USING EYE-TRACKING TO UNDERSTAND AND ESTIMATE RELIABILITY OF PREFERENCE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Model (LLM) alignment methods like Direct Preference Optimization (DPO) rely on availability of high quality preference annotations, where annotators are asked to choose which of typically two responses they prefer. Often, it is assumed that these human-annotated labels are reliable in that they represent preferences well. However, these annotations may not reflect underlying preferences due to factors like labeling errors, task difficulty, biases introduced by the experimental setup, etc. While examining the impact of noisy labels on downstream tasks has gained attention in the literature, less emphasis has been placed in understanding where the noise comes from—how annotators complete the annotation tasks. In this paper, we use eye-tracking to examine how preference annotation is done and collect an eye-tracking dataset EYETRACKING-HS2-800 using Apple Vision Pro (including 5 annotations for approximately 800 tasks from the HelpSteer2–Preference dataset completed by 100 participants). Our results point to positional biases in the distribution of visual attention over the samples, and two distinct strategies for completing the annotation tasks emerge: "readers" display an increasing task completion time and "scanners" a relatively constant task completion time with the length of the shown sample. We then use these observations to modulate weight of individual human-annotated preference labels so that more carefully annotated samples get a higher weight (measured by proportion of fixated text and reading speed). By increasing the weight of the "reader" annotations, we recover, and on average even outperform, the performance of all HelpSteer2–Preference labels despite having significantly fewer annotations, outperforming all other training configurations. Based on our results, we provide guidelines for conducting and reporting preference annotation studies. In particular, we suggest incorporating reading speed, or time spent per task, as a suitable alternative to improve LLM alignment in the case of annotator disagreement.

1 INTRODUCTION

Pairwise preference annotation tasks wherein an annotator is asked to indicate which of typically two responses they prefer are commonly used in Large Language Model (LLM) alignment to ensure compliance with human preferences (Wang et al., 2025a;b; Ethayarajh et al., 2022; Chiang et al., 2024b; Zheng et al., 2023). Studies collecting preference labels from such tasks consider a few labels for each of the thousands of examples, and aggregate them using methods like majority voting. In absence of better estimates, human labels are often considered "gold standard", and it is is well recognized that data quality is important for LLM alignment results (Wang et al., 2025a), especially that of the preferred responses (Pan et al., 2025). That said, human annotations are not perfect, which has been recognized in recent literature examining their subjective and biased nature (Hosking et al., 2024) and the impact of potential label noise (Im & Li, 2026). Less emphasis has been placed in understanding the origins of any biases or noise: the moment preference labels are collected.

In this paper, we study how annotators complete preference annotations and examine whether this information reflects reliability of preference labels. To illustrate, consider two annotators labeling the same sample and assume that one of the annotators is in a rush and therefore carelessly skims the texts, while the other spends a significant amount of time reading them. In decision-making scenarios, speed–accuracy tradeoff is a well recorded phenomena, wherein rushing may lead to an increase in judgment errors (Wickelgren, 1977; Heitz, 2014). Similarly, correlation between atten-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion and selection has been observed in decision-making research (Shimojo et al., 2003; Krajbich, 2019). Following these observations, we examine whether understanding of how preference annotation tasks are completed reflects preference label reliability.

To this end, we use eye-tracking to understand the strategies annotators use to complete preference annotation tasks. Eye-tracking is a method to understand visual attention by tracking what an individual is looking at at a given time (Schall & Romano Bergstrom, 2014), and therefore it is widely used in tasks like reading (McConkie et al., 1988; Kliegl et al., 2004; Hollenstein et al., 2018; Bai et al., 2026), visual search (Zelinsky, 1996; Shen et al., 2000; Bailly et al., 2014; Jokinen et al., 2020), free-viewing (Leiva et al., 2020; Jiang et al., 2023b), and decision-making (Lohse & Johnson, 1996; Krajbich & Rangel, 2011; Stewart et al., 2016) to examine information processing. To understand how preference annotation is done, we collect a novel eye-tracking dataset EYETRACKING-HS2-800 using Apple Vision Pro, containing annotators ($N = 100$) completing pairwise preference annotation tasks (797 tasks with 5 annotations each) from the HelpSteer2–Preference (Wang et al., 2025a) dataset, and analyze how the annotators approach these tasks. Two main approaches for completing preference annotation tasks emerge in our dataset: *scanning* and *reading*, where trial completion times stay constant or increase, respectively, with response heights. Our empirical results also point to differences in labels between the original HelpSteer2 annotations and our study.

We then use two proxies of how carefully an individual annotator completed a task: 1) percentage of the texts the annotator looked at (derived from eye-tracking), and 2) reading speed (derived from task completion times). We use these metrics as weights in calculating majority votes, wherein a more carefully completed task gets a higher weight. This approach complements existing approaches using eye-tracking data directly as an implicit learning signal (Bondar et al., 2025; Lopez-Cardona et al., 2025b; Papadopoulos et al., 2026; Lopez-Cardona et al., 2025a). We evaluate these modified majority vote labels by fine-tuning Qwen2.5 (0.5B, 3B, and 7B Instruct variants) (Qwen et al., 2025) and Mistral (7B-Instruct-v0.2) (Jiang et al., 2023a) with DPO (Rafailov et al., 2023) and using AlpacaEval (Li et al., 2023), MT-Bench (Zheng et al., 2023), ArenaHard (Li et al., 2024; Chiang et al., 2024a), and response lengths in evaluation. While the observed effects are small, we note that models fine-tuned on our modified majority labels ($N = 797$) recover or outperform the performance of the full HelpSteer2 dataset with thousands of labels.

Our contributions are three-fold: 1) we analyze how annotators approach preference annotation in EYETRACKING-HS2-800, 2) suggest how to use such data to estimate reliability of preference labels, and 3) propose guidelines for preference label collecting and reporting based on our findings.

2 RELATED WORK

2.1 PREFERENCE ANNOTATION DATASETS FOR LLM ALIGNMENT

Popular LLM alignment methods like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017), or Direct Preference Optimization (DPO) (Rafailov et al., 2023) rely on high quality preference annotations (Wang et al., 2025a; Pan et al., 2025). Several crowd-sourced datasets containing human preference data have been collected to support LLM alignment. Such datasets include HH-RLHF for helpfulness/harmlessness (Bai et al., 2022), OpenAssistant (OASST1 and OASST2) with annotated conversation trees (Köpf et al., 2023), and HelpSteer/HelpSteer2/HelpSteer3 for multi-attribute annotations (Wang et al., 2024; 2025a;b). The annotations are often collected in crowd sourcing platforms and usually the preference labels and potentially a justification for the responses are recorded. Given the often crowd-sourced nature of these datasets where quality control may be challenging, assessing reliability of the labels is relevant.

2.2 EYE-TRACKING IN LLM ALIGNMENT

There is increasing interest in using eye-tracking data as an implicit signal in LLM alignment, including RLHF and DPO (Lopez-Cardona et al., 2025a; Bondar et al., 2025; Papadopoulos et al., 2026; Lopez-Cardona et al., 2025b; Galliamov et al., 2025). It is hypothesized that eye-tracking data carries information about the decision-making process beyond binary labels typically used in LLM fine-tuning. Eye-tracking data has been collected in pairwise preference annotation tasks involving both textual (Lopez-Cardona et al., 2025a; Bondar et al., 2025) and image (Papadopoulos et al., 2026) stimuli. Such work has revealed that, consistent with multi-attribute choice tasks

(Shimojo et al., 2003), effects like the gaze cascade where the chosen option is glanced last exist (Papadopoulos et al., 2026). Given the expenses associated in collecting eye-tracking data, another approach implemented in integrating eye-tracking to LLM alignment protocols involves using synthetic data; generative models are used in simulating human eye movements in different tasks. While some works have suggested that involving such data in reward modeling or RLHF protocols shows promise (Lopez-Cardona et al., 2025b; Galliamov et al., 2025), others display conflicting results (Bondar et al., 2025). This observation may be attributed to the differences observed between humans and synthetically generated data (Lopez-Cardona et al., 2025a) and to the fact that eye movement models are sensitive to in-domain fine-tuning (Bondar et al., 2025). Following these challenges, we propose another way of using eye-tracking data in assessing how reliable a label is.

3 EyeTracking-HS2-800 dataset

To understand how preference annotation is done, we collected eye-tracking data on Apple Vision Pro. We report the method of this study and empirical results.

3.1 Method

Participants. A total of 100 participants were recruited using an internal recruitment procedure (see Appendix C, Figure 7 for a breakdown of participant details). All participants were over 18. The compensation for the study lasting for about 60 minutes was a $12 food voucher. Vision corrections were facilitated through the use of lens inserts. The study was conducted according to institutional ethical guidelines. Informed consent was obtained prior to completing the study.

