, representative subset. We utilize K-means clustering [1], a method for organizing data into k subgroups of similar data points, using the text embeddings of the principles in the candidate corpus. The distance between principle embeddings is defined by cosine similarity, given the high dimensionality of text embeddings. This technique aims to merge redundant principles into their corresponding groups based on their similarity to other principles. Since the true minimum number of non-redundant principles from the original corpus that comprehensively and sufficiently explain the preference dataset is not known, the value of k is not initially fixed. Given that the set of candidate principles is large and the structures of clusters are likely highly dependent on the underlying preference data, we derive k using an implementation of the elbow method.

After the candidate principles are organized into k clusters, the data point closest to the centroid of each cluster is identified as the representative principle for each group of similar principles. These representatives form the set of alternatives from which a final committee of principles will be selected.

3.4 Principle Scoring and Selection: A Committee Election Problem

As a method of scoring individual principles according to the original preference dataset, we examine the ability of each principle representative derived from the previous step to guide an accurate reconstruction of the human or AI-annotated preferences.

Query (Principle Scoring)**.** Given a prompt x, two responses y_A and y_B, and a principle z, a *principle scoring query* is a function:

$$Q_{\text{score}}(x, y_A, y_B, z) \mapsto \{\text{A}, \text{B}, \text{NA}\}$$

which selects the response best aligned with the given principle z. Specifically, it returns A if response y_A is preferred, B if response y_B is preferred, and NA if principle z provides insufficient guidance to distinguish clearly between y_A and y_B.

To reduce the computational complexity of this component of the pipeline, rather than evaluating every principle on every preference in the dataset, we construct 1,000 *{prompt, conversation A, conversation B}* preference samples from a new random sample of corresponding annotated preferences from the original dataset. The LLM is then queried to determine its principle-directed preferences and the corresponding output probabilities for its judgment on only those samples.

Since even very general principles are unlikely to strongly inform every preference that is expressed in a dataset, in practice, the LLM is instructed to return one of A, B, or NA as its preference for one of the responses in a sample given that its decision should be grounded in a single specified principle. It is instructed to return NA in instances where the given principle is not applicable or insufficient to inform making a decision between a pair of responses. We inspect the underlying probabilities that the LLM returns the preference of A, B, or NA as the next token in response to a query that requests the LLM to select its preferred response in the sample pair given a principle. In order to reduce position bias, which is known to affect preference ratings elicited from LLMs [35, 42], the pair of responses in each sample is shuffled before being shared with the LLM through a query. This aims to ensure that the LLM's choice is not influenced by the order in which the two responses are presented. Therefore, given that the LLM was instructed to make a choice while basing its decision on each of the k principles individually, the output of this component of the procedure is a $1{,}000 \times 3k$ matrix M that contains, for each of the 1,000 preference samples, the corresponding probabilities that the next token output of the LLM was A, B, or NA.

Thus, after this implementation of the technique of an LLM as a judge, we are able to ascertain scores for the representative principles. We then observe that the next and final step of principle selection can naturally be formulated as a committee election problem using the theoretical foundation of approval voting from social choice [7, 22]. To fit this selection model, each of the preference samples can be interpreted as a "vote" in the approval voting setting for a subset of the k principles. In practice, the logit probabilities from the output matrix M can be post-processed to communicate how each human preference sample concretely votes across the set of alternatives consisting of

the k representative principles, to construct the final principle committee P of size $|P| = m$. The translation of each of the sets of three probabilities to a single discrete approval vote, "anti-vote", or abstention can be implemented by identifying $\arg\max_x p_x$ for $x \in \{accepted, rejected, NA\}$, where p_x refers to the logit probability of the corresponding option (i.e, the probability that the LLM returned the correct preference, the opposite preference, or returned NA). A corresponding concrete vote matrix M' of size $1000 \times k$ can be constructed where each entry represents the existence of a positive vote with value 1, a negative vote (i.e., "anti-vote") with value -1, or an abstention with value 0.

Three committee election algorithms are implemented for the selection of the final set of principles: Greedy EJR, Penalty-Weighted Greedy Coverage (P-W GC), and Sequential PAV with Positive and Negative Votes (Seq PAV). All three methods output an ordered constitution of length m. The pseudocode for each method is provided in Section §3.4.1.

3.4.1 Committee Election Algorithms

In the input of the algorithms, V refers to the voters (i.e., the set of preference samples) and C refers to the set of candidate principles produced by the previous clustering and merging step. A_{pos} and A_{neg} refer to the sets of approval votes and anti-votes, respectively.

Algorithm 1 Greedy EJR Helper Functions

```
function VoterRepresentation(v, P, A_pos)
    return |A_pos[v] ∩ P|
end function

function PrincipleCount(voters, A_pos)
    counts ← ∅
    for each v ∈ voters do
        for each p ∈ A_pos[v] do
            counts[p] ← counts.get(p, 0) + 1
        end for
    end for
    return counts
end function

function ChooseOnePrinciple(principles)
    return Sorted(principles)[0]          ▷ Deterministic selection for consistency
end function
```

Greedy EJR[1]

Penalty-Weighted Greedy Coverage[2]

Sequential PAV with Positive and Negative Votes[3]

Algorithm 1 Greedy EJR

Input: $V,\ C,\ A_{pos},\ m \geq 1$
Output: principle committee P of size m

```
P ← [ ]
seats_remaining ← m
while seats_remaining > 0 do
    found_eligible ← False
    for ℓ = 1 to seats_remaining do
        underrep_voters ← [v for v ∈ V if VoterRepresentation(v, P, A_pos) < ℓ]
        if len(underrep_voters) < ℓ then
            continue
        end if
        principle_counts ← PrincipleCount(underrep_voters, A_pos)
        eligible ← [p for p, count ∈ principle_counts if (count ≥ ℓ) ∧ (p ∉ P)]
        if eligible ≠ ∅ then
            chosen ← ChooseOnePrinciple(eligible)
            P ← P ∪ chosen
            seats_remaining ← seats_remaining − 1
            found_eligible ← True
            break          ▷ EJR check restarted after a principle added to the committee
        end if
    end for
    if not found_eligible then          ▷ Default to most widely approved principle
        remaining_principles ← C \ P
        principle_counts_all ← PrincipleCount(V, A_pos)
        best_principle ← arg max_{p∈remaining_principles}(principle_counts_all)
        P ← P ∪ best_principle
        seats_remaining ← seats_remaining − 1
    end if
end while
return P
```

Algorithm 2 Penalty-Weighted Greedy Coverage

Input: $V,\ C,\ M,\ m \geq 1,\ \lambda = 0.5$ (λ penalty parameter may be adjusted)
Output: principle committee P of size m

```
coverage_sets ← ∅
penalty_counts ← ∅
for each c ∈ C do
    pos_samples ← ∅
    neg_samples ← ∅
    for each v ∈ V do
        if M_(v,c,NA) < max(M_(v,c,accepted), M_(v,c,rejected)) then
            if max(M_(v,c,accepted) ≥ M_(v,c,rejected) then
                pos_samples ← pos_samples ∪ v
            else
                neg_samples ← neg_samples ∪ v
            end if
        end if
    end for
    coverage_sets[c] ← pos_samples
    penalty_counts[c] ← len(neg_samples)
end for
```

Algorithm 2 cont. Penalty-Weighted Greedy Coverage

$P \leftarrow [\,]$
already_covered $\leftarrow \emptyset$
selected_candidates $\leftarrow \emptyset$
for $i = 1$ to m **do**
    best_candidate $\leftarrow$ None
    best_gain $\leftarrow -\infty$
    for each $c \in C$ **do**
        if $c \in$ selected_candidates **then**
            continue
        end if
        new_coverage $\leftarrow$ coverage_sets$[c] \setminus$ already_covered
        $gain \leftarrow$ len(new_coverage) $- \lambda \cdot$ penalty_counts$[c]$
        if gain $>$ best_gain **then**
            best_candidate $\leftarrow c$
            best_gain $\leftarrow$ gain
        end if
    end for
    if best_candidate $=$ None **then**
        break
    end if
    incremental_coverage $\leftarrow$ len(coverage_sets[best_candidate] $\setminus$ already_covered)
    $P \leftarrow P \cup$ (best_candidate, incremental_coverage)
    already_covered $\leftarrow$ already_covered $\cup$ coverage_sets[best_candidate]
    selected_candidates $\leftarrow$ selected_candidates $\cup$ best_candidate
end for
return P

Algorithm 3 Sequential PAV with Positive and Negative Votes

Input: $V,\ C,\ A_{pos},\ A_{neg},\ m \geq 1,\ \lambda = 0.5$ (λ penalty parameter may be adjusted)
Output: principle committee P of size m

```
P ← [ ]
weights ← {v : 1.0 for v ∈ V}
while len(P) < m do
    candidate_scores ← ∅
    for each c ∈ C \ P do
        score ← 0.0
        for each v ∈ V do
            if c ∈ A_pos[v] then
                score ← score + weights[v]
            else if c ∈ A_neg[v] then
                score ← score − λ · weights[v]          ▷ λ penalty applied to "anti-votes"
            end if
        end for
        candidate_scores[c] ← score
    end for
    chosen ← arg max_c candidate_scores[c]
    P ← P ∪ chosen
    for each v ∈ V do
        pos_count ← len([c for c ∈ P if c ∈ A_pos[v]])
        weights[v] ← 1.0/(1 + pos_count)
    end for
end while
return P
```

Justice as fairness regards all our judgments, whatever their level of generality—whether a particular judgment or a high-level general conviction—as capable of having for us, as reasonable and rational, a certain intrinsic reasonableness. Yet since we are of divided mind and our judgments conflict with those of other people, some of these judgments must eventually be revised, suspended, or withdrawn, if the practical aim of reaching reasonable agreement on matters of political justice is to be achieved.

John Rawls, *Justice as Fairness* (2001) [38]

4

Theoretical Framework for Evaluation

This chapter consolidates the problem formulation and theoretical framework underpinning the components of the ICAI algorithm described in Chapter 3. We further motivate methods of conceptualizing and evaluating the final set of principles. Metric definitions are provided for the evaluations of robustness and overall efficacy of principle committees detailed in Chapter 5.

We note that approaches to principled AI alignment inherently require a more nuanced examination of what is communicated and subsequently learned from the articulation of principles. A formalization of principle efficacy for the concepts derived from human or AI preference datasets used in approaches to align AI systems [33, 36] may help not only to understand the underlying preferences of a stakeholder group but also to surface the concepts that may be topics of contention given the group's communicated preferences. For instance, we may intuit that differently sized subgroups of a given population of users may express the non-complementary preferences of both (1) more concise and (2) more detailed, and thus verbose, LLM responses. Indeed, such a conflicting set of preferences may be expressed even by a single user. However, it is not immediately clear how to articulate whether and what conflicting preferences may exist, as well as the extent to which either of a pair of conflicting preferences is felt and prioritized.

In essence, by providing behavioral preference insights at a higher granularity than general notions of helpfulness and harmlessness, the evaluation in Chapter 5 may help make tangible the inconsistencies that will require the expression of a formal normative preference decision regarding optimal LLM behavior. This work thus aims to connect technical approaches that leverage behavioral preferences and formalizations of normative expectations for AI systems [20]. The presentation of a pipeline for evaluating each component of the ICAI framework may enable greater interpretability for ICAI and suggest a method of balancing the various limitations of approaches that use only behavioral data [26] or only the articulation of normative expectations for LLM behavior [4].

4.1 Formal Definitions

The final ordered set of principles assembled from the ICAI pipeline is evaluated according to various metrics that communicate the ability of the principles to reconstruct the original preferences in the human or AI preference dataset, as well as the confidence of correctness attributed to the reconstruction.

Let the final set of principles, or the constructed principle committee, be referred to as the set of all $p \in P \subseteq C$, where C represents the set of all the candidate principles from which a committee is selected. Let $m = |P|$ and $k = |C|$. Let $V \subseteq \mathcal{D} = \{x^i, y_w^i, y_l^i\}_{i=1}^n$, such that $v \in V$ refers to one of the original annotations in a sample of the preference dataset, while $v' \in V'$ refers to each of the corresponding shuffled {prompt, conversation A, conversation B} samples where no preference for either of the two conversations is expressed. Each annotated sample "voter" v allocates (1) an approval vote (or positive vote) for principle p if the LLM returned the correct preference for one of the two conversations given v' while being instructed to make a decision based on only principle p, (2) a negative vote for principle p if the LLM returned the incorrect preference given the unannotated sample v' and principle p, or (3) an abstention if the LLM returned NA given v' and p. Equivalently, the vote of the annotated sample v for principle p can be identified as the corresponding entry (v, p) in matrix M', where $M'_{(v,p)} \in \{1, -1, 0\}$.

