ConsEval: Illuminating and Improving the Consistency of LLM Evaluators

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown potential for data annotation and evaluation. Despite the evident benefits of speed and lower cost, we raise concerns about the reliability of LLMs when applied to this evaluation, especially within the ground of *consistency*. In this paper, we conduct extensive studies on the two different aspects of consistency in LLM evaluations, Self-Consistency (SC) and Inter-Consistency (IC), comparing the rating scale and criterion granularity. Additionally, we study the effects of inconsistency along with accuracy. We empirically observe that Llama-2-based evaluators are more consistent and accurate in general. Lastly, we present two effective methods: (1) Self-Consistency Evaluation (SCE) and (2) distilled In-Context Learning (*d*ICL) to jointly promote consistency and accuracy without further training. Along with accuracy-driven research, we insist on the importance of research towards additionally assessing the consistency in pursuit of safer LLM applications if we intend to exploit them as human evaluation proxies.

1 Introduction

Using large language models (LLMs) as annotators (Liu et al., 2022; Gilardi et al., 2023) for evaluating language generation across diverse traits (Chiang and Lee, 2023a; Liu et al., 2023; Fu et al., 2023) as alternatives to human evaluations can improve speed and reduce cost for developing models. With the combination of appropriate input prompts with a described trait of interest, LLMs have been shown to be sufficiently capable of outputting a feedback score given a specified scale (e.g., 1-5 Interval or Likert Scale). However, the sensitivity of models to the input prompts has become a major issue in natural language generation (NLG) (Liang et al., 2022; Sun et al., 2023), raising concerns about the reliability of LLMs (Jang and Lukasiewicz, 2023).

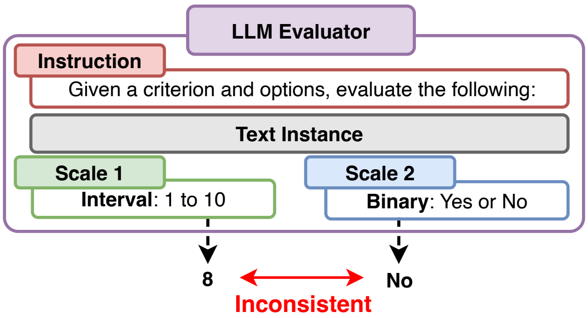

Figure 1: A simplified example of an LLM evaluator outputting inconsistent scores when given an identical input with only varying scales.

This observed inconsistency may be exacerbated in secondary applications of LLM, for instance, where AI feedback is replacing human feedback for preference alignment (Bai et al., 2022b; Lee et al., 2023; Tunstall et al., 2023), regarding that preference alignment algorithms are highly dependent on the dataset (Wang et al., 2024) or reward models (Shen et al., 2023). However, previous works on the LLM evaluators mainly focus on the alignment (i.e., the accuracy of evaluation) of instruction-following language models against human annotations (Liu et al., 2023; Fu et al., 2023; Chiang and Lee, 2023b; Kim et al., 2023), thereby leaving the question about consistency, robustness and credibility of LLM evaluators unresolved.

In this paper, we address an overlooked issue of using LLMs as evaluators: *consistency*. We demonstrate self-consistency and inter-consistency problems in model evaluators and suggest two methods to mitigate them along while maintaining accuracy. The summarized contribution of the paper are as the following:

1. We highlight two aspects of consistency in the context of LLM evaluator and conduct a comprehensive analysis along with accuracy.
2. We design an evaluation testbed, ConsEval, to

jointly analyze the consistency and accuracy and to promote further enhancements.

3. To mutually improve consistency and accuracy on ConsEval, we offer two methods: (1) Self-Consistency Evaluation (SCE) and (2) distilled In-Context Learning (dICL).

2 Related Works

LLMs as Evaluators Human evaluation has long served as the gold standard in training and evaluating model outputs. However, with sufficient training, model-based evaluation (Zhang et al., 2020; Sellam et al., 2020) has been demonstrated to be effective. With the introduction of large language models (LLMs) and their versatility, their potential to serve as annotators without the need for task-specific fine-tuning opens up a new alternative for replacing human annotations.

In the context for LLM evaluators, Gilardi et al., 2023 solely exploits LLM's zero-shot capability in data annotation in criteria of relevance, stance, etc. Chiang and Lee, 2023a points to the stability of using LLM evaluations on grammaticality, cohesiveness, and other text properties, assessing LLM evaluations to be reproducible and economical. Liu et al., 2023 and Fu et al., 2023 perform NLG evaluations on task-specific criteria, well showing the benefits of model zero-shot competence. Chiang and Lee, 2023b discusses the guidelines for LLM evaluations but only for text quality of natural language generation. Ye et al., 2023 suggests a protocol to produce fine-grained LLM evaluations on a skill set level. Kim et al., 2023 fine-tunes a language model to specialize in evaluating.

On Consistency of LLMs A model's ability to be consistent is an essentially desired trait in making a reliable tool. Numerous work highlight the consistency of different nuisances: logical reasoning consistency (Jung et al., 2022), semantic consistency (Raj et al., 2023), and consistency across LLMs in a debate setting (Xiong et al., 2023). Raj et al., 2023 raises a similar concern as our paper on the inconsistent generation of LLMs but focuses on semantic consistency only. Wang et al., 2023 takes further to introduce a confidence-based decoding strategy entitled, "Self-Consistency" to enhance the logical reasoning process over greedy decoding. Even so, within different contexts (e.g. LLM evaluators), more consideration on the ground of consistency needs to be contemplated in promoting more reliable and safer applications of LLMs.