Materials. Texts shown in the preference annotation tasks were obtained from the HelpSteer2–Preference dataset (Wang et al., 2025a). This dataset was selected as the annotations are relatively high quality as measured by inter-annotator agreement, the prompts are obtained from ShareGPT (ShareGPT, 2023) and are therefore fairly realistic, and the dataset is relatively well known in the community. We only considered responses that could fit on the display without scrolling. We used the original HelpSteer2 dataset (Wang et al., 2024)—containing annotations for helpfulness, correctness, coherence, complexity, and verbosity—to filter out responses with low correctness ($<$ 2 corresponding to the response containing mostly wrong information) as the participants could not search the internet during the study. We also filtered out responses with non-Latin characters (retaining glyphs and emojis). Our final dataset contains 800 individual tasks randomly sampled from the Helpsteer2–Preference dataset under the length and correctness constraints (sampled from $N \approx 1,310$ tasks). The rating instructions were adapted from the HelpSteer2–Preference setup to account for the filtering done for correctness and to make them shorter (see Appendix B for full instructions). In contrast to the original study, we did not enforce tie-breaks.

Task. Each task (Figure 1) contained three steps displayed on a virtual screen on Vision Pro: 1) a **Prompt** from HelpSteer2 is shown, followed by 2) two **Responses** (named A and B) displayed side-by-side horizontally, and 3) a seven-step **Rating** scale. Participants could navigate from one screen to the next (no backtracking allowed) by pressing a button labeled "Proceed" at the bottom of each screen using gaze-and-pinch gestures on Vision Pro. The gaze-based selection also fixed the starting position of each trial to bottom–center of the virtual screen as the navigation elements were shown in this location. The participants were allowed to look at each screen without a time limit.

Experiment design. We collected five annotations for each of the 800 tasks (4,000 annotations in total). These 800 annotations were divided into batches of 40, and each batch was shown to five participants in a within–subject manner. Within each batch, the order in which the tasks were presented to the five participants was randomized. The location of each response on screen was randomized within each task to avoid position biases. The batches were further divided into four blocks of 10 tasks with the intention of having a break after each batch to mitigate fatigue. At the start of every block, a centrally positioned fixation cross was shown for 10 s for calibration purposes.

Apparatus. The study was conducted using a Vision Pro device. Optical inserts were offered to participants who required vision corrections. The eye-tracking system in Vision Pro accounted

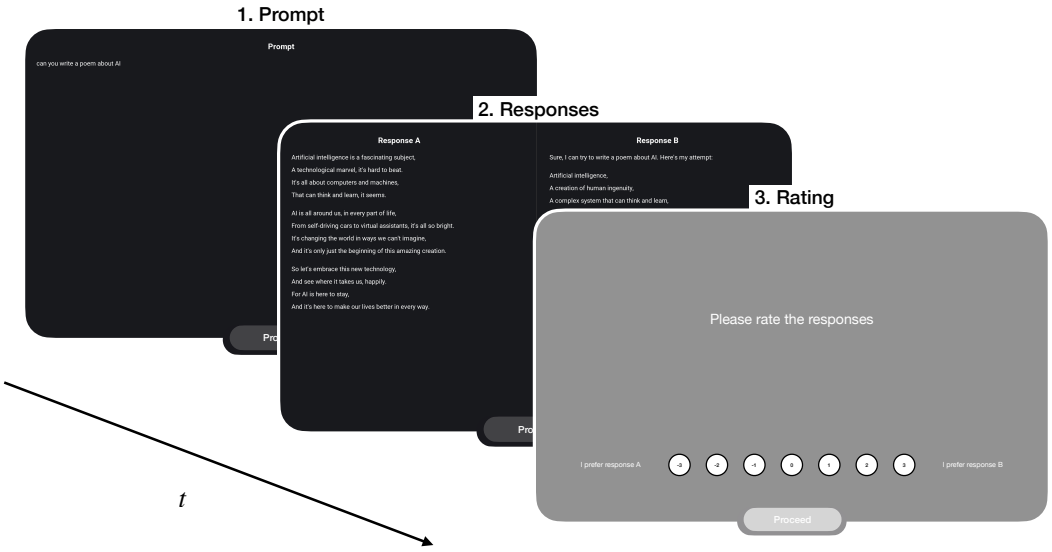


Figure 1: **Illustration of the preference annotation task shown on Vision Pro headset.** Each task began with a prompt shown to the participant (pane 1), followed by two responses (pane 2). Then, the participant indicated which response they preferred (pane 3). Participants navigated between screens using gaze-and-pinch gestures on Vision Pro.

for distortions caused by optical inserts when present. The participant was instructed to wear the headset while sitting on a chair that faced a wall at a 2 m distance. On this wall, a set of four markers displayed a 191 cm $\times$ 109 cm area the virtual screen on Vision Pro should be aligned to, in order to standardize the screen size across participants.

Procedure. Before the experiment, the participants signed an informed consent form. They were familiarized with the task and annotation instructions (see Appendix B for the full annotation instructions), as well as how to operate the Vision Pro device. Participants requiring vision corrections were given optical inserts. Each participant followed the standard Eyes & Hands setup procedure on Vision Pro in order to calibrate the eye-tracking and input gesture systems. After the calibration was completed, the participants were instructed to complete a test trial and to ask any remaining questions about the protocol. For each participant, eye-tracking data was collected in four blocks of 10 tasks. Each block started with a fixation screen where participants were asked to look at a centrally positioned cross on the screen for 10 s. The facilitator monitored each trial to ensure the participant successfully completed the tasks. They were also instructed to monitor task completion times to catch any participants who were not engaged with the task. After each block, the participants could self-administer a break and access the annotation instructions. At the end of the study, the participants were asked to fill a questionnaire, where they were asked to justify how they generally selected which response was better and to provide any other comments on the experiment.

Data processing and analysis. The (x, y, z) coordinates of the gaze stream obtained from Vision Pro were transformed to two dimensional gaze data using internal software. The (x, y) coordinates were rendered on a video of what the participant saw in their field of view at a given timestamp. Within this video, the virtual display showing the stimuli was visible in an orientation that depended on the participant's head position (i.e., it could appear as a trapezoid instead of a rectangle). Therefore, using a perspective transform from the `opencv-python` library (OpenCV Contributors, 2024), each gaze point was mapped to a location on the rectangular images displaying the tasks. The two-dimensional and perspective transformed gaze stream data was converted to fixations–periods when the eyes are still and visual information processed—using an I-VT fixation filter (Salvucci & Goldberg, 2000). Angular velocities were calculated based on the viewing distance and physical screen size of the virtual display rendered on the Vision Pro. A velocity threshold of 20° was used, and we assumed a minimum fixation duration of 45 ms given the preference annotation task is expected to involve reading. We estimated the fovea, area in the retina with the most

detailed vision, to cover a radius of 86 px based on the viewing distance of 2 m, and resolution of the images as 2300 px $\times$ 1400 px shown at a physical size of 191 cm $\times$ 109 cm (spanning 2° of visual angle). This information was used to estimate the percentage of the text a participant glanced.

Dataset summary. After data collection, we excluded seven participants who were unable to complete the data collection, or had formatting errors in the data. The final dataset consists of 3,982 trials (out of the 4000 targeted, leading to a data loss of 0.45 %), with approximately 570,000 fixations. The dataset then covered 797 of the 800 targeted responses.

3.2 Empirical results

Next, we report on how participants in EyeTracking-HS2-800 chose which response they prefer.

Preference labels. Our preference labels do not show significant bias towards either response in raw or majority vote preference labels, and there is also no apparent position bias (see Figure 8 in Appendix D). On the other hand, HelpSteer2–Preference labels displays a bias towards Response 2, which is the option presented second (Wang et al., 2025a).

Label agreement between HelpSteer2 and our data. We compare label agreement between EyeTracking-HS2-800 and HelpSteer2–Preference by matching the mean aggregated labels and majority vote labels by task. There is a moderate positive correlation between the mean labels (Pearson $r = 0.431$, Figure 2, pane A), while the majority vote labels display fair agreement (Cohen's $\kappa = 0.0644$, Figure 2, pane B).

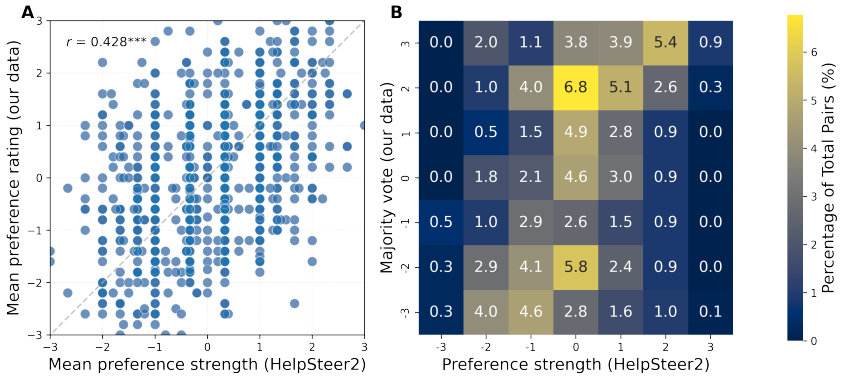


Figure 2: **Labels in HelpSteer2–Preference and EyeTracking-HS2-800.** Mean ratings are moderately correlated (pane A) and majority votes show fair agreement (pane B).

Adherence to annotation instructions. The participants were asked to justify how they approached the preference annotation tasks, generally. A qualitative analysis of the justifications indicates that participants highly valued instruction-following or correctness of the responses. Formatting played a large role in their preference selection as well. Participants frequently indicated a preference of short and concise responses over lengthy ones.

