Assessment and Mitigation of Inconsistencies in LLM-based Evaluations

Sarik Ghazarian
Amazon
USA
sghazari@amazon.com

Yidong Zou
Amazon
USA
yid@amazon.com

Swair Shah
Amazon
USA
shahswai@amazon.com

Nanyun Peng
UCLA University, Amazon
USA
violetpeng@cs.ucla.edu

Anurag Beniwal
Amazon
USA
beanurag@amazon.com

Christopher Potts
Stanford University, Amazon
USA
cgpotts@stanford.edu

Narayanan Sadagopan
Amazon
USA
sdgpn@amazon.com

ABSTRACT

Large Language Models (LLMs) offer a scalable approach to automatically evaluating generative models, but these evaluations are extremely sensitive to the nature of the guidelines and other instructions included in the LLM prompts. In this work, we comprehensively examine the effects of manipulating the position and length of the scoring guidelines on the results of the LLM-based evaluators. By design, these manipulations do not affect the prompt semantics, however we find that LLM-based evaluators do not respond to them consistently. We propose a simple yet cost-effective approach that in contrast to existing solutions does not rely on the frequent runs of LLMs to mitigate the inconsistency issue. We augment few-shot demonstrations of consistent scores under various perturbations of scoring guidelines to the input prompt to indirectly instruct the LLM the preferred behavior to follow. In binary and multi-class quality evaluations of generations by Claude, GPT3.5, and Mixtral on the SGD, MultiWOZ, and CI datasets, we find that LLM-based evaluators achieve up to 28% higher consistency by leveraging our proposed few-shot in-context examples of the manipulated guidelines which beat the existing baselines. This points to a central role of rich demonstrations in achieving reliable LLM-based evaluators.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**.

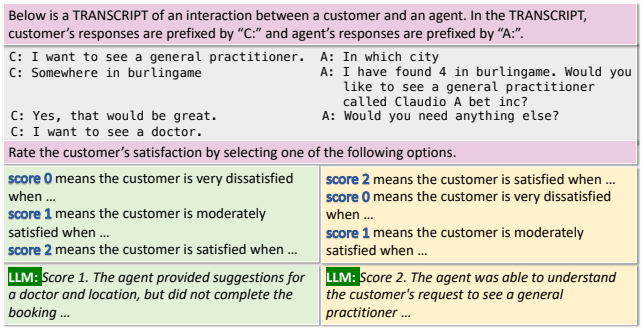

Figure 1: An example of the inconsistency issue in Claude-V2 model's evaluation scores due to manipulated order of assessment guidelines in the prompt. The green highlighted section is for the prompt with scoring options in the order of 0, 1, and 2 while the yellow part is related to the prompt with scores in the order of 2, 0, and 1.

KEYWORDS

Large Language Models, Consistency Evaluation, Few-shot In-context Learning

1 INTRODUCTION

Automatic evaluation metrics offer the potential for fast and iterative development of conversational agents, and Large Language Models (LLMs) are emerging as a powerful tool for such evaluations [3, 7]. LLM-based evaluators are typically defined by prompts that include task instructions, evaluation criteria, and guidance about how to generate either explicit or implicit scores resulting from a predefined classification label's confidence level [10].

Recent studies emphasize the quality of the provided prompts as a primary factor affecting LLM-based evaluators. Accurate description of the task and evaluation concepts is crucial, as is providing examples that can assist in steering the LLMs toward the desired

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

outcomes [1, 4, 16, 18]. Zheng et al. [18] provide a detailed empirical study of the limitations that LLMs have in automatic pairwise assessment tasks. They highlight **position bias** (the LLM's tendency to select responses in specific locations regardless of their quality) as a central limitation. Similarly, Ye et al. [16] show that LLM-based evaluations for multi-choice question answering can have **inconsistency** issues, where the assessment scores vary for inputs with the same meaning. They show that LLM evaluator performance often varies when the prompts are syntactically changed while their meanings are preserved or when the option orders are changed.

In the present paper, we build on this previous work by studying consistency issues for LLM-based evaluators in detail. Our focus is on following meaning-preserving changes to prompts: (1) manipulating the order of the score options in the guidelines, (2) perturbing the length of the explained options and their combination. We assess the consistency of Claude-V2,[1] GPT3.5-turbo,[2] and Mixtral-8x7B-v1.0[3] on Consistency-Inconsistency datasets and User Satisfaction Scores datasets for both binary and multi-class assessments. Figure 1 shows an instance in which Claude-V2's output scores and reasoning are significantly changed when the order of the guidelines for user satisfaction scores is changed. In our experiments, we find that such inconsistency is common for all three models.

The identification of consistency issues in LLM-based evaluators can lead to some doubts in leveraging them for evaluation and development purposes. Hence, providing a solution to resolve or mitigate this issue plays a critical role. Directly asking LLMs to have consistent predictions is useless as it is not comprehensible to them. Therefore, researchers got encouraged to explore this space. They propose to do majority voting or mean aggregation of output samples of LLMs to achieve more consistent reasoning by LLMs or fairer pairwise comparison of generations [13, 14], although such kind of approaches require LLMs to be sampled frequently which in general can not lead to an effective and quick solution.

