X-EVAL: Generalizable Multi-aspect Text Evaluation via Augmented Instruction Tuning with Auxiliary Evaluation Aspects

Minqian Liu[♠] **Ying Shen**[♠] **Zhiyang Xu**[♠] **Yixin Cao**[♣]
Eunah Cho[♡] **Vaibhav Kumar**[♡] **Reza Ghanadan**[♡] **Lifu Huang**[♠]
[♠]Virginia Tech [♡]Amazon Inc. [♣]Fudan University
{minqianliu, yings, zhiyangx, lifuh}@vt.edu
{eunahch, kvabh, ghanadan}@amazon.com caoyixin2011@gmail.com

Abstract

Natural Language Generation (NLG) typically involves evaluating the generated text in various aspects (e.g., consistency and naturalness) to obtain a comprehensive assessment. However, multi-aspect evaluation remains challenging as it may require the evaluator to generalize to any given evaluation aspect even if it's absent during training. In this paper, we introduce X-EVAL, a two-stage instruction tuning framework to evaluate text in both seen and unseen aspects customized by end users. X-EVAL consists of two learning stages: the vanilla instruction tuning stage that improves the model's ability to follow evaluation instructions, and an enhanced instruction tuning stage that exploits the connections between fine-grained evaluation aspects to better assess text quality. To support the training of X-EVAL, we collect ASPECTINSTRUCT, the first instruction tuning dataset tailored for multi-aspect NLG evaluation spanning 27 diverse evaluation aspects with 65 tasks. To enhance task diversity, we devise an augmentation strategy that converts human rating annotations into diverse forms of NLG evaluation tasks, including *scoring, comparison, ranking*, and *Boolean question answering*. Extensive experiments across three essential categories of NLG tasks: dialogue generation, summarization, and data-to-text coupled with 21 aspects in meta-evaluation, demonstrate that X-EVAL enables even a lightweight language model to achieve a comparable if not higher correlation with human judgments compared to the state-of-the-art NLG evaluators like GPT-4. [1]

1 Introduction

Recent advancements of pre-training (Chung et al., 2022; Touvron et al., 2023a,b), prompting (Brown et al., 2020; Wei et al., 2022b; Wang et al., 2023; Yao et al., 2023; Qi et al., 2023), and instruction tuning (Wei et al., 2022a) have improved the quality of machine generated texts by a significant degree. Nevertheless, the evaluation of various Natural Language Generation (NLG) tasks still lags far behind compared with the rapid progress of large language models (LLMs). Previous similarity-based metrics such as ROUGE (Lin, 2004), BLUE (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020) predominantly measures the similarity between the generated and reference text, failing to accurately reflect the quality of generated text (Gehrmann et al., 2023), especially for open-ended generation tasks.

To obtain a more comprehensive assessment of text quality, multi-aspect evaluation (Fabbri et al., 2021) has been proposed to evaluate the generated text from multiple fine-grained evaluation *aspects*, such as `fluency` and `consistency`. While most

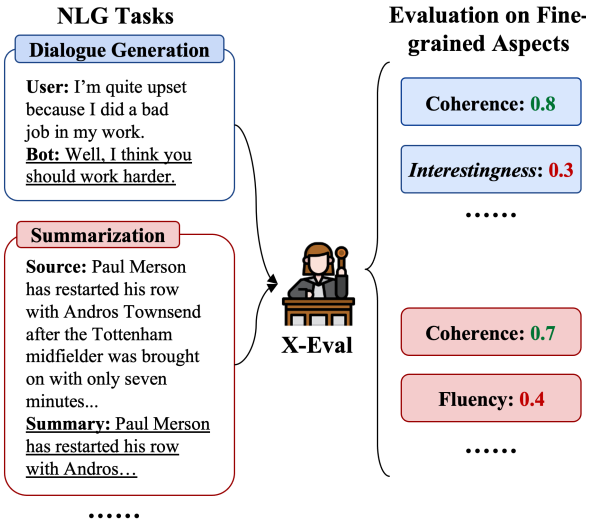

Figure 1: Illustration of X-EVAL for multiple seen and unseen fine-grained evaluation aspects across various NLG tasks. The unseen aspect (i.e., `Interestingness`) is highlighted in *italics*. The text to be evaluated is highlighted with underline. In this example, each evaluation score is from 0 to 1. The higher score indicates better quality.

