	Base LLM	BoN	Open-Assistant	SteamSHP	ChatGPT	L2Chat	Vicuna	WizardLM	AUTO-J
Selection	LLaMA-2-Chat-7B	8	8.17	8.02	8.20	8.13	8.09	7.93	**8.21**
		16	8.28	8.01	8.14	8.19	8.03	7.89	**8.33**
		32	8.25	7.84	8.14	8.16	8.05	7.94	**8.34**
	Vicuna-7B-v1.5	8	**7.51**	7.47	7.28	7.07	7.19	6.32	7.49
		16	7.69	7.74	7.29	7.02	7.53	6.46	**7.74**
		32	7.66	7.66	7.32	7.07	7.63	6.88	**7.97**
Correlation	Pearson		0.36	0.13	0.06	0.16	-0.05	0.41	**0.57**
	Spearman		0.42	0.13	0.06	0.24	-0.01†	0.35	**0.55**

Table 2: **Top half**: Average GPT-4 Rating on the Best-of-N (BoN) responses selected by different rating models. **Bottom half**: Correlations between different models and GPT-4 on all selected Best-of-N responses by different rating models, † means p-value >0.05. L2Chat: LLaMA-2-Chat-13B.

Based on the 1,993 query-response pairs with GPT-4 rating in the above best-of-N experiment, we calculate the response-level Spearman and Pearson correlations between model's rating and GPT-4 ratings. Results in Tab. 2 show a better correlation between AUTO-J and GPT-4 than all baselines.

6.4 ANALYSIS AND CASE STUDIES

System-level Ranking Besides response-level evaluation, and we also investigate the potential of AUTO-J on the system level, which is useful when we benchmark existing LLMs with leaderboard. We use the AlpacaEval leaderboard as it has archived complete outputs for each submitted model. We use AUTO-J in single-response evaluation protocol and calculate average ratings on the dataset for all open-source LLMs on the leaderboard. [3] The Spearman and Pearson correlations with GPT-4's ranking on the leaderboard are 0.97 and 0.96 respectively (Fig. 5), and we show detailed ranking in Tab. 24. This extremely strong correlation indicates that AUTO-J can also serve as a good system-level judge for ranking open-source LLMs.

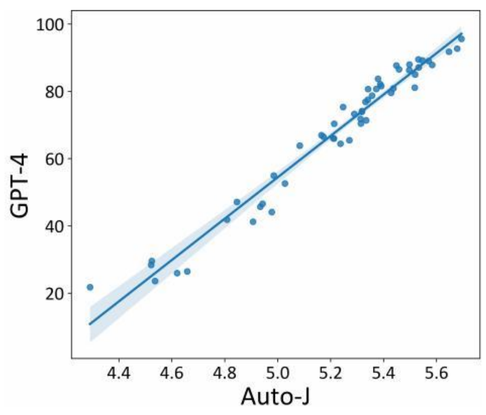

Figure 5: System-level correlation on AlpacaEval.

Ablation Studies (1) We train a model that outputs only the final decision using the same pairwise training data for AUTO-J. Its agreement rate with human on `Eval-P` is 55.0 (AUTO-J gets 54.8, in Tab. 1). We conclude that our model does not sacrifice the pairwise comparison performance for supporting multiple evaluation protocols and generating supporting explanations. (2) Using the same pairwise training data, we train a standard reward model to output a scalar rating for each query-response pair (its agreement rate on `Eval-P` is 54.5). We conduct best-of-32 response selection experiments. As shown in Tab. 3, despite not being directly optimized for a scalar output, AUTO-J achieves comparable performance to reward model. It also demonstrates higher correlation with GPT-4 ratings than the reward model trained solely for that purpose.

Case Studies We show a pairwise comparison case from the test set (`Eval-P`) in Tab. 4 (complete version in Tab. 25 and 26). This example shows only AUTO-J (and GPT-4) emphasize the advantages of the second response in terms of tone and interactivity for a family email, and make the correct choice.

We show a single-response evaluation case from the test set (`Eval-C`) in Tab. 5 (complete version in Tab. 27) shows that the critique given by AUTO-J is more aware of the user's status as a novice in cooking, and pinpoint more essential concerns on this.

The Best-of-N selection case from the test set (`Eval-R`) in Tab. 28 shows the usefulness of its rating in single-response evaluation. With more candidate responses given by the base LLM (Vicuna-7B-v1.5), AUTO-J is able to select a better response measured by both GPT-4 rating and human observation.

Base LLM	AUTO-J	ScalarRM
L2Chat7B	8.34	8.42
Vicuna7B	7.97	7.94
Correlation with GPT-4		
Pearson	0.57	0.39
Spearman	0.55	0.40

Table 3: Best-of-32 response selection for AUTO-J and a standard RM, and their correlation with GPT-4 ratings.

[3] 53 models or their variants on `https://tatsu-lab.github.io/alpaca_eval/` when the paper is written.