Task completion times. Annotation completion times are right-skewed (mean 57.1 s with $SD = 231.90$, see Appendix D, Figure 9, pane A). Most of this time is spent on looking at responses ($\mu = 35.33$ s, $SD = 24.82$), followed by prompts ($\mu = 18.37$ s, $SD = 21.54$). The preference rating is completed relatively fast ($\mu = 3.37$ s, $SD = 3.86$). Average time for completing all the blocks in the experiment (excluding breaks) is 37.7 minutes ($SD = 12.7$ min, Figure 9, pane B).

Distribution of attention over the samples. Our results point to at least two types of biases in the allocation of attention over the samples. Firstly, participants tend to look at the response presented on the left for longer (Figure 3). Additionally, the participants tend to allocate more attention towards the top-left corner of each response, which may relate to distribution of text over the virtual display.

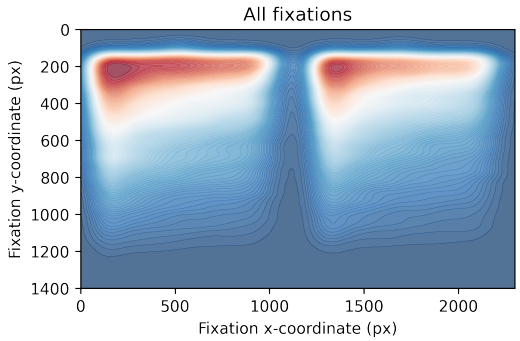


Figure 3: **Allocation of attention over all samples in EYETRACKING-HS2-800.** Participants tend to look at the response presented on the left, and the upper left corner of the screen more.

Readers and scanners. Given the wide range of task completion times, we inspect the relationship between the response heights and how long it takes the participants to complete a task. Our analysis suggests that some participants display a relatively constant trial completion time regardless of the maximum height of either of the responses, while others spend more time looking at tasks where either of the responses spans a larger space (Figure 10 in Appendix D, where 4.5 s / 100 px is used as the threshold for the categories to create approximately equally sized bins). This suggests that different participants may take different annotation approaches, some people scanning the responses in a constant manner, while others spend time reading the responses. Percentage of the text a participant foveated on is higher in the reader group (Figure 12 in Appendix D). Similarly, allocation of attention seems to be more evenly spread across lines of text and over the entire sample in the reader group (Figure 11 in Appendix D).

4 FINE-TUNING EXPERIMENTS

Our empirical results suggest that annotators allocate varying amount of time and look at different proportion of the responses when choosing. Therefore, we evaluate reliability of an individual preference label following the idea of speed-accuracy tradeoff (Wickelgren, 1977; Heitz, 2014), which suggests rushing a task decreases quality.

4.1 METHOD

We propose giving reliable annotations a higher weight when aggregating majority votes. Two sources for the weights are considered from empirical data: (1) the proportion of the text a participant looked at or *coverage* (obtained from eye-tracking) and (2) reading speed (obtained from task completion times and measured as characters per second). We assess instruction-following via AlpacaEval2.0 (Li et al., 2023; Dubois et al., 2024), MT-Bench (Zheng et al., 2023), Arena-Hard (Li et al., 2024; Chiang et al., 2024a), TruthfulQA (Lin et al., 2022), and MT-Bench response length across several datasets. We consider the following two variants of the HelpSteer2–Preference dataset: 1) the complete dataset with original labels ($N \approx 7,000$); and 2) the original labels corresponding to tasks used in our study ($N = 800$). We consider the following three variants of the EYETRACKING-HS2-800: 3) simple majority vote labels ($N = 797$), 4) coverage weighted majority vote labels ($N = 797$), and 5) reading speed weighted majority vote labels ($N = 797$). Coverage naturally falls between 0 and 1. We normalize a participant's reading speed relative to the average reading speed across the population of participants and then take the inverse. Therefore, a participant who reads slower than average will have a higher weight than someone who reads faster than the average.

Experiments are run with several different LLMs from the family of Qwen2.5 Instruct models (0.5B, 3B, 7B) along with Mistralv0.2 7B Instruct. Training was conducted using TRL (von Werra et al., 2020) (the learning rate, β, LORA r, and LORA alpha values were selected per model using grid search and are reported in Table 4 in Appendix E). The hyper-parameters found to have the highest accuracy on the HelpSteer2 validation set were selected, and the resulting models were evaluated.

Condition	AlpacaEval 2.0	Arena Hard	MT-Bench	**Rank (avg)**	Length (MT-Bench)
Base	61.7 ± 13.2	3.5% ± 1.9	6.3 ± 0.7	–	398.5 ± 25.9
HS2-full	-12.3 ± 3.7	+0.1% ± 0.4	+0.0 ± 0.1	3 (2.7)	+7.3 ± 7.6
HS2-800	-16.3 ± 2.6	-0.1% ± 0.2	-1.0 ± 0.1	5 (3.7)	-1.3 ± 2.3
ET-HS2	-10.9 ± 4.3	-0.4% ± 0.4	+0.1 ± 0.0	4 (3.0)	-3.7 ± 2.8
ET-HS2-coverage	-5.6 ± 3.2	-0.1% ± 0.3	-0.0 ± 0.1	2 (2.3)	-8.1 ± 4.3
ET-HS2-speed	-0.5 ± 1.0	-0.3% ± 0.4	+0.0 ± 0.1	**1** (2.0)	-7.5 ± 5.7

Table 1: Mean ± SE across LLMs per training condition relative to the base LLM. Cell shading reflects the magnitude of difference from Base, normalized per column, with sign encoded as orange (decrease) or blue (increase). The Rank column is not shaded by magnitude. AlpacaEval: LC win rate. Arena-Hard-Auto: win rate. MT-Bench: mean judge score. Length (MT-Bench): average MT-bench response length in tokens.

4.2 Fine-tuning results

We find that in general the LLMs learn to model the HelpSteer2–Preference data as indicated by test set accuracy values above 0.50. Comparing performance between the original HelpSteer2–Preference labels and those collected in this study, we see a general trend that test accuracy decreases (Table 3). This points to a mismatch in the preference distribution from our participants relative to the HelpSteer2 participants (also indicated in Figure 2). Based on the instruction-following performance reported in Table 1, the difference in preference distribution is not a limiting factor.

For all benchmarks that rely on a LLM judge, `gpt-4.1-2025-04-14` was used. All evaluated-LLM responses are sampled greedily ($\tau = 0$ and top-p=1.0). We found that scores depend heavily on sampling strategy, so we use this as a consistent sampling strategy to have a fair comparison across models, although it decreases benchmark performance overall. The mean and standard error in performance are reported across LLMs per benchmark in Table 1. Performance is reported as gains relative to the base model. The different dataset conditions do not impact TruthfulQA performance, so those results and the raw performance per LLM are reported in Appendix E, Table 5.

In general, the ET-HS2 preference labels improve instruction-following performance in comparison to the 800 labels from the HS2-800 condition, but there is no immediately obvious trend indicating which of our label aggregation strategies is best. However, when we rank each of the conditions based on average distance from the base model per task,[1] we find that the reading speed weighted label aggregation outperforms all other training configuration, including training on the full HelpSteer2-Preference data, despite relying on significantly fewer preference data points. Raw numbers, including corresponding ranks, are reported in Appendix E, Table 5. This result suggests that when the labels from participants who would be classified as readers are up weighted, the instruction-following quality of the DPO'ed model improves.

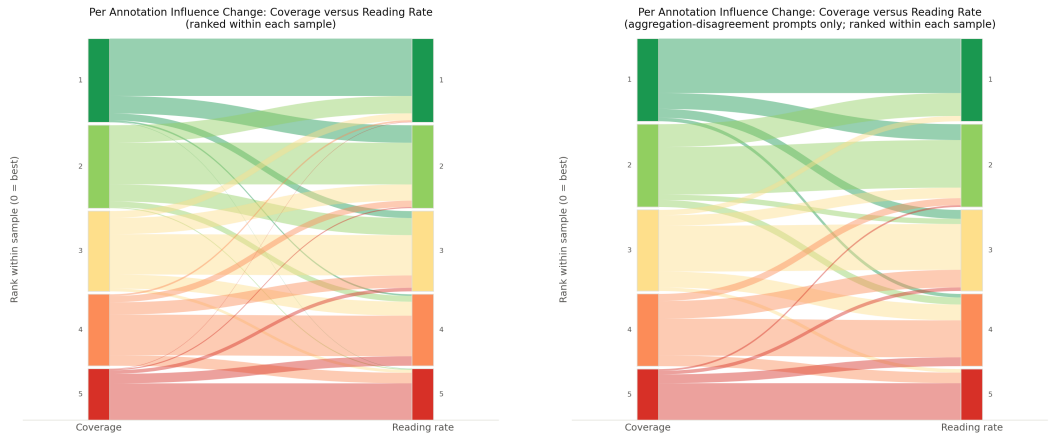
4.3 Preference label disagreements between weighting schemes

The fine-tuning results in Section 4.2 demonstrate that the strategy taken to weight the EyeTracking-HS2-800 preference labels (coverage vs. reading speed) impacts the instruction-following ability of the DPO'ed model. It is reasonable to assume that these differences stem from conflicting aggregated preference labels across the weighting schemes (what we call the "disagreement subset"). Generally the correlation between the aggregated preference labels across weighting schemes are strong, the highest correlation occurring between the unweighted and coverage schemes ($r = 0.97$), followed by coverage and reading speed ($r = 0.95$), and unweighted and reading speed (r = 0.92). Disagreements between the weighting schemes tend to occur in instances where inter-annotator agreement is lower ($\kappa \approx 0.25$ vs. $\kappa \approx -0.10$ in agreement and disagreement subsets, respectively, see Figure 14b in Appendix E). Intuitively, this makes sense as in cases where annota-

[1]We exclude the Length condition from our average, as it is unclear how "better" performance should be defined in this case.

tors largely agree on the appropriate label, the weighting strategy is irrelevant. The two weighting strategies, however, resolve disagreements differently.