3 Consistency of LLM Evaluators

3.1 Aspects of Consistency

For an estimator to be reliable, consistency stands as a strong prerequisite. Within the context of LLM evaluations, we are treating LLM evaluators as a potential alternative for costly human evaluations. Therefore, we highlight two aspects of consistency to be examined: (1) evaluation of confidence over equal prompt input (Self-Consistency) and (2) evaluation of cohesiveness across result-preserving structural changes (Inter-Consistency).

Self-Consistency (SC) The term *Self-Consistency* has been popularized by Wang et al., 2023, where the majority voting on the sampled forward passes led to a boost in logical reasoning performance. Though the mentioned work stresses the concept through a prompting methodology, we remark on the importance of Self-Consistency as an important testbed of consistency for LLM, especially in the application of evaluations to adequately output self-contained scores.

Inter-Consistency (IC) We also measure LLMs evaluators' consistency across variables. We entitle this "Inter-Consistency," aiming to quantify a model's evaluation cohesiveness across variables that are hypothesized not to massively disrupt the output trend but are suspected to be influential. We selected different scales and the criterion granularity as our main variable of interest for measuring inter-consistency.

3.2 Experimental Design

Prompt Design To facilitate a relatively more controlled setting for assessing consistency, we maintain a coherent prompt template with minimal changes across models and the variables of interest, as shown in Table 1.

Models We test seven instruction-following models to report the consistency in using LLMs as evaluators. We include Stable Vicuna[1] and three Llama-2-Chat models (Touvron et al., 2023) that are trained with reinforcement learning with human feedback (RLHF) are tested along with Zephyr (Tunstall et al., 2023) trained with DPO. Lastly,

[1] https://stability.ai/blog/stablevicuna-open-source-rlhf-chatbot

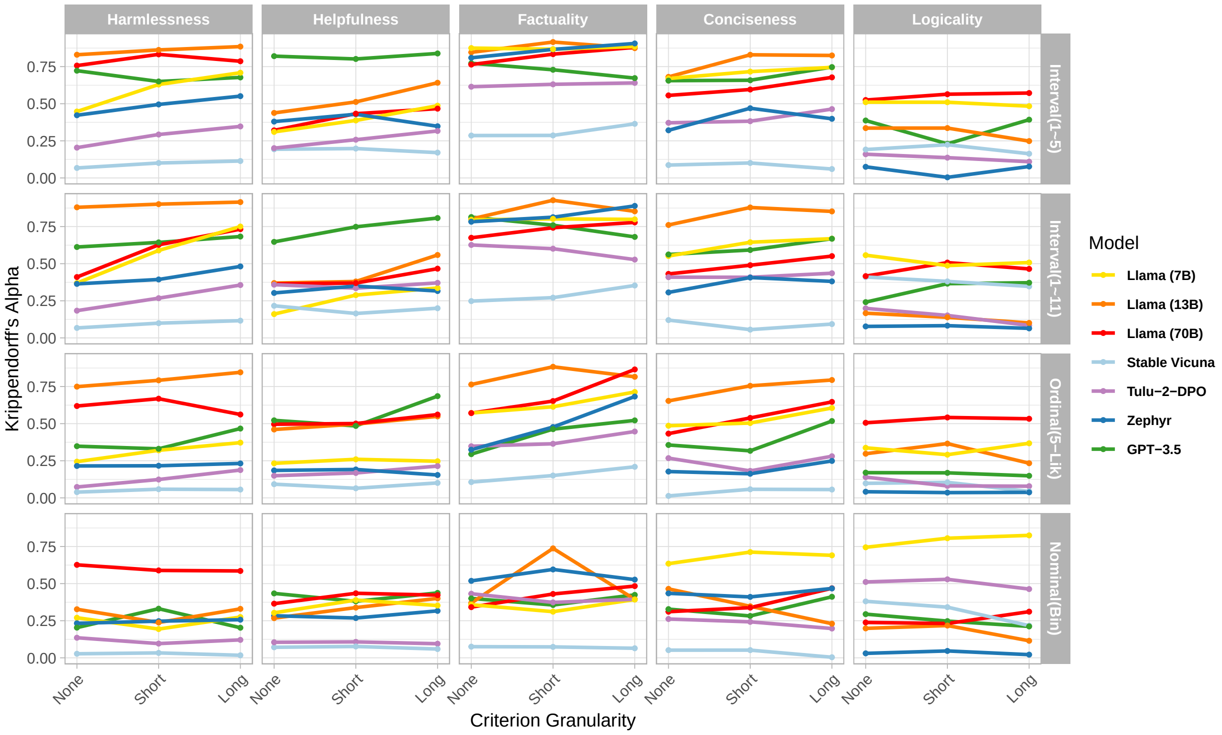

Figure 2: Self-Consistency evaluation results by five sampled evaluations for varying criterion granularity and rating scales. The evaluation is conducted on seven different models and five criteria of interest. Higher Krippendorff's α indicates the sampled responses to be more consistently similar.