[1]The source code, model checkpoints, and datasets are publicly available at https://github.com/VT-NLP/XEval for research purposes.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8560–8579
June 16-21, 2024

existing studies (Mehri and Eskenazi, 2020b; Yuan et al., 2021; Zhong et al., 2022) consider a closed set of aspects, in many realistic scenarios, the users may need to evaluate the text with their customized aspects and specifications, calling for building an evaluator that can be flexibly extended to any *unseen* aspects without the need of training data. Recent studies (Fu et al., 2023; Liu et al., 2023) propose to leverage large language models (LLMs) such as GPT-4 (OpenAI, 2023) as NLG evaluators, yielding promising zero-shot performance on unseen aspects. However, such evaluations, especially with proprietary LLMs, are cost-intensive, time-consuming, and pose concerns about data privacy and reproducibility.

In this work, we propose X-EVAL, an automatic evaluation framework that can conduct fine-grained evaluation on both seen and unseen aspects across various NLG tasks with a single model, as illustrated in Figure 1. X-EVAL follows a two-stage training paradigm: we first instruction-finetune an open-source language model to equip it with the capability of following human-written instructions for evaluation. Then, motivated by the observation that evaluation aspects usually exhibit interconnections (Fu et al., 2023) and thus their evaluations can benefit each other, we introduce an additional training stage to finetune the model on the instruction-tuning tasks enriched with the evaluations of a set of *auxiliary aspects*, which are expected to provide clues for evaluating the target aspect and encourage consistent evaluations across multiple aspects. During training, for each target aspect, we take all the remaining aspects defined in the corresponding dataset as auxiliary aspects and incorporate their gold evaluations into the instructions for the second-stage tuning. During inference, given the target aspect, we first select a set of auxiliary aspects based on the similarity of the aspect definitions and predict the evaluation result for each auxiliary aspect using the trained model. We then re-perform the evaluation for each target aspect by incorporating the results of auxiliary aspects.

To support our proposed two-stage training of X-EVAL, we construct ASPECTINSTRUCT, the first multi-aspect evaluation instruction tuning dataset spanning 27 diverse evaluation aspects over 65 tasks. This dataset is anchored around three core categories of NLG tasks: dialogue, summarization, and data-to-text. In light of insights from previous studies in instruction tuning (Wei et al., 2022a; Xu et al., 2023b), which emphasize the advantage of task diversity in enhancing zero-short generalization, we further augment the dataset by converting the original human rating data into diverse forms of NLG evaluation tasks, including *scoring, comparison, ranking* and *Boolean question answering*. In addition, to incorporate auxiliary aspects, we manually create templates that convert the numerical evaluation scores of each aspect into descriptions in natural language.

The main advantages of our approach are highlighted as follows: **(1) Generalization ability:** we introduce X-EVAL that can be flexibly generalized to evaluate the unseen NLG tasks or the aspects customized by user instructions in a zero-shot manner with a single model; **(2) Strong performance with high efficiency:** with significantly less amount of model parameters (780M), X-EVAL achieves strong performance compared to the state-of-the-art LLM-based evaluators (including GPT-4) demonstrated through comprehensive experiments; **(3) Reference-free and open-source:** our evaluator does not require gold reference to perform evaluation and it is more reliable and transparent thanks to its open-source nature.

2 Related Work

Similarity-based Metrics The previously dominant text evaluation paradigm is to predict a one evaluation score, where most of them are similarity-based metrics, including metrics that measure the surface overlap between the generated and reference text, such as ROUGE (Lin, 2004), BLUE (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005), as well as metrics measuring the distance between the contextualized embeddings of the generated text and the reference as the similarity score, such as BERTScore (Zhang* et al., 2020) and MoverScore (Zhao et al., 2019). Although these metrics are widely adopted, they often overlook fine-grained aspects and later study (Gehrmann et al., 2023) has proven that they fail to truly capture the quality of text with the coarse-grained score.