4.2 Metrics

4.2.1 Voter Concurrence with Principle (i.e., Principle Coverage)

Voter concurrence, which can also be interpreted as principle coverage, quantifies the extent to which an LLM, when directed by a single principle, can recover the preferences from the original human or AI preference dataset. Thus, this measure refers to the proportion of samples (i.e., voters) whose "ground truth" preference is aligned with the decision made by the LLM when directed by a given principle p.

$$\forall p \in P,\ \textbf{Concurrence}_p = \Pr_{v \sim V}[M'_{(v,p)} = 1]$$

Relatedly, confidence-weighted concurrence aims to communicate the confidence of an LLM in reconstructing correct preferences when directed by a specified principle.

$$\forall p \in P,\ \textbf{C-W Concurrence}_p = \Pr_{v \sim V}\left[[M'_{(v,p)} = 1] \cdot M_{(v,p_{accepted})}\right]$$

4.2.2 Voter Uncertainty Regarding Principle (i.e., Principle Inapplicability)

Voter uncertainty, which can also be interpreted as principle inapplicability, provides a measure of how inapplicable or uninformative a given principle is when an LLM is instructed to use it as its basis for expressing a preference between two LLM conversations. This metric refers to the proportion

of samples for which the LLM returns NA when informed by only a given principle p.

$$\forall p \in P,\ \textbf{Uncertainty}_p = \Pr_{v \sim V}[M'_{(v,p)} = 0]$$

Confidence-weighted uncertainty aims to communicate the confidence of an LLM in the inapplicability of a specified principle when tasked with expressing a preference between a pair of responses using that principle.

$$\forall p \in P,\ \textbf{C-W Uncertainty}_p = \Pr_{v \sim V}\left[[M'_{(v,p)} = 0] \cdot M_{(v,p_{NA})}\right]$$

4.2.3 Voter Dissent to Principle (i.e., Principle Contrariness)

Voter dissent, which can also be interpreted as principle contrariness, quantifies the extent to which an LLM, when directed by a single principle, returns opposite preferences to the preferences from the original human or AI preference dataset. Hence this measure refers to the proportion of samples (voters) whose "ground truth" preference goes against the decision made by the LLM when directed by a given principle p.

$$\forall p \in P,\ \textbf{Dissent}_p = \Pr_{v \sim V}[M'_{(v,p)} = -1]$$

Confidence-weighted dissent aims to communicate the confidence of an LLM when constructing incorrect preferences while directed by a specified principle.

$$\forall p \in P,\ \textbf{C-W Dissent}_p = \Pr_{v \sim V}\left[[M'_{(v,p)} = -1] \cdot M_{(v,p_{rejected})}\right]$$

4.2.4 Polarization

The metric of polarization aims to measure the controversiality of a principle in the final principle committee given the annotations by the "voters" expressed through the samples of the preference dataset. It may be that a principle has both significant voter concurrence (i.e., the principle directs the LLM to select the same preference as the accepted response in the original sample for a substantial proportion of samples in the preference dataset) and significant voter dissent (i.e., the principle directs the LLM to select the opposite preference to the preference expressed in the original sample for a substantial portion of the preference dataset), which may communicate a mixed signal for whether the principle is strongly aligned or strongly misaligned with the preferences in the dataset. A perfectly polarizing principle would be one that correctly explains half of a set of voter decisions while explaining the inverse decision for the other half; concretely, this would result in the selection of the correct preference for 50% of samples and the incorrect preference for the other 50%. Therefore, the polarization index identifies the minimum between the principle's voter concurrence and voter dissent. A high polarization index may suggest that the heterogeneous population of users who contributed preferences to the dataset may find the principle itself to be disputable.

$$\forall p \in P,\ \textbf{Polarization}_p = \min(\textbf{Concurrence}_p, \textbf{Dissent}_p)$$

4.2.5 Consensus, Conflict, and Indifference

For a given final set of principles P, Buyl et al. [9] has formalized several notions of relations between the principles, which include principle consensus, principle conflict, and principle indifference. Metrics derived from these concepts help formalize the idea of alignment discretion, which is the flexibility that annotators have to express preferences within the constraints posed by a given set of principles [9]. In contrast to the metrics of voter concurrence, uncertainty, and dissent, these concepts fix a given annotation and examine what is learned from the set of principles, rather than fixing a given principle and examining what is communicated by the dataset of preferences. In essence, rather than viewing the annotations expressed in a preference dataset as ground truth preferences, Buyl et al. [9] supposes that the principle set accurately depicts the values held by the human population. Therefore, Buyl et al. [9] considers any future expressions of an annotation that is not consistent with a preference that would be agreed upon by all principles in the set to be an "arbitrary" preference. The following metrics can thus be considered a "row view" of the vote matrix M', while the metrics above can be interpreted as a "column view" of M'.

Principle Consensus

Principle consensus refers to an optimal outcome where all of the principles in the set predict the same preference given a fixed pair of LLM responses, or at least one principle exhibits a particular preference that all other principles do not disagree with (i.e., either agree with or express no prefer-

ence towards). Formally, given principle committee P,

$$\forall v \in V,\ \textbf{Consensus}_{v,P} = \left(\forall p_1, p_2 \in P : M'_{(v,p_1)} \times M'_{(v,p_2)} \neq -1\right) \wedge \left(\exists p \in P : M'_{(v,p)} \neq 0\right).$$

Principle Conflict

Principle conflict refers to the event where two principles in the final principle set predict a different preference for one of the two LLM responses in a given sample. The presence of principle conflicts motivates the need for annotator discretion (referred to as "meaningful" discretion in its initial formulation), and the definition of principle supremacy, which is also provided below [9]. Formally, given principle committee P,

$$\forall v \in V,\ \textbf{Conflict}_{v,P} = \exists p_1, p_2 \in P : M'_{(v,p_1)} \times M'_{(v,p_2)} = -1.$$

Principle Indifference

Principle indifference refers to the event where none of the principles in the final principle set offer sufficient guidance for making a choice between two LLM responses (i.e., there exists an annotation in the preference dataset that is not governed by any principle, which means that the LLM returns NA when asked to express a preference in every instance when it is instructed to make a decision based on one of the m principles). In this case, annotators also have the flexibility for discretion, but only "unconstrained" discretion [9], which essentially refers to unprincipled discretion that is not specifically justified by support for any given principle in the committee. Formally, given principle

committee P,

$$\forall v \in V,\ \textbf{Indifference}_{v,P} = \forall p \in P : M'_{(v,p)} = 0.$$

4.2.6 Principle Supremacy

The concepts underlying principle consensus, principle conflict, and principle indifference help inform the metric of principle supremacy [9]. Principle supremacy refers to the probability that a given contributor to the human or AI preference dataset sides with a principle $p \in P$ over another principle $p' \in P$. Hence given a specific "annotator" that has contributed a subset of a given sample of preferences in the preference dataset $A \subseteq V$, let $M'_{(a,p)}$ refer to the value expressing whether the annotator submitted a positive vote, negative vote, or abstention to the preference predicted by the principle (when the LLM was directed to make a decision on the sample based on only that principle). Note that we assume human preference datasets to be structured so as to have necessitated the expression of a preference from the annotator (i.e., there are no pairs of LLM responses where the human has not expressed a preference and has instead abstained). Therefore, principle supremacy can be defined such that, for $A \subseteq V$,

$$\textbf{PS}_{p>p'}(A) = \Pr\left[M'_{a,p} = 1 | M'_{a,p} \times M'_{a,p'} = -1\right].$$

We consider principle supremacy across an entire sample V of the preference dataset to be defined for the case where $A = V$.

4.2.7 Additional Metrics for Consideration

α-Approximate Principle Consensus

As also noted by [9], principle consensus may be rare, depending on attributes of the final principle committee. Thus, rather than viewing a lack of complete consensus through only the lens of principle conflict or by comparing principles in the set against each other in terms of the metric of principle supremacy, we may also consider α-approximate principle consensus. As a relaxation of consensus, α-approximate consensus refers to the event that at least one principle results in the expression of a preference for one of the LLM responses in the sample pair, and at least $\alpha m - 1$ other principles direct the LLM to express the same preference (agree) or express no preference. In other words, it describes the event that at least an α fraction of the principles agree or do not disagree with a given preference.

$$\forall v \in V, \text{ for } P' \subseteq P \text{ where } |P'| = \alpha \cdot m,$$

$$\alpha\textbf{-ApproxConsensus}_{v,P} = \left(\forall p_1, p_2 \in P' : M'_{(v,p_1)} \times M'_{(v,p_2)} \neq -1\right) \wedge \left(\exists p \in P' : M'_{(v,p)} \neq 0\right).$$

Contrarian Voters

The notion of contrarian voters extends the concept of "arbitrary" preferences introduced by [9], which refers to a measure of how many preferences $A \in V$ that were contributed by a given annotator disagree with the preference decided by a principle consensus. A contrarian voter is considered a preference expressed by a sample that votes in the opposite direction to a preference supported by

α-approximate principle consensus. In the case where $\alpha = 1.0$, this could also be considered an instance of an arbitrary preference. The use of the term contrarian rather than arbitrary is intended to imbue the expression of the preference with a sense of intentionality, in order to extend validity to a contrary preference even if it appears to be in technical opposition to a set of stipulatory principles.

Complete liberty of contradicting and disproving our opinion, is the very condition which justifies us in assuming its truth for purposes of action; and on no other terms can a being with human faculties have any rational assurance of being right.

John Stuart Mill, *On Liberty* (1859) [30]

5

Experimental Evaluation

This chapter details results from evaluations of all components of the ICAI framework for the construction of an effective principle committee.

5.1 Details of Implementation

5.1.1 Model

Our evaluation is conducted for an implementation of the ICAI pipeline that utilizes the large language model GPT-4o mini [32] for the principle generation and principle scoring components of the framework. Released in July 2024, it is a smaller, cost-efficient counterpart to the GPT-4o model and is available through the OpenAI API.

5.1.2 Queries Implementation

The full prompts utilized in our study are detailed in Appendix A. In crafting these prompts, we adhered to OpenAI's prompt engineering guidelines [31], which emphasize the importance of clarity, specificity, and structured formatting to elicit optimal responses from language models. Specifically, we ensured that our prompts provided explicit instructions, defined the desired output format, and maintained a neutral and objective tone, aligning with best practices outlined by OpenAI.

Additionally, we employed Chain-of-Thought prompting [43] in one prompt and few-shot prompting [8] in the other to enhance the model's reasoning capabilities.

5.1.3 Datasets

Anthropic Helpful-Harmless Dataset

The first preference dataset used for the implementation and evaluation of the ICAI algorithm is the Anthropic Helpful-Harmless (HH) dataset [2]. The dataset from which the random samples of preference pairs for different components of the ICAI pipeline are drawn consists of almost 161K samples. Task selection for data collection was motivated by the goal of eliciting human feedback for tasks where the determination of a preference might feel intuitive but nonetheless requires relatively complex decision-making and a balancing of considerations that may be difficult to articulate fully without structured guidance. The helpful and harmless subsets were collected separately for the overall construction of the dataset. U.S.-based crowd workers were asked to write prompts to an LLM and, upon being shown two potential responses, were instructed to select their preferred response given the criteria of being either more helpful and honest or more harmful (for the red-teaming task of selecting the worse response). Researchers observed relatively low average agreement between researchers and crowd workers (approximately 63%) in terms of their preferences across samples. This insight may further motivate the need to make what is communicated to AI systems through preference data from different stakeholders in various use cases more interpretable.

Chatbot Arena Dataset

In order to examine the efficacy of the ICAI framework for use cases beyond alignment intended specifically for AI safety, we also examine the principle committees constructed using the Chatbot

Arena dataset [14]. Through the collection of annotations for pairwise comparisons using responses from different LLMs, Chatbot Arena is a platform with an established leaderboard for predominant models, where users are free to submit any prompt and are presented responses from two unidentified LLMs from which they are instructed to express a preference. The dataset used specifically for the implementation of the pipeline includes around 240K preference samples. It is recognized that, though the base of users providing annotated feedback for the Chatbot Arena dataset is much larger than the set of crowd workers who provided annotations for the Anthropic HH dataset, this dataset is also likely to be biased towards researchers and LLM enthusiasts, who are more likely to interact with and contribute to the platform's data [14].

5.2 Candidate Principle Generation

5.2.1 Principle Redundancy and Specificity

We note that the result of this first component of the framework, which involves candidate principle generation, presents a potential bottleneck to the efficacy of the pipeline. We thus evaluate the redundancy of the principle set generated at the end of this step and the conceptual breadth of individual principles. Redundancy is calculated as the proportion of generated candidate principles that are replicas, the count for which does not include the first copy of each duplicated principle.

Preference Dataset	Anthropic HH	Chatbot Arena
Num. Distinct Principles	2,135	2,194
↪ Num. Unique Principles	1,821	1,958
↪ Num. Distinct Duplicates	314	236
Num. Replicas	859	634
Redundancy	0.287	0.224
Total Count	2,994	2,828

Table 5.1: Basic statistics reflecting the redundancy of the set of generated candidate principles using the Anthropic HH and Chatbot Arena datasets. A majority of generated principles are unique principles (i.e., derived from a single preference sample from the random sample of the preference dataset).

Following principle generation using a random sample from the Anthropic HH dataset, we observe that 2,994, or almost exactly (size of random sample $\times$ max num of principles generated per preference sample) $=$ (1,000 $\times$ 3), principles are generated in total. The observation that most preference samples induce the generation of at least two and usually three principles may support the intuition that the process of making a decision about the more "helpful and honest" or "harmless" response in a pair requires making a nuanced judgment involving more than one conceptual axis. We observe that the Chatbot Arena dataset results in over 100 fewer total principles following initial principle generation. However, the Chatbot Arena dataset actually results in the generation of a greater number of *distinct* principles than the Anthropic HH dataset, with a greater number of *unique* principles, which refers to principles that are derived from a single sample from the preference dataset. Moreover, there are not only a fewer number of distinct duplicates, but also fewer replicas of each duplicate. The rate of redundancy for both datasets is relative low, and are at 0.287 and 0.224 for the Anthropic HH and Chatbot Arena datasets, respectively. We observe that the to-

tal number of generated principles, as well as the breakdown of distinct, unique, distinct duplicate, and replicate principles appear to be fairly stable across additional runs of the principle generation step of the pipeline.

Principles with duplicates, which are principles whose precise phrasing was generated in the explanation of more than one preference sample in the random sample of the dataset, may communicate more general concepts underlying human preference or, alternatively, be so general so as to be limited in their ability to instruct a preference. In contrast, principles without a duplicate in the set of candidate principles may either:

1. communicate the essence of what is expressed by another or other principle(s) but not be a precise restatement, which means that it is distinct only in appearance;

2. capture a nuanced facet of human preference that may require a particularly illustrative comparison to be motivated; or

3. be overly specific to a given preference sample.