As a step forward, we present evidence that we can increase consistency versus position/length perturbations by simply providing a set of conversations as demonstrations in the prompt that include manipulated orders for the guideline or perturbed length of the explanations with the same assessment scores. This successfully guides the LLM to keep its outputs constant for semantically similar input guidelines. Our contributions are as following:

- We examine the consistency of LLM-based evaluators in a more realistic setup; quality assessments with scoring guidelines. We apply position and length manipulations to the evaluation guidelines without touching the semantics of the input prompt and check the consistency of LLMs' judgements.
- We propose a simple yet cost-effective approach that mainly relies on in-context few-shot learning of LLMs.
- We survey the consistency of Claude, GPT3.5, and Mixtral on various Task Oriented Dialogue (TOD) datasets for both binary and multi-class quality evaluation of the generated responses from different perspectives. We show that our approach beats all the existing works and improves the LLM-based evaluators' consistency up to 28% without hurting the accuracy.

2 RELATED WORK

Researchers increasingly rely on LLMs for automatic evaluation of dialogue systems in zero-shot or few-shot setups [4]. Such LLM-based evaluators return explicit scores, implicit scores, or conduct pairwise comparisons [1, 10, 18]. Lin and Chen [9] employ the ability of LLMs to concurrently accomplish the assessment task by outputting different aspects all at once by replying to a single prompt.

Prompt format is very important in LLM-based evaluation; if the LLM misunderstands the prompt, it may result in unreliable scores [4, 9]. Zheng et al. [18] incorporate LLMs to conduct pairwise comparison of two chatbots based on the correctness of each candidate chatbot's replies. They also offer position bias, verbosity bias, and self-enhancement bias as core issues for LLM-based evaluators. These issues show that LLMs mostly prefer the first choice, longer replies, and their own generated replies over those given by other models regardless of the quality. Ye et al. [16] also study the consistency and robustness of LLM-based evaluation for multi-optioned question answering tasks. According to their findings, models have different performance when the prompts state the same meaning in various ways or have options with swapped orders.

Lately, Wang et al. [14] have pointed out to the discrepancy of the reasoning that LLMs generate when they are asked to accomplish complex arithmetic and commonsense reasoning tasks which end up to inconsistent outputs. Their suggestion is to sample a diverse set of reasoning paths through multiple runs of LLMs and return the most consistent answer from all the LLM generations. Wang et al. [13]'s approach relies on a similar idea of asking LLMs to repeatedly do pairwise comparison of responses by providing explanations. To alleviate the position bias of LLMs in pairwise comparisons, they propose to change the order of comparing options in the input prompts and return the average of the output scores. Both solutions are costly inefficient due to the need of frequent samples of LLMs. Zheng et al. [17] also pinpoint to the LLMs bias toward specific option IDs in multiple choice question tasks. Their approach is specifically for position bias and resorts to the probability distribution of output tokens to separate the LLM's prior bias for option IDs from the overall prediction distribution, which makes it limited and challenging for many of the closed API models.

3 INCONSISTENCY ASSESSMENT

The performance of LLM-based evaluators is directly affected by the quality of their input prompts. In this work, we focus on the consistency of these models' assessments given various semantically equivalent instructions. We study the consistency of LLMs' scoring for several perturbations applied to the guidelines that do not touch the semantics of the prompt.

The input prompts of all LLMs contain the task description, input dialogue to evaluate,[4] and scoring guidelines that explain each possible quality score. According to the task description, the

[1] https://www.anthropic.com/index/claude-2
[2] https://openai.com/
[3] https://mistral.ai/news/mixtral-of-experts/

[4] In one of the test datasets (CI dataset), we add knowledge-base as another input to the prompt.

Table 1: Statistics of human annotated quality scores for SGD, MultiWOZ2.1 and CI datasets including the number of samples in train/valid/test sets and the percentage of samples with 0, 1 or 2 labels.

Dataset	Train/Valid/Test	Classes 0/1/2
SGD	8674 / 1074 / 1085	22 / 30 / 48
MWOZ2.1	7648 / 952 / 953	27 / 39 / 34
CI	2553 / 319 / 318	65 / 35 / 0

LLMs are required to score the quality of the generation by taking into account the scoring options explained in the guidelines. Our goal is to check the consistency of these LLMs' output scores under three different applied perturbations to the input prompts that are discussed in the next section.

3.1 Perturbations

3.1.1 Position. Similar to [18] that study the position bias of LLMs in pairwise comparison of replies generated by two AI assistants, we examine the effect of position perturbations of guidelines on LLMs' outputs in a more general setup. We create prompts based on different orderings of the quality scores' categories, where all the categories are described roughly with the same length of tokens. In our experiments, we compute the consistency of LLMs' judgements for two and three category of evaluation labels where the guidelines can be presented in overall 4 and 6 distinct orders, respectively.

3.1.2 Length. Another significant attribute of the evaluation guidelines that we take into consideration is the number of tokens used to explain each scoring guideline. Zheng et al. [18] also point out to the length bias and how LLMs' favor in selecting longer responses in the pairwise comparison. Since scoring guidelines can be stated in various length and yet conveying same meaning, it is important to check whether the LLM evaluations remain consistent in such situations.

For length perturbations, we use the default order of the numerical labels but increase the length of one of the scoring options by adding more detailed explanation.

3.1.3 Position & Length. In the real world, it is plausible of having semantically equivalent prompts but with various order and length of scoring guidelines. To simulate such cases, we manipulate both the position and length of the guideline scores by preserving the prompts' meaning. To this end, we manipulate the order of guideline scores similar to the position only perturbations, and explain one of the labels more in details and provide a thorough guideline for that specific label.