Multi-Aspect Metrics To conduct a more holistic evaluation, recent studies (Wang et al., 2020a; Huang et al., 2020) propose to evaluate the NLG systems via multiple fine-grained aspects. UniEval (Zhong et al., 2022) proposes to re-frame NLG evaluation into a QA format and perform multi-aspect evaluation with a single model via continual learning (Madotto et al., 2021; Liu et al.,

2022; Liu and Huang, 2023). However, UniEval cannot maintain robust performance when generalizing to novel aspects. To obtain an evaluator that can be generalized to customized aspects, some recent studies (Fu et al., 2023; Liu et al., 2023) harness proprietary LLMs to perform fine-grained evaluation in a zero-shot manner. However, due to the closed-source nature, these evaluation metrics suffer from issues of reproducibility and are prohibitively expensive. More recently, some concurrent studies (Xu et al., 2023a; Jiang et al., 2023; Mehri and Shwartz, 2023) propose to extract instruction-following data from proprietary LLMs for finetuning a more lightweight model as the evaluator. Nevertheless, they still require high costs to call the APIs to obtain a large amount of training data and it is non-trivial to ensure the data are of high quality. In addition, to the best of our knowledge, we are the first to meticulously curate the instruction-tuning dataset and train an instruction-based evaluator for dialogue evaluation.

3 ASPECTINSTRUCT

3.1 Problem Definition

Multi-aspect automatic text evaluation aims to evaluate the quality of NLG system's output x given a set of evaluation aspects $\mathcal{A}$ (e.g., `coherence`, `naturalness` and so on), and optionally an additional set of texts $\mathcal{S}$ (e.g., the source documents for text summarization, or context for dialogue evaluation). The evaluation task can be formulated as:

$$c = f(x, \mathcal{S}, a)$$

where $a \in \mathcal{A}$ is the fine-grained aspect to be evaluated, and $f(\cdot)$ is the scoring function that provides an assessment c w.r.t. the aspect a.

3.2 Data Collection

We aim to build a unified automatic evaluation framework that can assess the text quality for both seen and unseen evaluation aspects across various NLG tasks via instruction tuning. To this end, we build an instruction-tuning dataset tailored for multi-aspect evaluation, namely ASPECTINSTRUCT, with the following steps:

Existing Dataset Collection We first collect 10 existing evaluation datasets with human annotations for 3 representative categories of NLG tasks, including dialogue generation (Sai et al., 2020; Gunasekara et al., 2020; Pang et al., 2020; Gopalakrishnan et al., 2019; Mehri and Eskenazi, 2020a), text summarization (Völske et al., 2017; Fabbri et al., 2021; Wang et al., 2020b; Zhong et al., 2022), and data-to-text (Wen et al., 2015).

Task Augmentation The original datasets we collect only contain numerical scores annotated by humans, which severely limits the diversity of instruction-tuning tasks. Thus, we further derive diverse forms of evaluation tasks from the original annotations to enhance task diversity. Denote the ground truth score for text x_i as y_i. We derive four types of tasks based on this annotation: **(1) Scoring:** we ask the model to directly predict a discrete score (e.g., in the Likert scale) where we map the continuous ground truth y_i into a discrete scale; **(2) Comparison:** we sample two texts x_i and x_j for an identical context, e.g., two versions of summaries for the same source document, and ask the model to select the text with the higher evaluation score; **(3) Ranking:** we further extend the comparison task into ranking by sampling three candidates under the same context and ask the model to predict the correct ranking of the candidates based on the text quality; **(4) Boolean Question Answering:** we also formulate evaluation as a Boolean QA task following (Zhong et al., 2022) by asking the model a question such as "*Is this response fluent?*" and let the model predict "*Yes*" or "*No*".

Instruction Creation Finally, we define a unified instructions format for tasks included in ASPECTINSTRUCT. Each instruction consists of three parts: (1) *task description* that briefly introduces the evaluation task, (2) *aspect definition*, and (3) *evaluation protocol* that details what the model should output to perform the evaluation. We present the detailed procedure for instruction annotation in Appendix A.1. We provide an example of the original annotation, and the derived evaluation tasks along with the curated instructions in Figure 6 in Appendix A.2. The full list of evaluation aspects and the collected instructions can be found in Appendix A.3.