Prompt Template

You will be given a pair of input query and response. Given a criterion and rating options, rate the response.
Evaluation Criterion:
`{criterion}:{detail}`
Options: `{options}`
Only output the evaluation score.
Query: `{query}`
Response: `{response}`
Answer:

Table 1: The default prompt template with only minor changes across different settings. `{options}` refers to the option of a range of the scale for scoring (e.g., 1-5, 1-10) or specific nominal options. `{criterion}` and `{detail}` are the criterion and the detailed definition for the selected criterion (e.g. *Logicality*, "correct and valid reasoning").

GPT-3.5-Turbo-16k[2] is tested as a large but relatively affordable proprietary LLM[3].

Datasets We test on five criteria by selecting the most representative and relatable datasets: Harmlessness (Bai et al., 2022a), Helpfulness (Zhou et al., 2023), Factuality (Lin et al., 2022) Logicality (Cobbe et al., 2021) and Conciseness. We sample out 1,000 instances for each criterion for representative dataset(s) for each trait. The details of the sampling process of the dataset can be found in Appendix A. To control the sensitivity of prompts, we mostly adopt a similar base prompt template in Table 1 across settings.

3.3 Experimental Variables

Evaluation Metric To jointly assess differing aspects of consistency, we select Krippendorff's α[4] (Hayes and Krippendorff, 2007) as our main metric in Sections 4 and 5 to assess inter-coder reliability. As it is applicable to varying types of variables (e.g. interval, nominal), the metric facilitates consistency comparison across variables. Additionally, we evaluate the accuracy with the Pearson correlation r with evaluation scores of GPT-4-Turbo[5] and Gemini-Pro (Team et al., 2023) in Section 6.

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] As of Feb. 2024 OpenAI Pricing for GPT-4 i 60x and 40x times the input and output tokens compared to GPT-3.5-Turbo

[4] https://github.com/pln-fing-udelar/fast-krippendorff

[5] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

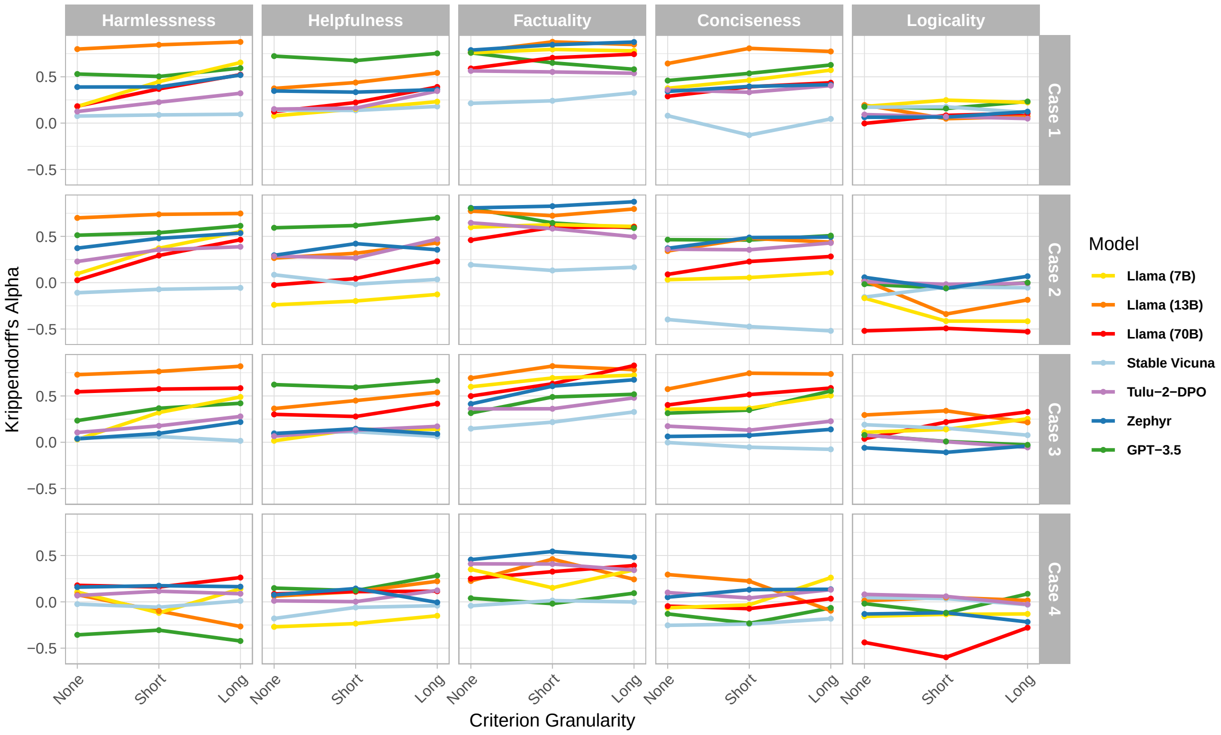

Figure 3: Inter-Consistency evaluation results for varying criterion granularity, rating scales, and decoding strategy. The detailed comparison sets of Cases 1, 2, 3, and 4 are in Table 4. Higher Krippendorff's α indicates the sampled responses to be more consistently similar.