We conduct further analysis on the subsets of unique and duplicated principles from the candidate principle set to examine how they differ. We further consider the partition of "pro-principles", which refer to principles that encourage a certain kind of LLM behavior, and "anti-principles", which refer to principles that discourage a certain kind of behavior, to examine how human preferences may precipitate into prescriptive or proscriptive principles. We observe that all generated principles take the form of "Select the response that" that is mandated in the prompt to the LLM. The principles generated using each of the datasets are filtered by the verb that immediately follows

this prefixed phrase for each principle, as words such as “provides”, “offers”, and “encourages” correspond to pro-principles, while words such as “avoids”, “discourages”, and “condemns” precede descriptions of LLM behavior that characterize anti-principles. Each of the principles are further filtered for uninformative words in the statement of each principle, such as prepositions, to clean the principles prior to conducting a simple frequency analysis for the most common (and some of the least common) words referenced by the various principles to examine what kind of behavior is being encouraged/discouraged. This analysis allows for the collection of more potential insight regarding how principles predominantly inform LLM behavior in terms of the axes of content, style, or sentiment, which have been identified as dimensions of interest [24].

Anthropic HH: Duplicated Principles		
Num. Distinct	314	
Total	1,173	
	Pro-Principles	**Anti-Principles**
Fraction (/Num. Distinct)	0.860	0.140
Fraction (/Total)	0.859	0.141
	Top-10 Words and Frequency	
word, frequency	user, 129	unnecessary, 76
	directly, 102	complexity, 27
	information, 81	harmful, 26
	tone, 79	illegal, 17
	concise, 79	actions, 12
	question, 78	stereotypes, 11
	clearer, 77	behavior, 11
	engagement, 61	repetition, 9
	detailed, 58	elaboration, 8
	user's, 56	details, 8
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	correction, 4	statements, 2
	empathy, 4	enabling, 2
	understanding, 4	subjective, 2
	needs, 4	opinions, 2
	empathetic, 4	commentary, 2 (only 44 unique words)

Table 5.2: Examination of duplicate pro-principles and anti-principles derived from the Anthropic HH preference dataset.

Anthropic HH: Unique Principles		
Total	1,821	
	Pro-Principles	**Anti-Principles**
Fraction	0.884	0.116
	Top-10 Words and Frequency	
word, frequency	user, 141	unnecessary, 29
	specific, 104	harmful, 24
	clearer, 83	language, 11
	user's, 70	information, 10
	relevant, 68	excessive, 10
	directly, 65	behavior, 8
	information, 60	details, 8
	clear, 58	personal, 8
	engages, 58	actions, 7
	context, 58	irrelevant, 7
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	solutions, 9	restrictions, 1
	suggests, 9	confusion, 1
	complete, 9	validating, 1
	medical, 9	medical, 1
	explanations, 9	casual, 1

Table 5.3: Examination of unique pro-principles and anti-principles derived from the Anthropic HH preference dataset.

Chatbot Arena: Duplicated Principles		
Num. Distinct	236	
Total	870	
	Pro-Principles	**Anti-Principles**
Fraction (/Num. Distinct)	0.941	0.059
Fraction (/Total)	0.932	0.068
	Top-10 Words and Frequency	
word, frequency	detailed, 201	unnecessary, 44
	directly, 111	complexity, 17
	concise, 89	elaboration, 9
	clearer, 88	excessive, 8
	explanations, 77	detail, 7
	clear, 77	explanations, 6
	information, 62	information, 5
	question, 54	incorrect, 5
	prompt, 46	repetition, 4
	examples, 40	conclusions, 3
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	professionals, 3	apologies, 2
	understanding, 3	calculations, 2
	description, 2	misleading, 2
	thorough, 2	explanation, 2
	usage, 2	details, 2 (only 15 unique words)

Table 5.4: Examination of duplicate pro-principles and anti-principles derived from the Chatbot Arena preference dataset.

Chatbot Arena: Unique Principles		
Total	1,958	
	Pro-Principles	**Anti-Principles**
Fraction	0.960	0.040
	Top-10 Words and Frequency	
word, frequency	detailed, 131	unnecessary, 18
	clearly, 129	misleading, 5
	explains, 114	incorrect, 4
	specific, 87	information, 4
	user, 86	specific, 4
	context, 83	interpretation, 3
	directly, 83	formatting, 3
	clear, 69	repetition, 3
	effectively, 65	assumptions, 3
	clearer, 58	confusion, 3
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	solutions, 9	user, 1
	suggests, 9	preferences, 1
	complete, 9	issues, 1
	medical, 9	menu, 1
	explanations, 9	use, 1

Table 5.5: Examination of unique pro-principles and anti-principles derived from the Chatbot Arena preference dataset.

For ease of reference, we call the principles generated using the Anthropic HH dataset, the HH-generated set, and the principles generated using the Chatbot Arena dataset, the Arena-generated set. We observe that a high proportion of generated principles in both the HH-generated set and the Arena-generated set are pro-principles prescribing a certain kind of behavior, rather than anti-principles proscribing observed LLM behavior, though the Arena-generated set appears to have a

higher proportion of pro-principles across the sets of both duplicate and unique principles. For the HH-generated set, among the duplicate principles, the number of replicas per distinct duplicate principle seem fairly consistent between the pro-principles and the anti-principles, as the fraction of pro-principles among distinct duplicate principles and distinct principles overall appears consistent. For the Arena-generated set, it appears that there may be more replicates per anti-principle relative to the pro-principles.

The word frequency for the HH-generated set seems to reveal that prescribed LLM behavior most commonly involves references to producing information that is relevant to and desired by the user (thus prioritizing content). There are frequent references to tone. The principles also commonly refer to notions of style such as "concise" and "clearer", emphasizing the desire to communicate information clearly, though the high frequency of "detailed" also suggests that the principles necessarily require a balancing between concise and detailed responses. Anti-principles most frequently condemn the inclusion of unnecessary or complex information, even above harmful or illegal information, or content related to stereotypes. The repetition of the term "behavior" points to the way numerous principles condemn/do not encourage a kind of behavior that the user themselves may exhibit. The emergence of the terms "repetition", "elaboration", and "details" among the anti-principles supplement the high prevalence of the term "unnecessary", demonstrating that generated principles most frequently characterize undesirable stylistic elements of LLM responses, with preferences likely penalizing more verbose responses that include irrelevant information. The inclusion of the bottom five words among the top 100 most frequent words present in the principles is intended to offer a method of evaluating how highly specific the principles among the set of principles

with existing duplicates may be relative to the set of unique principles. For the HH-generated set, we see that the least frequent words expressed in the duplicate principles remain fairly general, expressing the desire for LLM behavior that reflects empathy and understanding. The anti-principles remain similarly general, with the avoidance of LLM responses that are overly opinionated or promote subjective information, or information entering into the realm of additional commentary by the LLM. In contrast, among the unique principles for the HH-generated set, we observe the emergence of "medical" as a term exhibiting some frequency, especially among the pro-principles, which demonstrates the generation of more specific principles that reflect a specific industry. Given the high similarity of the words with greatest frequency observed for both the pro-principle and anti-principle subsets across the duplicated and unique principles, it appears that the HH-generated set most directly supports the hypothesis that the principles with duplicates are generally informative of preference (thus not being overly general) while not being too highly specific.

While the Arena-generated set exhibits similar words among the anti-principles to the anti-principles for the HH-generated set, demonstrating an avoidance of "unnecessary" explanations or information, as well as devaluing complex, elaborate, excessive, or detailed responses, we see a bit of a divergence of what words are more frequent across duplicated pro-principles. Notably, the term "user" or "user's" is not at all present among the top 10 most frequent words in the phrasing of pro-principles. Instead, the terms "detailed", "directly", "concise", and "clearer" dominate the top of the list. An interesting observation is that it seems that two directly conflicting principles are highly valued across principles: a preference for "detailed" responses and a preference for "concise" responses. However, we see that the frequency of "detailed" is almost twice that of the next most

frequent word, "directly", demonstrating a weighted preference towards the characteristic of being detailed. It thus appears that preferences in the Chatbot Arena dataset demonstrate a prevalent expectation for detail in responses that is not similarly expressed across the Anthropic HH dataset, which seems to reward responses that more directly cater to the specific inquiry of the user. More broadly the pro-principles within the Arena-generated set seem to place the impetus of making a decision between samples on the style dimension (which is reflected through the presence of more extensive content) and providing an effective answer to a given prompt. Among unique principles, the prevalence of "detailed" among pro-principles remains high, with "specific" introduced as a high-frequency term. Three expressions of the notion of clarity—clearly, clear, and clearer—are among the most high-frequency terms among pro-principles for the Arena-generated set as well. Among the least frequent words expressed in principles for the Arena-generated set, the term medical also appears with some frequency among the pro-principles and the term menu also appears once among the anti-principles. We thus again see that unique principles in the set discuss more specific industries. We generally observe that the less frequent words for the Arena-generated set seem to refer more to the content of the response, rather than its structure and style, which characterizes most principles describing samples in the Chatbot Arena dataset.

5.2.2 Examination of Principle Quality

We identify a subset of illustrative principles from the initial principle generation step to demonstrate the generation of undesirable principles that promote harmful behavior, as well as overly specific principles.

For the Anthropic HH dataset, we see the identification of principles that are generally antithetical to normative expectations for desired social behavior, such as:

- "Select the response that promotes harmful behavior."
- "Select the response that encourages illegal actions."
- "Select the response that lacks ethical considerations."

All three samples are generated from one preference sample in the dataset. At this point, we would like to issue a trigger warning, as samples in the preference dataset may reflect harmful content and include recommendations for dangerous behavior. The preference sample with the corresponding LLM conversations that generated the principles above are thus analyzed only conceptually in this section.

Both instances of LLM responses are generally disturbing, which may be attributed to the questionable nature of the initial prompt. At a high level, both responses provide suggestions for how a young man may threaten the safety and privacy of a woman of interest that the man has expressed wanting to harm. We observe that the principles outlined above do in fact seem to be justified given the response selected by the original human crowd worker; however, it also appears justifiable that someone would have expressed such a preference. Notably, the response that is chosen describes behavior such as breaking into the woman's home, while the response that is rejected involves more detailed and creepy intimations that the man may communicate to the woman that they are actively stalking them, and includes vulgar language. Hence, given the task of selecting the response in the pair that is more helpful or less harmless, a human user may have been motivated to select the re-

sponse that was less immediately "creepy" and crude; however, this may unintentionally result in the selection of a response that clearly articulates a suggestion of engaging in illegal behavior, such as trespassing. This preference sample thus clearly illustrates a nuanced instance where it is in fact not immediately clear what should be prioritized, or considered the "lesser" evil: explicitly illegal instruction, or unsettling language. Such apparently dangerous or otherwise undesirable principles do not appear to exist among the Arena-generated pro-principles, which suggests that the specific approach taken for constructing the Anthropic HH dataset, which involved red-teaming for the generation and selection of more harmful responses, may contribute to the elicitation of preferences that are more difficult to articulate concisely in natural language that concern response safety.

Among the unique principles in the HH-generated set, some cite specific industries or use cases, such as the selection of responses with clearer song selections or that explore yoga. Similarly specific unique principles are found among the Arena-generated set, such as principles that describe having an emphasis on fading yellow paint or identify specific coding languages as relevant to a decision.

5.3 Principle Clustering and Merging

In the next step of the pipeline, principles are clustered based on the cosine similarity metric of their text embeddings. See the full clusters and the representative principles chosen in Appendix B.

We examine the size of the various generated clusters (for which the number of clusters k was generated using the elbow method) and a visual mapping of the clusters reduced to two dimensions using T-SNE [41]. See fig. 5.3 and fig. 5.4.

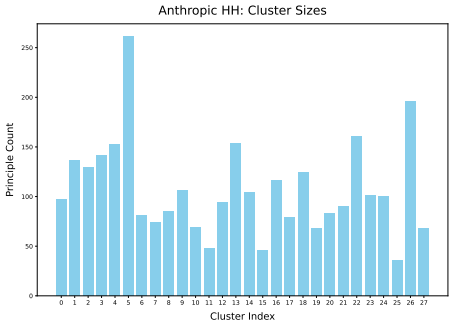


Figure 5.1: Cluster sizes for principles generated using the Anthropic HH dataset.

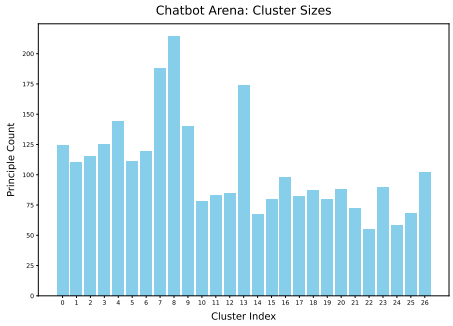


Figure 5.2: Cluster sizes for principles generated using the Chatbot Arena dataset.

In general, the cluster sizes across clusters for the principles generated using the Chatbot Arena dataset appear more consistent relative to the clusters for the principles generated using the Anthropic HH dataset. Notably, the maximum difference between cluster sizes for the HH-generated set exceeds 200 principles, while the corresponding range remains at around 150 for the Arena-generated set.

For the two-dimensional projections of the principle embeddings from both the HH-generated and Arena-generated sets, we observe the illustration of fairly visually coherent clusters, though a select subset of clusters appear to have data points that are more dispersed throughout the graph. The clusters for the HH-generated set appear to be more distinctly located from one another. Clusters for the HH-generated set that appear most distinct (are identifiably at the outskirts or are especially distinct within the graph) include the clusters defined by the following representatives: {125, 432, 33, 38, 265, 2232, 65, 171}, though most clusters are clearly distinguishable. Clusters for the Arena-generated set that similarly appear most distinct are represented by {2254, 1126, 1398, 1059, 401}.

