Mechanism Design for Alignment via Human Feedback

Julian Manyika [*1] **Michael Wooldridge** [1] **Jiarui Gan** [1]

Abstract

Ensuring the faithfulness of human feedback is crucial for effectively aligning large language models (LLMs) using reinforcement learning from human feedback (RLHF), as low-effort or dishonest reporting can significantly undermine the quality of this feedback and, consequently, the alignment process. We address the challenge of faithfully modeling pairwise feedback by framing it as a mechanism design problem. We introduce a new principal-agent model for preference elicitation that incorporates both effort and truthfulness as key aspects of annotator strategies, and mirrors the assumptions made in reward modeling for RLHF. We then define three incentive compatibility properties that desirable mechanism frameworks should be able to satisfy: Uninformed Equilibrium Incompatibility, ω-Bayes-Nash Incentive Compatibility, and Effort Competitiveness. We propose a novel mechanism framework called Acyclic Peer Agreement (APA), which we hope to prove can satisfy all three incentive compatibility frameworks. We conclude by discussing the next steps and outlining future research directions in the design of robust mechanisms for preference elicitation.

1. Introduction

Large language models (LLMs) have allowed for an enormous advance in the capabilities of AI systems. They have been able to achieve human and superhuman performance in a wide range of tasks, from translation, summarization and dialogue (Radford & Narasimhan, 2018; Devlin et al., 2019; Brown et al., 2020), to reasoning, coding and planning (OpenAI, 2024; Touvron et al., 2023; Google, 2024), and have had a massive impact on society at large (Chen et al., 2024b; Geng et al., 2024; Choi et al., 2024; Mishra et al., 2024). Although their capabilities and applications have grown, LLMs still exhibit some undesirable behaviors: They can fail to follow instructions, produce toxic outputs, spread misinformation and perpetuate harmful stereotypes and biases (Bender et al., 2021; Weidinger et al., 2022). The goal of mitigating these behaviors and getting LLMs to do what we want them to do is often referred to as *alignment* (Gabriel, 2020). One way to approach alignment is to use human preference as the standard for human interests, or "what we want," and this approach is captured by a training method called reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022; Christiano et al., 2017). In RLHF, a reward model is trained using human feedback on LLM outputs, in the form of pairwise comparisons, in order to be a generalizable proxy for human preference, and with it the LLM is made to optimize reward through reinforcement learning. In addition to being used for training LLMs (Nakano et al., 2022; Köpf et al., 2023; Bai et al., 2022), pairwise preferences are also used as an evaluation metric (Chang et al., 2024; Chiang et al., 2024). When assessing the effectiveness of new alignment techniques or algorithms compared to previous approaches, researchers often have humans provide their preferences among outputs from models augmented by the various techniques, and report the frequency with which a given method produced outputs that were preferred over those of another method, also referred to as their "win-rate". The credibility of pairwise preferences as a gold standard in the training and evaluation of LLMs is critically dependent on how well they can represent the underlying preferences of users or humans in general. In other words, the quality of human feedback is closely tied to how faithful it is to latent human preference. In practice, crowdsourcing quality feedback is very difficult (Casper et al., 2023; Pandey et al., 2022; Chmielewski & Kucker, 2020; Buening et al., 2025) since carefully comparing passages of text is cognitively demanding and time intensive. Low-effort or dishonest reporting of preferences undermines the efficacy of human feedback as a tool for alignment.

One way to improve the faithfulness of crowdsourced preferences might be to design a reward that incentivizes annotators to exert high effort and be honest when providing feedback on model outputs. This reverse-game theoretic approach, where a reward is designed to encourage a particular

[*]Equal contribution [1]Department of Computer Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Julian Manyika <julian.manyika@jesus.ox.ac.uk>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. PMLR 267, 2025.

strategy or outcome, is called *mechanism design* (Börgers et al., 2015). In this paper, we analyze payment structures that ensure that in order to maximize their payments, annotators are better off being truthful and giving their best effort.

Contributions. We first introduce a new principal-agent model of preference elicitation that mirrors the assumptions made for reward modeling and incorporates twofold agent strategies that consist of effort and truthfulness. We then define the desirable incentive compatibility properties for a pairwise preference elicitation mechanism, specific to our model. Additionally, we propose a new mechanism framework called *Acyclic Peer Agreement*, a modified version of simple peer agreement that takes advantage of the acyclic structure of private pairwise preference.

2. Preliminaries

2.1. Human Preference Reward Modeling

The goal of aligning an LLM to human feedback is to transform a pretrained statistical model of language into a statistical model of language that is consistent with human preferences. Pretrained LLMs are large neural networks that map sequences of tokens, often words or pieces of words, to probability distributions over the next token in the sequence (Radford & Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; Kaplan et al., 2020). While next token prediction has been very effective in modeling and generating natural language, it is not sufficient for ensuring that outputs are in accordance with human preferences. Reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020; Ouyang et al., 2022; Christiano et al., 2017) is a popular paradigm that attempts to align an LLM with human feedback on its own outputs by using a reward model, a proxy of the human feedback, to better produce preferred outputs.