Rating Scale We adopt both categorical and numerical rating systems for comprehensive observation, including five different interval scales and one ordinal Likert scale.

Scale	Notation	Range
Interval	5-P	[1, 5]
	7-P	[1, 7]
	10-P	[1, 10]
	11-P (POS)	[1, 11]
	11-P (NEG)	[-5, 5]
Ordinal	5-P Likert	{Strongly Disagree,...,Strongly Agree}
Nominal	Binary	{No, Yes}

Table 2: The types of rating scales used for the consistency assessment in Figure 2. 'Range' denotes a continuous ([]) or discrete ({}) range of scores given to the LLM evaluator.

Criterion Granularity We incorporate varying thoroughness of the criterion definitions. While assessing the evaluation consistency of models on five different criteria, we define each criterion with three different levels of granularity: no definition (None), single phrase definition (Short), and paragraphed definition (Long). The detailed definitions of each criterion can be found in Appendix B.

4 Self-Consistency

We measure Krippendorff's α across the five scores sampled with a temperature of 1.0 in Figure 2 with four scales selected from Table 2 for AC.

4.1 Results and Analysis

Higher criterion granularity generally aids SC The models tend to be more confident about their evaluation when the criterion definition is more detailed by having higher SC with high granularity. While this proclivity is clearly shown between no definition (None) and thorough definition (Long) setting in Figure 2, the tendency varies when we compare short definition and long definition.

Llama-Chat models are more self-consistent While the scaling effect is not clear in consistency, Llama-Chat models typically surpass other 7B models. Also, it is notable that they are frequently more consistent than GPT-3.5.

High variance in Harmlessness and Conciseness Mainly in Harmlessness and Conciseness, SC varied by the models. While Stable Vicuna (13B) and Tulu-2-DPO (7B) have nearly 0 Krippendorff's α with every scale, the two biggest Llama-Chat models and GPT-3.5 reached up to 0.9.

Model	Size	Harmlessness		Helpfulness		Factuality		Conciseness		Logicality	
		GPT-4	*Gemini*	*GPT-4*	*Gemini*	*GPT-4*	*Gemini*	*GPT-4*	*Gemini*	*GPT-4*	*Gemini*
Llama-Chat	7B	0.42	0.45	0.20	0.19	0.37	0.18	0.40	0.43	-0.04	0.03
Llama-Chat	13B	**0.54**	**0.57**	0.24	0.26	0.39	0.20	**0.42**	**0.48**	0.04	0.07
Llama-Chat	70B	**0.54**	**0.57**	0.29	0.26	0.51	0.25	0.38	0.40	0.07	0.10
Stable Vicuna	13B	0.13	0.14	0.11	0.10	0.09	0.05	0.11	0.09	0.01	0.10
Zephyr	7B	0.30	0.35	0.22	0.16	0.47	0.22	0.20	0.20	0.01	0.06
Tulu-2-DPO	7B	0.23	0.25	0.22	0.17	0.25	0.13	0.23	0.30	-0.03	0.01
GPT-3.5	-	0.44	0.46	**0.52**	**0.47**	**0.54**	**0.26**	**0.42**	0.44	**0.09**	**0.12**
Gemini	-	0.60	1.00	0.59	1.00	-0.03	1.00	0.02	1.00	0.45	1.00

Table 3: Accuracy Performance across models calculated by Pearson Correlation between representative proprietary LLMs (GPT-4, Gemini-Pro) with Interval Scale 1 to 5.

5 Inter-Consistency

To assess the Inter-Consistency, we calculate Krippendorff's α between the scores generated with different scales given the same input. Table 4 outlines the three cases of interest, and we jointly experiment on the effect of criterion granularity. The selected cases are aimed to configure the alignment of (1) three positive interval scales with different ranges, (2) a 5-point interval scale with a typical 5-point Likert scale, (3) a negative interval scale to a positive interval scale, and (4) binary scale to 10-point interval scale.

	Item 1	Item 2	Item 3	Detail
Case 1	5-P	7-P	11-P	Interval vs Interval
Case 2	5-P	5-P Likert	-	Interval vs Ordinal
Case 3	11-P (+)	11-P (-)	-	Interval(+) vs Interval(-)
Case 4	Binary	10-P	-	Nominal vs Interval

Table 4: Inter-Consistency comparison cases mapped to each row in Figure 3. The 'Detail' column describes the actual data type of evaluation scores that are being compared in each case.

5.1 Results and Analysis

Low IC in Logicality One of the most clear tendencies found in Figure 3 is that the models generally divagate in Logicality. We speculate the numeracy of language models is the main cause for this phenomenon (Petrak et al., 2023).

Higher IC in deterministic queries Recall that we have used TruthfulQA for the Factuality dataset; higher IC in comparison to other criteria implies higher inter-consistency in the queries that have deterministic standards.

Weak mapping between binary and interval scale evaluation In case-wise comparison, the IC was generally lower in Case 4, which is the consistency between binary evaluation and 10-P evaluation. It is noteworthy that even the proprietary model GPT-3.5 was not consistent in those settings.

6 Accuracy of LLM Evaluators

Although we stress the importance of the quality of being consistent, merely assessing consistency can lead to unexpected preference cases (e.g., consistently bad evaluations). Therefore, we also evaluate the models on the alignment towards proprietary LLMs, Gemini-Pro and GPT-4, which we will define as accuracy.

