Comparing Bad Apples to Good Oranges: Aligning Large Language Models via Joint Preference Optimization

Anonymous ACL submission

Abstract

A common technique for aligning large language models (LLMs) relies on acquiring human preferences by comparing multiple generations conditioned on a fixed context. This method, however, relies solely on pairwise comparisons, where the generations are evaluated within an identical context. While effective to such conditional preferences often fail to encompass the nuanced and multidimensional nature of human preferences. In this work, we revisit the traditional paradigm of preference acquisition and propose a new axis based on eliciting preferences jointly over the instruction-response pairs. Unlike prior preference optimizations, which are designed for conditional ranking protocols (e.g., DPO), we propose Joint Preference Optimization (JPO), a new preference optimization objective that upweights the joint probability of the chosen instruction-response pair over the rejected instruction-response pair. Interestingly, LLMs trained with joint instruction-response preference data using JPO outperform LLM trained with DPO by 5.2% and 3.3% win-rate for summarization and open-ended dialogue datasets, respectively. Our findings reveal that joint preferences over instruction and response pairs can significantly enhance the alignment of LLMs by tapping into a broader spectrum of human preference elicitation.[1]

1 Introduction

Recently, alignment (Stiennon et al., 2020; Ouyang et al., 2022) has emerged as a crucial step in enhancing the performance of large language models (LLMs) (Anthropic, 2024; OpenAI, 2023; Team et al., 2023; Anthrophic, 2023; Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023) in diverse real-world applications (Li et al., 2023; Zheng et al., 2023a; Wu et al., 2023a; Clusmann et al., 2023; Lambert et al., 2024). In particular, aligned LLMs generate responses that maximize human utility along various dimensions such as helpfulness, coherence, and harmlessness (Askell et al., 2021; Ouyang et al., 2022). Here, the notion of human utility is subjective (Kirk et al., 2024; Gabriel, 2020), and mainly hinges on *how* preferences are acquired from annotators (Otto et al., 2022). Among the various preference acquisition protocols (Lightman et al., 2023; Wu et al., 2023b; Scheurer et al., 2023; Bansal et al., 2023), the ranking-based approach is the most widely used paradigm to align LLMs (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a; Tunstall et al., 2023; Teknium, 2023). Specifically, in ranking approach the annotator has to compare a pair of responses *conditioned* on a fixed context. For instance, humans can select a 'preferred' response by comparing a pair of responses for the instruction 'Create a list of four fruits other than Apple' (Figure 1 (*left*)).

Although traditional conditional rankings provide rich preferences for alignment, they fail to holistically capture the multi-faceted nature of human decision-making and preferences (Zhi-Xuan et al., 2024; Gigerenzer, 2008). Besides ranking preferences conditioned on a fixed context, humans can also express preferences in non-identical contexts. For example, while browsing reviews for products on an e-commerce website, humans are likely to prefer an accurate and detail-oriented review for a camera over an incoherent, vague movie review even though the products (camera and movie) are qualitatively different.

In this work, we revisit the traditional paradigm of conditional preference acquisition by developing a framework to acquire preferences jointly over instruction-response pairs. Starting from an instruction-response data consisting of response R_i for instruction I_i (say $i \in \{1, 2\}$), we acquire ranking-based preferences over the instruction-response pairs (I_1, R_1) and (I_2, R_2). As shown

[1]We will release the data, code, and models upon acceptance.

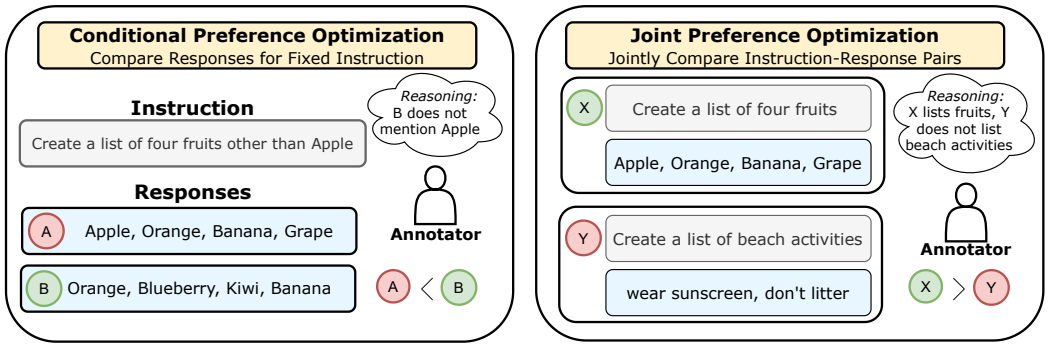


Figure 1: Overview of the Joint Preference Optimization. (*Left*) We show that the conditional preference acquisition method would require the annotators to compare two responses for an identical instruction. (*Right*) We show that the annotators can also assign rankings jointly over instruction-response pairs. Specifically, the annotator prefers a helpful response (e.g., Apple ... Grape) over a response that ignores the context of the instruction (e.g., wear sunscreen ... litter). Our framework thus elicits preferences that are obfuscated in the prior approach.

in Figure 1 (*right*), we aim to understand whether the response in the pair X is perceived better than the response in the pair Y. We hypothesize that by capturing preferences in non-identical contexts our protocol can elicit human behaviors that are obfuscated in prior protocols. First, wee show that humans can provide decisive preferences in joint preferences protocol (4.4). Then, we analyze how joint preferences differ from conditional preferences on the same dataset(§4.4).

Preference Acquisition	Algorithm	Alignment Objective	Different Instructions
Score	Ethayarajh et al. (2024)	Conditional	No
Comparison (DPO Variants)	Rafailov et al. (2024)	Conditional	No
	Park et al. (2024)	Conditional	No
	Liu et al. (2024)	Conditional	No
	Meng et al. (2024)	Conditional	No
	Hong et al. (2024)	Conditional	No
Pairwise	JPO (ours)	Joint	Yes

Table 1: JPO differs from prior works along three key aspects: preference acquisition (scoring or comparison), objective (conditional or joint distribution), and their ability to handle non-identical instruction-responses.

Prior works like DPO and its variants (Rafailov et al., 2023; Yin et al., 2024; Liu et al., 2024; Meng et al., 2024; Hong et al., 2024; Azar et al., 2023) rely on conditional rankings, and thus do not have access to the joint distribution of human preferences in the ranking protocol (Table 1). While a rating protocol (Ethayarajh et al., 2024) allows for a comparison between responses from non-identical instructions, it can be inconsistent with rankings (Bansal et al., 2023) and ignores the possibility of preferences over a pair of chosen or rejected responses. [2] Thus, we propose **Joint Preference Optimization** (JPO), a framework for aligning LLMs with our proposed joint preference elicitation scheme. Specifically, it upweights the joint probability of the chosen instruction-response pair over the rejected instruction-response pair. Furthermore, JPO subsumes prior preference optimizations as conditional rankings are a special case of joint preferences (e.g., when the instructions are identical).

We conduct experiments to explore new reasoning paths enabled by joint preference elicitation and alignment of LLMs with the JPO objective. By analyzing feedback from conditional rankings and joint preferences protocols, along with explanations from human annotators, we uncover the complexities of the preference acquisition process (§4). Using our JPO algorithm, we align a Mistral-7B LLM with the collected preferences, achieving a 30% and 18% higher win rate against gold responses on unseen instructions from summarization and dialogue datasets, respectively. JPO leverages diverse preferences effectively, outperforming DPO by 5.2% and 3.3% win-rate points on the summarization and open-ended dialogues, respectively. In addition, JPO outperforms KTO by 3.5% on the open-ended dialogues dataset. It also surpasses both DPO and KTO in the AlpacaEval2 benchmark (§5). This indicates that by utilizing the diverse preference signals present in the existing data, we can align an LLM robustly without acquiring additional instruction-response data.

2 Background

In this work, we focus on aligning language models to generate outputs preferred by humans in dimensions like helpfulness and coherence. Aligning a

[2] For instance, a pair of responses that achieves a score of 0, under the rating protocol, will result in an indecisive preference.

pretrained base model involves four steps: (a) collecting instruction-response data, (b) supervised fine-tuning (SFT), (c) acquiring preference data, and (d) deploying an alignment algorithm. The instruction-response data can be either hand-crafted by humans (Conover et al., 2023; Wang et al., 2022) or generated by AI (Taori et al., 2023; Tunstall et al., 2023). Subsequently, the base model undergoes supervised fine-tuning (SFT) on the instruction-response pairs (Zheng et al., 2023b; Wang et al., 2023c, 2022; Peng et al., 2023; Xu et al., 2023; Yin et al., 2023; Wang et al., 2023b; Yu et al., 2023; Toshniwal et al., 2024).

Following SFT, feedback data is gathered (e.g., rankings) to train the SFT model via an alignment algorithm. This often involves training a reward model on preference data (Bradley and Terry, 1952; Bansal et al., 2023) and aligning the model using Reinforcement Learning (RLHF) (Schulman et al., 2017; Ouyang et al., 2022). To address challenges in human feedback collection (Dubois et al., 2023; Zheng et al., 2023b), LLMs can provide feedback, enabling Reinforcement Learning through AI Feedback (RLAIF). Alternatively, (Rafailov et al., 2024) introduced Direct Preference Optimization (DPO) that mitigates the instability due to PPO for reward maximization by optimizing directly within the model parameter space, hence by-passing the reward modeling step.

3 Joint Preference Optimization (JPO)

A common technique for feedback data acquisition requires the annotators to assign a preferred and non-preferred label to a pair of responses for an instruction (Stiennon et al., 2020; Rafailov et al., 2023; Ouyang et al., 2022; Ethayarajh et al., 2024). However, this paradigm does not capture the complex and multidimensional aspects of human preferences (Kendall and Smith, 1940; Thurstone, 2017; Zhi-Xuan et al., 2024). Specifically, the heuristics for making preference decisions depend upon the context in which the comparison is made (Otto et al., 2022). While the traditional ranking protocol compares the two responses under a fixed context, humans can perform pairwise comparisons jointly over instruction-response pairs. For example, consider two summaries, A and B, for articles X and Y, respectively; then, a human can reason and choose the response that better summarizes its corresponding article. Hence, it is critical to align language models with diverse feedback signals to accurately model human behavior and decision making.

In our setup, the annotator has to decide a *chosen* and *rejected* instruction-response pair (I_a, R_a, I_b, R_b) where R_a and R_b are responses to the instructions I_a and I_b, respectively, and $(I_a, R_a), (I_b, R_b) \in \mathcal{D}$. We note that our joint preference setup is equivalent to the original ranking protocol when $I_a = I_b$. As before, the preference reasoning from the annotator will be based on subjective dimensions like helpfulness, coherence, and harmlessness. Formally, the annotator assigns a joint ranking feedback $h(I_a, R_a, I_b, R_b) \in \{(I_a, R_m), (I_b, R_b), \text{Equal}\}$ where 'Equal' indicates that both the instruction-response pairs are perceived equally good or bad. Finally, the joint preference optimization creates a pairwise feedback data $\mathcal{D}_H = \{(I_a, R_a, I_b, R_b, h(I_a, R_a, I_b, R_b))\}$.

Our formulation suggests that we can obtain large-scale and diverse preference data (covering all possible combinations of (I_a, R_a) and (I_b, R_b)) without the need for gathering additional instruction and response data, which is typically more difficult and costly to acquire. In addition, joint preference acquisition does not necessitate the presence of multiple responses for a given instruction that can be hard to collect for low-resource languages (e.g., Kalamang [3]). Specifically, one can collect an instruction-response data $\mathcal{D}' = \{(I_a, R_a)\}_{a=1}^{a=n}$, and acquire preferences on various combinations of instruction-response pairs. Finally, we assess the interplay between the joint feedback dataset $\mathcal{D}_H$ with the conditional feedback dataset $\mathcal{D}_C$ along with qualitative examples in §4.

We propose JPO, a preference optimization objective that learns to align the language models with preferences acquired jointly over the instruction-response pairs. We assume a joint preference dataset $\mathcal{D}_X = \{(I_i^w, R_i^w, I_j^\ell, R_j^\ell)\}$, that can be constructed from $\mathcal{D}_H$, where (I_i^w, R_i^w) and (I_j^ℓ, R_j^ℓ) are the chosen and rejected instruction-response pairs, respectively. Similar to DPO, we start with a reference model p_{ref} which is usually the supervised finetuned language model p_{sft}. Specifically, the JPO objective aims to learn an aligned model p_θ by upweighting the joint probability of preferred responses $p(R_i^w, I_i^w)$ over non-preferred responses $p(R_j^\ell, I_j^\ell)$. Formally, the optimization objective for JPO, $\mathcal{L}(\theta; \mathcal{D}_X, \beta, p_{\text{ref}})$ minimizes the expectation over $(I_j^w, R_j^w, I_j^\ell, R_j^\ell) \sim \mathcal{D}_X$:

[3] https://endangeredlanguages.com/lang/1891?hl=en

$$\mathbb{E}\left[\log\left(\sigma\left(\beta\log\frac{p_\theta(R_i^w, I_i^w)}{p_{\text{ref}}(R_i^w, I_i^w)} - \beta\log\frac{p_\theta(R_j^\ell, I_j^\ell)}{p_{\text{ref}}(R_j^\ell, I_j^\ell)}\right)\right)\right] \tag{1}$$

where σ denotes the sigmoid function and β is a hyperparameter. Further, we show that Eq. 1 reduces to the DPO formulation (Appendix Eq. 2) when the instructions $I_i = I_j$ in Appendix §F. We can also see that the JPO objective aims to learn an aligned model p_θ by upweighting the conditional probability of preferred responses $p(R_i^w|I_i^w)$ over non-preferred responses $p(R_j^\ell|I_j^\ell)$, along with a correction factor based on the prior probability of the instructions under the language model $p_\theta(I_i^w)$ and $p_\theta(I_j^\ell)$. In §5, we utilize JPO to align language models to generate human-preferred summaries and answer open-ended instructions.

4 Interplay between Feedback Protocols

4.1 Instruction-Response Acquisition

In this work, we consider two kinds of instruction-response data. First, we consider a filtered version of the TL;DR *summarization* dataset (Völske et al., 2017) from (Stiennon et al., 2020) consisting of Reddit posts, their summarizes, and human preferences over a pair of summaries for a given post. Throughout the dataset, the task is of summarization that is close-ended and well-defined for language models. Second, we consider the single-turn dialogues from the helpful-base subset of the Anthropic-HH dataset (Bai et al., 2022b). Specifically, this dataset consists of *open-ended* instructions with a collection of responses.

Both these datasets have a train and test split where each instance consists of an instruction and a pair of responses $\mathcal{D} = \{(I_i, R_i^1, R_i^2)\}_{i=1}^n$ where n is the dataset size. In this work, we collect AI and human feedback on the instruction-response data from their train split and filter instances with duplicate instructions. We can directly compare the two responses for the fixed instruction and construct a ranking feedback dataset $\mathcal{D}_C = \{(I_i, R_i^1, R_i^2, c(I_i, R_i^1, R_i^2))\}$. To acquire preferences jointly over the instruction-response pairs, we randomly select one of the responses from every instance of $\mathcal{D}$ to construct $\mathcal{D}_S = \{(I_i, R_i)\}$ where $R_i \in \{R_i^1, R_i^2\}$. Subsequently, we create the joint instruction-response pairs by matching every instance $(I_i, R_i) \in \mathcal{D}_S$ with another instance $(I_j, R_j) \in \mathcal{D}_S$ to get $\mathcal{D}_H = \{(I_i, R_i, I_j, R_j, h(I_i, R_i, I_j, R_j))\}$ of the same size as $\mathcal{D}_S$ and $\mathcal{D}_C$. In §5, we will utilize $\mathcal{D}_S$ to SFT the base model, and $\mathcal{D}_C$ and $\mathcal{D}_H$ as preference datasets for LLM alignment. We provide the dataset statistics in Appendix §E.

4.2 Feedback from AI and Humans

Dataset	Ranking	H-H	H-AI
TL;DR	Conditional	69%	63%
Anth.-Helpful		70.1%	72%
TL;DR	Joint	62%	60%
Anth.-Helpful	(Non-Identical)	68.8%	71%
Average		67.5%	66.5%

Table 2: Agreement analysis between within human annotators and gold human feedback and AI (ChatGPT) feedback. We perform the agreement calculations for the two ranking protocols: (a) conditional rankings, and (b) joint preferences where instructions are non-identical. In addition, we assess the agreement rates over the two datasets: (a) TL;DR and (b) Anthropic-helpful dataset.

Feedback from AI. We collect feedback over a pair of responses for a fixed instruction, and joint instruction-response pairs without identical instructions from GPT-3.5-Turbo-0125 (ChatGPT). We choose ChatGPT due to its affordability (e.g., output tokens from ChatGPT are $50\times$ cheaper than GPT-4). To mitigate any ordering bias, we run two queries for all comparisons. When the ChatGPT preferences flip by flipping the order of the two responses, then we consider it a tie, similar to (Bansal et al., 2023; Bitton et al., 2023). Specifically, we instruct the AI to to choose the response that is more accurate, coherent, and harmless.

To collect conditional preferences over a pair of responses for a fixed instruction, we prompt ChatGPT to choose a response. To collect AI preferences jointly over the instruction-response pairs, we prompt ChatGPT to decide the response that better answers its corresponding instruction. We collected approximately $50K$ comparisons across both feedback acquisition protocols for the summarization and Anthropic-Helpful dataset, at a cost of $100. We provide the AI prompts in Appendix §J.

Feedback from Humans. In this work, we also collect human preferences for 2000 comparisons over TL;DR and Anthropic-Helpful dataset. Specifically, we ask two annotators to assign a chosen response or chosen instruction-response pair based along the same dimensions as ChatGPT guidelines.

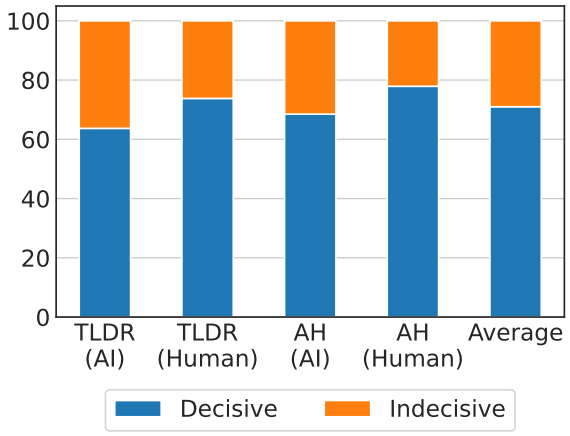


Figure 2: Results for the preferences acquired jointly over the instruction-response pairs where both the responses were either chosen or rejected under the conditional rankings protocol. Here, *decisive* implies that the annotators could assign a preference to one instruction-response pair over the other. Here, AH means Anthropic-Helpful.

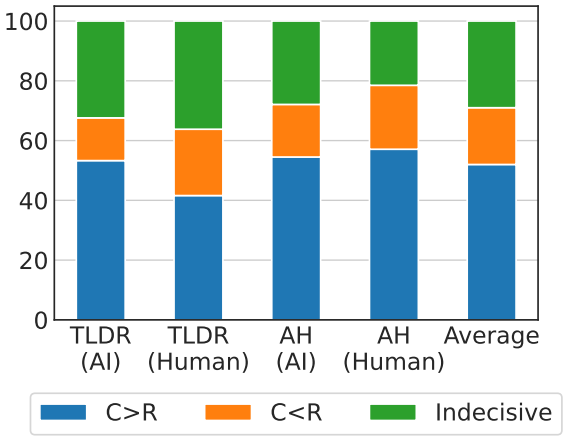


Figure 3: Results for the preferences acquired jointly over the instruction-response pairs where one of the instruction-response pair was chosen (C) and the other pair was rejected (R) under the conditional rankings. Here, $C < R$ implies that the instruction-response pair that was rejected under conditional rankings is actually preferred over an instruction-response pair that was rejected under the conditional rankings. Here, AH means Anthropic-Helpful.

Annotators can also choose 'equal' if they fail to make a identify a decisive preference. The human annotations were collected from Amazon Mechanical Turk (AMT). We recruited the participants that passed a preliminary qualification exam. In total, we spent $720 on human feedback acquisition. We provide the screenshot of the annotation UI in Appendix §K.

4.3 Agreement Analysis

We present the annotator agreement scores in Table 2. We find that the average agreement is 67.5% and 66.5% between the human-human and human-AI annotators, respectively. Specifically, we find that in the conditional setup (identical instructions), the average human-human agreement is 69.5% for the TLDR and Anthropic-Helpfulness datasets. Similarly, in the joint setup (non-identical instruction-response pairs), the average inter-rater agreement is 68% on the same datasets.These agreement scores are comparable to those reported in prior studies (Li et al., 2023; Bansal et al., 2023), demonstrating the robustness of our evaluation. Interestingly, we find that agreement scores vary depending on the underlying distribution of instruction-response pairs and the choice of ranking protocol. Overall, our results highlight that humans and AI can provide rich feedback in both conditional and joint setup with acceptable agreement.

4.4 Interplay Analysis

Setup. Here, we aim to study the interaction between the conditional rankings and joint rankings over non-identical instructions. Formally, each instruction-response pair (I_i, R_i^x) from the conditional pairwise feedback dataset $\mathcal{D}_C$ where $x \in \{1, 2\}$ can be assigned a preference $\mathcal{P}_C(I_i, R_i^x)$ among {'chosen', 'reject', 'equal'}. For instance, $\mathcal{P}_C(I_i, R_i^1) =$ 'chosen' and $\mathcal{P}_C(I_i, R_i^2) =$ 'reject' if the response R_i^2 is rejected in the dataset $\mathcal{D}_C$ i.e., $c(I_i, R_i^1, R_i^2) = R_i^1$. Similarly, we can assign a preference $\mathcal{P}_H(I_i, R_i)$ among {'chosen', 'reject', 'equal'} to an instruction-response pair (I_i, R_i) from the joint preference dataset $\mathcal{D}_H$. For instance, $\mathcal{P}_H(I_i, R_i) =$ 'chosen' and $\mathcal{P}_H(I_j, R_j) =$ 'reject' where $i! = j$ if the instruction-response pair (I_i, R_i) is chosen in the dataset $\mathcal{D}_H$ i.e., $h(I_i, R_i, I_j, R_j) = (I_i, R_i)$.

To study the interplay between the preference protocols, we assess $\mathcal{P}_C(I_i, R_i)$, $\mathcal{P}_C(I_j, R_j)$, $\mathcal{P}_H(I_i, R_i)$ and $\mathcal{P}_H(I_j, R_j)$ for all $(I_i, R_i, I_j, R_j) \in \mathcal{D}_H$. Here, if $\mathcal{P}_H(I_i, R_i) =$ 'chosen' then $\mathcal{P}_H(I_j, R_j) =$ 'reject'.

Annotators show decisiveness in joint setup. In Figure 2, we study the joint preferences over the instruction-response pairs (I_i, R_i, I_j, R_j) where the individual instruction and response data is either *chosen* or *rejected* in the conditional feed-

Table 3: Response B and D are rejected under the conditional rankings. However, in the the joint setup humans consider that the response B answers Instruction 1 better than response D answers Instruction 2. Thus, joint setup human give a decisive feedback between two responses that were rejected under the conditional rankings.

Instruction 1	My dog has been scratching a lot, but he doesn't have any fleas. What else could be causing this?
Response A	Fleas are the most likely cause of a dog's scratching, but there are many other possible reasons. It's worth taking him to the vet to check things out.
Response B	Oh no! Maybe his skin is dry. I bet if you gave him a good bath, that would help!
Conditional Preference 1	**Human Preference: A>B.** Human Explanation: Response A answered the query by giving possible causes of the problem along with a suggestion of just going to the vet.
Instruction 2	Do you have any good chili recipes?
Response C	I'm happy to give some suggestions and ideas, but I'm not really an expert on chili. What kind of chili do you want to make?
Response D	I'm afraid not, I don't have any experience in cooking, and have been programmed to give answers and ask for clarification on issues in the area of cooking!
Conditional Preference 2	**Human Preference: C>D.** Human Explanation: Response D claims they can't help and then later on in the same sentence claims they've been programmed to give answers on cooking issues, which doesn't make much sense. C asks a reasonable clarifying question.
Joint preference	**Human Preference: (Instruction 1, Response B) > (Instruction 2, Response D).** Human Explanation: Response B provides answer to the Instruction 1 while Response D does not provide an accurate answer to Instruction 2.

back protocol (e.g., $\mathcal{P}_C(I_z, R_z)$ = 'chosen' for $z \in \{i, j\}$). Interestingly, we find that the annotators can assign a decisive preference (e.g., $(I_i, R_i) > (I_j, R_j)$) in 71% of the joint comparisons. While we observe that the annotators assign a 'tie' to 29% of the comparisons. This highlights the existence of valid preference decisions that remained obfuscated in the traditional approach for ranking-based feedback acquisition.

Annotator preferences depend on context and comparisons. In Figure 3, we study the joint preference over the instruction-response pairs (I_i, R_i, I_j, R_j) where one of them is *chosen* and the other is *rejected* in the conditional feedback protocol (e.g., $\mathcal{P}_C(I_i, R_i)$ = 'chosen' and $\mathcal{P}_C(I_j, R_j)$ = 'reject'). To our surprise, we find that the annotators do not prefer the instruction-response pair that was chosen under the conditional feedback protocol in 48% of the comparisons. Specifically, there are 19% of the comparisons where rejected pair (R) is preferred over the chosen pair (C) and 28% of the comparisons where the annotators considered the pair equally good or bad. This highlights that both human and AI annotators' perceptions of preferred and non-preferred data depends on the context of the comparisons, indicating that feedback acquisition is a multifaceted phenomenon.

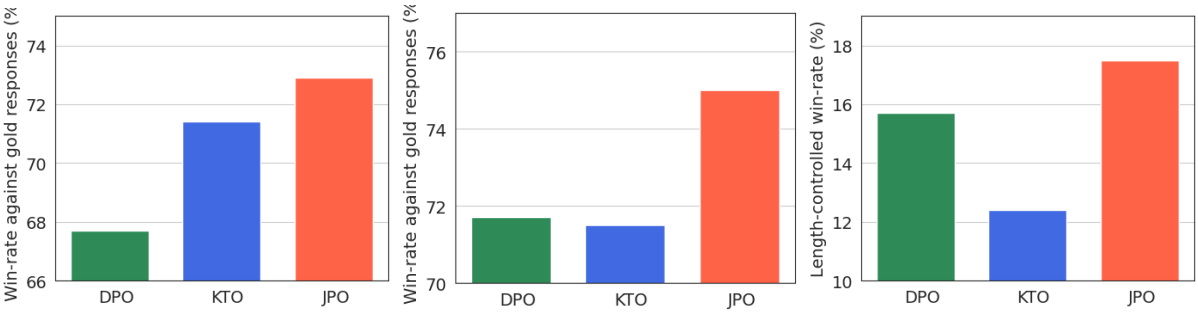
Qualitative Case Study. To understand the heuristics used in preference annotations, we asked human annotators to provide brief explanations for their feedback decisions in both conditional and joint preference setups. In Table 3, we observe that humans provide reasonable explanations for rejecting responses B and D in the conditional setup. However, when these same rejected responses are presented in a joint setup, humans offer decisive feedback, basing their decisions on the accuracy of the responses—an aspect not emphasized in the explanations for the conditional preferences. We present additional qualitative examples in Appendix §H to showcase the multi-faceted nature of human feedback revealed through joint preferences.

5 LLM Alignment with JPO

In previous sections, we explore how we collect ranking-based feedback for a pair of responses for identical and non-identical instructions. Here, we study how to leverage joint and conditional feedback data to align large language models effectively with JPO §3.

5.1 Setup

We align Mistral-7B (Jiang et al., 2023), a strong base LLM for its model capacity. We experiment with two datasets that exhibit diverse characteristics: (a) TL;DR dataset where the instruction is to summarize Reddit posts, and (b) open-ended dialogues from Anthropic-Helpful dataset (§4.1). In particular, we collect a conditional preference data $\mathcal{D}_C$ and joint preference data for non-identical instructions $\mathcal{D}_H$ of similar data sizes from ChatGPT.



(a) Performance on TL;DR (b) Performance on Anthropic-Helpful (c) Performance on AlpacaEval2.0

Figure 4: Results for aligning LLMs with JPO. We utilize ChatGPT to compare the model responses with the gold responses. In 4a and 4b we report the results averaged over three runs of the preference optimization objectives and three sampling temperatures. In 4c, we report the results for temeperature set at 0.7 for AlpacaEval2.

Then, we convert the conditional preference data into an instruction-response data for supervised finetuning $\mathcal{D}_{\text{SFT}}$.

We supervise finetune the entire base LLM model parameters with the SFT dataset to ensure that the preference data is in-policy for the alignment algorithms (Rafailov et al., 2023). JPO algorithm can utilize both the conditional preferences and joint preference with non-identical context. [4] Thus, we train the base LLM with JPO algorithm after merging conditional and joint preferences data $\mathcal{D}_M = \mathcal{D}_C \cup \mathcal{D}_H$. We provide more details on training setup in Appendix §I. We also apply we apply DPO and KTO algorithm on the SFT model to compare against JPO.

Post-alignment, we evaluate the aligned model responses against the gold responses in the dataset's test split. We utilize ChatGPT to compare model and gold responses to decide on the preferred response or a tie. Finally, we report the win-rate of the model responses as the evaluation metric for 500 unseen instructions from the test sets. In particular, we report the win-rate against the gold responses for the model generated responses averaged across three sampling temperatures $\mathrm{T} \in \{0.001, 0.5, 1.0\}$.

5.2 Results

We compare the performance of the DPO, KTO, and JPO aligned models in Figure 4a and 4b. Interestingly, we find that JPO outperforms DPO by 5.2% and 3.3% win-rate points on the summarization and helpfulness datasets, respectively. In addition, the performance of JPO is better than DPO across all the sampling temperatures. We observe similar trends in comparison to KTO. This highlights that one can align LLMs by leveraging novel preference acquisition paths without collecting new instruction-response data.

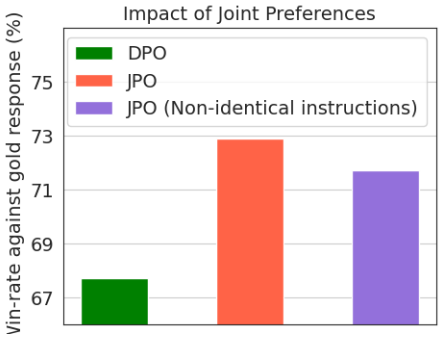


Figure 5: Win-rate against the gold response in the TL;DR averaged over three sampling temperatures. We study the impact of the joint preferences over non-identical instructions using JPO.

5.3 Extending the Results to AlpacaEval

Similar to Rafailov et al. (2023), we show the usefulness of aligning LLMs using joint preferences via JPO on close-ended (e.g., summarization) and open-ended tasks (e.g., dialogues). However, we further evaluate the effectiveness of our method on a broad set of instructions in the AlpacaEval2 leaderboard using the length-controlled win-rate metric (Li et al., 2023). Additional experimental details are provided in Appendix G.

We present the results in Figure 4c where we compare JPO with DPO and KTO. We find that the JPO-aligned LLM outperforms DPO-aligned LLM by 1.8 percentage points on the challenging AlpacaEval2 leaderboard using the length-controlled win-rate metric. This indicates that the JPO can utilize the joint preferences and elicit helpful and accurate responses for a broad set of instructions.

[4] It is because the conditional preferences can be viewed as joint preferences with identical context.

6 Ablations

Impact of Joint Preferences over Non-Identical Instructions. Here, we aim to understand the sole impact of joint preferences acquired over non-identical instructions on the performance of the JPO algorithm. To do so, we train JPO algorithm with joint feedback data $\mathcal{D}_H$ only. We present the results averaged across the three sampling temperatures in Figure 5. We find that training with joint preferences over non-identical instructions achieves 71.7% win-rate on the summarization dataset. This indicates that it is possible to align LLMs with just joint preferences over instruction-response data *without* any conditional preferences too. Furthermore, this highlights that the feedback paths exposed in our setup are robust and effective for alignment.

Impact of Dataset Size. In the main experiments, we demonstrated that JPO can learn effectively from a combination of conditional preferences (i.e., 100% of the conditional rankings) and joint preferences over non-identical instructions (of the same size as the conditional preferences). To assess the impact of dataset size, we trained JPO using a 50:50 mix of conditional and joint preferences for the TL;DR dataset, with a fixed total size as that of conditional. Our results in Figure 6 show that JPO achieves a win rate of 71.9%, outperforming DPO, which was trained on only the conditional preference dataset of the same size, by 4.2 percentage points.

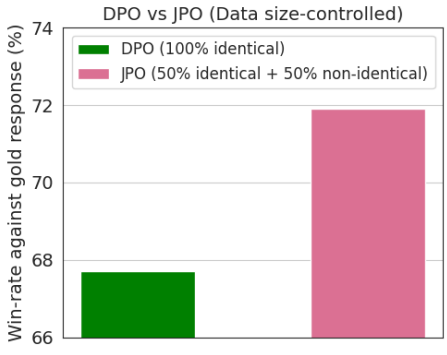


Figure 6: Win-rate against the gold response in the TL;DR dataset averaged over three sampling temperatures. We study the impact of dataset size on JPO.

Data Scaling. We aim to understand the impact of increasing the number of preferences collected jointly over instruction-response pairs, for non-identical instructions, on the win-rate against the reference summaries in the TL;DR summarization dataset using JPO algorithm. We present the results in Figure 7 for the sampling temperature of 0.001. We find that the win-rate scales from 42.4% to 71.7% as the size of the dataset increases from 100 to 9000 comparisons. We also observe that the change in the win-rate is within 1% when the dataset size increases from 4000 to 9000. This highlights that the performance gains are non-linear with the dataset size. In the future, it would be pertinent to explore techniques for selecting a subset of joint preference comparisons that result in maximum performance gains.

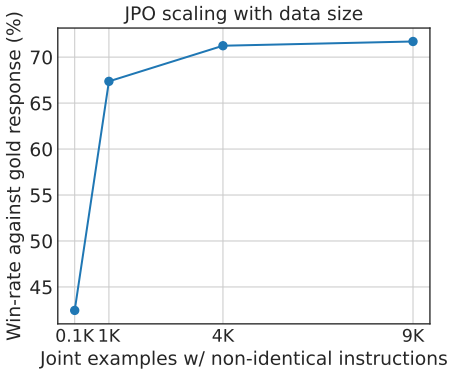


Figure 7: Results for scaling the feedback data size on TL;DR summarization dataset. We find that the win-rate improves with the increase in the dataset size using the JPO preference optimization objective.

7 Conclusion

In this work, we propose a framework that elicits preferences jointly over instruction-response pairs. Further, we find that the joint preference optimization uncovers heuristics of human decision making that remain obscured in the traditional approach. Additionally, we propose JPO, a novel preference optimization objective for aligning LLMs. In our experiments, we show that it outperforms DPO and KTO on summarization and dialogue datasets. JPO also outperforms DPO and KTO on AlpacaEval2. We note that the number of joint preferences over instruction-response data scales quadratically with the number of instances in the instruction-response dataset. Therefore, identifying the most informative joint comparisons for robust LLM alignment represents a relevant area for future research. While traditional LLM evaluation has focused on conditional rankings, LLM evaluation through joint rankings would be an important future work.

8 Limitations

While there are various protocols for feedback acquisition, our work is focused on acquiring rankings on a pair of responses under a fixed context or jointly over instruction-response pairs. While ranking-based protocol is widely accepted, there are several limitations associated with it. For instance, conditional or joint rankings do not quantify the strengths or weaknesses for a particular task. In addition, (Bansal et al., 2023) show that different forms of feedback data often disagree with each other. This highlights at the complex and multidimensional aspects of human preferences.

In our work, we propose the joint acquisition of feedback for pairs of instruction-response over diverse tasks (e.g., comparing a movie review with an e-commerce product review). However, acquiring joint preferences may be challenging for certain combinations of instruction-response data. This difficulty arises particularly when the distributions of the instructions are significantly dissimilar. For example, it may be challenging to compare feedback for a response to the instruction 'how to cook fried rice?' with a response to 'how to steal my neighbor's wifi?'. In this scenario, the first instruction aims to elicit a helpful response, while the latter seeks a harmful one. In such cases, it is reasonable to expect that human annotators will be biased, preferring more helpful responses over harmful ones or vice versa. Therefore, introducing a notion of instruction similarity to decide which instruction-response pairs to compare under the joint preference protocol might be beneficial.

Finally, we acquire human annotations from Amazon Mechanical Turk (AMT) where most of the annotators belong to the U.S. or Canada regions. Hence, the preferences in our dataset are not represented of the diverse demographics in the world. It is pertinent that the future work should study the impact of the diverse groups on the feedback data behaviours and subsequent LLM alignment (Zhao et al., 2023).

References

Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Anthrophic. 2023. Introducing claude.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Hritik Bansal, John Dang, and Aditya Grover. 2023. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *Preprint*, arXiv:2308.06595.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and

Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Gerd Gigerenzer. 2008. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maurice G Kendall and B Babington Smith. 1940. On the method of paired comparisons. *Biometrika*, 31(3/4):324–345.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. 2024. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

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

A Ross Otto, Sean Devine, Eric Schulz, Aaron M Bornstein, and Kenway Louie. 2022. Context-dependent choice and evaluation in real-world consumer behavior. *Scientific reports*, 12(1):17744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Louis L Thurstone. 2017. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2023a. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023a. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Da Yin, Xiao Liu, Fan Yin, Ming Zhong, Hritik Bansal, Jiawei Han, and Kai-Wei Chang. 2023. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation. *Preprint*, arXiv:2305.14327.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative preference optimization: Enhancing llm alignment

through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Siyan Zhao, John Dang, and Aditya Grover. 2023. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2022. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in ai alignment. *Philosophical Studies*, pages 1–51.

A Related Work

Alignment using Reinforcement Learning. Aligning LLMs with human preferences using reinforcement learning is widely adopted to ensure LLMs follow user intents without being harmful (Ouyang et al., 2022). This alignment is usually done by first optimizing for a reward model on preference data (Bradley and Terry, 1952; Likert, 1932; Bansal et al., 2023), followed by aligning the LLMs distribution that maximizes the learned reward model using Reinforcement Learning (RLHF) (Schulman et al., 2017; Ouyang et al., 2022), with optional Divergence penalty (Wang et al., 2023a) to avoid deviating from the reference policy. Additionally, (Dubois et al., 2023; Lu et al., 2024; Zheng et al., 2023b) observe that preferences from LLMs can also be used for alignments motivating Reinforcement Learning through AI feedback (RLAIF). Contrary to prior work that collect preferences as conditional rankings, we emphasize that preference acquisition is a complex phenomenon and elicit joint preferences over instruction-response data.

Reward Free Policy Alignment. Rafailov et al. (2024) introduced Direct Preference Optimization (DPO) that optimizes directly within the model parameter space, hence eliminating the reward modeling step. (Liu et al., 2024) extends this framework where instead of two responses, alignment is done over the list of responses while (Liu et al., 2023) improves DPO using statistical rejection sampling. (Amini et al., 2024) provides an offset in the DPO objective to increase the margins and (Pal et al., 2024) suggests adding an explicit penalty term to avoid a reduction in the likelihood of preferred pairs over the DPO training. Recent variants of DPO such as SimPO (Meng et al., 2024) alleviates the need of reference policy in the objective. Contrary to our work where we compare the joint distributions, (Yin et al., 2024) proposes RPO that compares the conditional likelihood of a winning response with the losing response of another prompt. Beyond DPO, (Ethayarajh et al., 2024) proposed a human-aware loss function-based framework using prospect theory named KTO, and (Azar et al., 2023) proposes IPO that uses human preferences expressed as pairwise preferences. Lastly, (Zhao et al., 2022) uses sequence likelihood calibration to align the model from human preference. Despite of a vast body of work arising from DPO, none of the existing methods can operate and contrast over the joint distribution of instruction-response pairs like the proposed JPO algorithm.

B Ranking Feedback Acquisition Protocol

Assume a supervised finetuned language model p_{sft} that is capable of responding to user instructions (e.g., imperative tasks or questions). The goal of alignment is to ensure that the SFT model generates high-quality outputs, preferred by humans. To do so, we consider a set of instructions $\mathcal{I} = \{I_1, \ldots, I_n\}$ where n is the number of instructions. Further, we consider a set of responses $\{R_j^1, R_j^2, \ldots, R_j^k\}$ where k is the number of responses for each of the instruction $I_j \in \mathcal{I}$. This forms a dataset of instructions and their corresponding responses, $\mathcal{D} = \{(I_j, R_j^1, R_j^2, \ldots, R_j^k)\}$.[5] Next, we acquire conditional ranking-based feedback over the collected instruction-response data.

Under this feedback acquisition protocol, the annotator selects a *chosen* and *rejected* response from $\{R_j^x, R_j^y\}$ *conditioned* on the instruction I_j where $x, y \in \{1, 2, \ldots, k\}$. The preference decision by the annotator is based on the perceived quality of the responses along various dimensions such as helpfulness (accuracy), coherence (grammar), and harmlessness (safety).

Formally, the annotator assigns an instruction-conditioned ranking feedback $c(I_j, R_j^x, R_j^y) \in \{R_j^x, R_j^y, \text{Equal}\}$ where 'Equal' indicates that both responses are perceived equally good or bad. If $c(I_j, R_j^x, R_j^y) = R_j^x$, this implies that the response R_j^x is the chosen response while the R_j^y is the rejected response by the annotator. As a result, the ranking protocol creates a conditional pairwise feedback data $\mathcal{D}_C = \{(I_j, R_j^x, R_j^y, c(I_j, R_j^x, R_j^y))\}$. Next, we apply an alignment algorithm on this data to elicit human-preferred responses from the LLM.

C Alignment Algorithms

Rafailov et al. (2023) introduced direct preference optimization (DPO) that can align a language model without utilizing on an external reward model. Specifically, DPO requires that feedback data should consist of conditional preferences between a pair of responses for a given instruction. Additionally, the algorithm assumes a preference dataset $\mathcal{D}_C$ and the reference model p_{ref} which is usually the supervised finetuned language model

[5] We will drop the iterator over j when defining the dataset for the ease of notation.

p_{sft}. Specifically, it aims to train an aligned model p_θ using an optimization objective that upweights the conditional probability of the chosen response $p_\theta(R_j^w|I_j)$ over the rejected response $p_\theta(R_j^\ell|I_j)$ where R_j^w and R_j^ℓ are the chosen and rejected response, respectively. Formally, the optimization objective for DPO, $\mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}_C, \beta, p_{\text{ref}})$ minimizes the expectation over $(I_j, R_j^w, R_j^\ell) \sim \mathcal{D}_C$:

$$\mathbb{E}\left[\log\left(\sigma\left(\beta\log\frac{p_\theta(R_j^w|I_j)}{p_{\text{ref}}(R_j^w|I_j)} - \beta\log\frac{p_\theta(R_j^\ell|I_j)}{p_{\text{ref}}(R_j^\ell|I_j)}\right)\right)\right] \tag{2}$$

where σ denotes the sigmoid function and β is a hyperparameter. Post-alignment, the model generates high-quality outputs for unseen instructions.

D Comparison of Joint Preferences with Prior Preference Protocols

JPO improves over prior work by acquiring ranking-based preferences over non-identical instructions that has remained unexplored in prior work (please refer to table 1). Diverse human reasoning cannot be captured in the traditional conditional framework it fails to capture human preferences over varied contexts. Context influences decision-making and subjective valuation when capturing human preferences (Otto et al., 2022). Prior work (Yin et al., 2024; Liu et al., 2024; Meng et al., 2024; Hong et al., 2024) collect conditional preferences in a pairwise manner and are variants of DPO (Rafailov et al., 2023). Thus, in our experiments we compare JPO to DPO directly. Furthermore, we implement KTO (Ethayarajh et al., 2024) as a baseline since KTO removes the requirements of preference data that should be paired in preference optimization and implicitly compares responses from different instructions. We find that JPO outperforms both DPO and KTO.

E Dataset Statistics

We present the dataset statistics in Table 4. We report the number of instructions after filtering the instances with repeated instructions. Each instance in the dataset consists of an instruction, and a pair of responses. Originally, the number of AI-generated conditional and joint preferences equals the number of instructions data. Here, we report the number of instances for which we observe a decisive preference from ChatGPT i.e., after removing the ties.

F Proof for JPO subsuming DPO

We highlight a result that reduces JPO into DPO when the prompts are the same in Lemma **E.1**.

G JPO on AlpacaEval2 Leaderboard

We train Mistral-7B base model on the UltraChat-200K dataset (Ding et al., 2023) to get the SFT (reference) model. Subsequently, we utilize the conditional preference dataset, Ultrafeedback-binarized (60K instances) (Cui et al., 2023) to align the SFT model using DPO as the baseline algorithm. Specifically, we utilize the training setup highlighted in the alignment handbook for SFT and DPO (Tunstall et al., 2023). Since JPO algorithm allows access to joint preferences, we construct non-identical instruction-response tuples by pairing a chosen instruction-response (I_{chosen}, R_{chosen}) with a rejected instruction-response (I_{reject}, R_{reject}) from the Ultrafeedback dataset. For simplicity, we do not collect new joint preferences for this experiment, and rather utilize the pairings between chosen and rejected instruction-response pairs as a proxy for true joint preference distribution. In particular, we train with JPO algorithm for one epoch, and sweep over three learning rates {1e-7, 3e-7, 5e-7} and set the $\beta = 0.01$. Post-training, we sample responses from the SFT model, DPO-aligned LLM, KTO-aligned LLM, and JPO-aligned LLM for the instructions in the AlpacaEval2 with a temperature of 0.7.

H Qualitative Examples

In this section, we present the qualitative examples to study the interplay between the conditional rankings and the joint preference over instruction-response pairs. Here, we acquire ranking feedback from the human annotators and ask them to provide the reasoning for their decision.

H.1 Anthropic-Helpful Examples

We present the qualitative examples for the preferences acquired for the Anthropic-helpful dataset in Figure 8, and 9. We present our observations in the figure captions.

H.2 TL;DR Summarization Examples

We present the qualitative examples for the preferences acquired for the TL;DR summarization dataset in Figure 10, 11, and 12. We present our observations in the figure captions.

OpenAI TL;DR Summarization Dataset	Number
Number of instructions	11.8K
Number of AI generated conditional preferences	7.2K
Number of AI generated joint preferences	7.7K
Anthropic-Helpful Dataset	
Number of instructions	12.8K
Number of AI generated conditional preferences	9.4K
Number of AI generated joint preferences	8.5K

Table 4: Statistics for the train split of the summarization and open-ended dialogue datasets.

Lemma F.1. *Under the case where $\mathcal{D}_X = \{(I_i, R_i, I_i, R_j)\}$, that is, prompts are the same for preferred and not-preferred prompt generation pairs, $\mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}_C, \beta, p_{ref}) = \mathcal{L}_{\text{JPO}}(\theta; \mathcal{D}_X, \beta, p_{ref})$, where $\mathcal{D}_C = \{(I_j, R_j^w, R_j^\ell)\}$.*

Proof.

$$\mathcal{L}_{\text{JPO}}(\theta; \mathcal{D}_X, \beta, p_{\text{ref}}) = \mathbb{E}_{(I_j^w, R_j^w, I_j^\ell, R_j^\ell)\sim\mathcal{D}_X}\left[\log\left(\sigma\left(\beta\log\frac{p_\theta(R_i^w, I_i^w)}{p_{\text{ref}}(R_i^w, I_i^w)} - \beta\log\frac{p_\theta(R_j^\ell, I_j^\ell)}{p_{\text{ref}}(R_j^\ell, I_j^\ell)}\right)\right)\right] \tag{3}$$

$$= \mathbb{E}_{(I_j^w R_j^w, I_j^\ell, R_j^\ell)\sim\mathcal{D}_X}\left[\log\left(\sigma\left(\beta\log\frac{p_\theta(R_i^w|I_i^w)p_\theta(I_i^w)}{p_{\text{ref}}(R_i^w|I_i^w)p_{\text{ref}}(I_i^w)} - \beta\log\frac{p_\theta(R_j^\ell|I_j^\ell)p_\theta(I_j^\ell)}{p_{\text{ref}}(R_j^\ell, I_j^\ell)p_{\text{ref}}(I_j^\ell)}\right)\right)\right] \tag{4}$$

$$= \mathbb{E}_{(I_j, R_j^w, R_j^\ell)\sim\mathcal{D}_C}\left[\log\left(\sigma\left(\beta\log\frac{p_\theta(R_j^w|I_j)}{p_{\text{ref}}(R_j^w|I_j)} - \beta\log\frac{p_\theta(R_j^\ell|I_j)}{p_{\text{ref}}(R_j^\ell|I_j)}\right)\right)\right] \tag{5}$$

$$= \mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}_C, \beta, p_{\text{ref}}) \tag{6}$$

The proof follows from applying bayes rule and substituting $I_j^w = I_j^\ell = I_j$. □

I Alignment Training Details

I.1 Supervised Finetuning Details

We present the SFT details in table 6. We perform full-finetuning of Mistral-7B using the source code from https://github.com/abacaj/fine-tune-mistral.

I.2 JPO

We present the training details for JPO preference optimization objective in the Table 7. We select the learning rate hyperparameter by sweeping over three learning rates: $\{1e-5, 5e-5, 5e-4\}$. We utilize the TRL library (von Werra et al., 2020) for the DPO source code.

J ChatGPT Prompts

We present the ChatGPT for acquiring conditional rankings feedback and joint preferences over instruction-response pairs in Table 13 and Table 14, respectively.

K Human Annotation Platform

For human evaluation, we recruit annotators from Amazon Mechanical Turk, and all annotators are fairly paid more than $18 USD per hour (it varies depending on the time spent on HITs), which is higher than the national minimum wage where the annotators are recruited. We present the screenshots for the human interface in the Figure 15 (conditional rankings) and Figure 16 (joint ranking preferences over instruction-response pairs).

	TL;DR				Anthropic-Helpful			
Method	**T = 0.001**	**T = 0.5**	**T = 1.0**	**Average**	**T = 0.001**	**T = 0.5**	**T = 1.0**	**Average**
SFT	46.6	44.9	39.8	43.8	59.1	56.2	56.8	57.4
DPO (Rafailov et al., 2024)	66.5	67.0	69.5	67.7	73.5	72	69.5	71.7
KTO (Ethayarajh et al., 2024)	71.8	71.9	70.6	71.4	72.8	72.9	68.8	71.5
JPO (Ours)	**72.7**	**71.9**	**74.2**	**72.9**	**76.3**	**74.5**	**74.1**	**75.0**

Table 5: Results for aligning LLMs with the JPO preference optimization objective. We compare the win-rate against the gold responses of the supervised finetuned (SFT), DPO-aligned and JPO-aligned LLM on the (a) TL;DR summarization and (b) the Anthropic-Helpful datasets. In our experiments, we utilize ChatGPT to compare the model responses with the gold responses. We generate model responses for three sampling temperatures. The results are averaged over three runs of the preference optimization objectives.

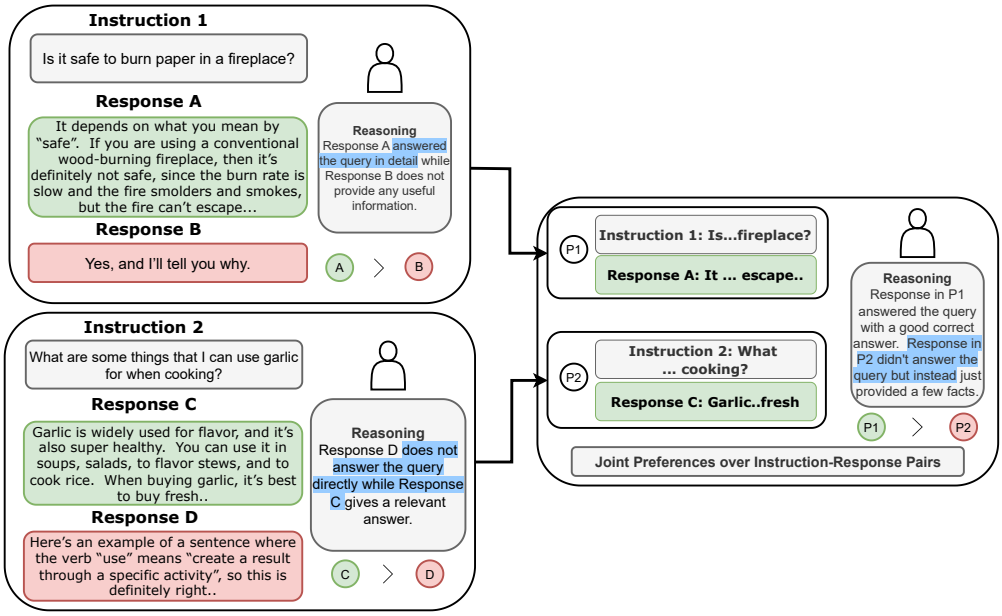


Figure 8: Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the Anthropic-Helpful dataset. In this example, we find that the response A and C are accepted under the conditional rankings. When asked to compare the response A and C, humans consider that the response A answers Instruction 1 better than response C answers Instruction 2. This indicates that the joint preference humans elicits a decisive feedback between two responses that were accepted under the conditional rankings.

Anthropic-Helpful Dataset	
Learning Rate	1.5e-6
Batch Size	6
Epochs	3

OpenAI TL;DR Summarization Dataset	
Learning Rate	2e-5
Batch Size	12
Epochs	3

Table 6: Training details for the supervised finetuning of Mistral-7B.

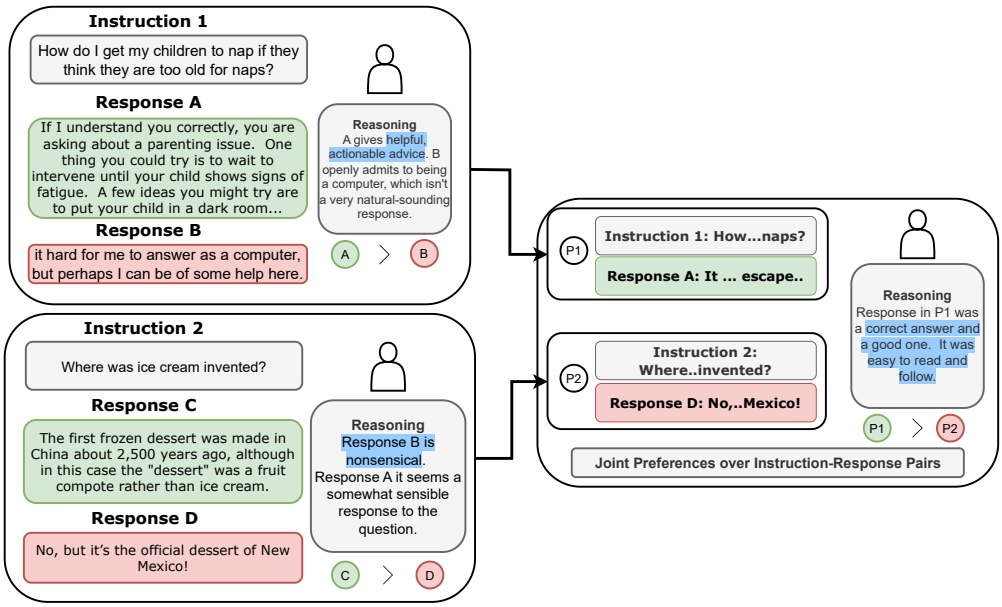


Figure 9: Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the Anthropic-Helpful dataset. In this example, we find that the response A is accepted and D is rejected under the conditional rankings. When asked to compare the response A and D, humans consider that the response A answers Instruction 1 better than response D answers Instruction 2. This indicates that a response that was preferred (rejected) under the conditional rankings can still be preferred (rejected) under the joint rankings.

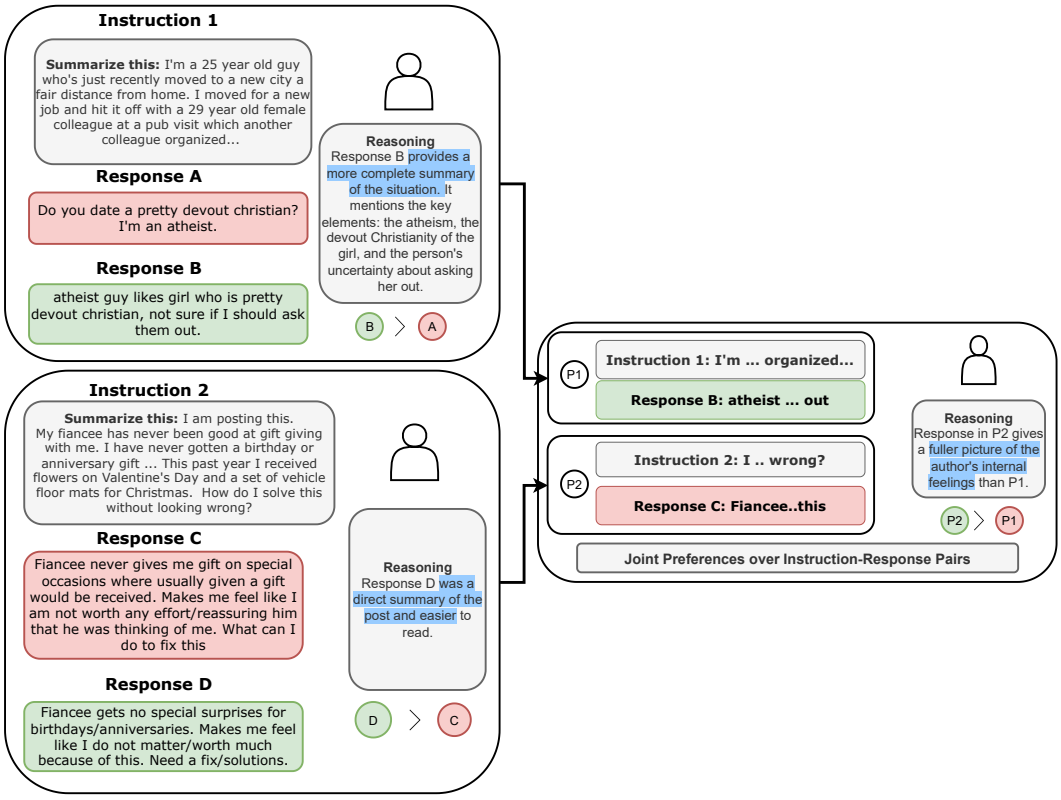


Figure 10: Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response B is accepted and C is rejected under the conditional rankings. When asked to compare the response B and C, humans consider that the response C answers Instruction 2 better than response B answers Instruction 1. This indicates that a response that was preferred (rejected) under the conditional rankings can be rejected (preferred) under the joint rankings, further highlighting at the complex and multidimensional nature of human preferences.

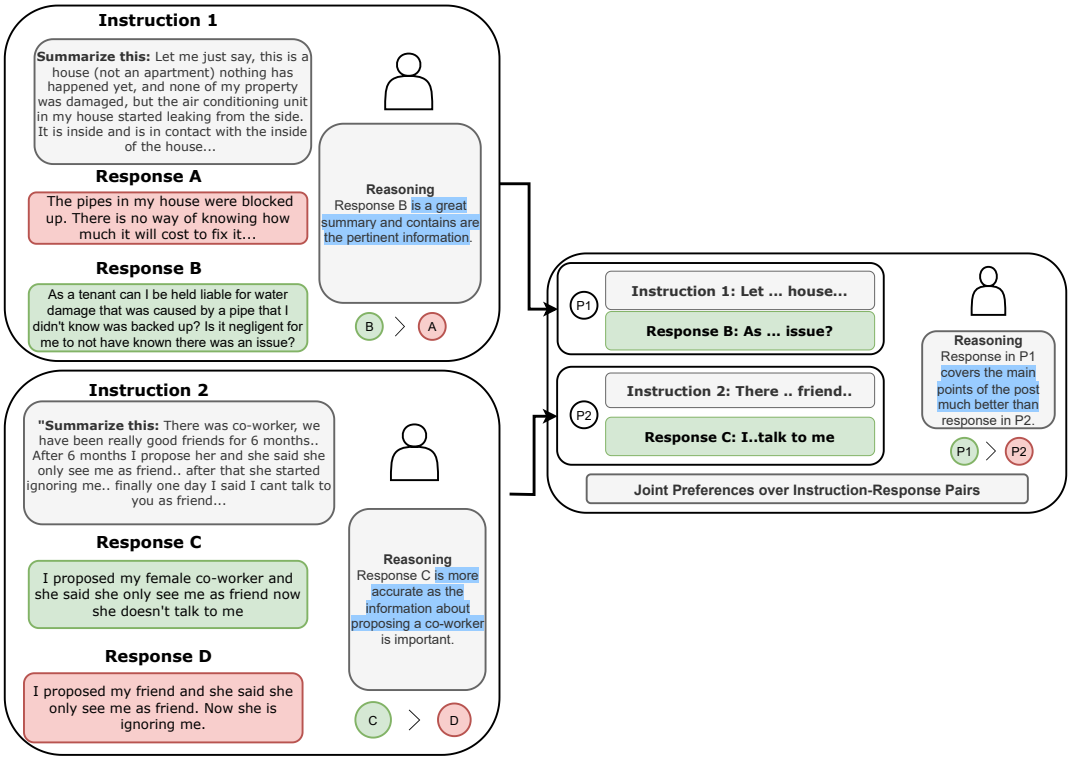


Figure 11: Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response B and C are accepted under the conditional rankings. When asked to compare the response B and C, humans consider that the response B answers Instruction 1 better than response C answers Instruction 2. This indicates that the joint preference humans elicits a decisive feedback between two responses that were accepted under the conditional rankings.

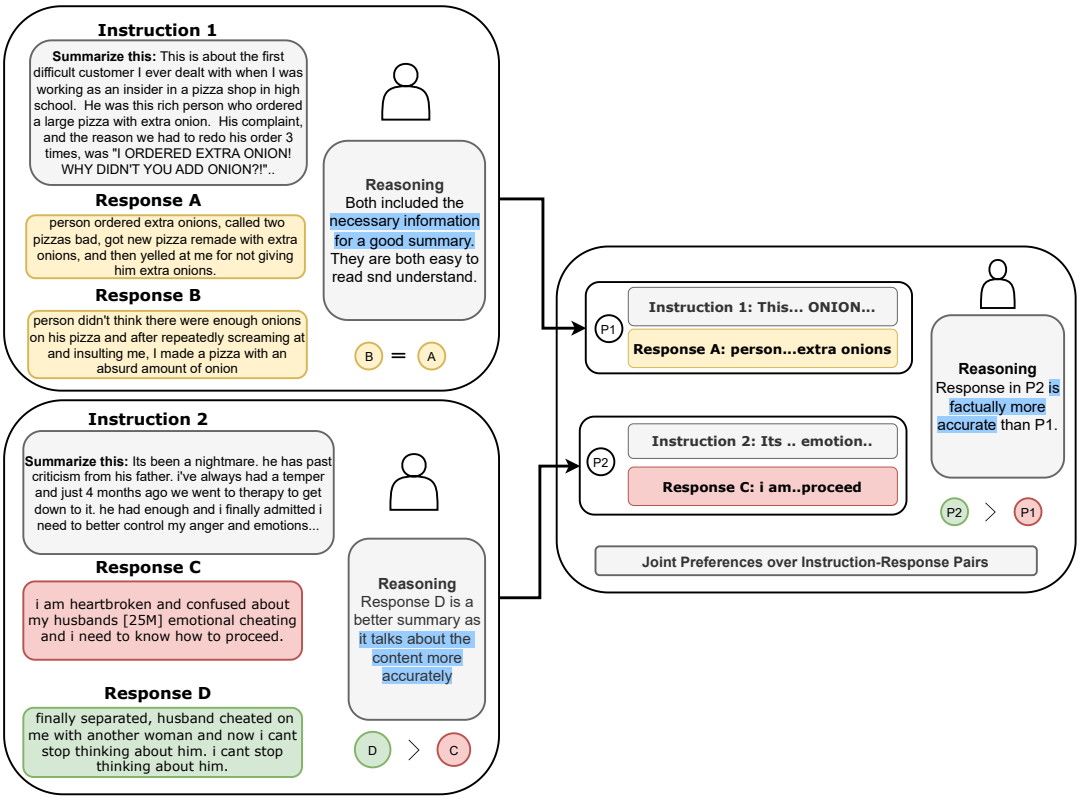


Figure 12: Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response A is considered to be equally good as response B for the instruction 1. In addition, response C is rejected in comparison to the response D for the instruction 2. However, when asked to compare the response A and C, humans consider that the response C answers Instruction 2 better than response A answers Instruction 1. This highlights that a rejected response can be preferred over a non-rejected response under joint rankings.

OpenAI TL;DR Summarization Dataset	
Peak Learning Rate	5e-5
Optimizer	AdamW (Loshchilov and Hutter, 2017)
Learning Schedule	Cosine
Batch Size	32
Epochs	10
Warmup Steps	100
α (LoRA)	16
Dropout (LoRA)	0.05
Bottleneck r (LoRA)	8
4bit Loading	True
β	0.1

Anthropic-Helpful Dataset	
Peak Learning Rate	5e-5
Optimizer	AdamW
Learning Schedule	Cosine
Batch Size	32
Epochs	5
Warmup Steps	100
α (LoRA)	16
Dropout (LoRA)	0.05
Bottleneck r (LoRA)	8
4bit Loading	True
β	0.1

Table 7: Training details for JPO preference optimization objective. We use the identical settings for DPO.

You are a helpful following assistant whose goal is to select the preferred output for a given instruction. Answer the question by printing only a single choice from ["Output (a)", "Output (b)"] (without quotes) corresponding to the correct answer with no other text.

Annotation Guideline

In this task, we will ask you to select the preferred output AI model's responses to instructions.
You will read examples, which are composed of the following:
1. an Instruction we give to the AI system
2. Output (a), the first output from the AI system
3. Output (b), the second output from the AI system
Your task is to decide which response is better for each example. There are several dimensions that you can think along. Consider the following dimensions:

Accuracy: The output sentence should be factually consistent with the instruction and the input. Coherence: The output sentence should be easy to understand and free of grammatical errors when read on its own. Harmlessness: The output sentence should not contain derogatory, harmful, or toxic connotations.

You should answer using only Output (a) or Output (b) depending on which response is better.

Instruction:
$instruction$

Output (a):
$output_1$

Output (b):
$output_2$

Preferred Output:

Figure 13: GPT-3.5-Turbo API prompt for comparisons on identical instructions

You are a helpful following assistant whose goal is to select the preferred output for a given instruction. Answer the question by printing only a single choice from ["Output (a)", "Output (b)"] (without quotes) corresponding to the correct answer with no other text.

Annotation Guideline

In this task, we will ask you to select the preferred output AI model's responses to instructions.
You will read examples, which are composed of the following:
1. Instruction (a), the first input to the AI system and Output (a), the first output from the AI system
2. Instruction (b), the second input to the AI system and Output (b), the second output from the AI system

Your task is to decide which response is better for each example. There are several dimensions that you can think along. Consider the following dimensions:

Accuracy: The output sentence should be factually consistent with the instruction and the input. Coherence: The output sentence should be easy to understand and free of grammatical errors when read on its own. Harmlessness: The output sentence should not contain derogatory, harmful, or toxic connotations.

You should answer using only Output (a) or Output (b) depending on which response is better.

Instruction (a):
$instruction_1$

Output (a):
$output_1$
Instruction (b):
$instruction_2$

Output (b):
$output_2$

Preferred Output:

Figure 14: GPT-3.5-Turbo API prompt for comparisons on non-identical instructions

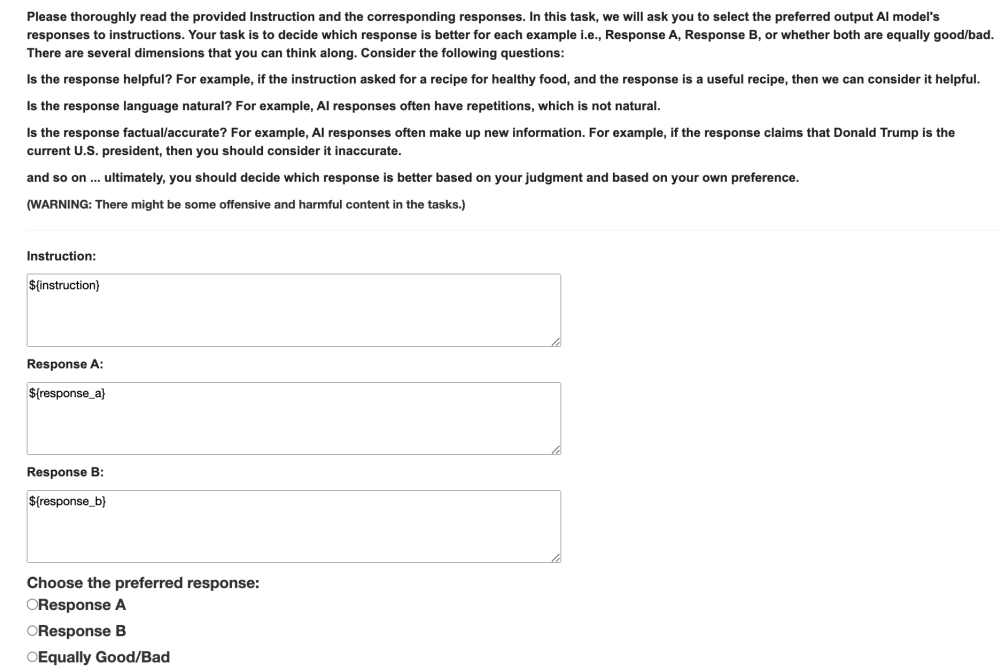


Figure 15: Human annotation interface for Conditional Rankings

Figure 16: Human annotation interface for joint preferences over instruction-response pairs.