The purpose of a reward model in RLHF is to be a faithful representative for human preference over sequences of tokens. Since learning preferences over text benefits from natural language understanding, the reward model π_{RM} is initialized as a pretrained LLM, often a copy of the model π that will be aligned. However, instead of producing a probability distribution over tokens, it instead provides a scalar reward r such that r represents how well a sequence of tokens is aligned with human preference. This mapping from sequence to reward is learned by training π_{RM} on crowdsourced human pairwise comparisons between outputs from π, where, for a given pair of outputs a and a', human annotators are determined to have collectively preferred a over a' or a' over a. If π has been trained to be a chatbot, for example, each item $a \in \mathcal{A}$ would be a prompt and corresponding response pair (x, y) where all a have the same prompt x. Each annotator provides their reports through direct pairwise comparisons in the form $a \succ a'$ or $a' \succ a$, and these pairwise judgments from all the annotators are then reduced to a single set of pairwise preferences, meant to be a consensus representation of the group's preferences, using an aggregation rule ϕ. The reward model π_{RM} is made to fit these aggregated pairwise preferences through linear regression. Higher reward r is assigned to preferred (chosen) outputs a_c and lower reward is assigned to dispreferred (rejected) outputs a_r. Preference-based training algorithms, including those that don't use a proxy reward models (Rafailov et al., 2023; Swamy et al., 2024; Ethayarajh et al., 2024; Xu et al., 2024) or RL for training π, still rely on both the scale and quality of the collected preferences in order to align LLMs.

2.2. Modeling Crowdsourced Probabilistic Choice

Learning to align to human preferences requires a model of how humans reveal their preferences. For reward modeling in particular, there are two key presumptions required for aggregated pairwise judgments to be a suitable representation of human preference: all human preferences can be quantified and measured, and preferences of multiple humans can be adequately represented by aggregating their respective inferred utilities (Lambert et al., 2023). RLHF and most other alignment approaches adopt the Bradley-Terry model (Bradley & Terry, 1952) to infer utility from observed pairwise preferences. The model, formalized in Definition 2.1, characterizes the likelihood of a preference for a over a' being observed as being a function of how much better a is compared to a' in terms their utilities θ_a and $\theta_{a'}$.

Definition 2.1. (Bradley-Terry Model) Let $\mathcal{A}$ be a set of items and $\theta \in \mathbb{R}^{|\mathcal{A}|}$ be a vector such that θ_a is a scalar quality associated with an item $a \in \mathcal{A}$ and independently and identically drawn from a fixed, non-atomic distribution on $\mathbb{R}$. For a fixed θ, the probability of a preference for a over a' being observed is

$$\Pr(a \succ a') = \frac{e^{\theta_a}}{e^{\theta_a} + e^{\theta_{a'}}} \quad \forall a, a' \in \mathcal{A} \tag{1}$$

$\Pr(a \succ_\theta a')$ can also be written as $\sigma(\theta_a - \theta_{a'})$, where $\sigma(x)$ is the sigmoid function $(1 + e^{-x})^{-1}$.

The Bradley-Terry model can also be extended to account for noise or precision in observing a pairwise judgment. The logit quantal response function (Goeree et al., 2020; McKelvey & Palfrey, 1998) accommodates this added complexity by introducing a precision parameter ω. The introduction of precision in this choice model satisfies some key desirable properties: When items are identically valued or the observation is uninformed ($\theta_a = \theta_{a'}$ or $\omega = 0$), $\Pr(a \succ a') = \Pr(a' \succ a)$, and as effort increases ($\omega \to \infty$), the observation approaches perfect accuracy for any $\theta_a > \theta_{a'}$. In particular, when ω is made to be an

endogenous variable for a particular observer i, then this choice model, shown in Equation 2 can fit scenarios where an individual has some control over their precision (Friedman, 2019).

$$\Pr(a \succ_i a') = \frac{1}{1 + e^{-\omega_i(\theta_a - \theta_{a'})}} \quad \forall a, a' \in \mathcal{A} \qquad (2)$$

The Bradley-Terry model is a special case of a more general choice model called Bayesian Strong Stochastic Transitivity (Bayesian SST). While the Bradley-Terry model presumes the existence of underlying utility for each item, the Bayesian SST model only requires that each individual's underlying preferences are transitive: $a \succ a'$ and $a' \succ a$ implies $a \succ a''$. The Bayesian SST model is a model for inferring the information structure entailed strong stochastic transitivity (SST) (Tversky & Edward Russo, 1969; Cavagnaro & Davis-Stober, 2014; Oliveira et al., 2018; Chen et al., 2024a). Described in Definition 2.2, SST requires that for an individual's private preference over a set of choices θ, when a is preferred over a', a' is preferred over a'', then a will be *even more* preferred to a''.