Examples of position and length perturbations are displayed in Table 2.

3.2 Datasets

We conduct our study on three publicly available datasets. Consistency-Inconsistency (CI) is a binary human-labeled dataset showing if the TOD's response contradicts its previous statements in the conversation's history, the user's query, or the provided knowledge base. The other two non-binary datasets include User Satisfaction Scores (USS) for turns and overall conversations between human and TODs [12]. We leverage the annotated data by Sun et al. [12] on 1000 samples of Schema Guided Dialogue (SGD) [11] and MultiWOZ 2.1 [2] datasets that both span multi-domain task-oriented dialogues. The output scores are classified as 0 (dissatisfied) / 1 (moderately satisfied) / 2 (satisfied) [3]. Dataset statistics are provided in Table 1. To examine the inconsistencies in LLMs, we randomly select 100 samples from each group of conversations of the test sets.

3.3 LLMs

We conduct our study on three recent LLMs. Claude-V2 is a 137B parameter model released by Anthropic that is able to process contexts containing up to 100k tokens. GPT-3.5-turbo is an OpenAI model that accepts shorter contexts with around 4k tokens. Mistral AI's Mixtral-8x7B [5] model has 47B parameters and handles contexts with 32k tokens. We run all models with temperature 0 and top_k of 1 and generate output with maximum 100 tokens.

3.4 Metrics

To the sake of comparison and assessment of models we use the following metrics: Consistency and Accuracy. In order to measure the consistency of each model's predictions after passing perturbed prompts, we count the number of conversations that LLM assigns the identical output scores for **all** different perturbed guidelines. To elaborate more, we say that a model is **consistent** for an example if it returns the same option (correct or incorrect) for all permutation orders or length of the output options, otherwise it is inconsistent. In the consistency score calculation, we only care about the similarity of the LLM outputs for different input prompts. Table 3 illustrates how this metric works for a toy dataset containing five conversations and three different quality labels (0,1,2). As it is clear, for conversations 1 and 2 no matter what is the order of the scoring guidelines in the prompt the LLM's output scores remain the same. In this example, LLM has a consistency score of 40 showing that for 2 out of 5 conversations its output predictions do not change.

The second metric is accuracy which takes into account the performance of the LLM with respect to the ground-truth labels. The accuracy is computed per perturbation meaning that for each perturbed guideline in the prompt we compute the accuracy of LLM. As an illustration, in Table 3 for the input prompt with scoring options in the order of 1, 2 and then 0 the assessment accuracy is 0.6 since LLM's judgements match with the ground-truth labels for only three of the conversations. In this paper, we show the mean aggregation of LLM's accuracy resulted from all the prompts with various perturbations.

3.5 Inconsistency in LLM-based Evaluators

The ZS ('zero-shot') columns in Table 4 summarize both the accuracy and consistency of the LLMs' predictions after taking perturbed prompts without demonstrations. The consistency scores show the severity of the issue that these models currently face with to generate consistent judgements. The lower consistency and accuracy scores for SGD and MWOZ datasets show that LLMs confront a more serious challenge in achieving both consistent and correct

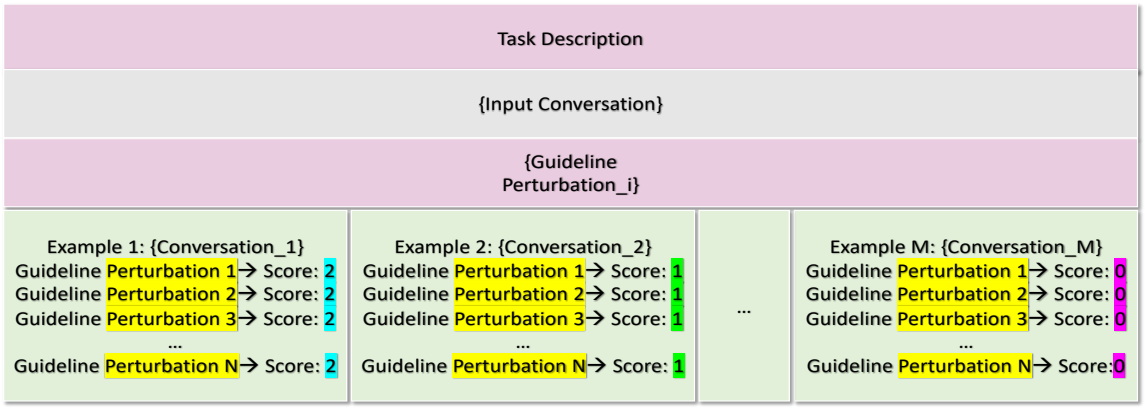

Figure 2: In-context learning for mitigating inconsistencies in LLM-based evaluations. The prompt includes the task description, the input conversation for the evaluation, the guideline, and a set of examples. Each example is accompanied by perturbed but semantically equal versions of guidelines and the same quality scores.

assessments when there are more number of plausible evaluation options.

We also find that Claude-V2 generates more inconsistent outputs with the length perturbed prompts, while GPT3.5 is mostly inconsistent with getting position perturbed prompts. A detailed breakdown of the LLMs' predictions can be found in Appendix B. This observation leads us to come up with an idea to tackle the issue. In the next section, we explain our proposed approach for this aim and compare it with existing baselines.