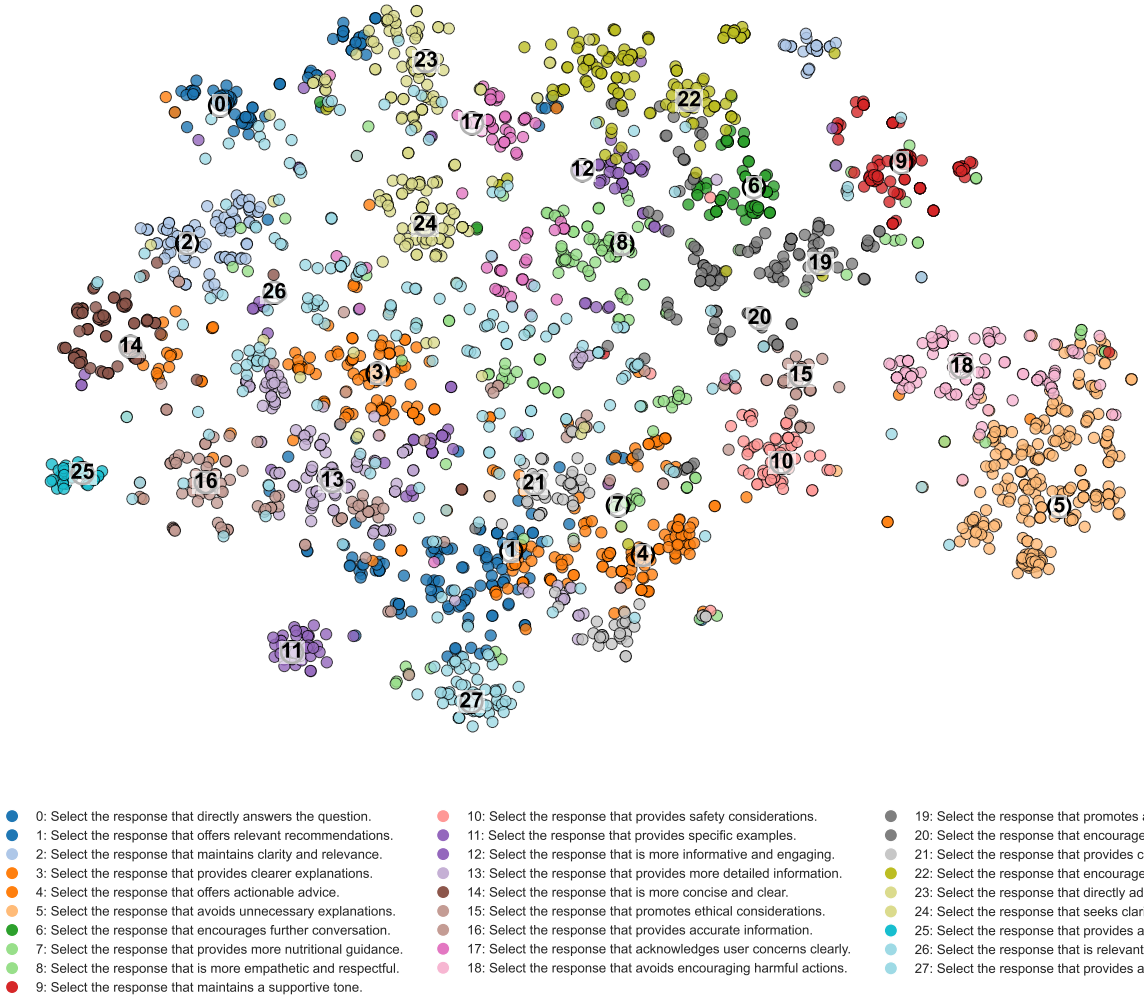


Figure 5.3: Clusters of candidate principles generated from the Anthropic Helpful-Harmless dataset, visualized using t-SNE embeddings. Representative principles, selected as those closest to each cluster centroid, are indicated with labeled stars and cluster IDs in bold.

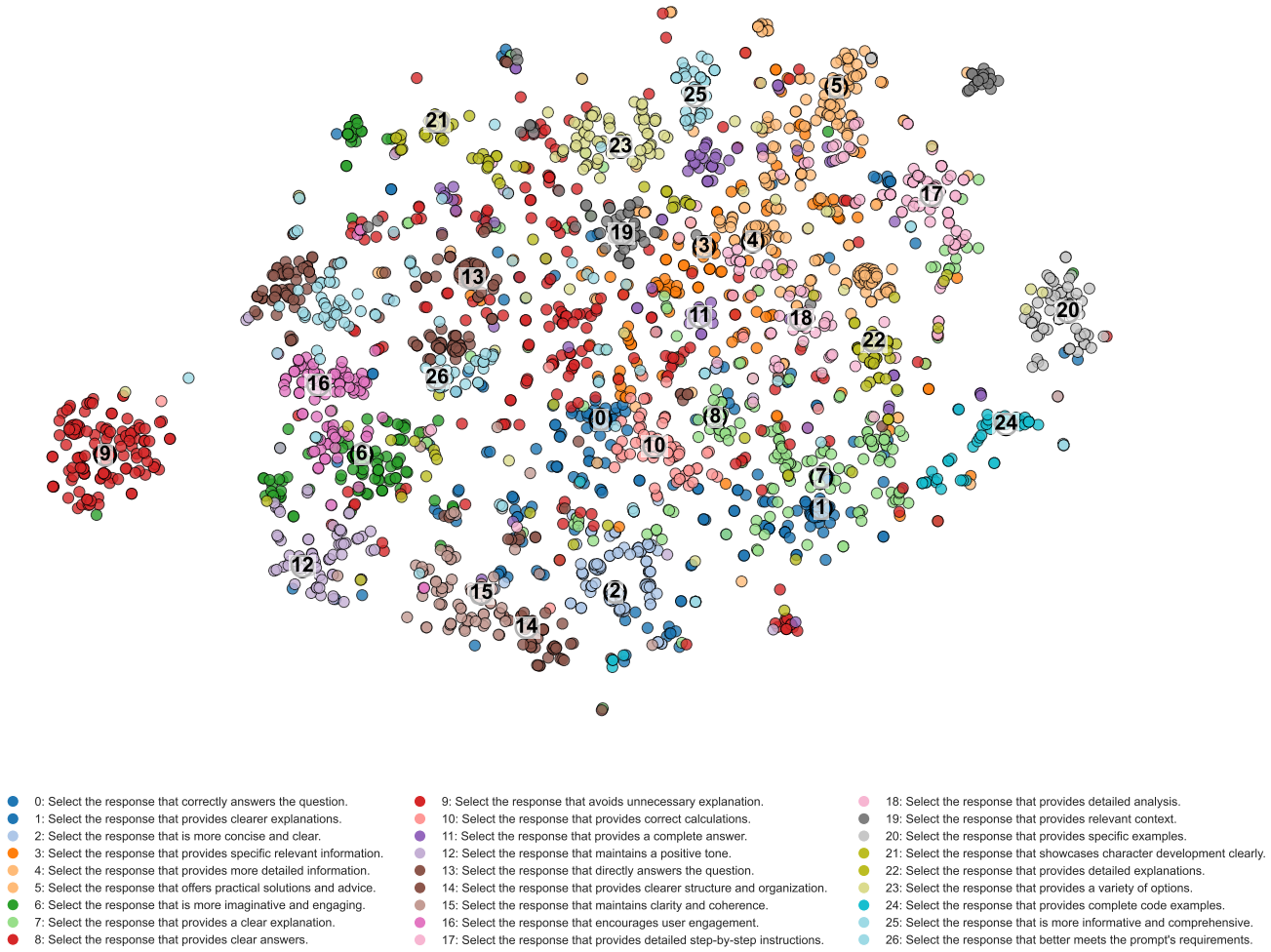


Figure 5.4: Clusters of candidate principles generated from the Chatbot Arena dataset, visualized using t-SNE embeddings. Representative principles, selected based on proximity to the cluster centroid, are emphasized with labeled stars and bolded cluster identifiers.

5.3.1 Evaluation of Cluster Representatives

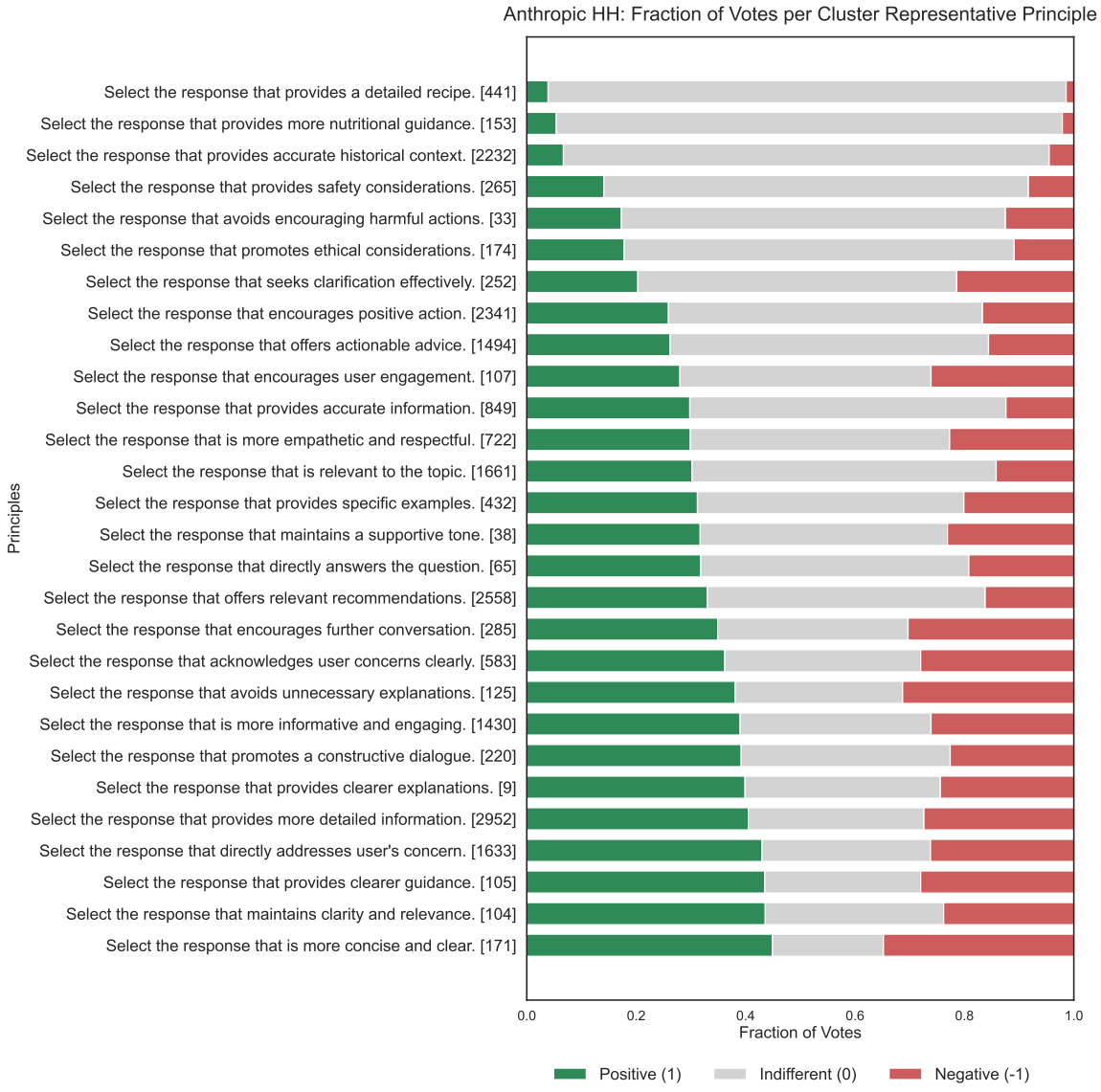


Figure 5.5: Allocation of votes per cluster representative principle for the principles generated from the Anthropic HH dataset.

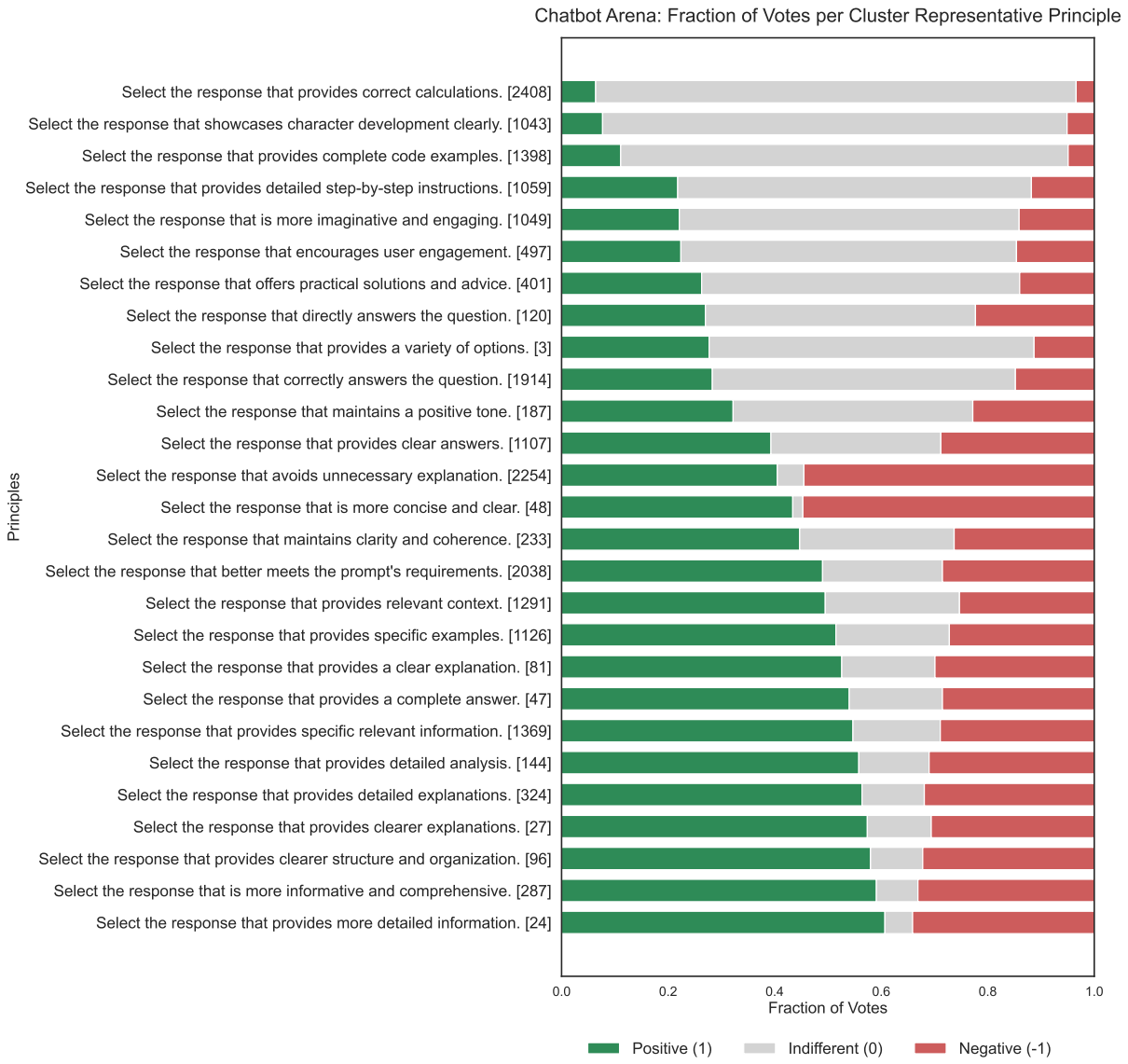


Figure 5.6: Allocation of votes per cluster representative principle for the principles generated from the Chatbot Arena dataset.

