sampled human preferences, whereby human annotators select which of two model outputs they prefer. This human feedback is subsequently used to train a reward model, which learns patterns in the preferences of the human annotators and can then automate preference decisions.

3.2.1 Human Preference Data Collection

Next, we collect human preference data for reward modeling. We chose a binary comparison protocol over other schemes, mainly because it enables us to maximize the diversity of collected prompts. Still, other strategies are worth considering, which we leave for future work.

Our annotation procedure proceeds as follows. We ask annotators to first write a prompt, then choose between two sampled model responses, based on provided criteria. In order to maximize the diversity, the two responses to a given prompt are sampled from two different model variants, and varying the temperature hyper-parameter. In addition to giving participants a forced choice, we also ask annotators to label the degree to which they prefer their chosen response over the alternative: either their choice is *significantly better*, *better*, *slightly better*, or *negligibly better/ unsure*.

For our collection of preference annotations, we focus on helpfulness and safety. Helpfulness refers to how well Llama 2-Chat responses fulfill users' requests and provide requested information; safety refers to whether Llama 2-Chat's responses are unsafe, e.g., *"giving detailed instructions on making a bomb"* could be considered helpful but is unsafe according to our safety guidelines. Separating the two allows us to apply specific guidelines to each and better guide annotators; for example, our safety annotations provide instructions to focus on adversarial prompts, among other guidance.

Apart from differences in annotation guidelines, we additionally collect a safety label during the safety stage. This additional information bins model responses into one of three categories: 1) the preferred response is safe and the other response is not, 2) both responses are safe, and 3) both responses are unsafe, with 18%, 47%, and 35% of the safety dataset falling into each bin, respectively. We do not include any examples where the chosen response was unsafe and the other response safe, as we believe safer responses will also be better/preferred by humans. Safety guidelines and more detailed information regarding safety annotations can be found in Section 4.2.1.

Human annotations were collected in batches on a weekly basis. As we collected more preference data, our reward models improved, and we were able to train progressively better versions for Llama 2-Chat (see the results in Section 5, Figure 20). Llama 2-Chat improvement also shifted the model's data distribution. Since reward model accuracy can quickly degrade if not exposed to this new sample distribution, i.e., from hyper-specialization (Scialom et al., 2020b), it is important before a new Llama 2-Chat tuning iteration to gather new preference data using the latest Llama 2-Chat iterations. This step helps keep the reward model on-distribution and maintain an accurate reward for the latest model.

In Table 6, we report the statistics of reward modeling data that we collected over time, and present them against multiple open-source preference datasets including Anthropic Helpful and Harmless (Bai et al., 2022a), OpenAI Summarize (Stiennon et al., 2020), OpenAI WebGPT (Nakano et al., 2021), StackExchange (Lambert et al., 2023), Stanford Human Preferences (Ethayarajh et al., 2022), and Synthetic GPT-J (Havrilla). We collected a large dataset of over 1 million binary comparisons based on humans applying our specified guidelines, which we refer to as *Meta* reward modeling data. Note that the number of tokens in prompts and answers differs depending on the text domain. Summarization and online forum data generally have longer prompts, while dialogue-style prompts are usually shorter. Compared to existing open-source datasets, our preference data features more conversation turns, and are longer, on average.

3.2.2 Reward Modeling

The reward model takes a model response and its corresponding prompt (including contexts from previous turns) as inputs and outputs a scalar score to indicate the quality (e.g., helpfulness and safety) of the model generation. Leveraging such response scores as rewards, we can optimize Llama 2-Chat during RLHF for better human preference alignment and improved helpfulness and safety.

Others have found that helpfulness and safety sometimes trade off (Bai et al., 2022a), which can make it challenging for a single reward model to perform well on both. To address this, we train two separate reward models, one optimized for helpfulness (referred to as *Helpfulness RM*) and another for safety (*Safety RM*).

We initialize our reward models from pretrained chat model checkpoints, as it ensures that both models benefit from knowledge acquired in pretraining. In short, the reward model "knows" what the chat model

Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

Table 6: Statistics of human preference data for reward modeling. We list both the open-source and internally collected human preference data used for reward modeling. Note that a binary human preference comparison contains 2 responses (chosen and rejected) sharing the same prompt (and previous dialogue). Each example consists of a prompt (including previous dialogue if available) and a response, which is the input of the reward model. We report the number of comparisons, the average number of turns per dialogue, the average number of tokens per example, per prompt and per response. More details on Meta helpfulness and safety data per batch can be found in Appendix A.3.1.

knows. This prevents cases where, for instance, the two models would have an information mismatch, which could result in favoring hallucinations. The model architecture and hyper-parameters are identical to those of the pretrained language models, except that the classification head for next-token prediction is replaced with a regression head for outputting a scalar reward.

Training Objectives. To train the reward model, we convert our collected pairwise human preference data into a binary ranking label format (i.e., chosen & rejected) and enforce the chosen response to have a higher score than its counterpart. We used a binary ranking loss consistent with Ouyang et al. (2022):

$$\mathcal{L}_{\text{ranking}} = -\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r))) \tag{1}$$

where $r_\theta(x, y)$ is the scalar score output for prompt x and completion y with model weights θ. y_c is the preferred response that annotators choose and y_r is the rejected counterpart.

Built on top of this binary ranking loss, we further modify it separately for better helpfulness and safety reward models as follows. Given that our preference ratings is decomposed as a scale of four points (e.g., *significantly better*), as presented in Section 3.2.1, it can be useful to leverage this information to explicitly teach the reward model to assign more discrepant scores to the generations that have more differences. To do so, we further add a margin component in the loss:

$$\mathcal{L}_{\text{ranking}} = -\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r) - m(r))) \tag{2}$$

where the margin $m(r)$ is a discrete function of the preference rating. Naturally, we use a large margin for pairs with distinct responses, and a smaller one for those with similar responses (shown in Table 27). We found this margin component can improve Helpfulness reward model accuracy especially on samples where two responses are more separable. More detailed ablation and analysis can be found in Table 28 in Appendix A.3.3.

Data Composition. We combine our newly collected data with existing open-source preference datasets to form a larger training dataset. Initially, open-source datasets were used to bootstrap our reward models while we were in the process of collecting preference annotation data. We note that in the context of RLHF in this study, the role of reward signals is to learn human preference for Llama 2-Chat outputs rather than *any model* outputs. However, in our experiments, we do not observe negative transfer from the open-source preference datasets. Thus, we have decided to keep them in our data mixture, as they could enable better generalization for the reward model and prevent reward hacking, i.e. Llama 2-Chat taking advantage of some weaknesses of our reward, and so artificially inflating the score despite performing less well.

With training data available from different sources, we experimented with different mixing recipes for both Helpfulness and Safety reward models to ascertain the best settings. After extensive experimentation, the