To understand these differences we examine the relationship between the coverage and reading speed signals. We find moderate correlation between all coverage and normalized reading speed weights ($r = 0.51$) and note that when a preference label is weighted high in coverage, the reading speed weight tends to also be high (Figure 4a and Figure 13a in Appendix E). However, it is also common for a label to decrease in influence when weighted with reading speed versus coverage ($\approx 40\%$ of samples) and vice versa, suggesting that the two metrics are measuring different aspects of annotation quality. Notably, looking back at our reader versus scanner annotator identities (Section 3.2), we find that scanners are weighted higher when coverage is the determining factor, whereas readers are weighted higher when reading speed is used. In particular, we ranked all annotations and find that according to coverage, scanners have a rank that is on average larger than readers by 0.3 and according to reading speed, readers have a rank that is larger on average by 0.4 (statistically significant at $t = 2.39; p < 0.05$). This suggests reading speed prioritizes readers and coverage scanners. However, jumps in influence are less prominent in the disagreement subset: for instance, no weight jumps from being the most influential in coverage to least influential in reading speed (Figure 4b and Figure 13b in Appendix E). This implies that in the disagreement subset, individual annotators do not display tendencies like looking at a small part of the responses (low coverage weight) for an extended period of time (high reading speed weight) relative to others, or vice versa. Combining these findings with our finetuning results in which we found that reading speed weighted label aggregation outperformed all other settings, we recommended reading speed (or time spent per task) as a method to resolve inter-annotator disagreements. After samples have been cleaned to account for instances of annotator disengagement with the task (e.g., stepping away for a break), majority vote should be computed relative to annotator reading speed, especially for samples with low inter-annotator agreement. We refer to Section 5.1 for an extensive list of recommendations for collecting preference labels based on our results.



(a) All Annotations

(b) Disagreement Subset

Figure 4: **The change in an annotation's rank within a preference sample (i.e., prompt + response pair) when its weight is determined by coverage vs. reading speed.** Moving from a lower bin on the left (coverage) to a higher bin on the right (reading speed) means an annotation became more influential when assessed by reading speed instead of coverage. The change in rank for all annotations (pane a) is compared to the change in rank for those preference samples (187 annotations) where the different weighting schemes induced different preference labels (pane b).

5 Discussion and Recommendations

Our results give important insights into how people complete preference labeling tasks: we find a clear distinction between readers and scanners, i.e., people who tend to read (most of) the response before deciding on their preference, and people who base their response on a more general overview

of the text. We find that readers' annotations are more reliable during fine-tuning. Based on these results, we discuss recommendations for collecting preference annotations in this section.

5.1 Guidelines for collecting preference labels

- **Task completion times should be collected:** Our results indicate that coverage and especially reading speed are useful indicators of preference label reliability. While collecting eye-tracking data for estimating coverage is costly, reading speeds are estimated from task completion times, which are typically available through crowd sourcing.
- **Ties and disagreement should be allowed:** While the original HelpSteer2 dataset does not allow ties in the preference labels, 31% of our collected preference labels are ties. Therefore, we recommend avoiding forcing preferences and collecting ties. Additionally, outliers are sometimes removed in preference annotation datasets. We propose exploring alternatives like the proposed weightings to resolve disagreements.
- **The layout used in the protocol should always be reported:** We note a difference in the preference labels across the original HelpSteer2 labels and our data. While this observation may be explained by individual differences, decision-making literature suggests this can result from differences in presentation of the options (Cataldo & Cohen, 2019). Therefore, illustration of the experimental setup should be reported for replication purposes.
- **Response presented on the left should be counterbalanced:** Our results point to a left bias in visual attention. Therefore, if the responses are presented side-by-side, the sample shown on the left hand side should vary from one annotator to another.

Our results are also important in settings where it is difficult to collect additional preference labels, but additional behavioral data can be accessed. This scenario may occur in, for instance, a personalization context, when labels on individual responses at scale may be unavailable. However, it could be possible to record task completion times, and even eye-tracking may be available in some cases (e.g., virtual reality headsets).

5.2 Future work and limitations

We choose a subsample from HelpSteer2 as a representative preference labeling set, but we encourage the replication of the study for more samples from HelpSteer2 or other preference annotation datasets to assess the generalizability of the results, also in screen-based environments in addition to the current augmented reality setup. Similarly, different ways to use the eye-tracking data to analyze reliability could be investigated. For example, the fixations were estimated using a relatively simple I-VT algorithm (Salvucci & Goldberg, 2000), whereas another filter could lead to better classifications of fixation points, especially if reading is involved.

6 Conclusion

In this work we used eye-tracking to understand and estimate the reliability of preference labels for LLM alignment. We find two patterns annotators display when completing the preference labeling task: *scanners* look at the responses for a relatively constant amount of time regardless of the response length, while *readers* display an increasing task completion time. We also note attentional biases, like tendency to allocate more visual attention on the left response. These patterns imply that some participants make choices with less information about the texts than others. We tested whether such attentional patterns, namely percentage of the texts an annotator looked at and reading speed, can be used to indicate reliability of individual labels when majority votes are aggregated. Taking these patterns into account, especially reading speed, we were able to recover and even improve fine-tuning performance of a significantly higher number of preference labels. As task completion time can be used as a proxy for reading speed, this serves as an accessible alternative for resolving disagreements and improving LLM alignment in future work.

AI USE STATEMENT

In this work, generative AI tools were used for designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping to develop theoretical models or conceptual frameworks, proposing or refining hypotheses, or cleaning and reformatting datasets. Formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for formulating questions for surveys or interviews, creating or modifying scientific figures or images, creating or editing software code, editing a research paper to improve readability, identifying relevant literature, and formatting references. All AI-assisted work has been reviewed by the authors and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The studies involving human subjects were conducted according to local and institutional procedure for ethics approval.

REFERENCES

Yunpeng Bai, Xiaofu Jin, Shengdong Zhao, and Antti Oulasvirta. Hierarchical resource rationality explains human reading behaviour. *Nature Human Behaviour*, pp. 1–17, 2026.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Gilles Bailly, Antti Oulasvirta, Duncan P. Brumby, and Andrew Howes. Model of visual search and selection time in linear menus. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 3865–3874, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450324731. doi: 10.1145/2556288.2557093. URL https://doi.org/10.1145/2556288.2557093.

Anna Bondar, David Robert Reich, and Lena Ann Jäger. AlEYEgnment: Leveraging Eye-Tracking-While-Reading to align language models with human preferences. In Cengiz Acarturk, Jamal Nasir, Burcu Can, and Cagrı Coltekin (eds.), *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pp. 58–70, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, BULGARIA. URL https://aclanthology.org/2025.gaze4nlp-1.8/.

Andrea M Cataldo and Andrew L Cohen. The comparison process as an account of variation in the attraction, compromise, and similarity effects. *Psychonomic Bulletin & Review*, 26:934–942, 2019.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024a.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024b. URL https://arxiv.org/abs/2403.04132.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Karim Galliamov, Ivan Titov, and Ilya Pershin. Enhancing RLHF with human gaze modeling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30637–30643, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1559. URL https://aclanthology.org/2025.emnlp-main.1559/.

Richard P. Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, Volume 8 - 2014, 2014. ISSN 1662-453X. doi: 10.3389/fnins.2014.00150. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2014.00150.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):180291, 2018.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 55864–55883, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f719b8ca81221f380836e573d479c223-Paper-Conference.pdf.

Shawn Im and Yixuan Li. How well can preference optimization generalize under noisy feedback? *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=8f5gRWwzDx.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a. URL https://arxiv.org/abs/2310.06825.

Yue Jiang, Luis A. Leiva, Hamed Rezazadegan Tavakoli, Paul R. B. Houssel, Julia Kylmälä, and Antti Oulasvirta. Ueyes: Understanding visual saliency across user interface types. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581096. URL https://doi.org/10.1145/3544548.3581096.

Jussi P.P. Jokinen, Zhenxin Wang, Sayan Sarcar, Antti Oulasvirta, and Xiangshi Ren. Adaptive feature guidance: Modelling visual search with graphical layouts. *International Journal of Human-Computer Studies*, 136:102376, 2020. ISSN 1071-5819. doi: https://doi.org/10.1016/j.ijhcs.2019.102376. URL https://www.sciencedirect.com/science/article/pii/S1071581919301429.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284, 2004. doi: 10.1080/09541440340000213. URL https://doi.org/10.1080/09541440340000213.