Definition 2.2. (Strong Stochastic Transitivity) For all items $a, a', a'' \in \mathcal{A}$ where $\Pr(a \succ_\theta a') \geq \Pr(a' \succ_\theta a)$ and $\Pr(a' \succ_\theta a'') \geq \Pr(a'' \succ_\theta a')$, then pairwise preference entailed by θ over items $a \in \mathcal{A}$ is *strongly stochastically transitive* if $\Pr(a \succ_\theta a'') \geq \max\left\{\Pr(a \succ_\theta a'), \Pr(a' \succ_\theta a'')\right\}$.

2.3. Mechanism Design for Crowdsourcing Preferences

An important aspect of aligning to human feedback is ensuring the feedback itself is faithful. Since they align LLMs to a single reward model or directly to single set of pairwise preferences, methods like RLHF apply to instances in which humans can be reasonably expected to share the same latent preference. In these kinds of settings, one approach to increase the likelihood of faithful aggregated preferences is to increase the pool of humans giving feedback. Research teams that train LLMs with RLHF tend to utilize large online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) and companies such as Scale, Surge AI, UpWork, and Prolific source annotators in order to collect pairwise comparisons of text segments at scale.

While online marketplaces and large annotator workforces can provide scale, they cannot directly ensure the quality of the collected data. As the domains in which a requester aims to align their LLM increase in complexity, so does the mental burden on crowd workers: A task as straightforward as reporting one's preferred response between two text outputs can require varying levels of effort from the worker, depending on factors such as the length of the responses, the topic, and any additional instructions from the designer on how the worker should determine their preference. Here, effort is analogous to precision, and can only be incentivized, not controlled, by the requester. To increase the likelihood of collecting more faithful pairwise preferences, rules and structures can be put in place to ensure that certain strategies, such as exerting maximum effort and truthfulness, are in the best interest of the workers. This process is known as *mechanism design*, where a principal (the requester) aims to secure what they consider to be a good outcome from a group of agents (the crowdworkers) (Hurwicz, 1977; Börgers et al., 2015). Requesters and crowdsourcing platforms alike tend to utilize mechanisms to increase the quality of their collected data, including annotator qualification requirements[1], selecting for annotators that agree with experts or the requesters themselves on a small set of examples (Ouyang et al., 2022; Touvron et al., 2023; Bai et al., 2022), and excluding reports that do not pass a minimum standard of quality. However, these screening and pre-selection methods aren't as relevant to the core mechanism design problem of incentivizing effort and truthfulness. They are mainly aimed at ensuring annotators possess relevant expertise and settings in which quality can be externally verified, and a key aspect of RLHF is that the humans giving the feedback are the gold standard and faithfulness cannot be externally validated.

3. Related Work

3.1. Peer Prediction Mechanisms for Preference Elicitation

Our contributions revolve around mechanism design for crowdsourcing preferences that utilize peer prediction (Miller et al., 2005), a mechanism in which an agent's reward depends on the extent to which their report predicts the reports of their peers. This mechanism framework is shown to be useful for a general class of preferences, including scalar ratings and written reviews. However, since our work is oriented towards learning and modeling pairwise preferences, the most comparable literature within this area consider settings in which ordinal data is collected. Much of the existing literature on preference elicitation mechanisms focus on designing mechanisms that incorporate peer prediction in order to singularly incentivize truthfulness (Chen et al., 2024a; Easley & Ghosh, 2015; Schoenebeck & Yu, 2023; Kong, 2024) or effort (Hartline et al., 2023). Our model and mechanism design approach are therefore most similar to Zhang & Schoenebeck (2023) and Dasgupta & Ghosh (2013), both of which also consider binary preference elicitation where an agent's strategy consists of their effort and their truthfulness. Both works differ from ours in scope: Zhang & Schoenebeck (2023) generalize to accommodate

[1] https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/IntroBestPractices.html

settings where verification is possible, and the mechanism from Dasgupta & Ghosh (2013) is only compatible with a multitask setting with access to reference or expert agents. We focus squarely on preference elicitation without verification, from anonymous agents with homogeneous types, and with the added goal of optimizing the faithfulness of the aggregated data.

3.2. Mechanism Design for LLMs

Since our work is orientated towards the downstream task of collecting informative preference data for aligning AI systems, we share a similar problem space with a number of works explicitly using mechanism design for RLHF (Sun et al., 2024; Soumalias et al., 2024; Park et al., 2024; Bergemann et al., 2024; Buening et al., 2025). Sun et al. (2024) and Soumalias et al. (2024) use reward models to represent agents and frame the preference elicitation process as an auction. In Sun et al. (2024) the principal is tasked with defining a mechanism consisting of a training objective and pricing rule for groups of agents, each group represented by a reward model. An agent's utility is a linear combination of the group's expected reward on the sequences generated by the LLM and the price they are charged by the principal according to the payment rule. Similarly, Park et al. (2024) design a mechanism to mitigate strategic misreporting agents by utilizing a mechanism that rewards agents for how close their preferences are to the aggregated preferences. Soumalias et al. (2024) develop an auction mechanism for LLM text generation that incorporates advertiser preferences. In their model, the user enters a query, and the advertisers want to influence the output given to the user. The user is represented by a reference LLM, which has been trained to produce useful outputs, and each advertiser's preferences are represented by a reward model. The goal of the auction mechanism is to generate the distribution over outputs that maximizes the aggregate reward for the advertisers without significantly diverging from the reference model. Similarly Dütting et al. (2024) utilize a token auction model for LLM agents representing advertisers, where the advertisers submit bids for each generated token. Unlike our model and mechanism framework, the agents in Soumalias et al. (2024) and Dütting et al. (2024) are advertisers represented by proxy LLM reward models, and are not assumed to have homogeneous types. Notably, in the domain without external rewards, Buening et al. (2025) point out the general impossibility of designing strategyproof mechanisms with small suboptimaility gaps. Our work leverages external rewards to tame agents' incentives, as an approach to overcoming such theoretical barriers.