4 INCONSISTENCY MITIGATION

4.1 baselines

We compare the consistency of LLM-based evaluators after applying our approach across following baselines:

4.1.1 Self-Consistency. Self-consistency originally is designed for increasing the consistency of LLMs for arithmetic and commonsense reasoning benchmarks [14]. Wang et al. [13] suggest a similar technique of generating and aggregating multiple samples. We apply this method to the evaluation task by sampling 5 set of reasoning sets (with temperature of 0.7 and top_k of 20) for each input conversation and subsequently generating 5 scores. At the end, the majority of the scores is returned as the assessment score.

4.1.2 Explicit-Consistency. Another reasonable method that has also been examined by [13, 17] is to explicitly hint the goal in the input prompt. We directly emphasize the LLM to have consistent evaluations regardless of the length or the position of the guidelines. An example of input prompt for Explicit-Consistency is depicted in Table 8.

4.1.3 Chain-of-Thought. We compare our approach versus Chain-of-Thought (CoT) prompting that its beneficence has been revealed in many tasks [8, 15]. We ask LLM to first state its thought about the input conversation's quality and subsequently generate the assessment score.

4.2 Proposed Approach

As it is shown in Figure 2, our proposed approach for mitigating inconsistency in LLM-based evaluators relies on in-context learning of LLMs. Each input prompt includes four sections. The first three sections of the input prompt are similar to the ones in any ordinary LLM-based evaluator. First, it starts with a task description section that explains the goal and the task to the LLM. Then it is followed by the target conversation. The evaluation guidelines that reveal the range of output scores come next.

The main characteristics of our proposed prompt is the *demonstrations section that contains some conversations with* ***perturbed guidelines*** *and* ***consistent output scores***. In other words, each prompt contains M examples accompanied with the same output scores regardless of N distinct applied perturbations to the guideline. This implicitly instructs the model to pay attention to the stability of the scores when the guidelines are semantically alike. Indeed, this approach can easily be generalized to any kind of consistency behavior that we are looking for in LLM-based evaluators by simply applying those perturbations to the guidelines and combining them with constant scores. A full example prompt with demonstrations is shown in Table 7 in the supplementary materials, where the illustrated example contains one conversation and a set of perturbed order of guidelines. Regardless of the order of the guidelines the score is 1 for all of them.

4.3 Few-Shot Learning for Inconsistency Mitigation

Following the zero-shot results in Table 4 showing that LLMs are often inconsistent evaluators, we check the influence of our proposed approach on LLMs' consistency by rerunning all the experiments with our proposed augmented prompts. Based on models' context size, we test 1, 2, 3 and 10 random demonstrations added to the prompts for Claude-V2, Mistral-8x7b-v0.1, and GPT3.5-turbo. Our best results are obtained with two demonstrations. These results are shown in the **'few-shot'** columns of Table 4. (For full results for other numbers of demonstrations, see Tables 15, 16 and 17.)

Table 2: An example of position and length perturbations on the scoring guidelines.

Position Perturbations
Order 0-1-2 Score 0 the customer is very dissatisfied when the agent fails to fulfill the customer's request. Score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. Score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task.
Order 1-2-0 Score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. Score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task. Score 0 the customer is very dissatisfied when the agent fails to fulfill the customer's request.
Length Perturbations
Score 0 as the longest described scoring option Score 0 means the customer is dissatisfied when the followings happens: the agent is unable to accomplish the customer's request, the agent does not understand the customer's request, the agent provides irrelevant information, the agent summarizes something new this is not the customer's request, the agent asks the customer to confirm something irrelevant, the agent does not attempt to complete the required task. Score 1 means the customer is moderately satisfied when the agent only provides summaries or suggestions and asks for confirmation and does not completely accomplish the task. Score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task.
Score 1 as the longest described scoring option Score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request. Score 1 means the customer is normally satisfied when the agent understands the request and one of the followings happens: the agent summarizes the customer's request, the agent provides new relevant information, the agent suggests options, the agent gathers necessary details by asking follow-up questions, the agent asks the customer to confirm but have not received the confirmation yet, the agent will attempt to complete the required task. Score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task.

Figure 3 depicts the impact of few-shot perturbed examples on the consistency of Claude model for different perturbations. We observe that even though in general few-shot learning increases the consistency of LLMs predictions, adding more demonstrations does not necessarily improve consistency and accuracy. It is possible that selected demonstrations are not optimal; automatic optimization of demonstration choices [6] might show larger and more consistent improvements.

Comparing the consistency scores in few-shot versus zero-shot learning, we find *a positive impact of augmenting in-context examples to the LLMs' prompts to increase the consistency of outputs across different semantically equivalent guidelines*. The accuracy of models in zero- and few-shot setups in Table 4 also shows that increasing inconsistency does not hurt the accuracy of the models.

Table 5 illustrates the comparison of Claude and Mixtral evaluators' consistency scores after applying our approach alongside the baselines. The results show that our proposed approach results in more consistent evaluations by LLMs versus baselines. According to this table, nor directly asking LLMs to generate reasoning and consistent results neither repeatedly sampling LLMs and performing majority voting does not increase the consistency of the models as our proposed approach does while being cost-effective.

We also conduct an ablation study to investigate whether the resulted higher consistency of LLMs judgements is because of our proposed specific version of input demonstrations with perturbed guidelines and persistent scores or in general the in-context learning of LLM is the main reason of this success. To this end, instead of adding all different perturbed guidelines and consistent scores to the conversations discussed in Figure 2, we randomly select *only one perturbed guideline* and add them to each demo. We run this ablation study for prompts with two demonstrations known as our best setup.