Ian Krajbich. Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29:6–11, 2019. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2018.10.008. URL https://www.sciencedirect.com/science/article/pii/S2352250X18301866. Attention & Perception.

Ian Krajbich and Antonio Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857, 2011. doi: 10.1073/pnas.1101328108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1101328108.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Luis A. Leiva, Yunfei Xue, Avya Bansal, Hamed R. Tavakoli, Tuğçe Köroğlu, Jingzhou Du, Niraj R. Dayama, and Antti Oulasvirta. Understanding visual saliency in mobile user interfaces. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375160. doi: 10.1145/3379503.3403557. URL https://doi.org/10.1145/3379503.3403557.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Gerald L. Lohse and Eric J. Johnson. A comparison of two process tracing methods for choice tasks. *Organizational Behavior and Human Decision Processes*, 68(1):28–43, 1996. ISSN 0749-5978. doi: https://doi.org/10.1006/obhd.1996.0087. URL https://www.sciencedirect.com/science/article/pii/S0749597896900874.

Angela Lopez-Cardona, Sebastian Idesis, Miguel Barreda-Ángeles, Sergi Abadal, and Ioannis Arapakis. Oasst-etc dataset: Alignment signals from eye-tracking analysis of llm responses. *Proc. ACM Hum.-Comput. Interact.*, 9(3), May 2025a. doi: 10.1145/3725840. URL https://doi.org/10.1145/3725840.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models, 2025b. URL https://arxiv.org/abs/2410.01532.

G.W. McConkie, P.W. Kerr, M.D. Reddix, and D. Zola. Eye movement control during reading: I. the location of initial eye fixations on words. *Vision Research*, 28(10):1107–1118, 1988. ISSN 0042-6989. doi: https://doi.org/10.1016/0042-6989(88)90137-X. URL https://www.sciencedirect.com/science/article/pii/004269898890137X.

OpenCV Contributors. opencv-python: Wrapper package for opencv python bindings. https://github.com/opencv/opencv-python, 2024. Accessed: 2026-04-09.

Yu Pan, Zhongze Cai, Huaiyang Zhong, Guanting Chen, and Chonghuan Wang. What matters in data for dpo? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 44689–44716. Curran Associates, Inc., 2025. doi: 10.52202/085713-1488. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3f37b8fbd43303106dd141a602838ad5-Paper-Conference.pdf.

Nikolas Papadopoulos, Shreenithi Navaneethan, Sheng Bai, Ankur Samanta, and Paul Sajda. Gaze patterns predict preference and confidence in pairwise AI image evaluation, 2026. URL https://arxiv.org/abs/2603.24849.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pp. 71–78, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132808. doi: 10.1145/355017.355028. URL https://doi.org/10.1145/355017.355028.

Andrew Schall and Jennifer Romano Bergstrom. 1 - introduction to eye tracking. In Jennifer Romano Bergstrom and Andrew Jonathan Schall (eds.), *Eye Tracking in User Experience Design*, pp. 3–26. Morgan Kaufmann, Boston, 2014. ISBN 978-0-12-408138-3. doi: https://doi.org/10.1016/B978-0-12-408138-3.00001-7. URL https://www.sciencedirect.com/science/article/pii/B9780124081383000017.

ShareGPT. Share your wildest chatgpt conversations with one click. https://github.com/domeccleston/sharegpt, 2023.

Jiye Shen, Eyal M Reingold, and Marc Pomplun. Distractor ratio influences patterns of eye movements during visual search. *Perception*, 29(2):241–250, 2000. doi: 10.1068/p2933. URL https://doi.org/10.1068/p2933. PMID: 10820604.

Shinsuke Shimojo, Claudiu Simion, Eiko Shimojo, and Christian Scheier. Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322, 12 2003. ISSN 1546-1726. doi: 10.1038/nn1150. URL https://doi.org/10.1038/nn1150.

Neil Stewart, Frouke Hermens, and William J. Matthews. Eye movements in risky choice. *Journal of Behavioral Decision Making*, 29(2-3):116–136, 2016. doi: https://doi.org/10.1002/bdm.1854. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.1854.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501. Curran Associates, Inc., 2024. doi: 10.52202/079017-0047. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/02fd91a387a6a5a5751e81b58a75af90-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=MnfHxPP5gs.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated

preference data across diverse tasks and languages, 2025b. URL https://arxiv.org/abs/2505.11475.

Wayne A. Wickelgren. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1):67–85, 1977. ISSN 0001-6918. doi: https://doi.org/10.1016/0001-6918(77)90012-9. URL https://www.sciencedirect.com/science/article/pii/0001691877900129.

Gregory J. Zelinsky. Using eye saccades to assess the selectivity of search movements. *Vision Research*, 36(14):2177–2187, 1996. ISSN 0042-6989. doi: https://doi.org/10.1016/0042-6989(95)00300-2. URL https://www.sciencedirect.com/science/article/pii/0042698995003002.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDICES

A COMPARISON OF EYETRACKING-HS2-800 SAMPLE TO HELPSTEER2–PREFERENCE

As described in Section 3.1–Materials, the prompt/response pairs selected for our experiment are length-restricted to fit on the virtual display without scrolling, and contain only responses that are rated to be correct (correctness value > 2 in the original HelpSteer2 dataset of Wang et al. (2024)). Therefore, in contrast to the samples in the original HelpSteer2–Preference dataset (Wang et al., 2025a), the ones presented in our study tend to be shorter (Figure 5). Likely attributed to this, the responses selected to our study also tend to have lower complexity scores (Figure 6).

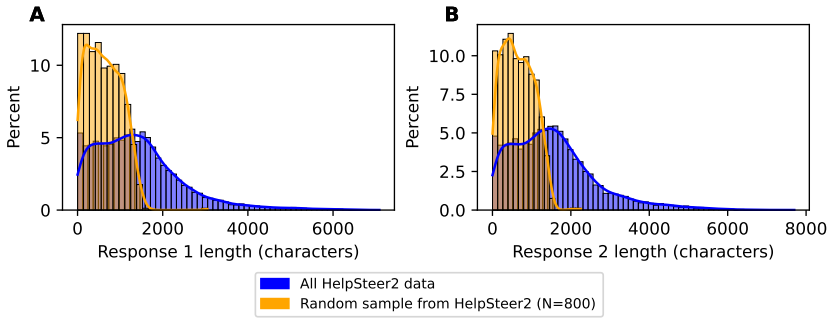


Figure 5: **Distributions of response character counts in all HelpSteer2–Preference responses (Wang et al., 2025a) and our random sample.** Our sample contains shorter responses to fit all of the text on the screen without scrolling.

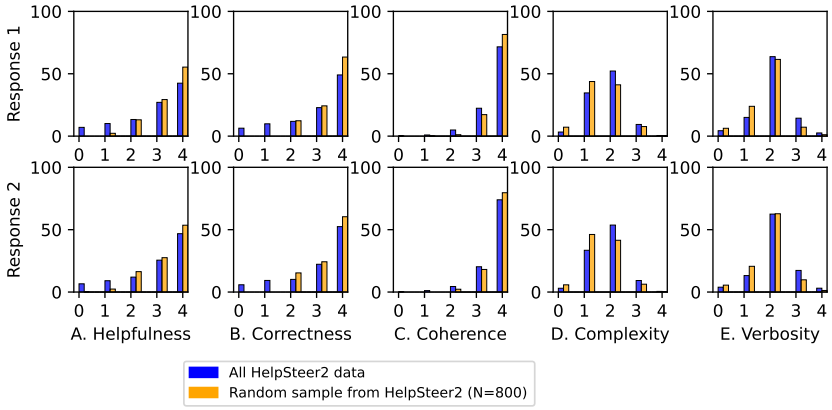


Figure 6: **Distributions of the HelpSteer2 attributes from Wang et al. (2024) in comparison to those of random sample.** The distributions are overall similar. Probably attributed to the length restrictions, responses in our sample are less complex. Additionally, responses with low correctness (< 2) are filtered out.

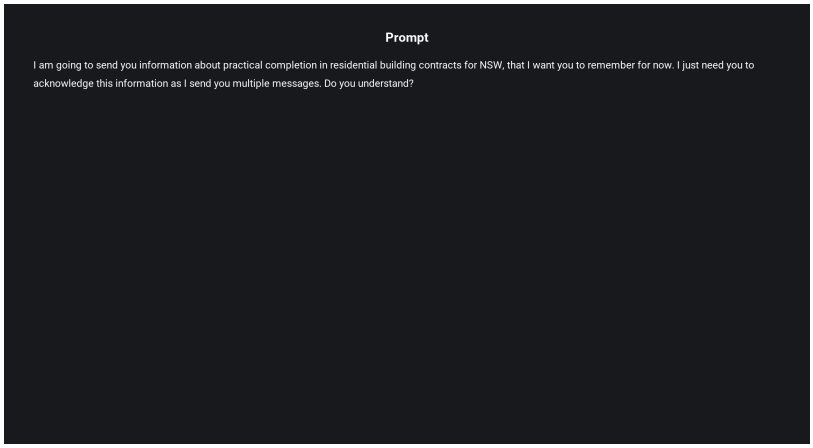
B ANNOTATION INSTRUCTIONS

The annotation instructions were presented to the participants at the beginning of the study, and they were able to access them between the blocks. The instructions were adapted from those presented in HelpSteer2–Preference (Wang et al., 2025a) by simplifying the language and dropping correctness as a feature to assess as the participants were not able to search the internet during the study.