The preference elicitation models adopted by these works consider LLM reward as a significant factor or the sole component of agent utility, whereas our model more closely mirrors the human crowdsourcing setting where it is reasonable to assume that agents are primarily or solely concerned with their monetary payment. There is also a growing literature on mechanism design using LLMs, ranging from auctions (Dubey et al., 2024) to peer review (Lu et al., 2024). However, in contrast to using AI systems to improve mechanism design, we focus on designing mechanisms that efficiently maximize the credibility of aggregated elicited preferences for the purpose of improving the training and evaluation of AI systems that use pairwise human feedback.

4. Model

In this section, we introduce our novel principal-agent model of pairwise preference elicitation. Crowdsourcing preference elicitation for learning a single reward or utility model, as in RLHF, presupposes a shared underlying preference captured by θ which assigns utility to items, so this model applies to settings for which it is reasonable to assume that humans have access to a latent utility model.

Principal. The principal aims to maximize the faithfulness of the preferences they elicit, which is captured by the likelihood that the aggregated reported preferences match the pairwise preferences captured in θ. Each agent i is presented pairwise comparisons between items $a, a' \in \mathcal{A}$ and report a judgment $\hat{S}_i \in \{-1, 1\}$ for each pair $a, a' \in \mathcal{A}$, where $\hat{S}_i = 1$ if i reports that they prefer a over a', and $\hat{S}_i = -1$ for the opposite preference. The reports are then aggregated with an aggregation rule ϕ. There are a range of proposals for the set of axioms a preference aggregation rule in RLHF should satisfy (Dai & Fleisig, 2024; Conitzer et al., 2024). Since the reward model in RLHF is made to learn the distribution of reward from pairwise comparisons, and elicited pairwise reports from each agent are not guaranteed to be complete or transitive, we consider axioms that apply to the direct, reported comparisons between items in $\mathcal{A}$. We specifically care that each agent is treated identically (anonymity), if all agents prefer a over a' then $a \succ_\phi a'$ (unanimity), and if all possible subsets of the agents prefer a over a' then $a \succ_\phi a'$ (consistency). Since optimizing the aggregation rule is out of scope for our problem we simply let ϕ be a pairwise majority rule. The aggregated report $\hat{s}_\phi$ for an ordered pair (a, a') is 1 if the majority of agents reported a over a' and -1 otherwise. For mechanism $\mathcal{M}$ and items $\mathcal{A}$, the principal's utility is shown in Equation 3, where K is a peer agreement function $K : \{-1, 1\}^2 \to \{1, 0\}$ that takes a value of 1 if the inputs agree, and 0 otherwise:

$$U_p(\mathcal{M}, \mathcal{A} \mid \phi, \theta) \propto \sum_{a,a' \in \mathcal{A}} K(\hat{s}_\phi, s_\theta) \tag{3}$$

The majority rule ϕ satisfies anonymity, unanimity and consistency, so for a given set of preferences specified by θ, the likelihood of a particular aggregated pairwise preference $\hat{S}_\phi$

matching s_θ for a pair of items (a, a') is proportional to the likelihood that a given agent i reports $\hat{S}_i = s_\theta$.

Agents. There are N agents, where each agent i represents a crowd worker. Agent i first strategizes over the effort ω_i that they will invest in revealing their preferred item in each pairwise comparison they are given. In expending ω_i, the agent incurs cost $c(\omega_i)$, and makes an observation s_i, which takes on a value of 1 when i prefers a to a', and -1 when i prefers a' to a. Agent i additionally strategizes over their reporting function $f_i : \{-1, 1\} \mapsto \{-1, 1\}$, where f is either honesty h ($f(s_i) = s_i$), or dishonesty $\neg h$ ($f(s_i) = -s_i$). After selecting f_i, i makes their observations s_i reports $\hat{s}_i = f_i(s_i)$ for each pair of items (a, a'). The agents strategize over their effort and reporting function in order to maximize their utility U_i, a linear combination of their payment and the cost of their exerted effort.