Statistics In total, we construct 65 tasks in ASPECTINSTRUCT, where we split 32 tasks and 14 seen aspects for instruction tuning and 33 tasks and 13 unseen aspects for meta-evaluation. We collect 72,637 instances in total with 55,602 instances for training and 17,035 instances for inference. Note that there is no overlap among the datasets used for training and inference. We consider two aspects that have identical aspect names but are in

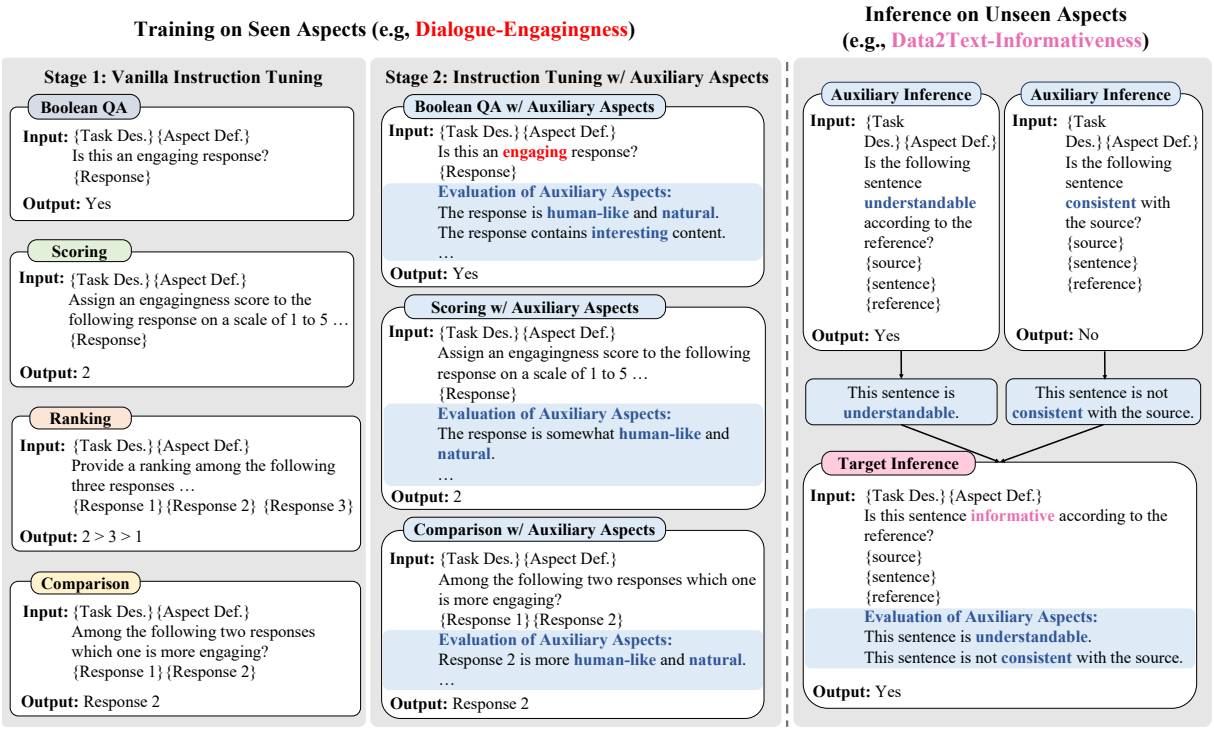

Figure 2: **Illustration of our X-EVAL framework.** The left section depicts our two-stage training approach: vanilla instruction tuning on diverse tasks and subsequent training on instruction tasks enriched with auxiliary aspects. The right section illustrates the inference pipeline with auxiliary aspects.

different NLG tasks as distinct aspects. We include more details about the source datasets, constructed instruction-tuning tasks, and the number of instances of each task in Appendix A.2.

4 X-EVAL

4.1 Two-Stage Instruction Tuning

Figure 2 presents an overview of X-EVAL, which consists of two stages of instruction tuning:

Vanilla Instruction Tuning The first training stage aims to equip the model with the ability to follow instructions to perform diverse evaluation tasks. We adopt Flan-T5 (Chung et al., 2022), an open-source language model as the base model for our evaluator. Based on Flan-T5, we further perform standard instruction tuning on the mixture of four types of tasks: *scoring, comparison, ranking*, and *Boolean QA*, as elaborated in Section 3.2.