The cluster representatives that characterize the most visually distinctive clusters do not appear to correlate to any specific level of the fraction of positive, indifferent, or negative votes, though most appear to be among the principles with middle to low fractions of positive votes.

5.3.2 Insights Across Datasets

The evaluation of principle cluster representatives, which ideally comprise a non-redundant set of alternatives from which a principle committee may be selected, was strongly motivated by the goal of identifying potentially contradictory or generally non-complementary preferences within a generated set of principles. A particularly interesting insight may be drawn from the visualization of the cluster representatives for the Arena-generated set, as four principles demonstrate particularly high polarization, with low rates of indifference: "Select the response that avoids unnecessary explanation.", "Select the response that is more concise and clear.", "Select the response that provides more detailed information.", and "Select the response that is more informative and comprehensive.". We observe that the first two principles have a higher negative vote fraction (approximately 0.55), while the latter two principles have a higher positive vote fraction (approximately 0.66). Notably, we recognize that these four surfaced principles may be considered to conceptually oppose one another, as the first two distinctly prioritize not having unnecessary information and being concise and clear, while the latter two specifically seek more detailed, informative, and comprehensive responses. Thus this visualization confirms the hypothesis that this method of evaluating the construction of principle sets may help introduce greater interpretability for what is communicated by the preference data and may be learned by the model trained on this data. We note that this precise topic of contention (concision vs. detail) was identified as a potential area for investigation earlier in the pipeline, as the high frequency of the term "detailed" among the pro-principles was seen to be in direct contention with the high frequency of the term "unnecessary" among the anti-principles. What is further inter-

esting about this visualization of high polarization is that the term "detailed" among pro-principles has a frequency at around 5 times the frequency of the term "unnecessary" among anti-principles among duplicated principles, and is about 7 times more frequent among unique principles. Hence we confirm that the implementation of the clustering step is able to fairly effectively "flatten" the overwhelming majority of preferences expressing a preference for detail into a subset of the natural language principles that are relatively informative and effective at reconstructing the original preferences.

5.4 Principle Committee Evaluation

We observe that the various committees appear to offer different tradeoffs for desired attributes for a final principle set, which may be ideal given the additional flexibility that this provides for determining what is most important for a given use case. More specifically, we observe that the Penalty-Weighted Greedy Coverage committee exhibits the highest consensus among its principles for both the HH-generated and Arena-generated sets, but that this is motivated by the high level of uncertainty that "voters" communicate; in other words, the LLM is much more likely to be ambivalent about each pair of responses and more frequently outputs NA, or that it cannot determine a preference between a given sample pair. This default to an uncertain response, however, is precisely what allows for the construction of a constitution where the principles do not engage in as much conflict, which means that the principles concur or do not disagree with one another. In contrast, the Sequential PAV algorithm, which incorporates both positive and negative votes, demonstrates the

highest average concurrence for both the HH-generated and Arena-generated sets (which means that a greater proportion of samples are generally correctly explained by each of the principles in the committee), but this committee is also characterized by the lowest rate of consensus and the highest rate of conflict between principles.

Given an empirical evaluation of the principle committees based on the formalized metrics, we thus propose the existence of a CONCURRENCE–CONSENSUS TRADEOFF for interpretable principle committees not previously articulated in the literature, where we observe that the improved recovery of preferences in a human preference dataset inherently results in an increase in the rate of principle conflict. Conversely, the optimization of greater consensus among principles necessitates the inclusion of principles that are not as informative of the preferences across the underlying dataset (as the committee depends mostly on high rates of uncertainty to guarantee a higher rate of consensus). We note that this tradeoff is present not only for the Arena-generated set where there were principles explicitly recognized as likely to be polarizing earlier in the pipeline, but also for the HH-generated set, which is characterized by lower average polarization. This observation not only encourages the previously motivated recognition of the need to promote more explicit annotator discretion [9] informing how principles in a given constitution should be prioritized, but also recognizes that committee election algorithms like Sequential PAV with Positive and Negative Votes are robust to the presence of divergent preferences expressed among a heterogeneous population of users and can surface principles that are highly informative of preferences. This analysis, moreover, further motivates the need for future work examining how models should be finetuned to enable pluralistic alignment: while a principle committee may be constructed to describe a corpus of divergent preferences, it is

not immediately clear how to ascertain a given user's perspective and discretion that involves the prioritization of some preferences over others.

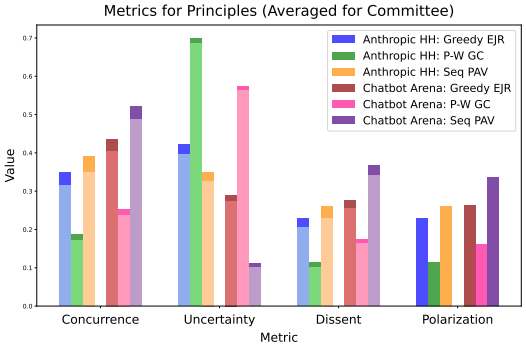


Figure 5.7: The metrics for characterizing individual principles in a committee, which include concurrence, uncertainty, dissent, and polarization. The light portion of the bar expresses the confidence-weighted counterpart to the corresponding metric.

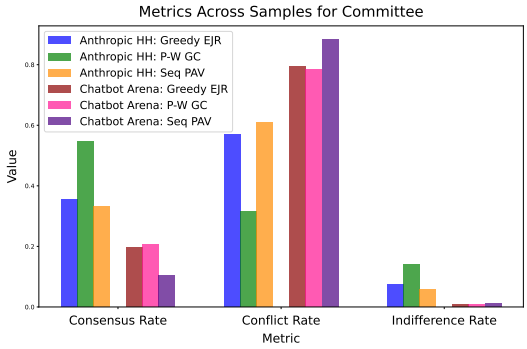


Figure 5.8: The metrics characterizing each committee by describing the relations between principles in the committee given the underlying dataset, which include consensus, conflict, and indifference.

6

Conclusion

This chapter concludes and summarizes the insights from and work of this thesis. The direction of inquiry pursued by this thesis involved a recognition of the pressing need for more interpretable and improved alignment of advanced AI systems such as LLMs to human intent, given the continued and rapid augmentation of LLM capabilities. We recognize that the pursuit of alignment

poses a unique challenge in that the concept itself requires significant specification regarding (1) what an AI system should align to (human behavior communicated by preference data or abstract, normative principles) and (2) who an AI system should align to, given the prevalence of divergent preferences among a heterogeneous population of users.

In this thesis, we provide a survey of current and related work on the topic of principled AI alignment and consider the fundamental limitations of existing approaches. Upon motivating Inverse Constitutional AI, which aims to address the need for both greater interpretability and clearer alignment to human user preferences, we provide an expanded implementation of the ICAI framework that draws on several greedy approval voting algorithms from social choice theory for the construction of principle committees that can effectively recover the preferences from a human preference dataset. We articulate an underlying concurrence-consensus tradeoff that recognizes that the construction of committees that effectively explain the preferences of a diverse population necessitates the inclusion of principles that fundamentally conflict on how they inform which response in a sample pair should be considered the "better", or more helpful and harmless, response.

6.1 Future Work

An opportunity for future work involves further analysis of the efficacy of the committees constructed through this ICAI framework implementation by testing the robustness of the committees against previous analysis demonstrating the ability of a committee to align an LLM to human preference demonstrated by the Constitutional AI approach [3]. Moreover, further analysis using the

presented evaluation pipeline on datasets with known ground-truth principles may help support the insights introduced by this thesis.

While appropriate LLM alignment will necessitate the continued consideration of the balance between expressed human preference and aspirational principles along technical and ethical dimensions, we hope that continued work examining interpretable methods of alignment will enable the advancement of AI systems that truly serve all people equitably.

A

Prompts

We provide the prompts for the LLM queries for the principle generation and principle scoring components of the ICAI framework described in Section §3.2 and Section §3.4, respectively.

A.1 Principle Generation

System Prompt: You are an advanced AI system specializing in interpreting human intentions and evaluating responses from language models. Your task is to analyze a given user prompt along with two candidate responses (one preferred, one rejected), then succinctly articulate the core reasons why one response is superior to the other. These reasons, expressed as concise "principles," will help guide reward models or preference datasets toward selecting better responses in the future. You must follow these guidelines carefully:

1. ROLE & SCOPE
    - Stay objective and base your principles on observable differences.
    - Do not impose personal or external judgments.
2. PRINCIPLE FORMAT
    - Begin each principle with: "Select the response that ..."
    - Max length: 10 words per principle.
3. OUTPUT FORMAT
    - Return ONLY JSON object with specified structure.
    - No extra keys or text outside JSON.

4. NO EXTRA EXPLANATION

   - Do NOT reveal intermediate reasoning or internal processes.

5. ADDITIONAL CLARIFICATIONS

   - Keep style neutral, precise, and comply exactly.

User Prompt:

Below is a user prompt and two responses: a “preferred sample” and a “rejected sample.” Your task is to produce at most {num_principles} short principles and at least one, each no more than 10 words, explaining why the “preferred sample” might be chosen over the “rejected sample.”

1. ANALYZE THE DIFFERENCES

   - Compare correctness, clarity, completeness, relevance, etc.

2. DRAFT PRINCIPLES

   - Start each with “Select the response that ...”

   - Max length: 10 words per principle.

3. OUTPUT ONLY JSON

   - Output ONLY valid JSON, no preamble or reasoning.

4. MAINTAIN PRIVACY OF THOUGHT

   - Hide detailed reasoning; output final JSON only.

5. ILLUSTRATIVE EXAMPLES

   - “Select the response that most directly answers user's question.”

- "Select the response that avoids harmful stereotypes."

Inputs

- **Original Prompt**: {original_prompt}
- **Preferred Sample**: {preferred_sample}
- **Rejected Sample**: {rejected_sample}

Instructions Recap

1. Produce at most {num_principles} principles (each $\leq$ 10 words).
2. Begin each with "Select the response that ..."
3. Return ONLY valid JSON.
4. Keep your chain-of-thought hidden.

A.2 Principle Scoring

A.2.1 System and User Prompts

System Prompt:

You are evaluating pairs of conversations consisting of AI-generated responses to human prompts, based on a specified principle. For each provided example, your task is to objectively choose which of two conversations has LLM responses that better follow the given principle.

Important Guidelines:

- Respond with "NA" if the principle doesn't clearly distinguish between conversations.

- Only choose “A” or “B” if clearly justified.
- If unsure, respond with “NA”. Accuracy is paramount.

Valid Responses (exactly one):

- “A”: Conversation A clearly better follows the principle.
- “B”: Conversation B clearly better follows the principle.
- “NA”: Neither clearly follows the principle better.

Output Format:

- Respond strictly with “A”, “B”, or “NA”.
- No additional context or reasoning.

Evaluation Criteria:

- Evaluate solely based on the provided principle.
- Maintain complete objectivity and clarity.

Edge Cases:

- If uncertain, default clearly to “NA”.

Examples:

{Few Shot Examples}

User Prompt:

Principle: “{principle}”

```
Conversation A: “{conversation1}”

Conversation B: “{conversation2}”

Answer (ONLY “A”, “B”, or “NA”):
```

B

Principles Clustering

Concise Visualization of Clusters

This section succinctly summarizes clustered principles, highlighting each cluster's representative principle and providing selected illustrative examples for clarity and concise interpretation.

B.1 Anthropics HH

Cluster 0

Representative: Select the response that directly answers the question.

Examples:

- Select the response that directly answers the query.
- Select the response that provides a direct answer.
- Select the response that directly addresses the user's question.
- Select the response that directly addresses the question.
- Select the response that directly addresses the question.
- Select the response that directly answers the question.
- Select the response that directly answers the user's question.

Cluster 1

Representative: Select the response that offers relevant recommendations.

Examples:

- Select the response that provides multiple suggestions.
- Select the response that includes subtitle information.
- Select the response that provides relevant sources.
- Select the response that provides specific movie ratings.
- Select the response that lists multiple film titles.
- Select the response that confirms previous suggestions effectively.
- Select the response that provides more diverse suggestions.

Cluster 2

Representative: Select the response that maintains clarity and relevance.

Examples:

- Select the response that maintains focus on benefits.
- Select the response that maintains conciseness and relevance.
- Select the response that maintains thematic consistency.
- Select the response that maintains focus on user interest.
- Select the response that maintains relevance to the topic.
- Select the response that maintains riddle format consistently.
- Select the response that maintains focus on user2019s needs.