Table 3: A toy example of showing the accuracy and consistency calculations. Each column in the Perturbations column shows the LLM's output after getting the corresponding perturbed order of three possible options in the prompt. In this example, for 2 out of 5 conversations the LLM's outputs remain consistent resulting a consistency score of 40%. The last row exhibits the accuracy score per LLM's predictions over various perturbed prompts. In Table 4, we only report the mean aggregation of the individual accuracy scores of LLMs predictions for various perturbed prompts.

	Perturbation Outputs						Ground	Consistent?
	012	021	102	120	201	210	truth	
Conv 1	2	2	2	2	2	2	0	✓
Conv 2	1	1	1	1	1	1	1	✓
Conv 3	0	1	1	0	0	2	2	✗
Conv 4	0	1	1	1	1	1	1	✗
Conv 5	1	0	2	2	2	2	2	✗
Accuracy	0.2	0.4	0.6	0.6	0.6	0.8		

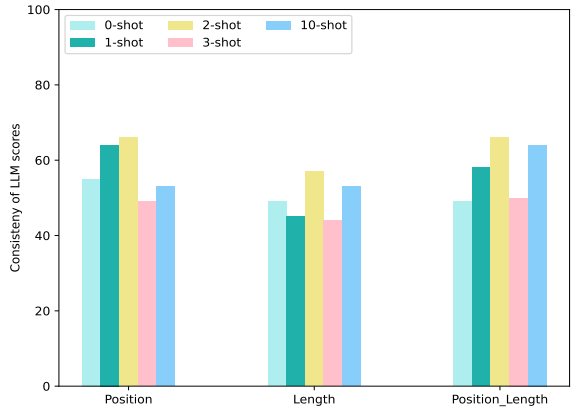

Figure 3: The Consistency of Claude-v2 outputs after applying position, length and both perturbations to the prompts with 0,1,2,3 and 10 augmented demonstrations.

Table 6 summarizes the importance of having perturbed guidelines and consistent judgements alongside the input conversations. According to this table, augmenting only demo conversations with one perturbed guideline does not improve the consistency of the LLM judgements. In other word, perturbed guidelines and constant judgements help the LLM to better perceive the consistency concept. This difference is more visible for both position and order perturbations of the guidelines as the changes become more complicated it is more challenging and necessary to let LLM know the overall goal.

5 CONCLUSION AND FUTURE WORK

The performance of LLMs in different applications, including automatic evaluation of TODs, is directly impacted by the input prompts and instructions. In this work, we explore the inconsistency of predictions by LLM-based evaluators. We show that in the case of having different syntactic versions of the same input guideline that result from perturbing the order of score options or the length of explanations, LLMs tend to generate inconsistent evaluation scores. We propose to apply in-context learning and augment examples of conversations with various manipulated guidelines and same assessment scores to the prompt to guide the model toward reaching the preferred consistent behavior.

In this work, we show that few-shot in-context learning does lead to more consistent LLM-based evaluators in comparison to baseline counterparts and it requires less time and cost sources. However, how to select optimum demonstrations and design a reliable prompt to reach the most consistency in LLM-based evaluations and concurrently preserve the accuracy of the metrics are still open problems. Hence, in future, we plan to conduct automatic prompt engineering in a more systematic way, with and without automatic optimization of prompts, to achieve these goals.

6 LIMITATIONS

In this work, we studied and showed the inconsistency of evaluations by LLM-based evaluators after manipulating the position and length of the guidelines in the prompts while keeping the semantics of the prompts similar. We mitigated the inconsistency issue by adding a few demonstrations of consistent outputs under various perturbations. While we showed that few-shot learning is capable of decreasing the inconsistency in LLM-based evaluators without hurting their accuracy, our work has some limitations:

First, we only examined the perturbations applied to the position and length of the guidelines. This study can be extended to improve consistency over other types of manipulations to make more robust LLM-based evaluators.

Second, another limitation of our work that should be addressed is to study a systematic way of adding the most optimized set of demonstrations to the prompt for solving the consistency issue in LLM-based evaluators.

REFERENCES

[1] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723* (2023).
[2] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669* (2019).
[3] Yue Feng, Yunlong Jiao, Animesh Prasad, Nikolaos Aletras, Emine Yilmaz, and Gabriella Kazai. 2023. Schema-Guided User Satisfaction Modeling for Task-Oriented Dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2079–2091. https://doi.org/10.18653/v1/2023.acl-long.116
[4] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166* (2023).
[5] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).
[6] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714* (2023).
[7] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023.

Table 4: Consistency and Accuracy of evaluations conducted by Claude-v2, GPT3.5 and Mixtral-8x7B models after getting different perturbed prompts. Bold numbers show higher consistency of LLMs outputs. Underlined numbers are for cases when few-shot (FS) learning does not hurt the accuracy of the LLM-based evaluators in comparison to zero-shot (ZS) setup.