Instruction Tuning with Auxiliary Aspects Through our study, we discern that certain evaluation aspects could be interrelated. As evidence, in dialogue evaluation (Gopalakrishnan et al., 2019) the aspect `naturalness` usually shows a notable correlation with `engagingness`. When a dialogue response is not `natural`, it is very likely that human considers the response to be not `engaging`. While these two aspects are not interchangeable given their different definitions, the evaluation of one aspect can offer useful clues for the evaluation of another potentially related aspect. Motivated by this, we enrich our training regimen with an additional instruction tuning stage to leverage potential connections to the target evaluation aspect.

More precisely, for each instruction-tuning task detailed in Section 3.2, we augment it based on the ground truth evaluation results of a predefined set of auxiliary aspects which are all other aspects collected in the source dataset. To convert the evaluation results of auxiliary aspects into natural language that can be fed into the input, we employ a template-based verbalizer, denoted as $v(\cdot)$, which takes in an aspect a and its evaluation score s for an instance, mapping it into a verbalized evaluation $h = v(s, a)$. For example, with the aspect `Consistency` on Data2Text and the evaluation score 0.9 out of 1.0, the verbalized result is phrased as "*This sentence is consistent with the source.*" (see more details in Appendix B). We construct the set of verbalized results $\mathcal{H}$ with the verbalizer for each auxiliary aspect (except for the target aspect). This set $\mathcal{H}$ is then concatenated into the additional

Algorithm 1: Inference Pipeline

Input: Set of evaluation aspects $\mathcal{A}$, Target aspect a_t, NLG system's output x, Additional set of texts $\mathcal{S}$, Scoring function $f(\cdot)$, Evaluation verbalizer $v(\cdot)$, Similarity measure $sim(\cdot)$, Sentence encoder $\mathcal{E}$

Output: Target score c_t

```
// Determine top-k auxiliary aspects
```

$L \leftarrow \{(sim(\mathcal{E}(a), \mathcal{E}(a_t)), a) \mid a \in \mathcal{A} \setminus \{a_t\}\}$

Sort L in descending order based on similarity

$\mathcal{A}^R \leftarrow$ first k aspects from sorted L

```
// Generate verbalized evaluation
   results for auxiliary aspects
```

Initialize an empty auxiliary evaluation set $\mathcal{H}$

for $a_r \in \mathcal{A}^R$ **do**

```
    // Score for auxiliary aspect
```

$c_r \leftarrow f(x, \mathcal{S}_r, a_r)$

```
    // Add verbalized evaluation to the
       auxiliary evaluation set
```

$\mathcal{H} \leftarrow [\mathcal{H}; v(c_r, a_r)]$

$\mathcal{S}_t \leftarrow [\mathcal{S}_t; \mathcal{H}]$

```
// Evaluate the target aspect
```

$c_t \leftarrow f(x, \mathcal{S}_t, a_t)$

return c_t

set of texts in the evaluator's input. The model then undergoes the second training stage on the instruction tasks enriched with these evaluation results.

4.2 Inference with Auxiliary Aspects

At the inference stage, we perform the following steps to evaluate the text on the target aspect: **First**, we select a set of auxiliary aspects for the target aspect. Based on the definitions of the target aspect and a pool of candidate aspects, we employ Sentence-T5 (Ni et al., 2022) to encode the definitions and measure the similarity between the sentence embeddings of target aspect definition and each candidate aspect definition. We select the aspects with top-k similarity scores as the auxiliary aspects to limit inference cost, where k is a hyperparameter. **Second**, we run an inference process using the Boolean QA task format, where the model predicts either *"Yes"* or *"No"*, as outlined in Section 3.2, on each auxiliary aspect. We convert the prediction into natural language results with the verbalizer. These verbalized results, denoted as $\mathcal{H}$, are subsequently integrated into the additional set of texts $\mathcal{S}$ for evaluating the target aspect. **Finally**, given the input enhanced by auxiliary aspects, we adopt the same Boolean QA format to compute the evaluation score c for the target aspect:

$$c = \frac{P(\text{“}Yes\text{”}|x, \mathcal{S}, a)}{P(\text{“}Yes\text{”}|x, \mathcal{S}, a) + P(\text{“}No\text{”}|x, \mathcal{S}, a)}$$

where $P(\cdot)$ denotes the probability of the model generating a specific word. The pseudo-code of our inference pipeline is in Algorithm 1.

