



Figure 3: Win-rate of AUTO-J against baselines on single-response critique generation task, judged by GPT-4 and Human. L2Chat refers to LLaMA-2-Chat-13B.

options to further optimize training and efficiency. The model is trained for 5 epochs (675 parameter update steps in total) and we save checkpoints for every 50 steps. We use AdamW (Loshchilov & Hutter, 2017) as our optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 0.1. We use a peak learning rate $1e-5$ with 3% warmup steps and cosine learning rate decay to 0, and set the batch size to 64 and maximum sequence length to 4,096. The loss is only calculated on the output end.

5 EVALUATION SETTING

5.1 TASK AND TEST SET

Task I: Pairwise Response Comparison (Eval-P) In this task, the evaluators will see a pair of generated responses for a given query and decide which is better or is tied. From each scenario defined in §3.1 we randomly sample 24 pairwise comparison samples from the data we collected in §3.2 and skip those that have been used as training data. For some scenarios, the number of paired samples with pre-existed human annotation is smaller than 24, so we extract queries from either ShareGPT or the brainstormed seed data for training scenario classifier in §B. Samples from these two sources have no annotated pairwise labels, so we only use the query for each sample, generate a new pair of responses from two random selected LLMs² and manually annotate them. In total, we have $58 \times 24 = 1,392$ testing samples, each with two responses generated by different LLMs and a human-annotated preference label. We refer to this test set as Eval-P, with the distribution on Win/Tie/Lose being 520/373/499.

Task II: Critique Generation for Single Response (Eval-C) In this task, we evaluate the quality of the generated critiques for single-response evaluation. The evaluators are required to write critiques for a response to pinpoint its shortcomings in addressing the query. We apply both GPT-4 and human evaluation to compare critiques generated by different models. In GPT-4 evaluation, we randomly shuffle the order of two critiques to mitigate the positional bias, and use the instruction in Tab. 16. In human evaluation, we recruit four expert-level annotators (graduate students) and guide them with the same instruction for GPT-4. We build the test set for this task on the basis of Eval-P by sampling 4 out of 24 queries for each scenario and pick the less preferred response for each query (if tie, we randomly pick one). We refer to this test set as Eval-C, with $58 \times 4 = 232$ query-response pairs.

Task III: Overall Rating for Single Response (Eval-R) In this task, we evaluate the usefulness of the final rating for single-response evaluation in two ways: (1) The first is to use the ratings as verbal “rewards” to help improve the base policy models through the Best-of- N selection (Lightman et al., 2023; Gao et al., 2023), i.e., selecting the best response among the first N candidates with the assigned rewards, and use GPT-4 to grade the selected response. Generally, a more reliable model will select a better response with a higher GPT-4 rating more often. (2) The second is to calculate the

²From LLaMA-2-chat family, Vicuna family, WizardLM family, Claude-2, ChatGPT and GPT-4