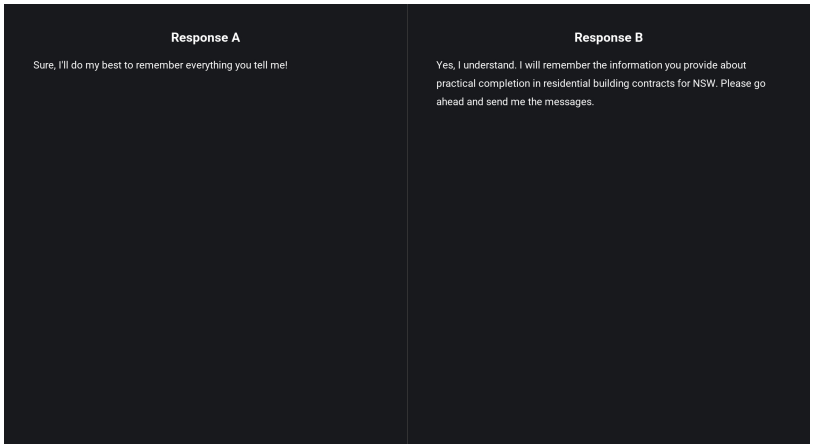
Task instructions

Each task you will complete consists of the following steps:

1. You are shown a Prompt that describes a task that a user wants a large language model to complete.



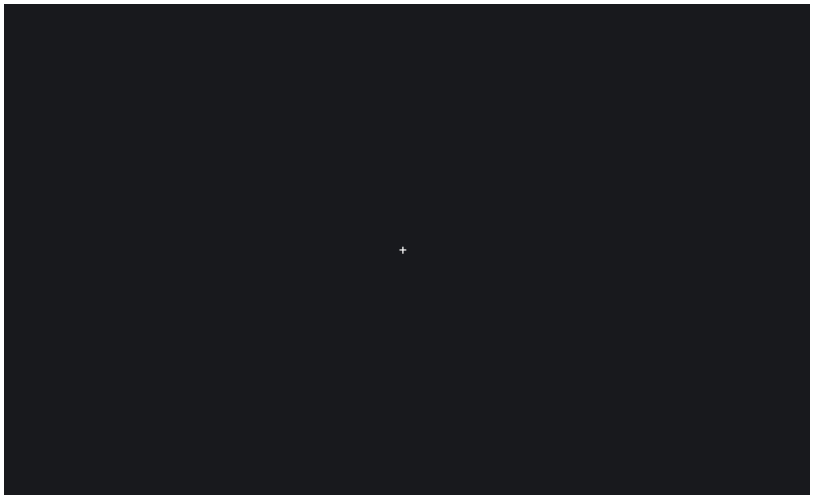
2. You are shown two responses (A and B) that contain a response to the prompt.



3. You have to rate which response you prefer on a scale from -3 (Response A is strongly preferred) to 3 (Response B is strongly preferred). Please refer to the instructions in the following pages that instruct how to choose the rankings.



4. You complete 10 of these tasks in one “block”.
5. You will navigate between the windows by fixating on a *Proceed* button and tapping your fingers together (default Apple Vision Pro input gesture).
6. After each block you can have a break or continue to the next block.
7. Before each block you are also asked to perform a calibration wherein you fixate on a cross placed in the middle of the window for 3 seconds. A fixation cross is illustrated below.



You will complete a test block of tasks to familiarize yourself with these instructions.

Ranking instructions

Evaluate characteristics of the responses in the following priority order. When responses differ across multiple characteristics, favor the one that wins on the **highest priority** characteristic.

You can assume all responses are factually correct.

Instruction Following **(Highest Priority)**

The response must address everything asked in the prompt — both explicit and implicit requirements.

Example: "*Plan a 5-day surfing trip to Portugal*" should be evaluated on:

- Is it an itinerary?
- Is it 5 days and in Portugal?
- Does it focus on surfing?

Note: If the prompt requests a *specific format* (e.g. JSON, a table), treat that as an instruction-following requirement, not a formatting one.

Formatting **(Medium Priority)**

When no specific format is requested, prefer responses that use formatting appropriately and helpfully. This includes:

- Splitting an itinerary by day.
- Using markdown tables or lists where they add value.
- Using code blocks for scripts or commands.

Clarity **(Lower Priority)**

Prefer responses that are well-written and easy to understand. Penalize responses that:

- Are unnecessarily verbose.
- Include irrelevant information.
- Lack explanation or context where it would help.

Ranking strengths

We would like to collect the following levels of preference, described below:

-3 Response A is **much better** than Response B.

-2 Response A is **better** than Response B.

-1 Response A is **slightly better** than Response B.

0 The responses are **equally** good.

1 Response B is **slightly better** than Response A.

2 Response B is **better** than Response A.

3 Response B is **much better** than Response A.

To determine whether a response is much better, better, slightly better or equal, please follow the following guidelines:

A response that is much better (level -3 or 3)
- Follows all instructions, while the worse response does not.
- Is clearly better in readability or formatting.

A response that is better (level -2 or 2)
- Misses some instructions, but has fewer flaws than the worse response.
- May be slightly misleading, but the worse response is more wrong or ignores instructions entirely.
- Is better phrased or structured, even though the worse response may be appropriate.
- Has clearer or better formatting in cases where both responses are correct and follow instructions.

A response that is slightly better (level -1 or 1)
- Displays no difference in instruction following to the worse response.
- Is slightly better in clarity or formatting only.

Select 0 ONLY if both responses are equally good and you genuinely cannot decide between them (e.g., if the answers are near identical).

C DEMOGRAPHIC DETAILS OF STUDY PARTICIPANTS

Figure 7 shows an overview of the characteristics of the participants who participated in the EYETRACKING-HS2-800 study. In summary:

- **Gender**: Of the participants, 88 % self identified as male and 12 % as female.
- **Age**: The participants were fairly evenly distributed across age brackets, except for the less represented categories of 18-25 and over 60. The age group with most participants was 41-45. All participants were over 18.
- **Familiarity with the Vision Pro**: Most participants (89%) had at least tested the Vision Pro device.
- **Familiarity with LLMs**: The most common use case of LLMs among the participants was using chat interfaces, followed by APIs. Some participants reported use of open-weight models, as well as fine-tuning or pre-training LLMs, or developing related algorithms.

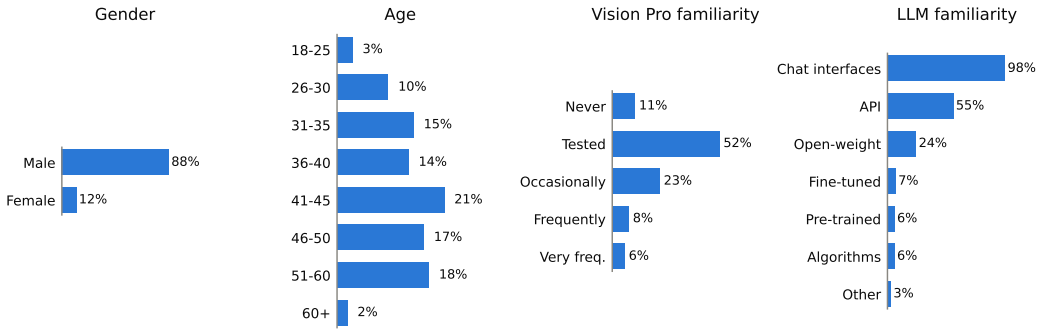


Figure 7: **Demographic details of study participants** ($N = 100$). LLM familiarity was multi-select and does not sum to 100%.

D SUPPLEMENTARY EMPIRICAL RESULTS

This section presents figures and tables that supplement results presented in Section 3.2.

Preference labels. Distribution of preference labels in EYETRACKING-HS2-800 are presented in Figure 8, including mean aggregated labels (pane A) and majority vote labels (pane B).

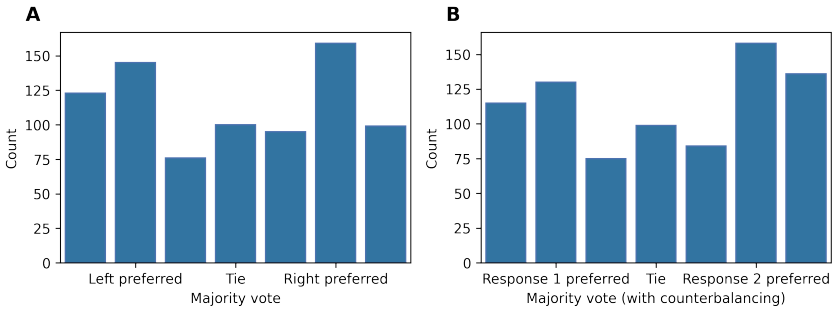


Figure 8: **Distribution of majority vote labels in EYETRACKING-HS2-800.** There is no statistically significant bias towards the option presented on left/right (pane A) or Response 1/Response 2 as named in Helpsteer2–Preference (pane B).