5 Experiment Setup

Meta Evaluation We meta-evaluate our X-EVAL on the test split of ASPECTINSTRUCT, where the details of the test set are introduced as follows. For text summarization, we adopt SummEval (Fabbri et al., 2021) and QAGS (Wang et al., 2020b). For dialogue generation, we employ Topical-Chat (Gopalakrishnan et al., 2019) and FED (Mehri and Eskenazi, 2020a). For data-to-text generation, we utilize SFHOT & SFRES (Wen et al., 2015). ASPECTINSTRUCT contains the following *unseen* aspects: `topic depth` (DEP), `likeability` (LIK), `understandability` (UND), `flexibility` (FLE), `informativeness` (INF), `inquisitiveness` (INQ), `interestingness` (INT), `specificity` (SPE), `correctness` (COR), and `semantic appropriateness` (SEM). More detailed descriptions of the test splits, as well as seen and unseen evaluation aspects, are be found in Appendix A.4.

Implementation Details We adopt Flan-T5-large (with ~780M parameters) as our base language model for subsequent finetuning. Without specification, we pick the top-1 aspect during inference, i.e., $k = 1$. More implementation details can be found in Appendix C.

Baselines We compare our X-EVAL with the following state-of-the-art NLG evaluation metrics: **(1) UniEval** (Zhong et al., 2022) is a unified multi-aspect evaluator that re-frames the evaluation process as a Boolean QA task; **(2) GPTScore** (Fu et al., 2023) is a multi-faceted and training-free evaluation framework that utilizes the output probabilities from LLMs to score generated texts; **(3) G-Eval** (Liu et al., 2023) proposes to leverage large language models such as GPT-3.5 or GPT-4 to assess the text quality with form-filling paradigm in a training-free manner; **(4) ROUGE-L** (Lin, 2004); **(5) DynaEval** (Zhang et al., 2021); **(6) BERTScore** (Zhang* et al., 2020); **(7) MoverScore** (Zhao et al., 2019); **(8) USR** (Mehri and Eskenazi, 2020b); **(9) BARTScore** (Yuan et al., 2021). We include more details of baselines (4)-(9) in Appendix C due to space limit.

Variants of X-EVAL We design several variants of X-EVAL for ablation studies: **(1) X-EVAL w/o**

Metrics	Dialogue-level							Turn-level					
	DEP	LIK	UND	FLE	INF	INQ	AVG	INT	SPE	COR	SEM	UND	AVG
BARTScore (Yuan et al., 2021)	0.082	0.099	-0.115	0.093	0.092	0.062	0.052	0.159	0.083	0.076	0.100	0.120	0.128
DynaEval (Zhang et al., 2021)	0.498	0.416	0.365	0.383	0.426	0.410	0.416	0.327	0.346	0.242	0.202	0.200	0.263
UniEval (Zhong et al., 2022)	0.046	0.009	-0.024	-0.003	-0.070	0.085	0.030	0.435	**0.381**	0.125	0.051	0.082	0.215
GPTScore (GPT-3-d01) (Fu et al., 2023)	**0.669**	**0.634**	0.524	0.515	**0.602**	0.503	**0.574**	0.501	0.214	0.434	**0.444**	0.365	0.392
GPTScore (GPT-3-d03) (Fu et al., 2023)	0.341	0.184	0.196	0.072	0.317	-0.101	0.168	0.224	0.151	0.428	0.405	0.311	0.304
G-Eval (GPT-3.5)† (Liu et al., 2023)	0.339	0.392	0.123	0.344	0.232	0.101	0.259	0.30	0.280	0.430	0.390	0.274	0.335
G-Eval (GPT-4)† (Liu et al., 2023)	0.583	0.614	**0.602**	**0.587**	0.510	**0.551**	0.573	**0.506**	0.368	**0.522**	0.443	**0.438**	**0.455**
X-EVAL (Ours)	0.583	0.436	0.588	0.324	0.480	0.497	0.485	0.421	0.370	0.492	0.376	0.332	0.398
- w/o Training	0.377	0.387	0.394	0.424	0.370	0.417	0.395	0.250	0.175	0.296	0.289	0.225	0.247
- w/o Instructions	0.350	0.333	0.495	0.355	0.425	0.435	0.399	0.477	0.353	0.203	0.255	0.211	0.300
- w/o Stage-Two Tuning	0.388	0.324	0.555	0.384	0.582	0.437	0.445	0.372	0.282	0.418	0.329	0.311	0.342