Dataset	Perturbation	Metric	Claude-v2		GPT-3.5		Mixtral-8x7B	
			ZS	FS	ZS	FS	ZS	FS
SGD	Position	Cons.	63	63	46	**70**	57	55
		Acc.	0.43	0.50	0.43	0.39	0.40	0.45
	Length	Cons.	51	**59**	63	**91**	66	**72**
		Acc.	0.43	0.43	0.40	0.39	0.39	0.42
	Position+Length	Cons.	49	**65**	50	**52**	56	**68**
		Acc.	0.44	0.49	0.43	0.43	0.41	0.41
MultiWOZ	Position	Cons.	55	**66**	40	**66**	53	**60**
		Acc.	0.43	0.48	0.38	0.38	0.39	0.43
	Length	Cons.	49	**57**	61	**84**	61	**70**
		Acc.	0.43	0.45	0.37	0.36	0.37	0.38
	Position+Length	Cons.	49	**66**	49	**55**	58	**71**
		Acc.	0.43	0.46	0.40	0.39	0.40	0.38
CI	Position	Cons.	84	**85**	93	93	96	**99**
		Acc.	0.78	0.83	0.79	0.82	0.78	0.79
	Length	Cons.	80	76	86	**89**	92	**95**
		Acc.	0.75	0.76	0.81	0.69	0.76	0.78
	Position+Length	Cons.	81	**83**	88	**91**	97	95
		Acc.	0.74	0.72	0.84	0.83	0.79	0.77

Table 5: Consistency of evaluations conducted by Claude-v2 and Mixtral-8x7B models after applying different baselines such as Self-Consistency (Self.), Explicit-Consistency (Exp.), Chain-of-Thought (COT) and our proposed approach based on Few-Shot learning (FS). Bold numbers show higher consistency of LLMs outputs.

Dataset	Perturbation	Claude-v2				Mixtral-8x7B			
		COT	Exp.	Self.	FS	COT	Exp.	Self.	FS
SGD	Position	52	60	61	**63**	15	**57**	**57**	55
	Length	54	53	52	**59**	34	68	66	**72**
	Position+Length	55	46	46	**65**	21	59	56	**68**
MultiWOZ	Position	44	55	55	**66**	14	57	53	**60**
	Length	50	46	48	**57**	24	60	61	**70**
	Position+Length	46	45	46	**66**	17	60	58	**71**

Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491* (2023).

[8] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35, 22199–22213.

[9] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Yun-Nung Chen and Abhinav Rastogi (Eds.). Association for Computational Linguistics, Toronto, Canada, 47–58. https://doi.org/10.18653/v1/2023.nlp4convai-1.5

[10] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2511–2522. https://doi.org/10.18653/v1/2023.emnlp-main.153

[11] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8689–8696.

[12] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2499–2506.

[13] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).

Table 6: An ablation study to check the effectiveness of prompts with or without perturbed (No Pert.) guidelines on the consistency of the LLM-based evaluations.

Dataset	Perturbation	Claude-v2		Mixtral-8x7B	
		FS (With Pert.)	FS (No Pert.)	FS (With Pert.)	FS (No Pert.)
SGD	Position	63	57	55	47
	Length	59	51	72	70
	Position+Length	65	37	68	44
MultiWOZ	Position	66	49	60	42
	Length	57	48	70	71
	Position+Length	66	31	71	53

[14] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. The Eleventh International Conference on Learning Representations (ICLR).

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, Vol. 35. 24824–24837.

[16] Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Junbo Zhao, et al. 2023. Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility. *arXiv preprint arXiv:2305.10235* (2023).

[17] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*.

[18] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the NeurIPS 2023 Datasets and Benchmarks Track*.

A THE STRUCTURE OF PROMPTS FOR IN-CONTEXT LEARNING

The structure of in-context learning prompts for mitigating inconsistency issue resulted from different perturbations (position, length, their combination) is demonstrated in Figure 2. According to this figure, the task description and provided examples are constant for conversations and their guidelines. Each prompt has overall M demonstrations each with N underlying applied perturbations and their constant output scores. As an example, in USS datasets with three assessment labels, the position perturbations on the order of the labels will cause N=6 different perturbed guidelines while in the CI dataset with binary labels N is 2.

Table 7 shows a sample of input prompt of SGD dataset for mitigating inconsistencies coming from position perturbations in the guidelines. The "<transcript_here>" is substituted with the conversations to evaluate. The guideline for evaluating the input transcripts includes three satisfaction scores in the order of 2, 1 and then 0. In this example, the prompt contains one conversation as the demonstration for in-context learning alongside its different guidelines with the same output score of 1. We have only shown three out of six perturbed guidelines to save the space.

B INCONSISTENT EVALUATION SCORES

We report the percentage of the individual labels predicted by LLMs under different enforced perturbations to the prompt in Tables 9, 10, 11, 12, 13 and 14. The unsteady numbers across different setups reveal the inconsistency issue in LLMs. In SGD and MWOZ datasets, LLMs identify very limited number of conversations as cases when the user is dissatisfied with the agent's replies (low percentage of predicted label 0). This means according to the provided guidelines LLM believes the user's goal are at least partially fulfilled. It is noteworthy to mention that the description of the user satisfaction labels are similar to those originally shared with human to collect the annotations [12], which means that human has annotated the data following some hidden or personal patterns that complicates the task for the models.

B.1 Few-shot Learning

Tables 15, 16 and 17 show a comprehensive set of experiments conducted by augmenting different number of our proposed perturbed demonstrations to the input prompts and their impact on the consistency of the LLMs' scores. The results show that mostly by applying only two demonstrations with a perturbed list of guidelines and consistent scores the LLM gets the chance to understand the main concept of consistency and achieves the highest consistency.