Task completion times. Figure 9 shows the distribution of the trial completion times across all participants. Additionally, mean number of fixations and durations across the different screens (prompt, responses, preference rating, see Figure 1) are outlined in Table 2.

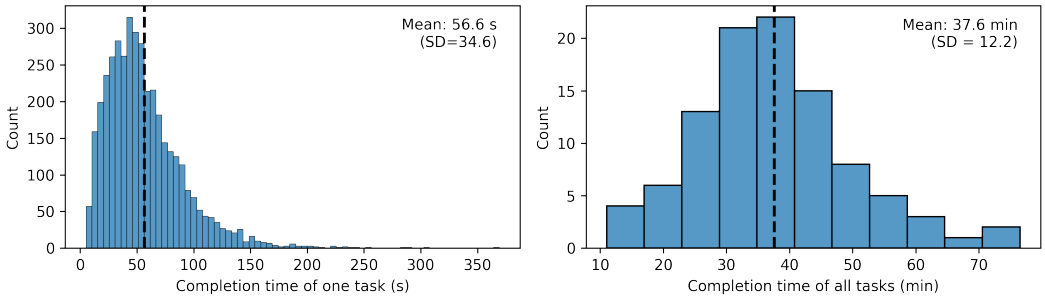


Figure 9: **Distributions of trial completion times (pane A) and completion times of all tasks excluding breaks (pane B) in EYETRACKING-HS2-800.**

	Prompt	Responses	Preference rating
Fixations (N)	206.28 (SD=222.92)	225.26 (SD=165.99)	30.25 (SD=16.18)
Duration (s)	18.37 (SD=21.54)	35.33 (SD=24.82)	3.37 (SD=3.86)

Table 2: Mean fixations and durations across screens displaying the prompt, responses, and preference ratings (see Figure 1).

Readers vs. scanners We categorized a participant as a "reader" if their task completion times increased more than 4.5 s / 100 px of height of either response (Figure 10), otherwise they were classified as a "scanner". Figure 12 shows the difference in how much of each response the two categories of annotators looked at: readers tend to look at a larger proportion of the texts generally, while both groups look at the left response more. Figure 11 also shows how the two groups allocated visual attention over the texts.

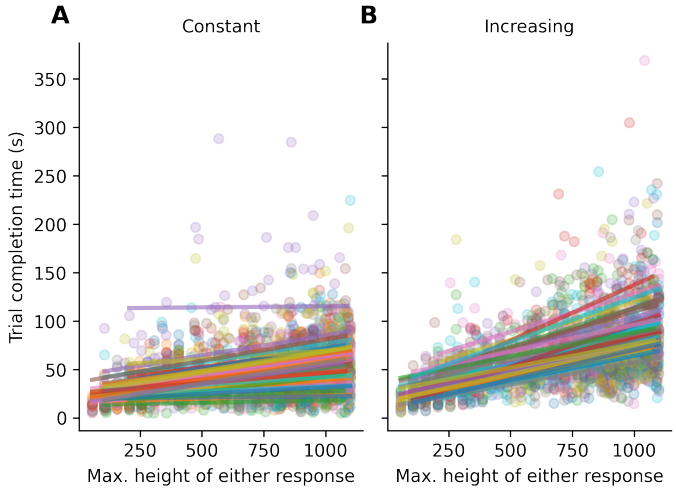


Figure 10: **Maximum height of either response vs. trial completion times split between a constant (pane A) and an increasing (pane B) trial completion time.** We use an increase of 4.5 s / 100 px as the threshold leading to approximately equal sized bins.

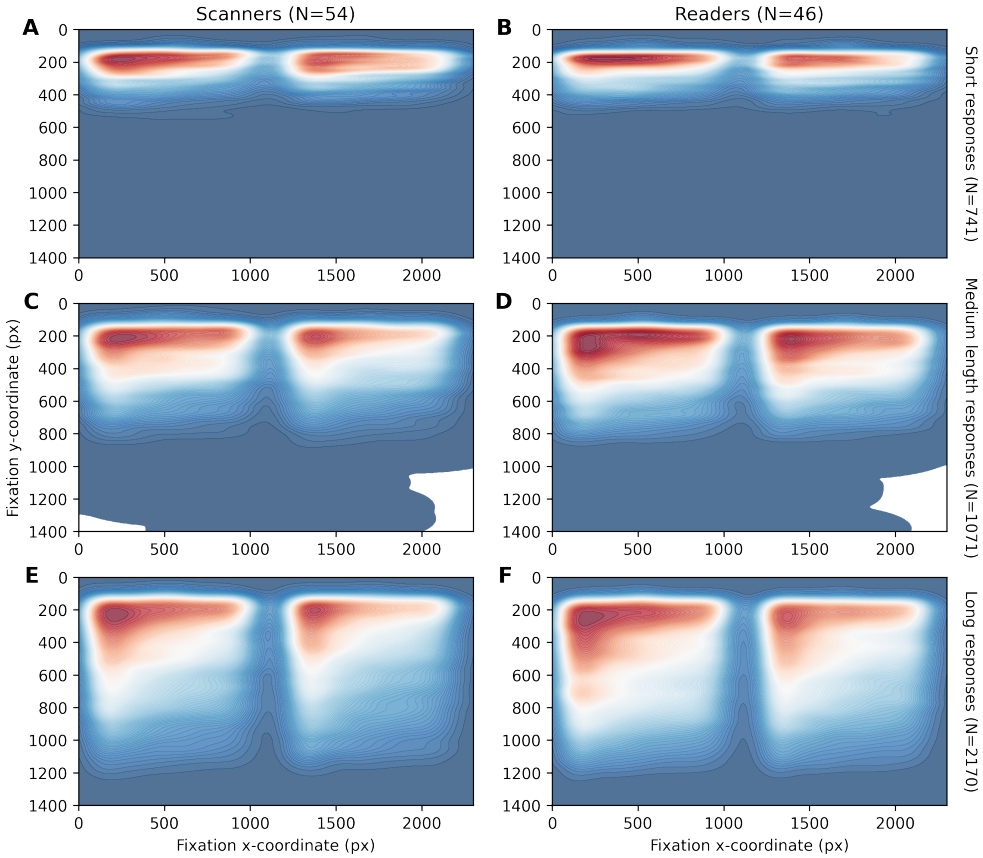


Figure 11: **Allocation of attention over the responses based on whether the participant was "a reader" (panes A, C, E) or "a scanner" (panes B, D, F).** The distinction between the groups is whether the task completion time stayed approximately constant (scanners) or increased (readers) with response length. The readers tend to allocate attention more evenly over rows of text, but also across the responses.

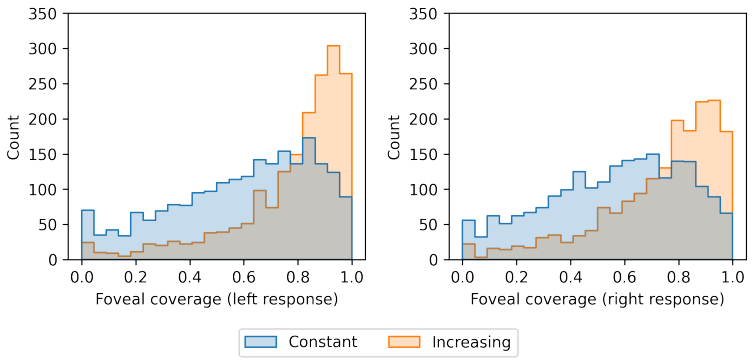


Figure 12: **Difference in proportion of the text covered by the foveal area based on whether the participant's task completion times were constant or increased with response length.** The participants with task completion times increasing with response left also tend to look at a larger proportion of the text.

E SUPPLEMENTARY FINE-TUNING RESULTS

DPO eval splits. DPO accuracy on the eval splits is reported in Table 3 for the different datasets.

LLM	1	2	3	4	5
Qwen/Qwen2.5-0.5B-Instruct	0.63	0.59	0.58	0.59	0.59
Qwen/Qwen2.5-3B-Instruct	0.64	0.58	0.58	0.58	0.56
Qwen/Qwen2.5-7B-Instruct	0.69	0.63	0.61	0.59	0.60
mistralai/Mistral-7B-Instruct-v0.2	0.72	0.64	0.64	0.64	0.66

Table 3: DPO accuracy on the eval splits. The eval split is the same for all conditions while the specific train data varies. 1: Full HelpSteer2; 2: HelpSteer2-800-subset (original HelpSteer2 preference labels); 3: HelpSteer2-800-subset ours with simple majority vote; 4: HelpSteer2-800-subset ours weighted majority vote; 5: HelpSteer2-800-subset ours weighted majority vote with reading speed.

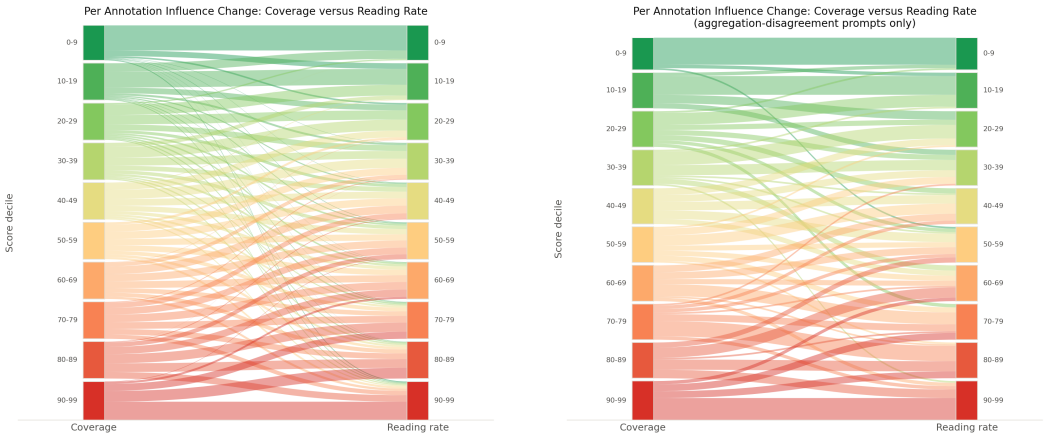
DPO hyperparameters. Table 4 reports the hyperparameters used in fine-tuning with DPO.

LLM	learning rate	β	LORA-α	LORA-r
HelpSteer2 (original): $\sim$10,000 samples				
Qwen/Qwen2.5-0.5B-Instruct	1e$-$5	0.3	64	256
Qwen/Qwen2.5-3B-Instruct	1e$-$5	0.3	128	16
Qwen/Qwen2.5-7B-Instruct	1e$-$5	0.1	128	256
mistralai/Mistral-7B-Instruct-v0.2	6e$-$6	0.07	256	16
HelpSteer2 (subset): 800 samples				
Qwen/Qwen2.5-0.5B-Instruct	6e$-$6	0.1	256	64
Qwen/Qwen2.5-3B-Instruct	1e$-$5	0.5	256	16
Qwen/Qwen2.5-7B-Instruct	1e$-$5	0.5	256	256
mistralai/Mistral-7B-Instruct-v0.2	4e$-$6	0.2	256	16
ET-HS2 (simple majority): 800 samples				
Qwen/Qwen2.5-0.5B-Instruct	4e$-$6	0.2	64	256
Qwen/Qwen2.5-3B-Instruct	6e$-$6	0.06	256	64
Qwen/Qwen2.5-7B-Instruct	6e$-$6	0.07	256	128
mistralai/Mistral-7B-Instruct-v0.2	1e$-$5	0.3	128	16
ET-HS2 (coverage weighted): 800 samples				
Qwen/Qwen2.5-0.5B-Instruct	1e$-$5	0.07	128	256
Qwen/Qwen2.5-3B-Instruct	1e$-$5	0.3	256	128
Qwen/Qwen2.5-7B-Instruct	1e$-$5	0.14	256	256
mistralai/Mistral-7B-Instruct-v0.2	4e$-$6	0.1	128	256
ET-HS2 (speed weighted): 800 samples				
Qwen/Qwen2.5-0.5B-Instruct	1e$-$5	0.3	64	128
Qwen/Qwen2.5-3B-Instruct	4e$-$6	0.1	256	128
Qwen/Qwen2.5-7B-Instruct	1e$-$7	0.01	16	16
mistralai/Mistral-7B-Instruct-v0.2	1e$-$5	0.3	128	16

Table 4: The DPO hyper-parameters per LLM and training dataset. Selected to maximize accuracy on the HelpSteer2 validation set.

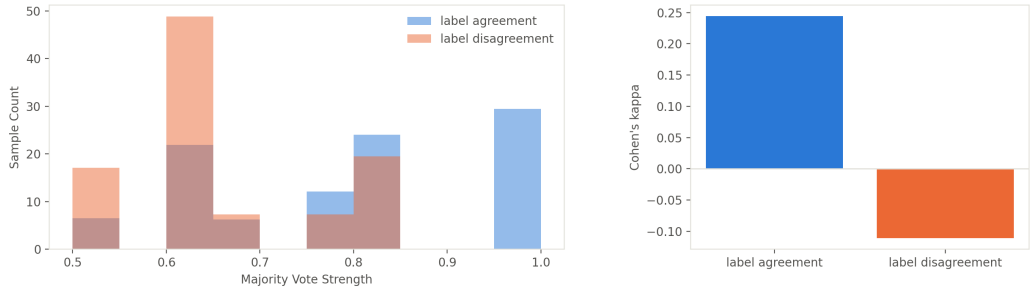
Detailed evaluation results for each benchmark. This section reports evaluation results for each benchmark, model, and dataset combination with the AlpacaEval (Li et al., 2023), Arena Hard (Li et al., 2024), MT Bench, (Zheng et al., 2023), and TruthfulQA (Lin et al., 2022) evaluation results shown in Table 5.

Preference label agreement/disagreement. Differences in coverage and reading speed ranks for each annotator-annotation combination are shown in Figure 13 for all annotations (a) and the disagreement subset (b). Figure 14 reports inter-annotator agreement for the samples where coverage and reading speed weights lead to agreement and disagreement between the preference labels.



(a) All Annotations (b) Disagreement Subset

Figure 13: **The change in an annotation's rank across all preference samples (i.e. prompt + response pair) and annotations when determined by coverage versus reading speed.** Moving from a lower bin on the left (coverage) to a higher bin on the right (reading speed) means an annotation became more influential when assessed by reading speed instead of coverage. The change in rank for all annotations (left) is compared to the change in rank for those preference samples (187 annotations) where the different weighting schemes induced different preference labels (right).



(a) Majority Vote Strength (b) Cohen's κ

Figure 14: **Inter-annotator agreement (IAA) for the samples where the coverage versus reading speed weighting schemes induce the same (agreement) versus different (disagreement) aggregated preference labels.** IAA is higher with label agreement, over disagreement as measured by majority vote strength (fraction of annotations matching the majority votes, pane a) and Cohen's κ (pane b).

Condition	Qwen2.5-0.5B	Qwen2.5-3B	Qwen2.5-7B	Mistral-7B	Mean Δ	**Rank**
AlpacaEval LC Win Rate (Win Rate)						
Base	22.91 (30.56)	66.92 (84.22)	76.72 (91.93)	80.25 (89.94)	–	–
HS2-full	12.20 (15.90)	44.61 (66.83)	65.21 (83.06)	75.61 (89.07)	-12.3	4
HS2-800	12.31 (16.52)	44.48 (65.67)	63.27 (82.58)	61.63 (79.30)	-16.3	5
ET-HS2	12.32 (16.54)	46.19 (66.54)	64.62 (82.75)	80.24 (90.30)	-10.9	3
ET-HS2-coverage	12.15 (16.54)	67.71 (83.48)	65.41 (83.35)	79.18 (89.57)	-5.6	2
ET-HS2-speed	25.41 (32.80)	65.18 (84.60)	75.79 (91.80)	78.41 (89.57)	-0.5	1
Arena Hard						
Base	0.50%	4.10%	8.80%	0.70%	–	–
HS2-full	0.50%	3.00%	9.70%	1.20%	+0.1	1
HS2-800	0.60%	3.60%	8.40%	0.90%	-0.1	2
ET-HS2	0.60%	3.40%	7.40%	1.00%	-0.4	4
ET-HS2-coverage	0.50%	3.20%	9.40%	0.60%	-0.1	2
ET-HS2-speed	0.40%	3.90%	7.40%	1.20%	-0.3	3
MT Bench (Mean Judge Score)						
Base	4.22	6.79	7.68	6.62	–	–
HS2-full	4.04	6.97	7.73	6.64	+0.0	2
HS2-800	2.99	5.84	6.80	5.58	-1.0	3
ET-HS2	4.23	6.92	7.79	6.67	+0.1	1
ET-HS2-coverage	3.87	6.91	7.68	6.71	-0.0	2
ET-HS2-speed	4.05	6.97	7.75	6.55	+0.0	2
TruthfulQA (mc2 accuracy)						
Base	0.42	0.59	0.65	0.68	–	–
HS2-full	0.42	0.59	0.65	0.71	+0.0	1
HS2-800	0.42	0.59	0.65	0.69	+0.0	1
ET-HS2	0.42	0.59	0.65	0.68	-0.0	1
ET-HS2-coverage	0.42	0.59	0.65	0.68	-0.0	1
ET-HS2-speed	0.42	0.59	0.65	0.69	+0.0	1
MT Bench (Mean Response Length)						
Base	379.20	438.80	441.80	334.10	–	–
HS2-full	387.90	432.20	440.90	362.00	+7.3	N/A
HS2-800	381.10	434.10	436.00	337.50	-1.3	N/A
ET-HS2	373.50	433.00	433.90	338.70	-3.7	N/A
ET-HS2-coverage	371.30	420.10	433.40	336.60	-8.1	N/A
ET-HS2-speed	388.10	423.20	427.20	325.50	-7.5	N/A

Table 5: AlpacaEval, Arena Hard, MT Bench, and TruthfulQA mc2 Accuracy results across LLMs and training conditions. Cell shading reflects the magnitude of difference from Base, normalized per column within each benchmark block, with sign encoded as orange (decrease) or blue (increase); the Base row is the uncolored reference. For all benchmarks, higher is better.