Table 1: **Meta-evaluation on dialogue** based on *unseen* aspects in terms of dialogue-level and turn-level Spearman (ρ) correlations on FED. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with underline. †: our re-implementation, where we adopt our annotated instructions and aspect definitions as inputs to OpenAI's API to obtain the performance of G-Eval on FED.

Metrics	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L (Lin, 2004)	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BERTScore (Zhang* et al., 2020)	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR (Mehri and Eskenazi, 2020b)	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval (Zhong et al., 2022)	0.480	0.512	0.518	0.609	0.544	0.563	0.462	0.456	0.501	0.535
G-Eval (GPT-3.5) (Liu et al., 2023)	0.532	0.539	0.519	0.544	**0.660**	**0.691**	0.586	0.567	0.574	0.585
G-Eval (GPT-4) (Liu et al., 2023)	**0.549**	**0.565**	**0.594**	0.605	0.627	0.631	0.531	0.551	**0.575**	0.588
X-EVAL (Ours)	0.417	0.478	0.558	0.622	0.449	0.593	**0.734**	**0.728**	0.540	**0.605**
- w/o Training	0.054	0.051	0.063	0.073	0.258	0.298	0.427	0.436	0.200	0.214
- w/o Instructions	0.415	0.452	0.560	0.574	0.397	0.532	0.690	0.701	0.515	0.565
- w/o Stage-Two Tuning	0.396	0.446	0.581	**0.642**	0.408	0.569	0.725	0.706	0.528	0.592

Table 2: Turn-level Pearson (r) and Spearman (ρ) correlations on *seen* aspects on Topical-Chat. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with underline.

Training denotes the original Flan-T5 (without any further finetuning on our proposed ASPECTINSTRUCT); **(2) X-EVAL w/o Instructions**: based on Flan-T5, we only conduct prompt-based multi-task training and inference in the same way as (Zhong et al., 2022) where we ask the model to answer Boolean questions without using aspect definitions; **(3) X-EVAL w/o Stage-Two Tuning**: for this variant, we only conduct vanilla instruction tuning in Stage 1 based on Flan-T5. During inference, we directly perform evaluation based on the instructions without using auxiliary aspects.

6 Main Results

We report the main results of dialogue evaluation in Table 1 and Table 2, summarization in Table 3 and Table 9 , and data-to-text in Table 4. Each table is divided into three sections: the top section delineates the performance of traditional metrics and evaluators based on lightweight language models. The middle section shows the performance of the evaluators based on GPTs (Brown et al., 2020; OpenAI, 2023) that are proprietary and much larger than our approach. The bottom section shows the performance of X-EVAL and its variants.

Results of Dialogue Evaluation on FED To assess X-EVAL's ability to generalize to *unseen* aspects, we present the Spearman correlation on FED in Table 1. X-EVAL surpasses the baselines in the top section. Also, X-EVAL matches the performance of GPT-based baselines with much fewer parameters. The bottom section of the table highlights the improvement achieved by two-stage tuning, incorporating instructions, and integrating auxiliary aspects. It is worth noting that UniEval achieves notably poor performance on dialogue-level evaluation on FED, which is probably due to UniEval being overfitted to turn-level evaluation and failing to generalize to dialogue-level evaluation.