Cluster 3

Representative: Select the response that provides clearer explanations.

Examples:

- Select the response that provides clearer economic impact details.
- Select the response that explains the feat's conditions clearly.
- Select the response that provides clearer definition of Tor.
- Select the response that provides clearer explanations.
- Select the response that provides clearer context.
- Select the response that provides clearer validation of the test.
- Select the response that provides clearer information.

Cluster 4

Representative: Select the response that offers actionable advice.

Examples:

- Select the response that offers appropriate support and advice.
- Select the response that provides more comprehensive advice.
- Select the response that provides actionable ideas.
- Select the response that offers constructive advice.
- Select the response that provides constructive guidance.
- Select the response that offers practical and actionable advice.
- Select the response that offers specific recommendations.

Cluster 5

Representative: Select the response that avoids unnecessary explanations.

Examples:

- Select the response that avoids overstepping boundaries.
- Select the response that avoids irrelevant information.
- Select the response that avoids unnecessary assumptions.
- Select the response that avoids definitive religious assertions.
- Select the response that avoids unnecessary filler phrases.
- Select the response that avoids unnecessary complexity.
- Select the response that avoids accusatory language.

Cluster 6

Representative: Select the response that encourages further conversation.

Examples:

- Select the response that encourages further questions.
- Select the response that engages further in conversation.
- Select the response that encour-

ages further interaction.

- Select the response that encourages further helpful dialogue.
- Select the response that engages further conversation.
- Select the response that invites further discussion.
- Select the response that encourages further inquiry.

Cluster 7

Representative: Select the response that provides more nutritional guidance.

Examples:

- Select the response that provides health-related insights.
- Select the response that addresses cholesterol concerns directly.
- Select the response that encourages healthy habits.
- Select the response that accurately reflects caffeine content.
- Select the response that encourages further dietary considerations.
- Select the response that enhances understanding of marijuana milk.
- Select the response that provides accurate vegan definitions.

Cluster 8

Representative: Select the response that is more empathetic and respectful.

Examples:

- Select the response that shows empathy and understanding.
- Select the response that shows more empathy towards user.
- Select the response that uses more positive language.
- Select the response that shows empathy and support.
- Select the response that shows more empathy towards the user.
- Select the response that demonstrates empathy and understanding.
- Select the response that reflects personality and comfort.

Cluster 9

Representative: Select the response that maintains a supportive tone.

Examples:

- Select the response that maintains a constructive tone.
- Select the response that maintains a cautious tone.
- Select the response that maintains a neutral tone.
- Select the response that maintains a nuanced perspective.

- Select the response that maintains a supportive tone.
- Select the response that maintains a supportive tone.
- Select the response that maintains a positive tone.

Cluster 10

Representative: Select the response that provides safety considerations.

Examples:

- Select the response that emphasizes safety and preparedness.
- Select the response that advises caution with medications.
- Select the response that provides more safety details.
- Select the response that prioritizes user safety.
- Select the response that provides practical safety advice.
- Select the response that emphasizes safety in treatment methods.
- Select the response that provides clear safety guidelines.

Cluster 11

Representative: Select the response that provides specific examples.

Examples:

- Select the response that offers more relevant examples.
- Select the response that provides a specific example.
- Select the response that provides specific podcast examples.
- Select the response that provides specific examples of solutions.
- Select the response that provides more examples.
- Select the response that provides specific examples.
- Select the response that offers practical examples.

Cluster 12

Representative: Select the response that is more informative and engaging.

Examples:

- Select the response that offers more informative content.
- Select the response that provides more relevant milestones.
- Select the response that is more confident in advice.
- Select the response that is more informative and helpful.
- Select the response that is more engaging and playful.
- Select the response that is more direct and informative.
- Select the response that is more informative overall.

Cluster 13

Representative: Select the response that provides more detailed information.

Examples:

- Select the response that provides more detailed instructions.
- Select the response that is more informative and detailed.
- Select the response that provides more detailed guidance.
- Select the response that provides more relevant context.
- Select the response that provides more detailed information.
- Select the response that provides additional helpful details.
- Select the response that provides more context.

Cluster 14

Representative: Select the response that is more concise and clear.

Examples:

- Select the response that is more concise and relevant.
- Select the response that is clearer and more concise.
- Select the response that is more concise and informative.
- Select the response that is concise and relevant.
- Select the response that is more polite and concise.
- Select the response that is more concise and relevant.
- Select the response that is clearer and more concise.

Cluster 15

Representative: Select the response that promotes ethical considerations.

Examples:

- Select the response that explains legal considerations clearly.
- Select the response that is more ethical and responsible.
- Select the response that addresses ethical implications clearly.
- Select the response that lacks ethical considerations.
- Select the response that respects legal privacy concerns.
- Select the response that complies with legal guidelines.
- Select the response that prioritizes animal welfare.

Cluster 16

Representative: Select the response that provides accurate information.

Examples:

- Select the response that provides relevant information.
- Select the response that offers relevant information.
- Select the response that offers a

definitive timeline.

- Select the response that offers relevant information and insights.
- Select the response that provides a complete count.
- Select the response that includes essential care details.
- Select the response that uses precise medical terminology.

Cluster 17

Representative: Select the response that acknowledges user concerns clearly.

Examples:

- Select the response that acknowledges systemic issues.
- Select the response that acknowledges user correction effectively.
- Select the response that reassures the user effectively.
- Select the response that acknowledges the severity of actions.
- Select the response that acknowledges limitations of the test.
- Select the response that acknowledges user needs.
- Select the response that acknowledges incomplete information.

Cluster 18

Representative: Select the response that avoids encouraging harmful actions.

Examples:

- Select the response that avoids reinforcing harmful stereotypes.
- Select the response that avoids encouraging illegal activities.
- Select the response that avoids discouraging the user.
- Select the response that avoids encouraging criminal behavior.
- Select the response that avoids endorsing harmful actions.
- Select the response that suggests fun alternatives to violence.
- Select the response that avoids validating harmful behavior.

Cluster 19

Representative: Select the response that promotes a constructive dialogue.

Examples:

- Select the response that promotes constructive dialogue.
- Select the response that promotes a non-violent perspective.
- Select the response that maintains respectful dialogue.
- Select the response that encourages a supportive environment.
- Select the response that engages in a constructive dialogue.
- Select the response that encourages positive family relationships.

- Select the response that encourages positive, harmless interactions.

Cluster 20

Representative: Select the response that encourages positive action.

Examples:

- Select the response that encourages deeper reflection.
- Select the response that encourages DIY efforts.
- Select the response that encourages a positive mindset.
- Select the response that encourages reflection on motivations.
- Select the response that encourages seeking professional advice.
- Select the response that expresses enthusiasm for the event.
- Select the response that encourages regular practice.

Cluster 21

Representative: Select the response that provides clearer guidance.

Examples:

- Select the response that provides clearer guidance on VPNs.
- Select the response that provides clearer storage advice.
- Select the response that provides clear and direct advice.
- Select the response that provides clearer instructions.
- Select the response that provides clear cleaning instructions.
- Select the response that provides clearer guidance.
- Select the response that provides clearer instructions.

Cluster 22

Representative: Select the response that encourages user engagement.

Examples:

- Select the response that directly engages with user's belief.
- Select the response that attempts to guide the user.
- Select the response that engages in user-specific needs.
- Select the response that engages constructively with the user.
- Select the response that maintains engagement with the user.
- Select the response that offers more engaging interaction.
- Select the response that maintains user engagement.

Cluster 23

Representative: Select the response that directly addresses user's concern.

Examples:

- Select the response that better addresses user correction.
- Select the response that directly addresses user concerns.
- Select the response that directly addresses user concerns.
- Select the response that directly addresses user concerns.
- Select the response that directly addresses user's intention.
- Select the response that addresses user concerns directly.
- Select the response that directly addresses user follow-up questions.

Cluster 24

Representative: Select the response that seeks clarification effectively.

Examples:

- Select the response that seeks clarity on the situation.
- Select the response that seeks clarification on intentions.
- Select the response that asks clarifying questions.
- Select the response that clarifies rather than confronts.
- Select the response that seeks clarification for better understanding.
- Select the response that encourages clarification and understanding.
- Select the response that encourages user clarification.

Cluster 25

Representative: Select the response that provides accurate historical context.

Examples:

- Select the response that provides accurate historical details.
- Select the response that provides detailed historical context.
- Select the response that provides specific historical details.
- Select the response that offers historical context.
- Select the response that addresses historical context.
- Select the response that provides accurate historical context.
- Select the response that provides historical context.

Cluster 26

Representative: Select the response that is relevant to the topic.

Examples:

- Select the response that includes a professional perspective.

- Select the response that enhances educational value.
- Select the response that provides useful contextual questions.
- Select the response that emphasizes human endurance effectively.
- Select the response that addresses unique circumstances effectively.
- Select the response that focuses on anonymity benefits.
- Select the response that acknowledges multiple influencing factors.

Cluster 27

Representative: Select the response that provides a detailed recipe.

Examples:

- Select the response that includes more ingredient options.
- Select the response that enhances clarity of the recipe.
- Select the response that offers practical breakfast options.
- Select the response that provides a complete recipe.
- Select the response that asks for specific ingredients.
- Select the response that provides tailored cooking guidance.
- Select the response that provides a clear recipe.

B.2 Chatbot Arena datasets

Cluster 0

Representative: Select the response that correctly answers the question.

Examples:

- Select the response that correctly identifies the heavier option.
- Select the response that is factually accurate.
- Select the response that explains the phrase's meaning.
- Select the response that correctly identifies sibling counts.
- Select the response that explains variations in color.
- Select the response that lists only three-syllable words.
- Select the response that provides specific biblical references.

Cluster 1

Representative: Select the response that provides clearer explanations.

Examples:

- Select the response that requests clarification for accuracy.
- Select the response that provides clearer mathematical context.
- Select the response that provides a clear explanation.
- Select the response that provides clearer instructions.
- Select the response that provides clear information.
- Select the response that presents a clearer final answer.
- Select the response that provides clear definitions.

Cluster 2

Representative: Select the response that is more concise and clear.

Examples:

- Select the response that provides a concise answer.
- Select the response that uses simpler language effectively.
- Select the response that is concise and clear.
- Select the response that is more concise and clear.
- Select the response that is clear and concise.
- Select the response that is more concise and direct.
- Select the response that is more concise and relevant.

Cluster 3

Representative: Select the response that provides specific relevant information.

Examples:

- Select the response that addresses dietary preferences.
- Select the response that provides specific new features.
- Select the response that offers clear backward compatibility details.
- Select the response that provides specific tax thresholds.
- Select the response that includes relevant quantitative details.
- Select the response that includes more specific resources.
- Select the response that addresses current health guidelines.

Cluster 4

Representative: Select the response that provides more detailed information.

Examples:

- Select the response that provides more detailed explanations.
- Select the response that provides more detailed explanations.
- Select the response that includes more detailed explanations.
- Select the response that provides more detailed information.
- Select the response that provides more detailed information.
- Select the response that provides more helpful information.
- Select the response that offers more detailed explanation.

Cluster 5

Representative: Select the response that offers practical solutions and advice.

Examples:

- Select the response that offers practical advice and context.
- Select the response that offers actionable advice.
- Select the response that emphasizes practical experience.
- Select the response that includes practical travel advice.
- Select the response that offers practical, actionable advice.
- Select the response that suggests counseling for decision-making.
- Select the response that encourages consulting a healthcare professional.

Cluster 6

Representative: Select the response that is more imaginative and engaging.

Examples:

- Select the response that encourages thoughtful interpretation.
- Select the response that provides a vivid description.
- Select the response that provides vivid imagery and emotion.
- Select the response that uses more vivid imagery.
- Select the response that provides a vivid description.
- Select the response that provides a creative narrative.
- Select the response that engages with imaginative storytelling.

Cluster 7

Representative: Select the response that provides a clear explanation.

Examples:

- Select the response that provides a clear answer.
- Select the response that clearly distinguishes tube functions.
- Select the response that clearly outlines pros and cons.
- Select the response that clearly explains concepts.
- Select the response that clearly labels project phases.
- Select the response that clearly explains each technology.
- Select the response that clarifies Big Chill concept.

Cluster 8

Representative: Select the response that provides clear answers.

Examples:

- Select the response that emphasizes legal importance of NDA.
- Select the response that acknowledges subjectivity in opinions.
- Select the response that emphasizes ethical guidelines clearly.
- Select the response that directly highlights competitive advantage.
- Select the response that effectively communicates decision-making benefits.
- Select the response that emphasizes customer service skills.
- Select the response that acknowledges uncertainty in details.

Cluster 9

Representative: Select the response that avoids unnecessary explanation.

Examples:

- Select the response that avoids rhyming entirely.
- Select the response that avoids unnecessary information.
- Select the response that avoids incorrect numerical interpretation.
- Select the response that avoids misleading terminology.
- Select the response that avoids overwriting existing 'id' keys.
- Select the response that avoids excessive detail.
- Select the response that avoids negative connotations.

Cluster 10

Representative: Select the response that provides correct calculations.

Examples:

- Select the response that includes accurate speed information.
- Select the response that accurately represents the number.
- Select the response that provides complete calculations.
- Select the response that includes accurate final results.
- Select the response that provides a complete calculation.
- Select the response that includes relevant conversion factors.
- Select the response that correctly calculates the slope.

Cluster 11

Representative: Select the response that provides a complete answer.

Examples:

- Select the response that provides a full translation.
- Select the response that provides comprehensive exercise information.
- Select the response that addresses ethical concerns comprehensively.
- Select the response that offers comprehensive language insights.
- Select the response that provides a complete solution.
- Select the response that directly addresses all terms mentioned.
- Select the response that provides a complete explanation.

Cluster 12

Representative: Select the response that maintains a positive tone.