We utilize the logit quantal response function in Equation 2 for probabilistic choice (Goeree et al., 2020; McKelvey & Palfrey, 1998; Friedman, 2019). To capture the notion that effort can improve the likelihood of an agent observing the true preference for pair (a, a') determined by their underlying utilities, we set the precision parameter ω to be endogenous effort, a non-negative real number, where $\omega = 0$ results in an uninformed strategy. Given the underlying preference $s_\theta = \text{sgn}(\theta_a - \theta_{a'})$ with respect to items a and a', the likelihood of i's preference aligning with that of θ is the Bradley-Terry model parameterized by ω_i such that $\Pr(a \succ_i a', \theta_a > \theta_{a'} | \omega_i = 0) = \frac{1}{2}$ and $\lim_{\omega \to \infty} \Pr(a \succ_i a', \theta_a > \theta_{a'} | \omega_i = \omega) = 1$. We refer to this model formally as the Bradley Terry Effort model:

Definition 4.1. (Bradley-Terry Effort Model) Let $\mathcal{A}$ be a set of items and $\theta \in \mathbb{R}^{|\mathcal{A}|}$ be a vector such that θ_a is a scalar quality associated with an item $a \in \mathcal{A}$ and independently and identically drawn from a fixed, non-atomic distribution on $\mathbb{R}$. Additionally, let $\omega \in \mathbb{R}$ be a scalar value for the effort exerted by an agent i when arriving at a pairwise preference signal $S_i \in \{-1, 1\}$, a random variable that has a value 1 if i has determined that $a \succ a'$, and -1 if i has determined that $a' \succ a$. For a fixed θ,

$$\Pr(S_i = s_\theta) = \frac{1}{1 + e^{-\omega_i(\theta_a - \theta_{a'})}} = \sigma\big(\omega_i(\theta_a - \theta_{a'})\big) \quad (4)$$

We then define the cost for an agent exerting an effort of ω to be a continuous, strictly increasing concave function $c(\omega_i)$, where $c(0) = 0$. As in Easley & Ghosh (2015); Chen et al. (2024a); Zhang & Schoenebeck (2023), we assume a homogeneous population of agents with identical private types, particularly due to the self-selecting nature of crowd work and the common use of preselection filtering. We also assume that the principal knows the cost functions of all the agents. Given the mechanism $\mathcal{M}$ determined by the principal where i is paid $M_i(\vec{s})$ based on the reported pairwise preferences $\vec{s}$ from all agents about items in $\mathcal{A}$, let i's strategy σ be the tuple (ω_i, f_i), and $\vec{\sigma}$ be a vector of all of the agents' strategies. For a strategy profile $\vec{\sigma}$, agent i's utility given $\mathcal{M}$ and θ is their payment subtracted by the cost of their exerted effort, shown in Equation 5:

$$U_i(\vec{\sigma} \mid \mathcal{M}, \theta) = M_i(\vec{s}) - c(\omega_i) \quad (5)$$

5. Incentive Compatibility Properties

A mechanism is referred to as *incentive compatible* when agents are never worse off by being truthful about their preferences. However in our model, the principal's utility is proportional to the likelihood that a given agent i reports $\hat{S}_i = s_\theta$, which is a function of effort and honesty; the more informed and truthful the agents in reporting their preferences, the more faithful the aggregated preferences are likely to be. A mechanism framework under our model are *incentive compatible* (IC) if it satisfies three particular properties:

- Uninformed Equilibrium Incompatibility (UEI)
- Symmetric ω-Bayesian-Nash Incentive Compatibility (ω-BNIC)
- Effort-Competitiveness (EC)

A mechanism is *uninformed equilibrium incompatible* if a collective lack of effort is not an equilibrium. A mechanism that encourages effort ideally ensures that when all other agents j exert $\omega_j = 0$ effort, an agent i is always better off deviating to a strategy with nonzero effort $\omega_i > 0$.

A mechanism is *symmetrically ω-BNIC* if, when every agent exerts non-zero effort (informed) at a particular effort value ω and reports there preferences honestly, no agent is better off deviating to dishonesty or a different effort level, and the expected payment for every agent is no less than that of any other symmetric equilibrium. In other words, there should exist some $\omega > 0$ such that the symmetric strategy profile where agents exert ω effort and select an honest reporting function is not only an equilibrium, but the equilibrium with the greatest reward for each agent.

Effort competitiveness simply means that if an agent increases their effort, other agents can increase their expected reward by doing so too. The formal definitions of these qualities are given below:

Definition 5.1. Let $\vec{\sigma} = (0, f)$ be the symmetric strategy profile where all agents exert zero effort, and σ_i denote agent i's effort and reporting strategy (ω_i, f_i). A mechanism $\mathcal{M}$ is *uninformed equilibrium incompatible* (UEI) if $\arg\max_{\sigma_i} \mathbb{E}_\theta[U_i(\vec{s}) \mid \vec{\sigma}_{-i},\ \sigma_i] \notin \{(0, h), (0, \neg h)\}$.

Definition 5.2. A mechanism $\mathcal{M}$ is *symmetrically ω-Bayesian-Nash incentive compatible* (ω-BNIC) if there exists $\omega > 0$ such that the symmetric strategy profile $\vec{\sigma}$ where all agents have strategy $\sigma = (\omega, h)$ is a symmetric Bayes-Nash Equilibrium, and the expected utility of each agent is no less than that of any other symmetric equilibrium.

