Model	Eval-P		Eval-C	
	Seen	Unseen	Seen	Unseen
Unseen scenarios: Randomly select one scenario from each group.				
Auto-J (Complete Version)	54.5	56.8	146/200	25/32
Auto-J (Training w/o unseen scenarios)	53.5	55.7	143/200	25/32
Unseen scenarios: Scenarios of NLP Tasks group.				
Auto-J (Complete Version)	54.2	57.6	136/188	35/44
Auto-J (Training w/o unseen scenarios)	54.2	54.9	130/188	38/44

Table 8: Auto-J's generality on unseen scenarios. We train two new variants by removing a set of scenarios from the training data, and compare their performance with the complete version of Auto-J that has been trained on all data. We report the agreement rate with human label on the pairwise response comparison task (`Eval-P`), and the winrate against ChatGPT judged by GPT-4 on the critique generation task (`Eval-C`).

You are given the criteria to craft good responses for this type of query from users: - {scenario description} The criteria are as follows: [Criteria start] {criteria for the scenario} [Criteria end]

Table 9: Scenario criteria as system message in prompt.

C AUTO-J'S GENERALITY ON UNSEEN SCENARIOS

It is quite important to see how Auto-J performs on the scenarios that are not included in its training data. To investigate this research problem, we retrain two variants of Auto-J by holding out two sets of unseen scenarios.

1. We randomly select one scenario from each scenario group as the unseen scenarios, and retrain Auto-J with the remaining scenarios. This in total leads to 8 unseen scenarios in testing.

2. We take the complete "NLP Tasks" group as the unseen scenarios, and retrain Auto-J with the remaining scenarios. This in total leads to 11 unseen scenarios in testing.

Under both setting, we select the complete version of Auto-J as the baseline to compare with. We report the agreement with human annotattion labels on the pairwise response selection task (`Eval-P`) and the win rate against ChatGPT in the critique generation task (`Eval-C`) judged by GPT-4. The results are shown in Tab. 8.

Compared with the complete version of Auto-J, the two re-trained variants only show slightly degraded performance on the two evaluated tasks both on the seen and unseen scenarios, which indicates that Auto-J can generalize well to scenarios unseen during training.

D PROMPTS

Tab. 9-16 shows different prompts. Tab. 9-13 guide GPT-4 to generate training data (§3.2). Tab. 9 and 10 provide GPT-4 system messages, where the scenario and the criteria are defined. Tab. 11-13 show GPT-4 user messages, providing the instance-related information. Tab. 14-15 elaborate the prompts (§5.2), which all baseline models use to generate the testing results. Tab. 16 is used for GPT-4 evaluation that conducts a pairwise comparison between our AUTO-J with one baseline.