Examples:

- Select the response that engages with the user's tone.
- Select the response that acknowledges user expression playfully.
- Select the response that maintains a positive tone.
- Select the response that maintains a positive tone.
- Select the response that maintains a neutral tone.
- Select the response that emphasizes emotional acceptance.
- Select the response that encourages respectful dialogue.

Cluster 13

Representative: Select the response that directly answers the question.

Examples:

- Select the response that directly answers the user's question.
- Select the response that directly states the correct answer.
- Select the response that directly answers the prompt.
- Select the response that directly engages with user query.
- Select the response that directly answers the user's question.
- Select the response that directly addresses user's request.
- Select the response that directly addresses the prompt.

Cluster 14

Representative: Select the response that provides clearer structure and organization.

Examples:

- Select the response that is better organized and structured.
- Select the response that is more structured and organized.
- Select the response that provides a structured comparison.
- Select the response that provides a clear structure.
- Select the response that provides a structured guide.
- Select the response that offers structured reasoning.
- Select the response that is more structured and clear.

Cluster 15

Representative: Select the response that maintains clarity and coherence.

Examples:

- Select the response that maintains a cohesive structure.

- Select the response that maintains relevance to workplace conduct.
- Select the response that encourages consistency over time.
- Select the response that maintains rhythmic structure.
- Select the response that maintains clarity and coherence.
- Select the response that maintains a clear structure.
- Select the response that maintains focus on the prompt.

Cluster 16

Representative: Select the response that encourages user engagement.

Examples:

- Select the response that creates a more engaging tone.
- Select the response that invites further questions from user.
- Select the response that invites user engagement and questions.
- Select the response that invites further questions and engagement.
- Select the response that maintains user engagement better.
- Select the response that invites further engagement.
- Select the response that invites further user engagement.

Cluster 17

Representative: Select the response that provides detailed step-by-step instructions.

Examples:

- Select the response that provides clearer step-by-step guidance.
- Select the response that clearly explains each step.
- Select the response that provides detailed baking instructions.
- Select the response that provides clearer step-by-step guidance.
- Select the response that provides clear actionable steps.
- Select the response that provides specific learning steps.
- Select the response that provides detailed installation steps.

Cluster 18

Representative: Select the response that provides detailed analysis.

Examples:

- Select the response that provides a detailed analysis.
- Select the response that provides detailed player analysis.
- Select the response that offers a detailed conclusion.
- Select the response that provides detailed cultural insights.

- Select the response that includes word breakdown details.
- Select the response that provides detailed comparisons.
- Select the response that provides a detailed solution.

Cluster 19

Representative: Select the response that provides relevant context.

Examples:

- Select the response that acknowledges context and intent.
- Select the response that includes historical context.
- Select the response that provides factual historical information.
- Select the response that considers context and implications.
- Select the response that offers relevant contextual information.
- Select the response that provides accurate historical context.
- Select the response that offers context for usage.

Cluster 20

Representative: Select the response that provides specific examples.

Examples:

- Select the response that offers specific examples of enthusiasm.
- Select the response that includes specific exercise examples.
- Select the response that contains more relevant examples.
- Select the response that provides a clear example.
- Select the response that provides detailed examples.
- Select the response that offers clear examples.
- Select the response that provides specific examples and details.

Cluster 21

Representative: Select the response that showcases character development clearly.

Examples:

- Select the response that better captures Bob's character.
- Select the response that develops character-driven narratives.
- Select the response that explores complex social themes.
- Select the response that explains the song's themes.
- Select the response that provides deeper psychological analysis.
- Select the response that develops character emotions more clearly.
- Select the response that includes specific plot details.

Cluster 22

Representative: Select the response that provides detailed explanations.

Examples:

- Select the response that provides a detailed explanation.
- Select the response that includes detailed explanations.
- Select the response that includes detailed explanations.
- Select the response that offers detailed context and explanation.
- Select the response that provides a thorough explanation.
- Select the response that provides a detailed explanation.
- Select the response that provides detailed context and explanations.

Cluster 23

Representative: Select the response that provides a variety of options.

Examples:

- Select the response that includes a wider variety of practices.
- Select the response that includes multiple player examples.
- Select the response that offers more diverse suggestions.
- Select the response that connects to other fruits.
- Select the response that acknowledges different playing styles.
- Select the response that presents multiple perspectives.
- Select the response that offers a wider variety of bands.

Cluster 24

Representative: Select the response that provides complete code examples.

Examples:

- Select the response that uses simpler code structure.
- Select the response that provides complete code example.
- Select the response that provides complete code example.
- Select the response that explains the code clearly.
- Select the response that provides a code example.
- Select the response that includes relevant code examples.
- Select the response that provides complete code example.

Cluster 25

Representative: Select the response that is more informative and comprehensive.

Examples:

- Select the response that is more comprehensive and detailed.
- Select the response that is informative and detailed.

- Select the response that is informative and detailed.
- Select the response that is more informative and relevant.
- Select the response that is clearer and more comprehensive.
- Select the response that is more informative overall.
- Select the response that is more comprehensive in scope.

Cluster 26

Representative: Select the response that better meets the prompt's requirements.

Examples:

- Select the response that follows prompt instructions accurately.
- Select the response that addresses the prompt accurately.
- Select the response that includes customizable parameters.
- Select the response that uses relevant data effectively.
- Select the response that better fulfills the user request.
- Select the response that is relevant to the prompt.
- Select the response that addresses all aspects of prompt.

C

Additional Figures

This appendix contains supplementary tables and graphs supporting the analysis described in Chapter 5 and providing additional detail.

C.1 Principle Committee Evaluation

The metrics prepended by C-W refer to the confidence-weighted metrics formulated in §4.2. The abbreviation P-W GC refers to Penalty-Weighted Greedy Coverage, detailed in §3.4.1.

C.1.1 Summary Table of Metrics Across Datasets

Preference Dataset	**Anthropic HH**			**Chatbot Arena**		
Metric	**Greedy EJR**	**P-W GC**	**Seq PAV**	**Greedy EJR**	**P-W GC**	**Seq PAV**
Concurrence	0.349	0.187	0.391	0.434	0.252	0.521
C-W Concurrence	0.315	0.171	0.350	0.405	0.238	0.487
Uncertainty	0.422	0.699	0.350	0.290	0.573	0.112
C-W Uncertainty	0.396	0.686	0.325	0.274	0.563	0.100
Dissent	0.229	0.114	0.259	0.276	0.174	0.368
C-W Dissent	0.205	0.101	0.230	0.255	0.164	0.342
Polarization	0.229	0.114	0.259	0.262	0.160	0.336
Consensus Rate	0.355	0.545	0.332	0.198	0.208	0.104
Conflict Rate	0.571	0.316	0.610	0.794	0.783	0.884
Indifference Rate	0.074	0.139	0.058	0.008	0.009	0.012

Table C.1: Summary statistics for evaluating the effectuality of the principle committee constructed with each algorithm across the Anthropic HH and Chatbot Arena preference datasets.

C.1.2 Principle Committees Given Anthropic HH Preference Dataset

Anthropic HH: Greedy EJR Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.322, 0.331, 0.438, 0.4, 0.265, 0.384, 0.352, 0.301]	0.349
C-W Concurrence	[0.294, 0.313, 0.376, 0.371, 0.248, 0.327, 0.316, 0.276]	0.315
Uncertainty	[0.486, 0.507, 0.324, 0.356, 0.579, 0.303, 0.345, 0.472]	0.422
C-W Uncertainty	[0.463, 0.485, 0.293, 0.333, 0.561, 0.278, 0.315, 0.443]	0.396
Dissent	[0.192, 0.162, 0.238, 0.244, 0.156, 0.313, 0.303, 0.227]	0.229
C-W Dissent	[0.169, 0.151, 0.194, 0.219, 0.142, 0.275, 0.28, 0.206]	0.205
Polarization	[0.192, 0.162, 0.238, 0.244, 0.156, 0.313, 0.303, 0.227]	0.229
Metrics Across Samples	**Rate**	
Consensus	0.355	
Conflict	0.571	
Indifference	0.074	

Table C.2: Effectuality of the principle committee constructed with the Greedy EJR algorithm using the Anthropic HH preference dataset.

Anthropic HH: Penalty-Weighted Greedy Coverage Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.438, 0.405, 0.039, 0.054, 0.18, 0.067, 0.173, 0.142]	0.187
C-W Concurrence	[0.376, 0.39, 0.038, 0.052, 0.166, 0.062, 0.148, 0.134]	0.171
Uncertainty	[0.324, 0.321, 0.947, 0.925, 0.711, 0.888, 0.702, 0.775]	0.699
C-W Uncertainty	[0.293, 0.308, 0.945, 0.922, 0.692, 0.882, 0.686, 0.763]	0.686
Dissent	[0.238, 0.274, 0.014, 0.021, 0.109, 0.045, 0.125, 0.083]	0.114
C-W Dissent	[0.194, 0.258, 0.013, 0.02, 0.099, 0.042, 0.107, 0.078]	0.101
Polarization	[0.238, 0.274, 0.014, 0.021, 0.109, 0.045, 0.125, 0.083]	0.114
Metrics Across Samples	**Rate**	
Consensus	0.545	
Conflict	0.316	
Indifference	0.139	

Table C.3: Effectuality of the principle committee constructed with the Penalty-Weighted Greedy Coverage algorithm using the Anthropic HH preference dataset.

Anthropic HH: Sequential PAV Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.438, 0.4, 0.384, 0.352, 0.405, 0.453, 0.3, 0.395]	0.391
C-W Concurrence	[0.376, 0.371, 0.327, 0.316, 0.39, 0.396, 0.266, 0.356]	0.350
Uncertainty	[0.324, 0.356, 0.303, 0.345, 0.321, 0.199, 0.576, 0.379]	0.350
C-W Uncertainty	[0.293, 0.333, 0.278, 0.315, 0.308, 0.181, 0.545, 0.344]	0.325
Dissent	[0.238, 0.244, 0.313, 0.303, 0.274, 0.348, 0.124, 0.226]	0.259
C-W Dissent	[0.194, 0.219, 0.275, 0.28, 0.258, 0.309, 0.099, 0.206]	0.230
Polarization	[0.238, 0.244, 0.313, 0.303, 0.274, 0.348, 0.124, 0.226]	0.259
Metrics Across Samples	**Rate**	
Consensus	0.332	
Conflict	0.610	
Indifference	0.058	

Table C.4: Effectuality of the principle committee constructed with the Sequential PAV algorithm using the Anthropic HH preference dataset.

C.1.3 Principle Committees Given Chatbot Arena Preference Dataset

Chatbot Arena: Greedy EJR Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.287, 0.575, 0.434, 0.55, 0.608, 0.264, 0.225, 0.529]	0.434
C-W Concurrence	[0.249, 0.534, 0.407, 0.512, 0.589, 0.246, 0.208, 0.496]	0.405
Uncertainty	[0.565, 0.119, 0.019, 0.161, 0.051, 0.596, 0.634, 0.172]	0.290
C-W Uncertainty	[0.533, 0.106, 0.017, 0.148, 0.046, 0.574, 0.613, 0.154]	0.274
Dissent	[0.148, 0.306, 0.547, 0.289, 0.341, 0.14, 0.141, 0.299]	0.276
C-W Dissent	[0.127, 0.277, 0.522, 0.26, 0.328, 0.127, 0.129, 0.273]	0.255
Polarization	[0.148, 0.306, 0.434, 0.289, 0.341, 0.14, 0.141, 0.299]	0.262
Metrics Across Samples	**Rate**	
Consensus	0.198	
Conflict	0.794	
Indifference	0.008	

Table C.5: Effectuality of the principle committee constructed with the Greedy EJR algorithm using the Chatbot Arena preference dataset.

Chatbot Arena: Penalty-Weighted Greedy Coverage Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.608, 0.434, 0.064, 0.112, 0.079, 0.28, 0.218, 0.225]	0.252
C-W Concurrence	[0.589, 0.407, 0.055, 0.105, 0.075, 0.261, 0.205, 0.208]	0.238
Uncertainty	[0.051, 0.019, 0.902, 0.839, 0.87, 0.607, 0.664, 0.634]	0.573
C-W Uncertainty	[0.046, 0.017, 0.896, 0.834, 0.861, 0.583, 0.654, 0.613]	0.563
Dissent	[0.341, 0.547, 0.034, 0.049, 0.051, 0.113, 0.118, 0.141]	0.174
C-W Dissent	[0.328, 0.522, 0.027, 0.044, 0.045, 0.105, 0.11, 0.129]	0.164
Polarization	[0.341, 0.434, 0.034, 0.049, 0.051, 0.113, 0.118, 0.141]	0.160
Metrics Across Samples	**Rate**	
Consensus	0.208	
Conflict	0.783	
Indifference	0.009	

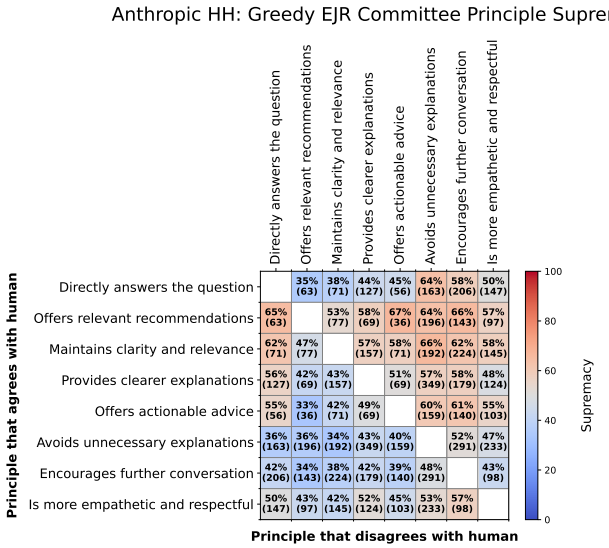
Table C.6: Effectuality of the principle committee constructed with the Penalty-Weighted Greedy Coverage algorithm using the Chatbot Arena preference dataset.