Definition 5.3. Let $\vec{\sigma}_{-i}$ be the strategy profile of all other agents j, and $(\omega, h) \in \arg\max_{\sigma_i \sim (\omega_i, f_i)} \mathbb{E}_\theta[U_i(\vec{s}) \mid \vec{\sigma}_{-i}, \sigma_i]$. Let $\vec{\sigma}'_{-i}$ be the strategy profile of other agents resulting from a deviation by an honest agent ι, where ι changes their strategy (ω_ι, h) to a (ω'_ι, h) where $\omega'_\iota > \omega_\iota$. A mechanism $\mathcal{M}$ is *effort competitive* (EC) if there $\mathbb{E}_\theta[U_i(\vec{s}) \mid \vec{\sigma}'_{-i},\ f_i = h] - \mathbb{E}_\theta[U_i(\vec{s}) \mid \vec{\sigma}'_{-i},\ \omega_i = \omega,\ f_i = h]$ is non-negative and non-decreasing for all $\omega_i \in [\omega, \omega_\delta]$, where $\omega_\delta = \omega + (\omega'_\iota - \omega_\iota)$.

We then evaluate mechanism frameworks by determining which of the three properties they can satisfy, and for the properties they can satisfy, we identify the necessary conditions on the mechanism parameters. In order for a parameter constraint to ensure that a property is satisfied, it must be exogenous: It cannot be a function of the agent strategies.

6. Mechanism Frameworks

We consider mechanism frameworks that leverage peer-prediction and properties entailed by the Bradley-Terry effort model in Definition 4.1. Additionally, since workers are averse to high rejection rates and desire for more consistent pay and simple payment policies (Huang et al., 2023), we require that mechanisms must be simple and explainable to the agents, and the agents do not receive negative payment. Lastly, we only consider mechanisms that treat agents as anonymous and allow for a small, finite number of agent reports per comparison. In sum, we only examine peer-prediction mechanism frameworks with single-item reports of pairwise judgments, require only a minimum three agents per item comparison, and ensure limited liability for each agent.

Peer Pairwise Agreement Framework. A peer pairwise agreement (PPA) mechanism framework is a basic agreement mechanism that rewards agents for giving the same reports:

$$M_i^{PPA}(\vec{s}) = \sum_{a,a'} \sum_{j \neq i} v_{ij} \tag{6}$$

where we denote by v_{ij} the agreement between agents i and j for reports $\hat{s}_i$ and $\hat{s}_j$ with respect to a pair of items a and a', such that $v_{ij} = z$ if $\hat{s}_i = \hat{s}_j$, and $v_{ij} = 0$ otherwise.

Bayesian Strong Stochastic Transitivity Framework. The Bayesian SST mechanism, developed by Chen et al. (2024a), rewards adherence to an information structure implied by stochastic transitivity called uniform dominance:

Definition 6.1. (Uniform Dominance) Let $S_i = S(a, a')$, $S_j = S(a'', a')$, and $S_k = S(a'', a)$ be random variables such that $(S_i, S_j, S_k) \in \{-1, 1\}^3$ and $S(a, a') = 1$ if $a \succ a'$ and -1 if $a' \succ a$. S_j *uniformly dominates* S_k if $\Pr(S_j = s_i | S_i = s_i) > \Pr(S_k = s_i | S_i = s_i)$ for all $s_i \in \{-1, 1\}$.

Intuitively, uniform dominance means that if a preference $a \succ a'$ is observed, the preference $a'' \succ a'$ is more likely to be observed than $a'' \succ a$.

Where for $(\hat{s}_i, \hat{s}_j, \hat{s}_k)$ i is reporting a preference for $\hat{S}_i = \hat{S}_i(a, a')$ for (a, a'), j is reporting a preference $\hat{S}_j = \hat{S}_j(a'', a')$ for (a'', a') and k is reporting a preference $\hat{S}_k = \hat{S}_k(a'', a)$ for (a'', a). The general bonus payment $v(\hat{s}_i, \hat{s}_j, \hat{s}_k)$ from Chen et al. (2024a) is shown below, where $z > y > 0$. For simplicity of notation, we denote $\hat{s}_i$ to be i's reported preference between a and a', $\hat{s}_j$ to be j's reported preference between a'' and a' and $\hat{s}_k$ to be k's reported preference between a'' and a.

$$M_i^{BSST}(\vec{s}) = \sum_{a,a',a''} \sum_{j,k \neq i} v\big(\hat{s}_i, \hat{s}_j, \hat{s}_k\big) \tag{7}$$

$$v(\hat{s}_i, \hat{s}_j, \hat{s}_k) = \begin{cases} z & \text{if } (\hat{s}_j = \hat{s}_i) \wedge (\hat{s}_k = -\hat{s}_i) \\ y & \text{if } (\hat{s}_j = \hat{s}_i) \oplus (\hat{s}_k = -\hat{s}_i) \\ 0 & \text{otherwise} \end{cases}$$

Acyclic Peer Agreement Mechanism Framework. An acyclic peer agreement mechanism (APA) takes into account the information structure across an agent's reported preferences and their peer agreement with other agents. The motivation behind the mechanism framework is to penalize the incoherence of an agent's preferences by only allowing the agent to accrue peer agreement reward for the subset of pairwise judgments that were coherent. We refer to a set of reports from agent i as coherent if each pairwise judgment $S_i(a, a')$ is such that there does not exist a'' such that $S_i(a, a') = S_i(a', a'') = S_i(a'', a)$. In other words, coherent reports are not in cycles with one another.