By applying in-context learning with only one example the predictions of GPT3.5 become less accurate showing that output scores mostly are the same as the provided demo's label. To test whether GPT3.5 output relies on the in-context example's label, we change the demo example to other conversation with a different quality label. GPT3.5 predictions move toward the newly provided label. This shows that in comparison to other LLMs GPT3.5 is more on the influence of the only provided demo's label as it is repeated for different perturbed guidelines it could mislead LLM to predict the same label regardless of the input. When we incorporate more number of conversations with distinct labels the model's accuracy increases as the LLM understands that the provided repeated labels are to show the consistency concept.

Table 7: A prompt with an in-context demonstration for mitigating inconsistency in LLM's evaluation scores due to perturbed order of guidelines.

Example of in-context learning prompts
Below is a TRANSCRIPT of an interaction between a customer and an agent. In the TRANSCRIPT, customer's responses are prefixed by "customer:" and agent's responses are prefixed by "agent:". TRANSCRIPT: "<transcript_here>" Imagine you are the quality evaluator of the agent, considering this TRANSCRIPT and following the GUIDELINE please rate the customer's satisfaction mostly paying attention to the end of the transcript by selecting one of the following options. GUIDELINE: score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task. score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request or customer changes his or her mind at the end or don't agree or like the agent's suggestions. Please provide your rating starting with "Score:" and then the score value at the beginning line, followed by 'Reason:' and then your reasoning in one or two sentences to set right expectations. Example 1: TRANSCRIPT: " CUSTOMER: Find me a Gynecologist in San Jose. AGENT: Anjali Tate, M.D. is a Gynecologist in San Jose. CUSTOMER: What is their physical location and contact number? AGENT: You can contact them at 510-845-8035 and their physical location is 2400 Samaritan Drive #105. CUSTOMER: That is good for me. I need to see a doctor. AGENT: What day are you going? CUSTOMER: I want to go this Sunday. " GUIDELINE: score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request or customer changes his or her mind at the end or don't agree or like the agent's suggestions. score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task. Score: 1 GUIDELINE: score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request or customer changes his or her mind at the end or don't agree or like the agent's suggestions. score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task. Score: 1 GUIDELINE: score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task. score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task. score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request or customer changes his or her mind at the end or don't agree or like the agent's suggestions. Score: 1

Table 8: An example of prompt with a hint of having Explicit-Consistency versus position perturbed guidelines.

Example of in-context learning prompts

Below is a TRANSCRIPT of an interaction between a customer and an agent. In the TRANSCRIPT, customer's responses are prefixed by "customer:" and agent's responses are prefixed by "agent:".
TRANSCRIPT: "<transcript_here>"
Imagine you are the quality evaluator of the agent, considering this TRANSCRIPT and following the
GUIDELINE please rate the customer's satisfaction mostly paying attention to the end of the transcript by selecting one of the following options.
GUIDELINE:
score 0 means the customer is dissatisfied when the agent is unable to accomplish the customer's request or customer changes his or her mind at the end or don't agree or like the agent's suggestions.
score 1 means the customer is moderately satisfied when the agent only provides summary or suggestions and asks for confirmation and does not completely accomplish the task.
score 2 means the customer is satisfied when the agent accomplishes the customer's request and is able to COMPLETE the initial task.

Hint: The order of the descriptions in the scoring options does not matter. The scoring is based on the meaning, not the order. Please provide your rating starting with "Score:" and then the score value at the beginning line, followed by 'Reason:' and then your reasoning in one or two sentences to set right expectations.

Table 9: The distribution of satisfaction labels for prompts with perturbed order of guidelines.

Dataset	Model	0-1-2	0-2-1	1-0-2	1-2-0	2-0-1	2-1-0
SGD	Claude	4/43/53	8/32/60	5/22/73	4/27/69	9/21/70	8/19/73
SGD	GPT3.5	6/39/55	6/58/36	6/18/77	7/12/81	4/54/42	4/45/51
SGD	Mixtral	1/46/53	2/41/56	4/44/52	11/36/53	7/45/48	4/68/28
MWOZ	Claude	6/54/40	9/42/49	1/25/64	7/33/60	13/28/59	9/28/63
MWOZ	GPT3.5	6/55/39	4/68/28	8/25/67	11/17/72	5/61/34	6/56/38
MWOZ	Mixtral	0/50/50	1/45/52	5/43/52	10/37/53	9/44/47	6/61/33

Table 10: The distribution of consistency labels for prompts with perturbed order of guidelines.

Dataset	Model	Order 01	Order 10
CI	Claude	66 / 34	56 / 44
CI	GPT3.5	45 / 56	42 / 58
CI	Mixtral	28 / 72	31.5 / 68.5

Table 11: The distribution of satisfaction labels for prompts with different length guidelines.

Dataset	Model	same length	0 the longest	1 the longest	2 the longest
SGD	Claude	4/ 43/ 53	7/ 22/ 71	4/ 63/ 33	5/ 30/ 65
SGD	GPT3.5	6/ 39/ 55	5/21/ 74	5/ 44/ 51	5/ 10/ 85
SGD	Mixtral	0/50/50	1/46/53	2/16/82	1/22/77
MWOZ	Claude	6/ 54/ 40	9/33/58	5/78/17	7/36/57
MWOZ	GPT3.5	6/ 55/ 39	6/34/ 60	3/ 53/ 44	5/19/76
MWOZ	Mixtral	0/50/50	0/22/78	0/51/49	0/32/68

Table 12: The distribution of consistency labels for prompts with different length guidelines and the consistency of the output scores.

