response-level correlations between model-generated ratings and GPT-4 ratings. To save cost, we only collect the GPT-4 ratings on the previous "best-of-N" responses. The test set for this task is built on the basis of `Eval-C` by sampling 2 out of 4 queries for each scenario. We ask two different base LLMs (LLaMA-2-chat-7B and Vicuna-7B-v1.5) to generate 32 responses for each query through uniform sampling (temperature set as 1.0). We refer to this test set as `Eval-R`, with 58×2=116 queries and 116×32=3,712 query-response pairs for each base LLM.

5.2 BASELINES

General-purpose models: We use LLaMA-2-Chat-13B (Touvron et al., 2023b), Vicuna-13B-v1.5 (Chiang et al., 2023), WizardLM-13B-v1.2 (Xu et al., 2023), and ChatGPT (`GPT-3.5-turbo-0613`). We also use GPT-4 (`GPT-4-0613`) in the pairwise comparison and critique generation, and Claude-2 and LLaMA-2-Chat-70B in pairwise comparison. These models are used with corresponding prompt for each task: pairwise comparison prompt in Tab. 14, critique generation prompt in Tab. 18 (the same input format for AUTO-J's single-response evaluation), and rating prompt in Tab. 15. **Evaluation-specific models**: We use SelFee (Ye et al., 2023) in critique generation, SteamSHP (Ethayarajh et al., 2022) in pairwise comparison and overall rating, Open-Assistant's reward model (Köpf et al., 2023) in overall rating, and PandaLM (Wang et al., 2023c) in pairwise comparison.

6 EXPERIMENTS

6.1 PAIRWISE RESPONSE COMPARISON

A common problem in pairwise response comparison is positional bias (Wang et al., 2023a), where an LLM may tend to favor specific positions, causing inconsistency in comparison results when response orders are swapped. To pursue stable and reliable results, we conduct two comparisons for each sample by swapping the order of the two responses in the prompt. We consider a model's judgment to agree with human only when the two comparison results are consistent and align with the human judgment.

The agreement rates for AUTO-J and the baselines on `Eval-P` are in Tab. 1. AUTO-J achieves a significantly higher agreement rate than all baselines except GPT-4 on every scenario group. We also plot the prediction consistency for each model in Fig. 4. AUTO-J has a similar consistency rate to GPT-4 and is far more consistent than all other baselines, which makes it a more reliable and robust judge for pairwise comparison.

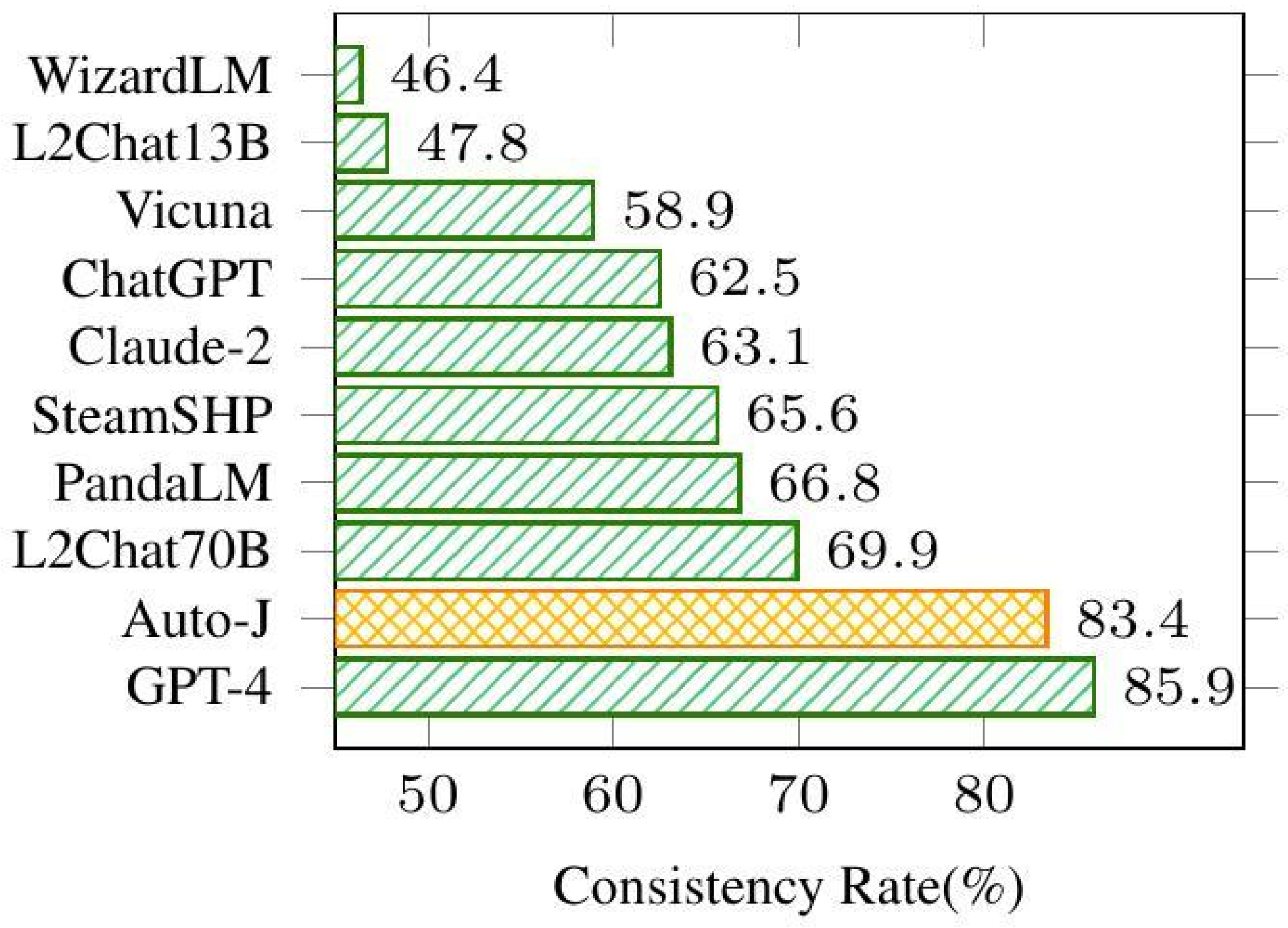

Figure 4: Consistency of prediction when swapping the response order.

6.2 CRITIQUE GENERATION FOR SINGLE-RESPONSE

The comparison results on `Eval-C` given by GPT-4 and human are in Fig. 3, and the complete comparison results for different scenario groups are in Tab. 23. In both evaluation settings, AUTO-J performs significantly than all baselines, including GPT-4, reflecting the strong ability to criticize other LLMs' outputs. We also observe that GPT-4 tends to provide judgments with very few ties, whereas humans often give tie judgments in comparisons, sometimes even exceeding 30%. One possible explanation is that the critique from AUTO-J exhibit a clearer structure and readability, which leads GPT-4 to pay less attention to the content when making comparisons, while humans are able to read more attentively and discern subtle differences between two critiques.

6.3 OVERALL RATING FOR SINGLE-RESPONSE

We conduct experiments on `Eval-R` with the N in Best-of-N selection set as 8, 16, and 32. In practice, if two responses share a common model rating, we choose the one with a higher output probability. Results in Tab. 2 show that responses selected by AUTO-J generally get higher GPT-4 ratings than those selected by baselines on different N.