6.1 Result and Analysis

Llama-Chats are accurate in Harmlessness and Conciseness Out of seven models that we test, Llama-Chat (13B) tends to have the highest correlation with GPT-4 and Gemini-Pro. However, it is notable that GPT-3.5 has distantly higher accuracy in Helpfulness and Factuality, while Llama-Chat (13B) and (70B) have marginally higher accuracy in Harmlessness and Conciseness.

Low Accuracy in Logicality Along with low IC for Logicality in Section 5.1, the models show a similar tendency in accuracy for Logicality. While GPT-3.5 showed the highest correlation against both Gemini-Pro and GPT-4, it stays around 0.1, which is a low value for Pearson Correlation. Again, we infer the main reason would come from the low numeracy of open-source language models as discussed in Section 5.1.

Gemini and GPT-4 are partially aligned While both Gemini-Pro and GPT-4 are set as oracles, they were not aligned in a few criteria. Their evaluations were especially not aligned in Factuality and Con-

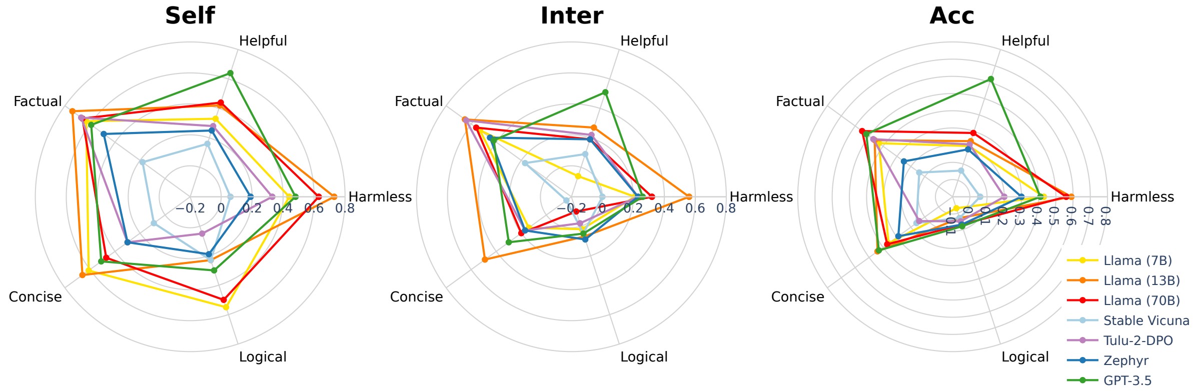

Figure 4: Radar Chart of ConsEval over the five criteria: Harmlessness, Helpfulness, Factuality, Conciseness, Logicality and seven models. For simplicity of the trend, the Inter-Consistency Krippendorff's αs and the Accuracy (Pearson r) are averaged, respectively.

ciseness, resulting in Pearson Correlation of -0.03 and 0.02, respectively.

7 ConsEval: Assessing the Reliability of LLM Evaluators

Reflecting upon the analyses carried out, we design and report results from an integrated, light evaluation testbed: ConsEval, for evaluating the reliability of LLM evaluators in the mentioned aspects: Self-Consistency, Inter-Consistency, and Accuracy.

7.1 Evaluation Design

Dataset We split the original split dataset per criterion in an 8:2 split to enable training or attending a portion of the dataset for performance increase. Additionally, as criterion granularity is shown to be influential in consistency, we separate instances per criterion granularity, which triples the original dataset size.

Evaluation Metrics Most of the evaluation metrics are identical to the settings laid out in Section 3.3 with the exception of Self-Consistency. There is a visible difference in the consistency scores (Figure 2), but it is not significant enough to report all of the values per scale. Instead, we integrate the scores from the four scales to calculate a single representative metric for Self-Consistency.

Remarks Figure 4 outlines the ConsEval results of the five criteria and the models referred to in the previous sections. The overall trends on Self-Consistency, Inter-Consistency, and accuracy mentioned in the previous sections are effectively captured concisely, even with a smaller subset of the data and breaking down the instances over criterion granularity. From the results, we can observe a positive association in consistency and accuracy in the aspect of LLM evaluators. The trend of Inter-Consistency and Accuracy are very similar in their performance pattern through criteria. Meanwhile, the Self-Consistency radar specifies the *consistently bad* cases where the models are confident to the scores, which are off to a large margin from the best performing proprietary LLMs (GPT-4, Gemini-Pro).

8 Towards Improving LLM Evaluators

Even though the correlation between accuracy and consistency seems to be present in Figure 4, the causal relationship cannot be assumed. In light of this, we suggest two methods to induce consistency and accuracy simultaneously without any training procedure.

8.1 Methods

On top of the baseline prompting method, we suggest two training-free methods and one training method that can be utilized simultaneously.