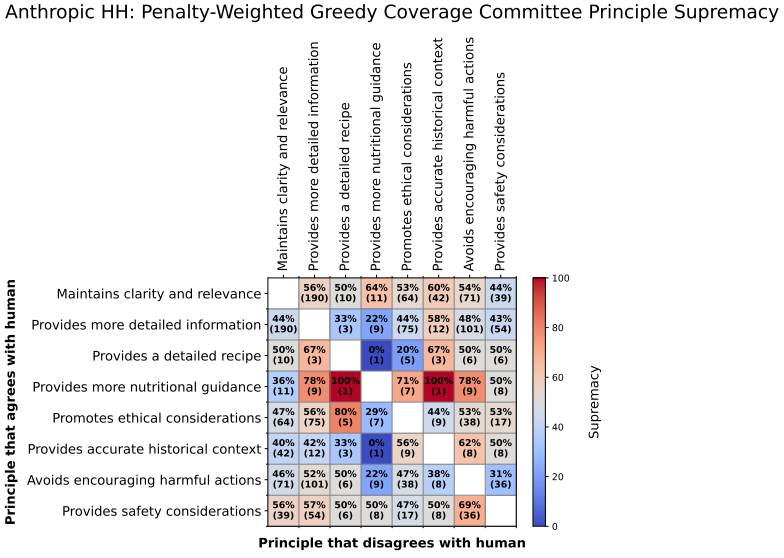
Chatbot Arena: Sequential PAV Principle Committee		
Metrics for Principles	**Value for Each Principle in Committee**	**Avg**
Concurrence	[0.434, 0.608, 0.407, 0.58, 0.56, 0.517, 0.566, 0.494]	0.521
C-W Concurrence	[0.407, 0.589, 0.358, 0.545, 0.539, 0.486, 0.543, 0.426]	0.487
Uncertainty	[0.019, 0.051, 0.048, 0.098, 0.13, 0.211, 0.115, 0.221]	0.112
C-W Uncertainty	[0.017, 0.046, 0.043, 0.087, 0.121, 0.191, 0.107, 0.186]	0.100
Dissent	[0.547, 0.341, 0.545, 0.322, 0.31, 0.272, 0.319, 0.285]	0.368
C-W Dissent	[0.522, 0.328, 0.506, 0.293, 0.293, 0.251, 0.306, 0.238]	0.342
Polarization	[0.434, 0.341, 0.407, 0.322, 0.31, 0.272, 0.319, 0.285]	0.336
Metrics Across Samples	**Rate**	
Consensus	0.104	
Conflict	0.884	
Indifference	0.012	

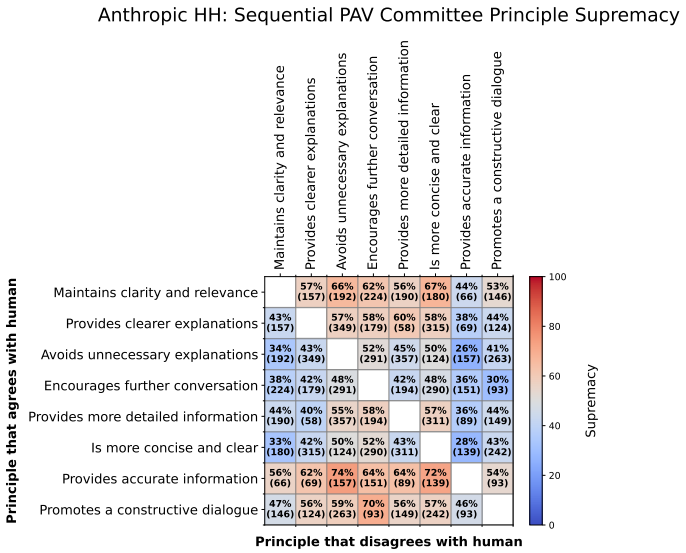
Table C.7: Effectuality of the principle committee constructed with the Sequential PAV algorithm using the Chatbot Arena preference dataset.

C.2 Principle Supremacy

C.2.1 Anthropic HH



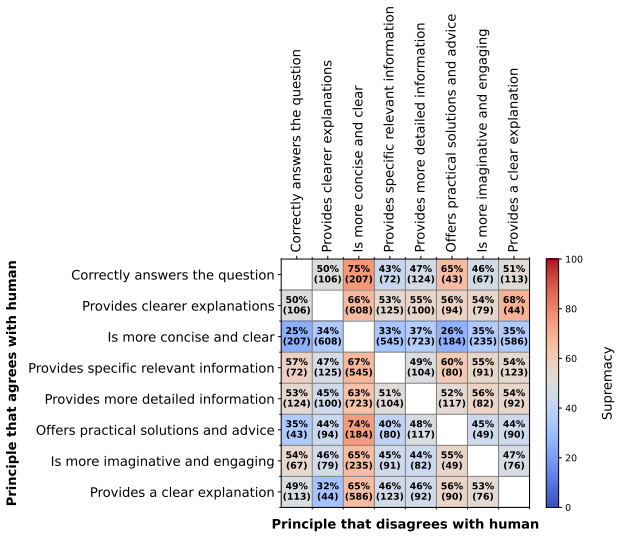




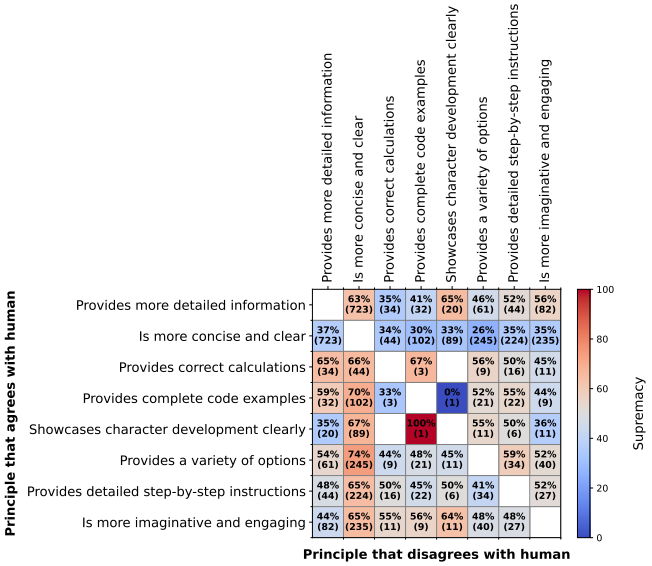
C.2.2 Chatbot Arena

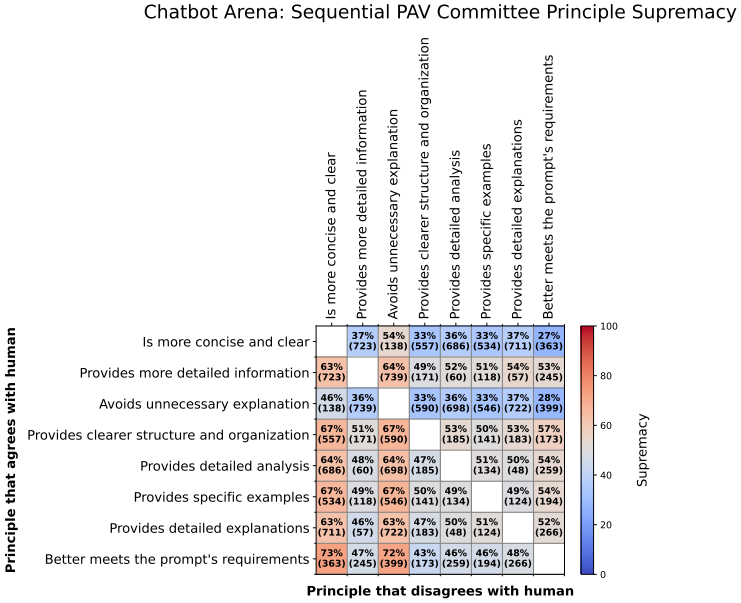
Among the final principle committees from the Arena-generated set, we observe that all committees consist of the principle recommending the response that is more concise and clear and the principle recommending the response that provides more detailed information.

Chatbot Arena: Greedy EJR Committee Principle Supremacy



Chatbot Arena: Penalty-Weighted Greedy Coverage Committee Principle Supremacy





C.3 Data from Additional Trials

C.3.1 Anthropic HH: Principle Generation

Num. Distinct Principles	2,127
↪ Num. Unique Principles	1,812
↪ Num. Distinct Duplicates	315
Num. Replicas	865
Redundancy	0.289
Total Count	2992

Table C.8: Basic statistics reflecting the redundancy of the set of generated candidate principles. A majority of generated principles are unique principles (i.e., derived from a single preference sample from the random sample of the preference dataset).

Similar to the analysis presented in §5.2.1, we make the observation that 2,992, or almost exactly (size of random sample $\times$ max num of principles generated per preference sample) $= (1{,}000 \times 3)$, principles are generated in total.

Duplicated Principles		
Num. Distinct	315	
Total	1,180	
	Pro-Principles	**Anti-Principles**
Fraction (/Num. Distinct)	0.854	0.146
Fraction (/Total)	0.860	0.140
	Top-10 Words and Frequency	
word, frequency	user, 154	unnecessary, 81
	directly, 122	harmful, 26
	clearer, 92	complexity, 25
	concise, 86	illegal, 19
	tone, 86	stereotypes, 10
	question, 81	actions, 10
	engagement, 73	behavior, 10
	information, 64	questions, 10
	detailed, 59	information, 10
	user's, 57	elaboration, 8
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	thinking, 4	instructions, 2
	clarifies, 4	advice, 2
	preferences, 4	speculation, 2
	asks, 4	sharing, 2
	steps, 4	sensitive, 2

Table C.9: Examination of duplicated pro-principles and anti-principles derived from the Anthropic HH preference dataset in a previous trial.

Unique Principles		
Total	1,812	
	Pro-Principles	**Anti-Principles**
Fraction	0.882	0.118
	Top-10 Words and Frequency	
word, frequency	user, 169	unnecessary, 32
	specific, 120	language, 17
	clearer, 79	harmful, 13
	information, 78	vague, 11
	relevant, 76	behavior, 11
	user's, 74	negative, 10
	context, 66	information, 10
	directly, 62	excessive, 9
	clear, 57	irrelevant, 9
	engages, 52	illegal, 8
	Bottom 5 of Top-100 Words and Frequency	
word, frequency	cleaning, 10	overload, 1
	alternatives, 10	comments, 1
	focuses, 9	options, 1
	conversation, 9	misinformation, 1
	language, 9	origins, 1

Table C.10: Examination of unique pro-principles and anti-principles derived from the Anthropic HH preference dataset in a previous trial.

References

[1] Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

[2] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022a). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

[3] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022b). Constitutional ai: Harmlessness from ai feedback.

[4] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022c). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs].

[5] Bereska, L. and Gavves, E. (2024). Mechanistic interpretability for ai safety – a review.

[6] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

[7] Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of computational social choice*. Cambridge University Press.

[8] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[9] Buyl, M., Khalaf, H., Verdun, C. M., Paes, L. M., Machado, C. C. V., and Calmon, F. d. P. (2025). AI Alignment at Your Discretion. arXiv:2502.10441 [cs].

[10] Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. (2024a). Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*.

[11] Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. (2024b). Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences.

[12] Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., Wattenberg, M., and Viégas, F. (2024a). Designing a dashboard for transparency and control of conversational ai.

[13] Chen, Z., Subhash, V., Havasi, M., Pan, W., and Doshi-Velez, F. (2024b). What makes a good explanation?: A harmonized view of properties of explanations.

[14] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference.

[15] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2023). Deep reinforcement learning from human preferences.

[16] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

[17] Findeis, A., Kaufmann, T., Hüllermeier, E., Albanie, S., and Mullins, R. (2024). Inverse Constitutional AI: Compressing Preferences into Principles. arXiv:2406.06560 [cs].

[18] Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., and Wüthrich, M. (2023). Generative social choice. *arXiv preprint arXiv:2309.01291*.

[19] Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., and Wüthrich, M. (2025). Generative Social Choice. arXiv:2309.01291 [cs].

[20] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Research Publication 2020-1, Berkman Klein Center for Internet Society. Available at SSRN.

[21] Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

[22] Ge, L., Halpern, D., Micha, E., Procaccia, A. D., Shapira, I., Vorobeychik, Y., and Wu, J. (2024). Axioms for ai alignment from human feedback. *arXiv preprint arXiv:2405.14758*.

[23] Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., Chung, H. W., Toyer, S., Heidecke, J., Beutel, A., and Glaese, A. (2025). Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339 [cs].

[24] Henneking, C.-L. and Beger, C. (2025). Unlocking Transparent Alignment Through Enhanced Inverse Constitutional AI for Principle Extraction. arXiv:2501.17112 [cs].

[25] Hosking, T., Blunsom, P., and Bartolo, M. (2023). Human feedback is not gold standard. *arXiv preprint arXiv:2309.16349*.

[26] Hosking, T., Blunsom, P., and Bartolo, M. (2024). Human feedback is not gold standard.

[27] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2023). Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

[28] Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. (2024). A survey of reinforcement learning from human feedback.

[29] Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. (2024). Inference-time intervention: Eliciting truthful answers from a language model.

[30] Mill, J. S. (1859). *On Liberty*. Broadview Press.

[31] OpenAI (2023). Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering. Accessed: 2025-03-28.

[32] OpenAI (2024). Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-03-28.

[33] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022a). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

[34] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022b). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

[35] Pezeshkpour, P. and Hruschka, E. (2023). Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

[36] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model.

[37] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

[38] Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.

[39] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms.

[40] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2022). Learning to summarize from human feedback.

[41] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[42] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

[43] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

[44] Wiener, N. (2021). Some moral and technical consequences of automation (1960). In *Ideas That Created the Future: Classic Papers of Computer Science*. The MIT Press.

[45] Wu, M. and Aji, A. F. (2023). Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

[46] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

[47] Zhu, B., Jiao, J., and Jordan, M. I. (2024a). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons.

[48] Zhu, B., Jiao, J., and Jordan, M. I. (2024b). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons.