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

Hritik Bansal [*1] **Ashima Suvarna** [*1] **Gantavya Bhatt** [*2] **Nanyun Peng** [1] **Kai-Wei Chang** [1] **Aditya Grover** [1]

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 only leverages the pairwise comparisons when the generations are placed in an identical context. However, such conditional rankings often fail to capture the complex and multidimensional aspects of human preferences. In this work, we revisit the traditional paradigm of preference acquisition and propose a new axis that is based on eliciting preferences jointly over the instruction-response pairs. While prior preference optimizations are designed for conditional ranking protocols (e.g., DPO), our proposed preference acquisition protocol introduces DOVE, a new preference optimization objective that upweights the joint probability of the chosen instruction-response pair over the rejected instruction-response pair. Interestingly, we find that the LLM trained with joint instruction-response preference data using DOVE outperforms the LLM trained with DPO by 5.2% and 3.3% win-rate for the 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. We provide the code, models and data at https://dove-alignment.github.io/.

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) (The; 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, the aligned LLMs can produce high-quality outputs that satisfy human values and intents, including helpfulness, coherence, and harmlessness (Askell et al., 2021; Ouyang et al., 2022). LLM alignment hinges on the quality of preferences acquired from human or AI annotators (Ouyang 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 for aligning LLMs(Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a; Tunstall et al., 2023; Teknium, 2023). Specifically, in this approach the annotator has to compare a pair of responses *conditioned* on a fixed context. For instance, human 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*)).

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. Although the traditional conditional rankings provide rich preference for alignment, they fail to holistically capture the various dimensions of reasoning of human preferences. In this work, we revisit the traditional paradigm of conditional preference acquisition and propose a new approach for jointly eliciting preferences over instruction-response pairs. This method aims to uncover diverse reasoning paths in the process of acquiring feedback.

In this work, we develop 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 in Figure 1 (*right*), we aim to under-

[*]Equal contribution [1]University of California, Los Angeles [2]University of Washington. Correspondence to: Hritik Bansal <hbansal@ucla.edu>, Ashima Suvarna <asuvarna31@ucla.edu>, Gantavya Bhatt <gbhatt2@uw.edu>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment

stand whether the response in the pair X is perceived better than the response in the pair Y. For instance, humans would prefer a helpful response to the instruction 'Create a list of four fruits' over a response that completely ignores the instruction 'Create a list of beach activities'. This suggests that we can reveal preference axes like adherence to instructions, grammatical fluency, and clarity, even when following a joint preference optimization protocol. In addition, our protocol can elicit human preference behaviours that are obfuscated in the prior protocols, and redefines conditional preference elicitation as a special case where the instructions are identical.

Next, we propose DOVE, 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. This differs from the other frameworks that assume conditional rankings in their feedback data, such as DPO (Radford et al., 2019; Azar et al., 2023) and preference optimizations that train a separate reward model such as PPO and rejection sampling (Schulman et al., 2017; Nakano et al., 2021). We further point that DOVE subsumes the prior preference optimizations as conditional rankings are a special case of joint preferences (e.g., when $I_1 = I_2$). In our experiments, we focus on extending and comparing against DPO because of their simplicity, stability, and high-performance. However, our framework can be easily applied to PPO-like approaches by training a reward model on the joint preferences.

Finally, we conduct experiments to explore the new reasoning paths enabled by joint preference elicitation, followed by aligning LLMs with the DOVE objective. To do so, we explore the interplay between the feedback data collected under conditional rankings and joint preferences protocol. In addition, we ask human annotators to explain their preference decisions, uncovering new reasoning paths that highlight the complexities of the preference acquisition process (§4). After feedback acquisition, we aim to investigate the impact of diverse preferences collected from conditional and joint preferences on LLM alignment. In our experiments, we align a Mistral-7B LLM with the preferences acquired from the conditional rankings and joint preferences, using our DOVE algorithm. We find that the DOVE outperforms the supervised finetuned LLM by 30% and 18% win-rate against the gold responses on the unseen instructions from the summarization and open-ended dialogues datasets, respectively. Surprisingly, we find that DOVE can effectively tap into the diverse preferences in the conditional and joint feedback data and outperforms DPO by 5.2% and 3.3% win-rate points on the summarization and open-ended dialogues, respectively. 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, our aim is to align language models to generate outputs that are preferred by humans across various dimensions such as helpfulness and coherence. The process of aligning a base model, which is pretrained on a large corpus of text (Commoncrawl; Raffel et al., 2020; Soldaini et al., 2024; Penedo et al., 2023), involves multiple steps: (a) instruction-response data collection, (b) supervised fine-tuning, (c) preference data acquisition, and (d) deployment of 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 machines (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; Geng et al., 2023; Yin et al., 2023; Wang et al., 2023b; Yu et al., 2023; Toshniwal et al., 2024). Following SFT, feedback data is acquired under a specific acquisition protocol (e.g., rankings) from the annotators (§2.1). Finally, an alignment algorithm trains the SFT model on the feedback data (§2.2).

2.1. 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, $\mathcal{D} = \{(I_j, R_j^1, R_j^2, \ldots, R_j^k)\}$.[1] 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

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

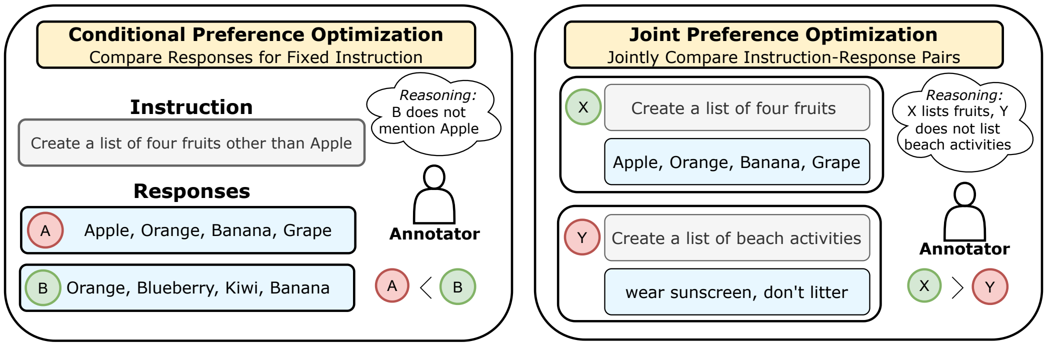


Figure 1. Overview of the proposed Joint Preference Optimization paradigm. (*Left*) We show that the conditional preference acquisition method would require the annotators to compare two responses for an identical instruction. (*Right*) We show that the annotators can also assign rankings jointly over 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.

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.

2.2. Alignment Algorithm

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 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 $(I_j, R_j^w, R_j^\ell) \sim \mathcal{D}_C$:

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

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

3. Joint Preference Optimization using DOVE

3.1. Joint Preference Acquisition Protocol

In §2.1, we describe a common technique for feedback data acquisition that requires the annotators to assign a preferred and non-preferred label to a pair of responses for an instruction. However, this paradigm does not capture the complex and multidimensional aspects of human preferences (Kendall & Smith, 1940; Thurstone, 2017). Specifically, the reasoning paths for making preference decisions depend upon the context in which the comparison is made. 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 elicit high-quality responses that humans prefer under various contexts.

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 [2]). 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.

3.2. DOVE

Here, we propose DOVE, a preference optimization objective that learns to align the language models with the 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 DOVE 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 DOVE, $\mathcal{L}(\theta; \mathcal{D}_X, \beta, p_{\text{ref}})$ minimizes the expectation over $(I_j^w, R_j^w, I_j^\ell, R_j^\ell) \sim \mathcal{D}_X$:

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

where σ denotes the sigmoid function and β is a hyperparameter. Further, we show that Eq. 3 reduces to the DPO formulation (Eq. 1) when the instructions $I_i = I_j$ in Appendix §C. We can also see that the DOVE 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 DOVE to align language models to generate human-preferred summaries and answer open-ended instructions.

4. Interplay between Feedback Protocols

4.1. Instruction-Response Acquisition

The instruction-response data is a collection of real-world queries that are presented to the text AI assistants. 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 ranging from 'Which coffee bean is better for a morning roast?' to 'How do I attract more hummingbirds in my yard?'.

Both these datasets come with the 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 where instructions are repeated. 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 select one of the responses, at random, 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 §B.

4.2. Feedback from AI and Humans

Feedback from AI. Prior work (Dubois et al., 2023; Bai et al., 2022b) has shown that AI feedback can be leveraged to align language models to generate helpful and harmless responses to unseen instructions. In addition, acquiring AI feedback at large-scale is more accessible and cheaper in comparison to human feedback. To this end, 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). The choice of ChatGPT was motivated by its affordability (e.g., output tokens from ChatGPT are $50\times$ cheaper than GPT-4).

To collect ranking feedback over a pair of responses for a fixed instruction, we prompt ChatGPT to assign a chosen response. To mitigate any bias from the ordering of the two responses, we run two queries for all comparisons to cover all possible orderings. 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, the AI is instructed to provide its preference based on the accuracy, coherence, and harmlessness of the

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

responses.

To collect AI preferences jointly over the instruction-response pairs, we prompt ChatGPT to decide the response that better answers its corresponding instruction. Similar to the previous scenario, we run two queries for all comparisons to mitigate any ordering bias and provide guidelines to choose the response that is more accurate, coherent, and harmless. 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 §D.

Feedback from Humans. In this work, we also collect human preferences for 2000 comparisons over summarization and Anthropic-Helpful dataset. Such a data is useful for providing insights into the human behavior under different preference acquisition protocols (§4.3). In addition, this data aids in agreement between the ChatGPT and human decisions.

Specifically, we ask two annotators to assign a chosen response or choose 'equal' after comparing the quality of the responses along the same dimensions as ChatGPT guidelines. The human annotations were collected from Amazon Mechanical Turk (AMT) from 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 §E. We find that the AI achieves an agreement score of 64% with the human annotators in our agreement analysis, compared to the 69% inter-annotator agreement (Section A). This highlights the high quality of the AI feedback across datasets and preference protocols (i.e., conditional and joint preferences over non-identical instructions).

4.3. 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'. For instance, if $\mathcal{P}_C(I_i, R_i) =$ 'chosen' and $\mathcal{P}_C(I_j, R_j) =$ 'chosen' then it implies that the annotators can reason about the joint preferences over a pair of instruction-response pairs that are originally preferred under the conditional ranking feedback protocol. We quantitatively study the interplay between the two ranking-based feedback from AI and Human annotators over summarization and open-ended Anthropic-Helpful datasets.

Data (Annotator)	Decisive	Indecisive
TL;DR (AI)	63.7%	36.2%
TL;DR (Human)	73.8%	25.7%
Anthropic-Helpful (AI)	68.5%	31.5%
Anthropic-Helpful (Human)	77.9%	22.0%
Average	**71.0%**	**29.0%**

Table 1. 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.In total, we compare $48K$ and $1K$ annotations from the AI and humans, respectively.

Data (Annotator)	C > R	C < R	Indecisive
TL;DR (AI)	53.3%	14.3%	30.4%
TL;DR (Human)	41.6%	22.2%	36.1%
Anthropic-Helpful (AI)	54.5%	17.6%	27.8%
Anthropic-Helpful (Human)	57.1%	21.4%	21.4%
Average	**52.0%**	**19.0%**	**29.0%**

Table 2. 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.In total, we compare $48K$ and $1K$ annotations from the AI and humans, respectively.

Results. We present the results for the interaction analysis in Table 2 and Table 1. In Table 1, 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 feedback 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.

In Table 2, 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 Examples. To probe the reasoning paths of the human annotators used for decision making, we ask them to provide brief explanations for their feedback decisions regarding a few conditional and joint preferences. We provide a list of qualitative examples consisting of instructions, responses, and respective preferences in Appendix §F. In Figure 7, we discovered that human annotators provided decisive feedback when comparing instruction-response pairs, basing their decisions on the accuracy of the responses. In Figure 10, we find that the human annotators preferred a instruction-summary pair, that was rejected under the conditional preference, because it provides a fuller picture of the original reddit post. In summary, we expose the multi-faceted reasoning paths of humans in joint instruction-response feedback acquisition that would have been concealed in the conditional feedback acquisition paradigm.

5. LLM Alignment

In the previous sections, we show that the humans and AI are capable of providing ranking-based feedback for a pair of responses for identical and non-identical instructions in the context. Here, we aim to study the impact of feedback data on eliciting high-quality responses from the large language models.

5.1. Setup

Here, we aim to 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 on both the datasets from ChatGPT. Then, we convert the conditional preference data into an instruction-response data for supervised finetuning $\mathcal{D}_{\text{SFT}}$.

First, 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). Subsequently, we apply DPO algorithm on the SFT model using the conditional preference data for 10 epochs and 5 epochs for the summarization and Anthropic-helpful data, respectively. Specifically, we use low-rank adaptation (Hu et al., 2021) of SFT model during DPO alignment. The DPO optimization was trained on a single GPU Nvidia A6000 with a batch size of 32.

We note that our proposed DOVE algorithm can utilize both the conditional preferences and joint preference with non-identical context. It is because the conditional preferences can be viewed as joint preferences with identical context. As a result, we train the base LLM with DOVE algorithm after merging conditional and joint preferences data $\mathcal{D}_M = \mathcal{D}_C \cup \mathcal{D}_H$. We keep the hyperparameters (e.g., β), number of epochs, and the batch size identical to the DPO algorithm. In our experiments, we also train DOVE algorithm on the joint preferences with non-identical instructions and highlight their usefulness for LLM alignment. We provide more details on training setup in Appendix §G.

Post-alignment, we evaluate the aligned model responses against the gold responses in the dataset's test split. Specifically, both datasets come with a human-preferred response for an instruction, which is treated as the gold response. 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 (Rafailov et al., 2023).

5.2. Results

We compare the performance of the SFT, DPO, and DOVE aligned models in Table 3. In particular, we report the win-rate against the gold responses for the model responses generated from various sampling temperatures $\text{T} \in \{0.001, 0.5, 1.0\}$.

DOVE outperforms SFT model. We find that the DOVE achieves high win-rates across all sampling temperatures. Specifically, we observe that DOVE outperforms the SFT model by 29.1% and 18% on the close-ended summarization and open-ended dialogue dataset, respectively, averaged across the sampling temperatures. This indicates that DOVE can utilize the diverse set of feedback from the conditional and joint preferences to align LLMs.

DOVE outperforms DPO and KTO. Further, we aim to understand whether DOVE is able to tease out useful feedback signals from the combination of the conditional preferences and joint preferences over instruction-response pairs. Surprisingly, we find that DOVE outperforms DPO by

	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
DOVE (Ours)	**72.7**	**71.9**	**74.2**	**72.9**	**76.3**	**74.5**	**74.1**	**75.0**

Table 3. Results for aligning LLMs with the DOVE preference optimization objective. We compare the win-rate against the gold responses of the supervised finetuned (SFT), DPO-aligned and DOVE-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.

5.2% and 3.3% win-rate points on the summarization and helpfulness datasets, respectively. In addition, the performance of DOVE is better than DOVE across all the sampling temperatures. This highlights that one can improve the alignment of the LLMs by leveraging novel preference acquisition paths without collecting new instruction-response data. We observe the similar trends in comparison to KTO. Hence, our results indicate that DOVE is a robust alignment algorithm that can elicit high-quality outputs by learning from diverse ranking-based preferences.

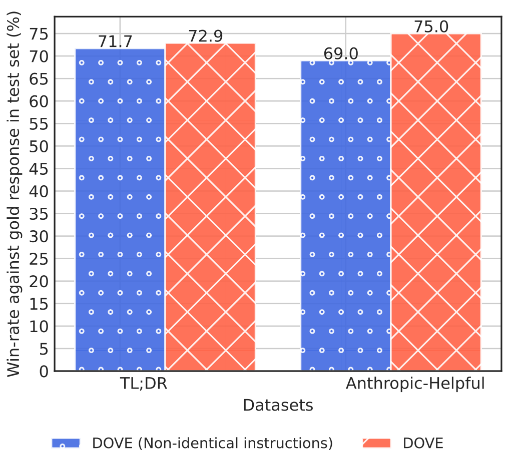


Figure 2. Win-rate against the gold response in the TL;DR and Anthropic-Helpful datasets averaged over three sampling temperatures. We study the impact of the joint preferences over non-identical instructions using DOVE.

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 DOVE algorithm. To do so, we train DOVE algorithm with joint feedback data $\mathcal{D}_H$ only. We present the results averaged across the three sampling temperatures in Figure 2. We find that training with joint preferences over non-identical instructions achieves 71.7% and 69% win-rate on the summarization and anthropic-helpful datasets, respectively. 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 eliciting preferred model responses. We also show that the training with joint preferences scales with the amount of feedback data using the DOVE algorithm in Appendix §H.

6. Related Work

Alignment using Reinforcement Learning. Large Language Models (LLMs) showcase impressive zero-shot and few-shot performance (Brown et al., 2020) with instruction tuning further improving their downstream performance (Victor et al., 2022). Not all responses for a prompt are preferable due to a variety of reasons, thus aligning them with human preference is now a common step in fine-tuning the pipeline. This alignment is usually done by first optimizing for a reward model on preference data (Bradley & 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. Despite the success of RLHF, collecting human feedback can be challenging, however, (Dubois et al., 2023; Lu et al., 2024; Zheng et al., 2023b) observe that LLMs can provide feedback motivating Reinforcement Learning through AI feedback (RLAIF). In this work, we show that preference acquisition is a complex phenomenon and elicit joint preferences over instruction-response data. Further, we study the interactions between joint preferences and traditional approach of conditional rankings for preference elicitation.

Reward Free Policy Alignment. (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. (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. Contrary to our work where 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, more recently (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. Here, we introduce DOVE, a novel preference optimization, complementary to existing works, that learns from joint preferences over instruction-response data and achieves good performance on diverse tasks.

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 new paths of human reasoning that remain obscured in the traditional approach. Additionally, we propose DOVE, a novel preference optimization objective for aligning LLMs. In our experiments, we show that it outperforms DPO on summarization and dialogue datasets. 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).

9. Acknowledgement

We thank the anonymous reviewers for their feedback, as well as members of MINT Lab and UCLANLP for their thoughts. Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380.

References

The claude 3 model family: Opus, sonnet, haiku. URL https://api.semanticscholar.org/CorpusID:268232499.

Amini, A., Vieira, T., and Cotterell, R. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Anthrophic. Introducing claude. 2023. URL https://www.anthropic.com/index/introducing-claude.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bansal, H., Dang, J., and Grover, A. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*, 2023.

Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., and Schimdt, L. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023.

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.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J.-N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S.-C., Unger, M., Veldhuizen, G. P., et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

Commoncrawl. Common crawl. https://commoncrawl.org. Accessed on March 23, 2024.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., and Song, D. Koala: A dialogue model for academic research. Blog post, April 2023. URL https://bair.berkeley.edu/blog/2023/04/03/koala/.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

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

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Liu, T., Qin, Z., Wu, J., Shen, J., Khalman, M., Joshi, R., Zhao, Y., Saleh, M., Baumgartner, S., Liu, J., et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, Y., Yang, X., Li, X., Wang, X. E., and Wang, W. Y. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S., and White, C. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. *arXiv preprint arXiv:2305.18290*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Scheurer, J., Campos, J. A., Korbak, T., Chan, J. S., Chen, A., Cho, K., and Perez, E. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

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

Toshniwal, S., Moshkov, I., Narenthiran, S., Gitman, D., Jia, F., and Gitman, I. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023.

Victor, S., Albert, W., Colin, R., Stephen, B., Lintang, S., Zaid, A., Antoine, C., Arnaud, S., Arun, R., Manan, D., et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., and Huang, S. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023a.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*, 2022.

Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023b.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions, 2023c.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023a.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023b.

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Yin, D., Liu, X., Yin, F., Zhong, M., Bansal, H., Han, J., and Chang, K.-W. Dynosaur: A dynamic growth paradigm for instruction-tuning data curation, 2023.

Yin, Y., Wang, Z., Gu, Y., Huang, H., Chen, W., and Zhou, M. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*, 2024.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Zhao, S., Dang, J., and Grover, A. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023.

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2022.

Zheng, L., Chiang, W.-L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Lin, Z., Xing, E., et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023a.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b.

A. Agreement Analysis.

Here, we compare the agreement between the human and AI preference decisions for the summarization and open-ended tasks under the two acquisition protocols. To establish the gold chosen response in the human annotations, we compare the preference from the two annotators. If the two annotators agree in their response (e.g., Response $R_1 >$ Response R_2), then it is considered the gold chosen response. If the annotators disagree in their preferences (Response $R_1 >$ Response R_2 for annotator 1 and A<B for annotator 2) then we consider it as tie. If an annotator is indecisive while the other annotator makes a decisive preference (e.g., Response $R_1 =$ Response R_2 for annotator 1 and A>B for annotator 2), then the gold response is the one for which a decisive feedback has been provided i.e., Response R_1 . In addition, the inter-annotator agreement is computed by comparing the preference from the two annotators on the identical comparison. In case of disagreements where one of the annotator chooses 'equal', the agreement score is considered as 0.5. We apply similar guidelines to compute human-AI agreement.

We present the agreement results in Table 4. We find that the average agreement 69% and 64% between the human-human and human-AI annotators, respectively. These agreement scores are close to the annotation agreements presented in prior works (Li et al., 2023; Bansal et al., 2023). Interestingly, the agreement scores vary based on the underlying distribution of the instruction-response pairs and the choice of ranking protocol. Overall, our results highlight that AI provides high-quality feedback across summarization and open-ended dataset for both the feedback protocols.

B. Dataset Statistics

We present the dataset statistics in Table 5. 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.

C. Proof for DOVE subsuming DPO

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

D. ChatGPT Prompts

We present the ChatGPT for acquiring conditional rankings feedback and joint preferences over instruction-response pairs in Table 3 and Table 4, respectively.

E. Human Annotation Platform

We present the screenshots for the human interface in the Figure 5 (conditional rankings) and Figure 6 (joint ranking preferences over instruction-response pairs).

F. 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.

F.1. Anthropic-Helpful Examples

We present the qualitative examples for the preferences acquired for the Anthropic-helpful dataset in Figure 7, 8, and 9. We present our observations in the figure captions.

F.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.

G. Alignment Training Details

G.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`.

G.2. DOVE

We present the training details for DOVE 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.

H. Trends with 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 DOVE algorithm. We present the results in Figure 13 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

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 3. 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 4. GPT-3.5-Turbo API prompt for comparisons on non-identical instructions

Dataset	Ranking Protocol	Human-Human	Human-AI
Summarization (TL;DR)	Two Responses under	69%	63%
Anthropic-Helpful	Fixed Instruction	70.1%	62%
Summarization (TL;DR)	Joint Instruction and	62%	60%
Anthropic-Helpful	Response Pairs	74.8%	71%
Average		69%	64%

Table 4. 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 over instruction-response pairs where instructions are non-identical. In addition, we assess the agreement rates over the two datasets: (a) summarization and (b) Anthropic-helpful dataset.

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 5. Statistics for the train split of the summarization and open-ended dialogue datasets.

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.

Lemma C.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, $\mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}_C, \beta, p_{ref}) = \mathcal{L}_{\text{DOVE}}(\theta; \mathcal{D}_X, \beta, p_{ref})$, where $\mathcal{D}_C = \{(I_j, R_j^w, R_j^\ell)\}$.*

Proof.

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

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

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

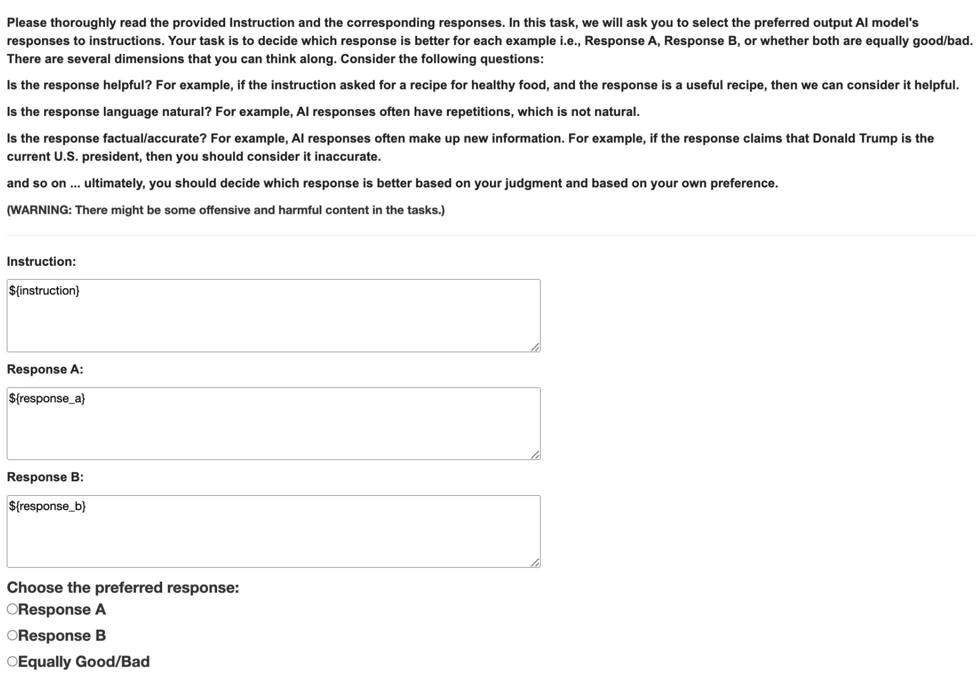
$$= \mathcal{L}_{\text{DPO}}(\theta; \mathcal{D}_C, \beta, p_{\text{ref}}) \tag{6}$$

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

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.



Please thoroughly read the provided Instruction and the corresponding responses. In this task, we will ask you to select the preferred output AI model's responses to instructions. Your task is to decide which response is better for each example i.e., Response A, Response B, or whether both are equally good/bad. There are several dimensions that you can think along. Consider the following questions:

Is the response helpful? For example, if the instruction asked for a recipe for healthy food, and the response is a useful recipe, then we can consider it helpful.

Is the response language natural? For example, AI responses often have repetitions, which is not natural.

Is the response factual/accurate? For example, AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate.

and so on ... ultimately, you should decide which response is better based on your judgment and based on your own preference.

(WARNING: There might be some offensive and harmful content in the tasks.)

Instruction:

${instruction}

Response A:

${response_a}

Response B:

${response_b}

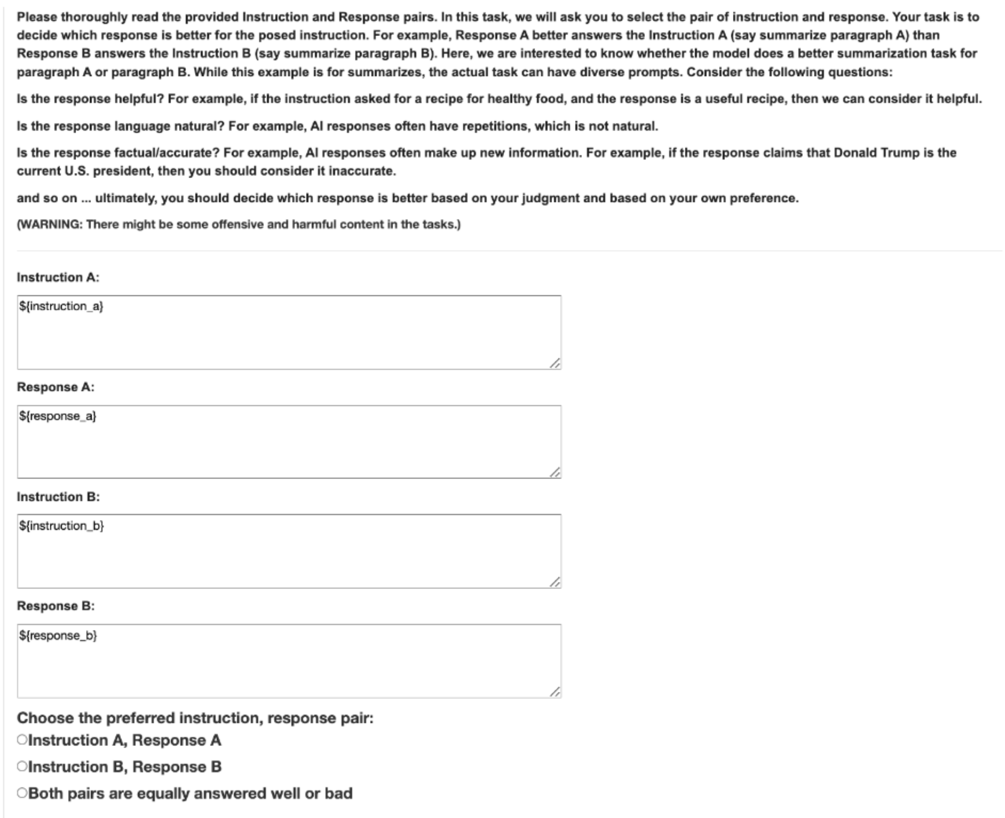
Choose the preferred response:

○**Response A**

○**Response B**

○**Equally Good/Bad**

Figure 5. Human annotation interface for Conditional Rankings



Please thoroughly read the provided Instruction and Response pairs. In this task, we will ask you to select the pair of instruction and response. Your task is to decide which response is better for the posed instruction. For example, Response A better answers the Instruction A (say summarize paragraph A) than Response B answers the Instruction B (say summarize paragraph B). Here, we are interested to know whether the model does a better summarization task for paragraph A or paragraph B. While this example is for summarizes, the actual task can have diverse prompts. Consider the following questions:

Is the response helpful? For example, if the instruction asked for a recipe for healthy food, and the response is a useful recipe, then we can consider it helpful.

Is the response language natural? For example, AI responses often have repetitions, which is not natural.

Is the response factual/accurate? For example, AI responses often make up new information. For example, if the response claims that Donald Trump is the current U.S. president, then you should consider it inaccurate.

and so on ... ultimately, you should decide which response is better based on your judgment and based on your own preference.

(WARNING: There might be some offensive and harmful content in the tasks.)

Instruction A:

${instruction_a}

Response A:

${response_a}

Instruction B:

${instruction_b}

Response B:

${response_b}

Choose the preferred instruction, response pair:

○Instruction A, Response A

○Instruction B, Response B

○Both pairs are equally answered well or bad

Figure 6. Human annotation interface for joint preferences over instruction-response pairs.

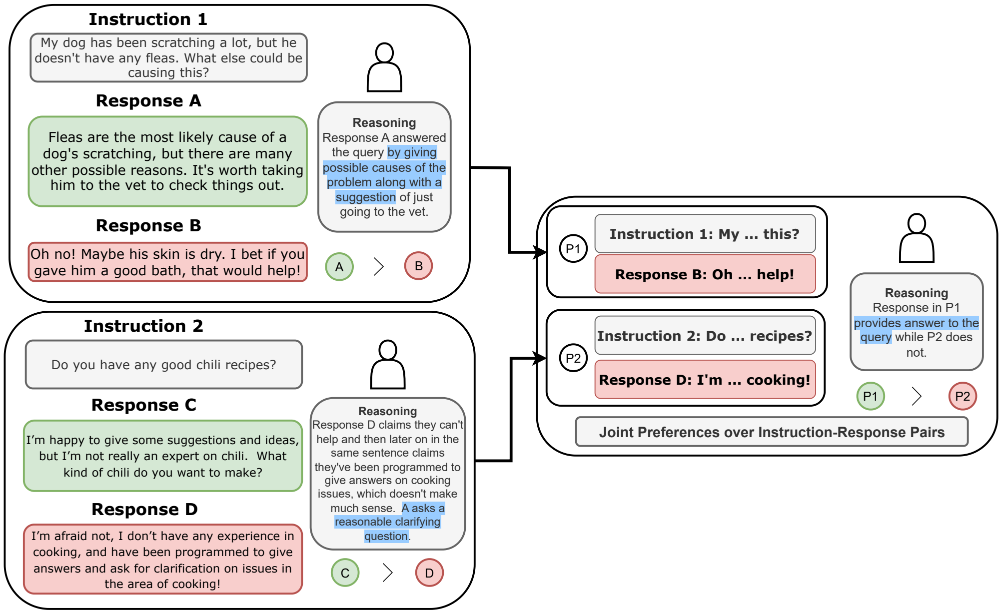


Figure 7. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the Anthropic-Helpful dataset. In this example, we find that the response B and D are rejected under the conditional rankings. When asked to compare the response B and D, humans consider that the response B answers Instruction 1 better than response D answers Instruction 2. This indicates that the joint preference humans elicits a decisive feedback between two responses that were rejected under the conditional rankings.

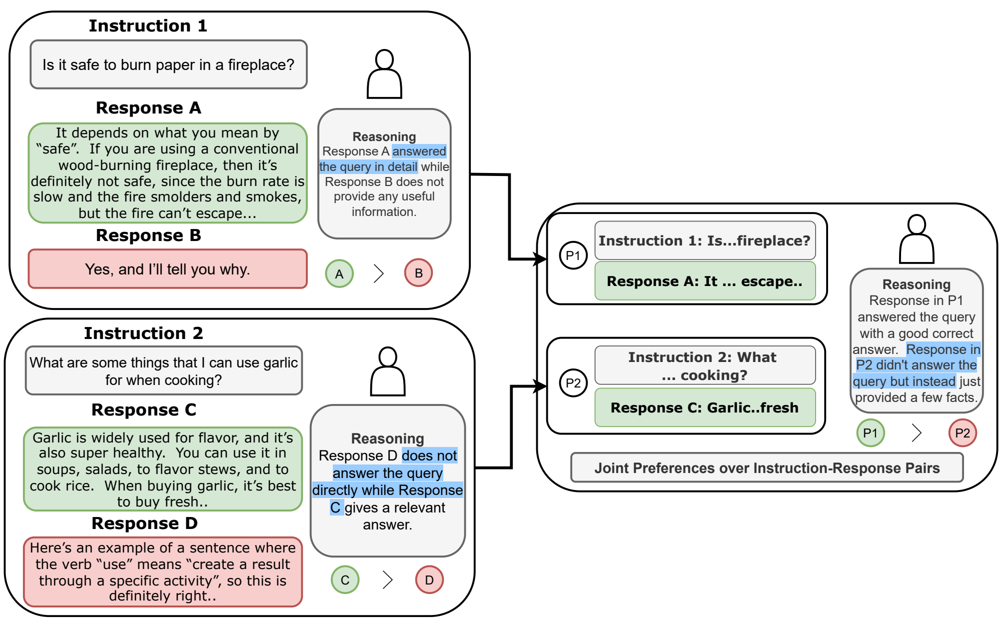


Figure 8. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the Anthropic-Helpful dataset. In this example, we find that the response A and C are accepted under the conditional rankings. When asked to compare the response A and C, humans 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.

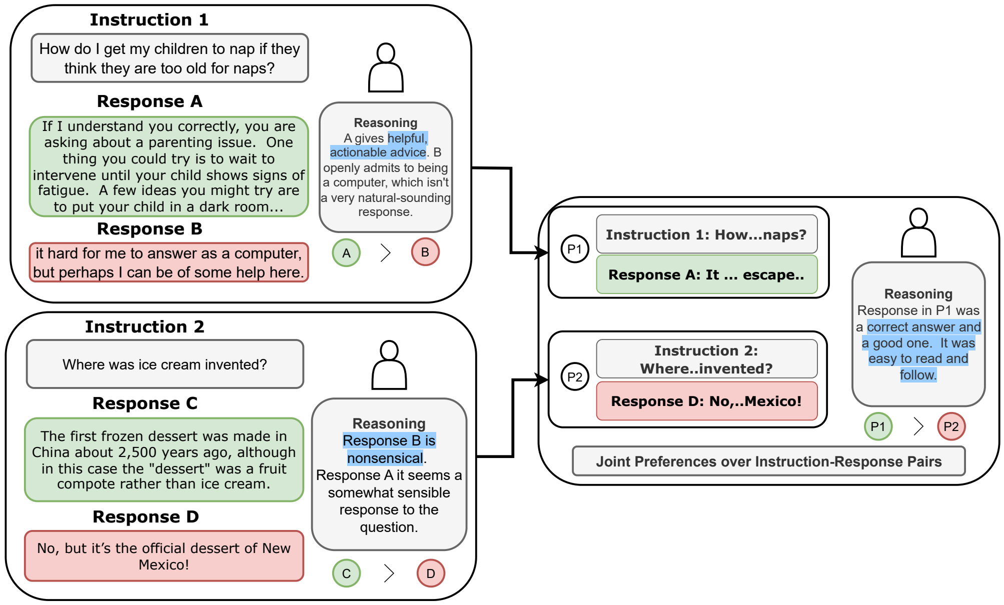


Figure 9. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the Anthropic-Helpful dataset. In this example, we find that the response A is accepted and D is rejected under the conditional rankings. When asked to compare the response A and 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.

OpenAI TL;DR Summarization Dataset	
Peak Learning Rate	5e-5
Optimizer	AdamW (Loshchilov & 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 DOVE preference optimization objective. We use the identical settings for DPO.

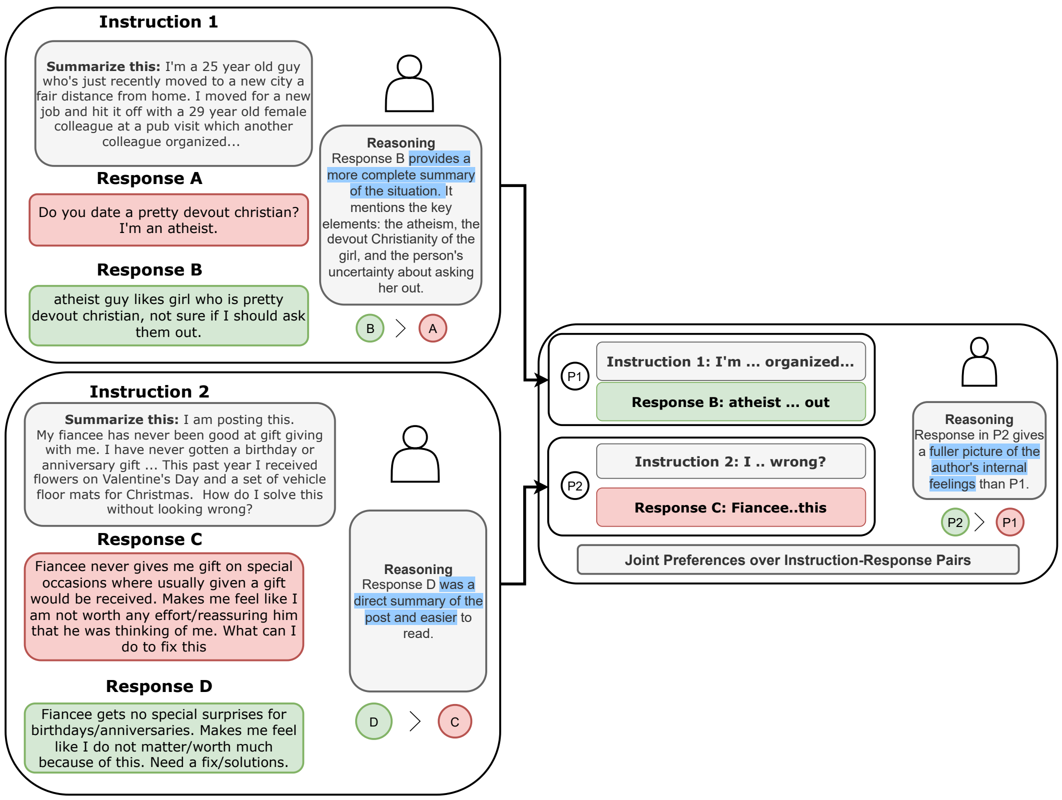


Figure 10. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response B is accepted and C is rejected under the conditional rankings. When asked to compare the response 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.

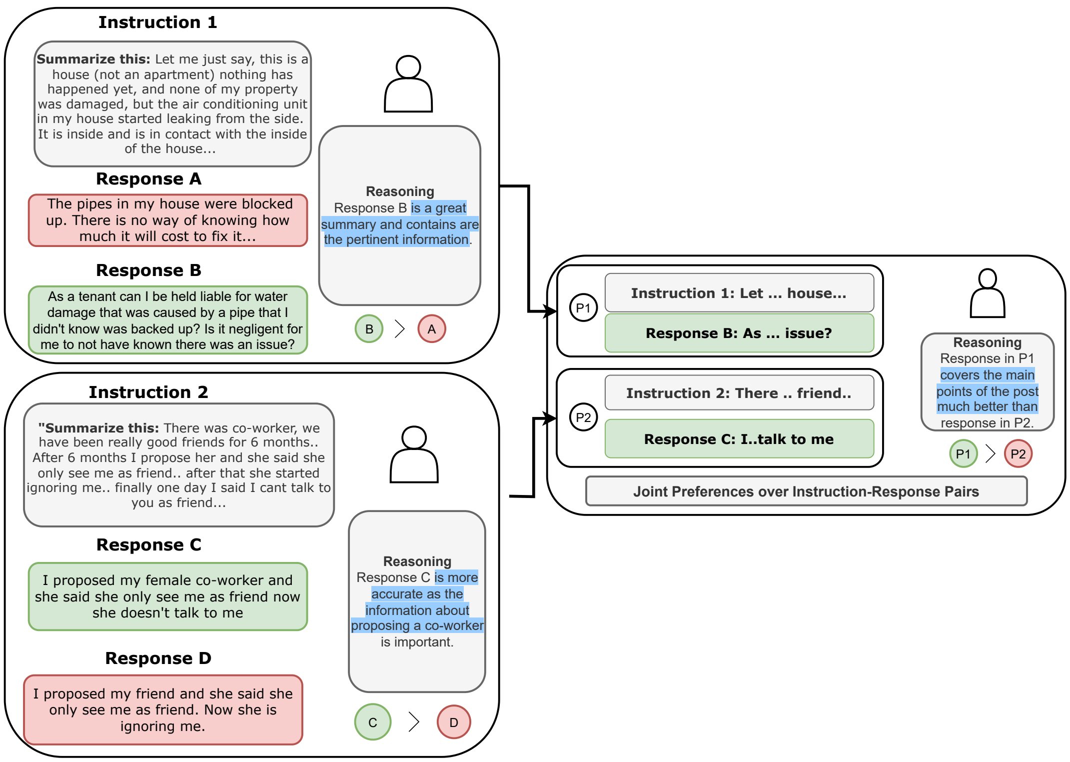


Figure 11. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response B and C are accepted under the conditional rankings. When asked to compare the response B and C, humans 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.

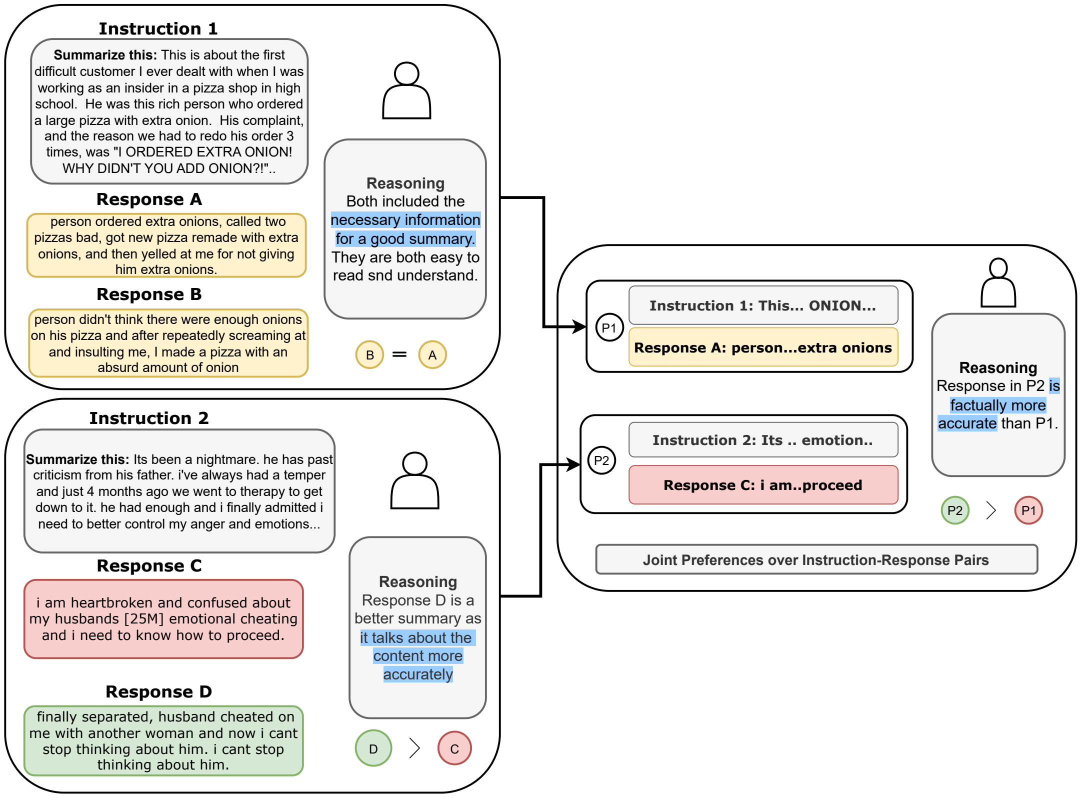


Figure 12. Interplay between the conditional rankings and joint rankings and reasoning acquired from the human annotators for the TL;DR summarization dataset. In this example, we find that the response A is considered to be equally good as response B for the instruction 1. In addition, response C is 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.

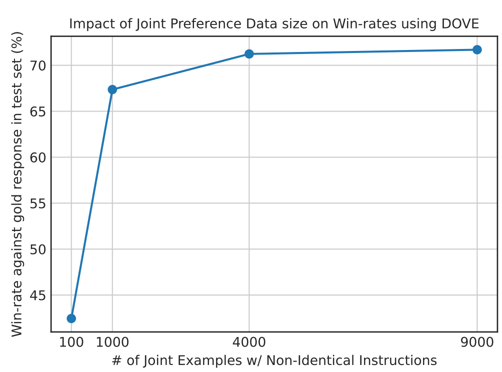


Figure 13. Results for scaling the feedback data size on TL;DR summarization dataset. We find that the win-rate improves with the increase in the dataset size using the DOVE preference optimization objective.