The APA mechanism is shown in Equation 8, where $g(\vec{s}_i)$ is the number of coherent reports, $G(\vec{s}_i)$ is the set of coherent reported item pairs such that $|G(\vec{s}_i)| = g(\vec{s}_i)$, and v_{ij} is the peer agreement term from the PPA mechanism framework:

$$M_i^{APA}(\vec{s}) = \sum_{a,a' \in G(\vec{s}_i)} \sum_{j \neq i} v_{ij} \tag{8}$$

7. Discussion and Future Work

We present Acyclic Peer Agreement as a novel mechanism framework for preference elicitation. Our next step is to analyze each of the mechanism frameworks on the UEI, ω-BNIC and EC properties. We aim to support our intuition

that APA can uniquely satisfy all three properties, and discover if it can accommodate a more sophisticated model of preference elicitation where agents have heterogenous cost functions. We also seek to follow our theoretical analysis with experimental results.

The problem of designing mechanisms for high effort and truthful preference elicitation extends beyond just human feedback in RLHF. Mechanisms that encourage faithful feedback are needed in a world where aligned agentic AI systems are deployed on behalf of organizations and individuals. In addition to the payment structure, the principal could have more control over the ways in which pairwise preferences are elicited, or have control over what tasks the agents are given. The principal's utility could also extend beyond faithfulness, and include qualities such as budget efficiency. While we opted to focus on analyzing the incentive compatibility of mechanism frameworks, there is also an opportunity to consider the space of all possible mechanisms that map outcomes to payments, and find the one that optimizes the principal's utility. We also have an opportunity to explore a more sophisticated model in which the agents do not have homogeneous types, and the principal utility is dependent on additional or different qualities, such as fairness.

Lastly, the difficulty of ensuring truthfulness and effort from human agents, in both the real world and an idealized model, raises open questions about how to scale supervision in a way that is less reliant on direct human feedback, but remains effective in ensuring that AI systems "do what we want them to do." With a more sophisticated model of synthetic agents, our model and mechanism design approach could be extended to fit alignment processes that involve a mix of human and non-human supervision.

Impact Statement

This paper presents work whose goal is to advance the fields of Machine Learning and Mechanism Design. There are many potential societal consequences of contributions being further developed, particularly related to the payment structures for online crowdsourced annotations. However, since we only propose a method in order to later prove some theoretical guarantees, we do not feel that these consequences need to be highlighted or elaborated upon.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Bergemann, D., Bojko, M., DŸtting, P., Leme, R. P., Xu, H., and Zuo, S. Data-Driven Mechanism Design: Jointly Eliciting Preferences and Information. Cowles Foundation Discussion Papers 2418, Cowles Foundation for Research in Economics, Yale University, December 2024. URL https://ideas.repec.org/p/cwl/cwldpp/2418.html.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Buening, T. K., Gan, J., Mandal, D., and Kwiatkowska, M. Strategyproof reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2503.09561.

Börgers, T., Krähmer, D., and Strausz, R. *An Introduction to the Theory of Mechanism Design*. Oxford University Press, 07 2015. ISBN 9780199734023. doi: 10.1093/acprof:oso/9780199734023.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199734023.001.0001.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification.

Cavagnaro, D. and Davis-Stober, C. Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, 1, 04 2014. doi: 10.1037/dec0000011.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Chen, Y., Feng, S., and Yu, F.-Y. Carrot and stick: Eliciting comparison data and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=ofjTu2ktxO.

Chen, Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L. R., and Wang, W. Y. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=upAWnMgpnH. Survey Certification.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Chmielewski, M. and Kucker, S. C. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473, 2020. doi: 10.1177/1948550619875149. URL https://doi.org/10.1177/1948550619875149.

Choi, A., Akter, S. S., Singh, J., and Anastasopoulos, A. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22032–22054, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1230. URL https://aclanthology.org/2024.emnlp-main.1230/.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mosse, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Dai, J. and Fleisig, E. Mapping social choice theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL https://openreview.net/forum?id=iYthn3WATc.

Dasgupta, A. and Ghosh, A. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pp. 319–330, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488417. URL https://doi.org/10.1145/2488388.2488417.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Dubey, A., Feng, Z., Kidambi, R., Mehta, A., and Wang, D. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 713–722, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3672022. URL https://doi.org/10.1145/3637528.3672022.

Dütting, P., Mirrokni, V., Paes Leme, R., Xu, H., and Zuo, S. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 144–155, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645511. URL https://doi.org/10.1145/3589334.3645511.