Dataset	Model	same length	0 the longest	1 the longest
CI	Claude	65 / 35	72 / 28	67 / 33
CI	GPT3.5	45 / 56	50 / 50	43 / 58
CI	Mixtral	28 / 72	31 / 69	23 / 77

Table 13: The distribution of satisfaction labels for prompts with perturbed order and score 1 as the most lengthy score described in the guidelines.

Dataset	Model	0-1-2	0-2-1	1-0-2	1-2-0	2-0-1	2-1-0
SGD	Claude	5/63/33	4/40/55	3/63/34	5/67/28	6/28/66	2/48/49
SGD	GPT3.5	4/44/52	3/75/22	8/68/24	7/44/49	3/62/35	3/64/33
SGD	Mixtral	0/62/38	0/54/46	3/74/23	5/55/40	5/62/33	1/67/32
MWOZ	Claude	5/78/17	4/58/38	4/70/26	4/77/19	7/37/56	5/53/42
MWOZ	GPT3.5	3/54/43	1/87/12	11/67/22	11/62/27	5/77/19	6/76/18
MWOZ	Mixtral	0/70/30	0/60/40	2/76/22	4/59/37	4/70/26	2/66/32

Table 14: The distribution of consistency labels for prompts with perturbed order and class 0 as the lengthy score described in the guidelines.

Dataset	Model	Order 01	Order 10
CI	Claude	64 / 36	80 / 20
CI	GPT3.5	53 / 47	43 / 58
CI	Mixtral	30.5 / 69.5	27.5 / 72.5

Table 15: Consistency (Cons.) and Accuracy (Acc.) of Claude-v2 after applying different perturbations to the prompts with 0,1,2,3 and 10 augmented demonstrations.

Dataset	Perturbation	Metric	Claude-v2				
			zero-shot	one-shot	two-shot	three-shot	ten-shot
SGD	Position	Cons.	63	63	63	52	51
		Acc.	0.43	0.52	0.50	0.53	0.35
	Length	Cons.	51	46	**59**	54	52
		Acc.	0.43	0.44	0.43	0.49	0.42
	Position+Length	Cons.	49	57	65	55	**70**
		Acc.	0.44	0.49	0.49	0.50	0.33
MultiWOZ	Position	Cons.	55	64	**66**	49	53
		Acc.	0.43	0.53	0.48	0.35	0.38
	Length	Cons.	49	45	**57**	44	53
		Acc.	0.43	0.44	0.45	0.46	0.44
	Position+Length	Cons.	49	58	**66**	50	64
		Acc.	0.43	0.46	0.46	0.44	0.28
CI	Position	Cons.	84	84	**85**	76	75
		Acc.	0.78	0.74	0.83	0.79	0.68
	Length	Cons.	80	**84**	76	74	82
		Acc.	0.75	0.58	0.76	0.64	0.56
	Position+Length	Cons.	81	84	83	85	**87**
		Acc.	0.74	0.64	0.72	0.69	0.73

Table 16: Consistency (Cons.) and Accuracy (Acc.) of GPT3.5 after applying different perturbations to the prompts with 0,1 and 2 augmented demonstrations.

Dataset	Perturbation	Metric	GPT3.5		
			zero-shot	one-shot	two-shot
SGD	Position	Cons.	46	62	**70**
		Acc.	0.43	0.44	0.39
	Length	Cons.	63	52	**91**
		Acc.	0.40	0.41	0.39
	Position+Length	Cons.	50	**53**	52
		Acc.	0.43	0.45	0.43
MultiWOZ	Position	Cons.	40	60	**66**
		Acc.	0.38	0.31	0.38
	Length	Cons.	61	49	**84**
		Acc.	0.37	0.30	0.36
	Position+Length	Cons.	49	52	**55**
		Acc.	0.40	0.26	0.39
CI	Position	Cons.	93	88	93
		Acc.	0.79	0.79	0.82
	Length	Cons.	86	78	**89**
		Acc.	0.81	0.78	0.69
	Position+Length	Cons.	88	84	**91**
		Acc.	0.84	0.73	0.83

Table 17: Consistency (Cons.) and Accuracy (Acc.) of Mixtral-8x7B-v0.1 after applying different perturbations to the prompts with 0,1,2,3 and 10 augmented demonstrations.

Dataset	Perturbation	Metric	Mixtral-8x7B-v0.1				
			zero-shot	one-shot	two-shot	three-shot	ten-shot
SGD	Position	Cons.	57	57	55	53	**59**
		Acc.	0.40	0.45	0.45	0.43	0.39
	Length	Cons.	66	69	**72**	62	63
		Acc.	0.39	0.41	0.42	0.42	0.42
	Position+Length	Cons.	56	59	68	71	**76**
		Acc.	0.41	0.43	0.41	0.38	0.37
MultiWOZ	Position	Cons.	53	54	60	55	**66**
		Acc.	0.39	0.43	0.43	0.42	0.39
	Length	Cons.	61	67	**70**	68	66
		Acc.	0.37	0.38	0.38	0.38	0.37
	Position+Length	Cons.	58	65	71	74	**78**
		Acc.	0.40	0.41	0.38	0.39	0.37
CI	Position	Cons.	96	70	**99**	98	97
		Acc.	0.78	0.64	0.79	0.79	0.80
	Length	Cons.	92	93	**95**	**95**	91
		Acc.	0.76	0.77	0.78	0.79	0.81
	Position+Length	Cons.	97	78	95	97	95
		Acc.	0.79	0.71	0.77	0.79	0.81