8.1.1 Self-Consistency Evaluation (SCE)

Self-consistency (SC) is utilized as a main metric for ConsEval. However, resembling Wang et al., 2023, we further exploit it as an evaluation mechanism that can potentially aid consistency and accuracy at the same time. We sample five scores with a temperature of 1.0 and average them to form a single score. [6]

[6] We distinguish Self-Consistency (SC) and Self-Consistency Evaluation (SCE) in which SC is used as a

	HARMLESSNESS							CONCISENESS						
		Consistency				Accuracy			Consistency				Accuracy	
Models	*SC*	*IC1*	*IC2*	*IC3*	*IC4*	*GPT-4*	*Gemini*	*SC*	*IC1*	*IC2*	*IC3*	*IC4*	*GPT-4*	*Gemini*
Llama-Chat	0.44	0.35	0.36	0.18	-0.04	0.39	0.47	0.61	0.34	-0.11	0.33	0.02	0.26	0.46
w/ *SCE*	-	0.58	0.50	0.41	0.10	0.50	0.61	-	0.53	-0.12	0.53	0.20	0.35	0.59
w/ $dICL_r$	0.34	-0.01	-0.27	-0.03	0.14	0.03	0.01	0.44	-0.01	-0.19	0.06	0.17	0.18	0.17
w/ $dICL_r$ &*SCE*	-	-0.04	-0.29	-0.03	0.11	0.12	0.01	-	0.20	-0.17	0.17	0.12	0.17	0.18
w/ $dICL_{sm}$	0.35	-0.03	0.06	-0.17	0.09	0.01	0.08	0.49	0.13	0.01	0.05	0.20	0.18	0.12
w/ $dICL_{sm}$ &*SCE*	-	0.05	0.19	-0.05	0.14	0.03	0.04	-	0.35	0.18	0.20	0.25	0.14	0.13
Tulu-2-DPO	0.19	0.28	0.33	0.22	0.09	0.28	0.33	0.30	0.24	0.23	0.17	0.04	0.24	0.34
w/ *SCE*	-	0.48	0.55	0.52	0.24	0.31	0.44	-	0.63	0.53	0.38	0.15	0.27	0.50
w/ $dICL_r$	0.33	0.20	0.20	0.13	0.13	0.32	0.34	0.34	0.18	0.15	0.09	0.06	0.06	0.12
w/ $dICL_r$ &*SCE*	-	0.40	0.42	0.27	0.13	0.33	0.35	-	0.34	0.42	0.35	0.28	0.13	0.23
w/ $dICL_{sm}$	0.48	0.23	0.23	0.24	0.19	0.31	0.27	0.48	0.22	0.23	0.10	0.19	0.06	0.15
w/ $dICL_{sm}$ &*SCE*	-	0.35	0.36	0.37	0.16	0.35	0.39	-	0.37	0.38	0.36	0.31	0.11	0.21
GPT-3.5	0.48	0.54	0.58	0.31	-0.40	0.44	0.39	0.51	0.53	0.45	0.38	-0.15	0.38	0.49
w/ *SCE*	-	**0.75**	**0.73**	0.52	-0.48	0.52	0.50	-	**0.72**	**0.61**	**0.59**	-0.12	0.43	**0.57**
w/ $dICL_r$	0.70	0.52	0.54	0.39	-0.19	0.49	0.48	0.57	0.36	0.29	0.24	-0.24	0.33	0.32
w/ $dICL_r$ &*SCE*	-	0.62	0.62	0.52	-0.23	0.53	0.56	-	0.47	0.42	0.37	-0.32	0.38	0.38
w/ $dICL_{sm}$	**0.78**	0.60	0.65	0.47	-0.04	0.54	0.61	**0.69**	0.48	0.44	0.39	**0.07**	0.41	0.37
w/ $dICL_{sm}$ &*SCE*	-	0.69	0.72	**0.54**	**-0.01**	**0.60**	**0.65**	-	0.58	0.50	0.49	**0.07**	**0.44**	0.43

Table 5: ConsEval Performance of Llama-Chat (7B), Tulu-2-DPO (7B), and GPT-3.5 Turbo on Harmlessness and Conciseness with different evaluation methods. *SCE* refers to Self-Consistency Evaluation while *SC* refers to the consistency metric. $dICL_r$ refers to distilled In-Context Learning from randomly sampled examples and $dICL_{sm}$ refers to sampling from examples with same target scale.

8.1.2 Distilled In-Context Learning (dICL)

To bridge the gap between large proprietary models and smaller LLMs, we propose a prompting method called Distilled In-Context Learning (dICL). The idea is rooted in distilling the best proprietary LLM evaluation capability to increase accuracy and robustness to scales (consistency) of smaller evaluators. We generate evaluations from GPT-4-Turbo and Gemini-Pro and sample out five examples to form an example set of $\{(x_1, y_1), \ldots, (x_k, y_k)\}$ that will be included in the input for LLM evaluation[7]. We test on two sampling strategies: random and scale_matched. Random sampling literally means selecting examples from the entire set of distilled evaluations. Scale_Matched Sampling refers to a filtered sampling process that only samples from the examples with the same scale as the target instance.

8.2 Experimental Design

Models We assess ConsEval with additional methods on two 7B models, Llama-2-Chat (7B) and Tulu-2-DPO (7B), as they have shown the highest and lowest overall consistency and accuracy in Section 3, respectively. In addition, we also test GPT-3.5, which had high consistency and accuracy overall, to improve the limits of performance.

Criteria Selection We select two criteria, Harmlessness and Conciseness, regarding the high variance of consistency across the models in Figures 2 and 3. To effectively capture the impact of each method, two criteria that have distant consistencies between the two models were selected.

8.3 Experimental Result

LLM evaluators are generally inconsistent in IC4 Recall to Table 4, the LLM evaluators were highly inconsistent in binary evaluations. As discussed in Section 5, the models generally show low Krippendorff's α and Pearson Correlation in IC4. Even with the methods proposed in Section 8.1, consistency in IC4 did not increase in most cases for every model.

SCE boosts both consistency and accuracy SC enhances the consistency and accuracy of both Llama-Chat and Tulu-2-DPO. Especially in Harmlessness, Llama-Chat surpasses the plain GPT-3.5 with SC. It is notable that even though the SC is relatively low (as in Tulu-2-DPO), SCE improves both consistency and accuracy by a large margin. While it will be further discussed, simple SCE is mostly the best way of prompting LLMs as evalua-

metric of model confidence, and SCE is used as an evaluation strategy.

[7] Refer to Appendix 6 for the actual template

tors in terms of both consistency and accuracy.

dICL makes the models more self-consistent As shown in the first row of Harmlessness and Conciseness in Table 5, dICL improves the SC of Tulu-2-DPO and GPT-3.5. On the other hand, the SC of Llama-Chat is degraded with dICL in Conciseness and improved in Harmlessness.

dICL improves GPT-3.5 for both accuracy and consistency Compared to the default prompting style, dICL better aligns the GPT-3.5 as an evaluator model to the oracle models that generated the in-context examples. This implies that intrinsic evaluation skills can also be distilled through not only straight fine-tuning but also through in-context learning. Especially, $dICL_{sm}$ yields extra improvements in both accuracy and consistency, in comparison to $dICL_r$, which provides the example set with random scales.

Comprehensive use of SCE and dICL can further benefit GPT-3.5 With the previous observations, jointly using $dICL_{sm}$ and SCE maximizes the consistency and accuracy of GPT-3.5 more than any other settings. Meanwhile, the effect of $dICL_{sm}$ + SCE varied for Llama-Chat and Tulu-2-DPO. For instance, Tulu-2-DPO is best aligned to GPT-4 with $dICL_{sm}$ + SCE in Harmlessness, but the accuracy was degraded in Conciseness with $dICL_{sm}$ + SCE. Moreover, the accuracy of Llama-Chat was critically lowered with $dICL_{sm}$ + SCE, which Llama-Chat models are not capable of learning the evaluation skills in context.

8.4 Future Directions

Despite evident gains from the suggested methods, more research can be conducted from in inducing further capabilities. For example, Kim et al., 2023 introduces a fine-grained LLM evaluator solely trained for evaluations generated by GPT-4. However, due to the discrepancy of the prompt configurations (e.g., Score Rubric, Explanation), we were not able to generate effective evaluations of the models as expected with our prompt template. Furthermore, Chiang and Lee, 2023b highlights the advantage of generating a rationale of the evaluation, introducing performance benefits beyond generation interpretability. We expect more consistency-driven research to be conducted as it closely ties in with its reliability. Lastly, we leave for future researches to unveil the drawbacks of excessive reliance on top proprietary LLMs and the blind-spots that they fail to excel, especially on the aspect of evaluation.

Conclusion

By highlighting two aspects of the inconsistency of large language model (LLM) evaluators, we surface the lack of credibility as reliable alternatives to human evaluation. We report a comprehensive analysis of each type of consistency and its implications across criterion granularity and rating scale using Krippendorff's α agreement measure. We find that using Llama2 models with fine-grained positive rating scales and criterion definitions generally leads to more consistent evaluations. Then, we present ConsEval, a light testbed to assess diverse aspects of consistency along with accuracy. Finally, we propose two novel methods: (1) Self-Consistency Evaluation (SCE) and (2) distilled In-Context Learning (dICL) to mutually improve accuracy and consistency. Considering the ubiquitous scope of LLM applications, we assert the need to thoroughly investigate the extent of inconsistency in the LLM evaluation pipeline and disclose how it is manifested throughout subsequent utilization. The inconsistency is likely to persist throughout the succeeding stages (e.g. preference alignment).

Limitations

This work assesses the consistency of LLM evaluations across variables of interest. Although we are trying to control other effective variables, as we are not directly mitigating the prompt sensitivity of the language models, minor prompt shifting may have an impact on the results. Additionally, though many paper present top proprietary LLMs such as GPT-4 and Gemini-Pro as oracles and sufficient alternatives to human evaluations, an in-depth human evaluation process will benefit the research proposed.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron

McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Myeongjun Erik Jang and Thomas Lukasiewicz. 2023. Consistency analysis of chatgpt.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Dominic Petrak, Nafise Sadat Moosavi, and Iryna Gurevych. 2023. Arithmetic-based pretraining – improving numeracy of pretrained language models.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. 2023. The trickle-down impact of reward (in-) consistency on rlhf. *arXiv preprint arXiv:2309.16155*.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. Secrets of rlhf in large language models part ii: Reward modeling.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining the inter-consistency of large language models: An in-depth analysis via debate. *arXiv e-prints*, pages arXiv–2305.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Details on Dataset

As mentioned in the paper, we manually select relevant dataset(s) and sample out instances in the following manner:

Harmlessness We sample out from 500 instances Anthropic's HH-RLHF Dataset (Bai et al., 2022a) in which a assistant response is either chosen or rejected. We use both of the chosen and rejected responses to sum to a total of 1000 instances.