Easley, D. A. and Ghosh, A. Behavioral mechanism design: Optimal crowdsourcing contracts and prospect theory. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 2015. URL https://api.semanticscholar.org/CorpusID:9551077.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Friedman, E. Endogenous quantal response equilibrium. *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*, 2019. URL https://dx.doi.org/10.2139/ssrn.2800364.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437, 09 2020. doi: 10.1007/s11023-020-09539-2.

Geng, M., Chen, C., Wu, Y., Chen, D., Wan, Y., and Zhou, P. The impact of large language models in academia: from writing to speaking, 2024. URL https://arxiv.org/abs/2409.13686.

Goeree, J. K., Holt, C. A., and Palfrey, T. R. Chapter 1 stochastic game theory for social science: a primer on quantal response equilibrium. In *Handbook of Experimental Game Theory*. Edward Elgar Publishing, Cheltenham, UK, 2020. ISBN 9781785363320. doi: 10.4337/9781785363337.00008. URL https://china.elgaronline.com/view/edcoll/9781785363320/9781785363320.00008.xml.

Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Hartline, J. D., Shan, L., Li, Y., and Wu, Y. Optimal scoring rules for multi-dimensional effort. In Neu, G. and Rosasco, L. (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2624–2650. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/hartline23a.html.

Huang, O., Fleisig, E., and Klein, D. Incorporating worker perspectives into MTurk annotation practices for NLP. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1010–1028, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.64. URL https://aclanthology.org/2023.emnlp-main.64.

Hurwicz, L. *Optimality and informational efficiency in resource allocation processes*, pp. 393–460. Cambridge University Press, 1977.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020. URL https://api.semanticscholar.org/CorpusID:210861095.

Kong, Y. Dominantly truthful peer prediction mechanisms with a finite number of tasks. *J. ACM*, 71(2), April 2024. ISSN 0004-5411. doi: 10.1145/3638239. URL https://doi.org/10.1145/3638239.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Nguyen, D. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D. A., Dantuluri, A. V., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. J. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=VSJotgbPHF.

Lambert, N., Gilbert, T. K., and Zick, T. The history and risks of reinforcement learning and human feedback, 2023. URL https://arxiv.org/abs/2310.13595.

Lu, Y., Xu, S., Zhang, Y., Kong, Y., and Schoenebeck, G. Eliciting informative text evaluations with large language models. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pp. 582–612, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673532. URL https://doi.org/10.1145/3670865.3673532.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for extensive form games. *Experimental Economics*, 1(1):9–41, 1998. doi: 10.1023/A:1009905800005.

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 09 2005. URL http://tricountycc.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/eliciting-informative-feedback-peer-prediction/

docview/213170454/se-2. Copyright - Copyright Institute for Operations Research and the Management Sciences Sep 2005; Document feature - references; equations; tables; Last updated - 2024-12-04; CODEN - MNSCDI.

Mishra, T., Sutanto, E., Rossanti, R., Pant, N., Ashraf, A., Raut, A., Uwabareze, G., Ajayi, O., and Zeeshan, B. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Scientific Reports*, 14, 12 2024. doi: 10.1038/s41598-024-81370-6.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

Oliveira, I., Zehavi, S., and Davidov, O. Stochastic transitivity: Axioms and models. *Journal of Mathematical Psychology*, 85:25–35, 2018. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2018.06.002. URL https://www.sciencedirect.com/science/article/pii/S0022249617301839.

OpenAI. Gpt-4 technical report, 2024.

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. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Pandey, R., Purohit, H., Castillo, C., and Shalin, V. L. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772, 2022. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2022.102772. URL https://www.sciencedirect.com/science/article/pii/S1071581922000015.

Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. E. RLHF from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024. URL https://openreview.net/forum?id=CaZCBbeHua.

Radford, A. and Narasimhan, K. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Schoenebeck, G. and Yu, F.-Y. Two strongly truthful mechanisms for three heterogeneous agents answering one question. *ACM Trans. Econ. Comput.*, 10(4), February 2023. ISSN 2167-8375. doi: 10.1145/3565560. URL https://doi.org/10.1145/3565560.

Soumalias, E., Curry, M., and Seuken, S. Truthful aggregation of LLMs\\ with an application to online advertising. In *Agentic Markets Workshop at ICML 2024*, 2024. URL https://openreview.net/forum?id=Pp6483Ma1m.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sun, H., Chen, Y., Wang, S., Chen, W., and Deng, X. Mechanism design for LLM fine-tuning with multiple reward models. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=kYyu2ToEg5.

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W.,

Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Tversky, A. and Edward Russo, J. Substitutability and similarity in binary choices. *Journal of Mathematical Psychology*, 6(1):1–12, 1969. ISSN 0022-2496. doi: https://doi.org/10.1016/0022-2496(69)90027-3. URL https://www.sciencedirect.com/science/article/pii/0022249669900273.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zhang, Y. and Schoenebeck, G. High-effort crowds: Limited liability via tournaments. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 3467–3477, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583334. URL https://doi.org/10.1145/3543507.3583334.