Helpfulness (Zhou et al., 2023)'s training set is composed of exactly 1000 high quality instances to train human alignment. We use the entire training set for evaluation.

Factuality The factuality dataset in evenly sampled from the TruthfulQA (Lin et al., 2022) dataset. From the original TruthfulQA dataset, we sample out 500 questions and select the correct and a sampled incorrect answer choice to add up to 1000 instances.

Logicality We sample out 1000 instances of the GSM8K (Cobbe et al., 2021) test set from the 1395 instances.

Conciseness We find conciseness a central characteristic that is favored for most cases of LLM generations. Thus, we evenly sample from the four datasets above (250 instances each) to sample a total of 1000 instances to evaluate.

B Instruction Granularity: Thoroughness of definition

This section provides the details of three different definitions used for elaborating each evaluation metric to the language model. The definition of each skill is generated with GPT-3.5.

We prompted "*You are a word dictionary. Please tell me the definition of {skill_name} in a short phrase.*" and "*Please tell me the definition of {skill_name} within 100 words.*" to acquire the short and long definition of each skill respectively.

Harmlessness

Short Definition

a state or quality of being free from causing harm or damage to others.

Long Definition

a term that denotes the quality or state of not causing harm, injury, or damage to individuals, living beings, or entities. It embodies a deliberate and conscientious approach to actions, behaviors, and intentions, aiming to avoid any adverse consequences or negative effects on others. This concept is often associated with ethical and moral principles, such as non-violence and empathy, and is integral to various philosophies and belief systems, including pacifism and some religious teachings. Practicing harmlessness involves considering the well-being and rights of others, promoting peace and non-aggression, and seeking constructive and non-destructive solutions to conflicts and challenges in interpersonal and societal contexts.

Logicality

Short Definition

the quality of being logical or the extent to which something adheres to logical principles and reasoning.

Long Definition

a noun that refers to the quality of being logical or conforming to principles of sound reasoning and coherence. It pertains to the extent to which an argument, statement, or action follows a rational and consistent thought process. In discussions, debates, and problem-solving, logicality plays a crucial role in ensuring that conclusions are derived from valid premises, and the connections between ideas are clear and reasonable. It encompasses the ability to think logically, analyze information, and draw well-founded conclusions. The concept of logicality is fundamental in philosophy, mathematics, science, and everyday decision-making, as it promotes clarity, consistency, and the avoidance of fallacious or irrational thinking.

Factuality

Short Definition

the quality or state of being based on factual information or truth.

Long Definition

term that denotes the degree to which something is grounded in facts or reality. It relates to the accuracy and truthfulness of a statement, assertion, or information. When information or claims are described as having a high level of factuality, it signifies that they are supported by objective evidence, data, or verifiable sources, making them reliable and trustworthy. Conversely, low factuality implies a lack of factual basis, often indicating a reliance on speculation, opinion, or falsehoods. Factuality is essential in critical thinking, journalism, and decision-making processes, as it helps distinguish between information that can be relied upon and that which should be viewed skeptically.

Helpfulness

Short Definition

the quality of being willing and able to assist or support others when needed.

Long Definition

the characteristic of being inclined and capable of providing aid, support, or assistance to others. It entails a genuine willingness to offer guidance, information, or resources in order to make tasks, challenges, or situations easier for someone else. Helpfulness is often associated with empathy, compassion, and a positive attitude toward helping others achieve their goals or overcome difficulties. It fosters cooperation, teamwork, and a sense of community, making it an essential trait in building strong relationships, both personally and professionally. People who exhibit helpfulness are often seen as dependable and reliable contributors to the well-being of those around them.

Conciseness

Short Definition

the quality of being clear and brief, expressing ideas in a succinct manner.

Long Definition

a fundamental aspect of effective communication. It refers to the quality of expressing thoughts, ideas, or information clearly and succinctly, without unnecessary elaboration or wordiness. Conciseness aims to convey a message in the most efficient and direct way possible, eliminating superfluous words or details that might confuse or bore the audience. Concise writing or speech gets straight to the point, making it easier for readers or listeners to grasp the intended message quickly. It enhances clarity, maintaining the audience's attention and interest while avoiding ambiguity or misunderstanding. Achieving conciseness requires careful editing and choosing words judiciously to convey the essential information without unnecessary clutter or verbosity.

Five Shot Prompt Template

```
You will be given a pair of input query and response. Given a criterion and rating options, rate the response.
Evaluation Criterion:
{criterion}:{detail}

### Example
Query: {query}
Response: {response}
Options: {options}
Answer: {Answer}

### Example Query: {query}
Response: {response}
Options: {options}
Answer: {Answer}

### Example
Query: {query}
Response: {response}
Options: {options}
Answer: {Answer}

### Example
Query: {query}
Response: {response}
Options: {options}
Answer: {Answer}

### Example
Query: {query}
Response: {response}
Options: {options}
Answer: {Answer}

### Target
Only output the evaluation score.
Options: {options}
Answer:
```

Table 6: The default prompt template for a few shot setting for *d*ICL