Results of Dialogue Evaluation on Topical-Chat We also evaluate the performance for the *seen* aspects on Topical-Chat and report the results in Table 2. Notably, in addition to the superior performance over lightweight baselines, X-EVAL also

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L (Lin, 2004)	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
MOVERSscore (Zhao et al., 2019)	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BERTScore (Zhang* et al., 2020)	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
BARTScore (Yuan et al., 2021)	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval (Zhong et al., 2022)	0.495	0.374	0.435	0.365	0.419	0.346	0.424	0.327	0.443	0.353
GPTScore (Fu et al., 2023)	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-Eval (GPT-3.5) (Liu et al., 2023)	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
G-Eval (GPT-4) (Liu et al., 2023)	**0.582**	**0.457**	**0.507**	**0.425**	0.455	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
X-EVAL (Ours)	0.530	0.382	0.428	0.340	**0.461**	0.365	0.500	0.361	0.480	0.362
- w/o Training	0.187	0.131	0.193	0.152	0.135	0.104	0.444	0.325	0.240	0.178
- w/o Instructions	0.458	0.333	0.414	0.328	0.395	0.309	0.496	0.359	0.441	0.333
- w/o Stage-Two Tuning	0.536	0.385	0.413	0.326	0.455	0.360	0.503	0.363	0.476	0.359

Table 3: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval. All aspects are *seen* aspects. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with underline.

Metrics	SFRES		SFHOT		AVG
	NAT	INFO	NAT	INFO	
ROUGE-L	0.169	0.103	0.186	0.110	0.142
BERTScore	0.219	0.156	0.178	0.135	0.172
MOVERScore	0.190	0.153	0.242	0.172	0.189
BARTScore	0.289	0.238	0.288	0.235	0.263
UniEval (Summ)	0.333	0.225	0.320	0.249	0.282
GPTScore	0.190	0.232	0.036	0.184	0.161
G-Eval (GPT-3.5)†	0.144	0.118	0.072	0.102	0.109
G-Eval (GPT-4)†	**0.351**	0.189	**0.338**	0.198	0.269
X-EVAL (Ours)	0.316	**0.265**	0.322	**0.310**	**0.303**
- w/o Training	0.240	0.192	0.207	0.262	0.225
- w/o Instructions	0.303	0.255	0.297	0.277	0.283
- w/o Stage-Two Tuning	0.322	0.257	0.311	0.292	0.295

Table 4: Spearman correlation on the data-to-text NLG task. NAT and INFO indicate `Naturalness` and `Informativeness`, respectively. The best results are highlighted in **bold**. †: our re-implementation.

surpasses all GPT-based metrics in averaged Spearman correlation. We notice that the correlation of X-EVAL on `groundedness` is notably higher than other baselines. One plausible reason is that Flan-T5 has been finetuned on related tasks such as natural language inference (Chung et al., 2022), as X-EVAL w/o Training has achieved decent performance without finetuning on ASPECTINSTRUCT.

Results of Summarization Evaluation We use summary-level Spearman and Kendall-Tau correlation to assess various evaluators on SummEval. Note that all the aspects in SummEval are seen aspects. From Table 3, X-EVAL surpasses lightweight evaluators in averaged Spearman correlation and outperforms both GPTScore and G-Eval (GPT-3.5). G-Eval (GPT-4) consistently excels across all aspects. We speculate this may stem from

Metrics	Topic.	FED	Summ.	D2T	AVG
X-EVAL (w/o STT)	**0.592**	**0.375**	**0.480**	0.295	**0.436**
- w/o Scoring	0.547	0.281	0.438	**0.300**	0.392
- w/o Comparison	0.554	0.347	0.448	0.293	0.411
- w/o Ranking	0.591	0.354	0.433	0.252	0.408
- w/o QA	0.579	0.357	0.418	0.284	0.410

Table 5: Ablation study on stage one instruction tuning task type (Spearman correlation). "w/o STT" denotes the model does not use Stage-Two Tuning. The best results are highlighted in **bold**.

GPT-4's strong ability to handle long input contexts. In addition, we report the results on QAGS in Table 9 in Appendix due to the space limit.

Results of Unseen NLG Task Evaluation In this experiment, we evaluate X-EVAL on the unseen data-to-text generation task. Table 4 shows that while X-EVAL experiences a slight performance loss in `naturalness` compared to G-Eval (GPT-4), it consistently excels over all other baselines across all aspects. This underscores the generalization capability of X-EVAL on unseen NLG tasks.

7 Discussions

Ablation Study of Instruction Tuning Tasks We conduct ablation studies to investigate the contribution of incorporating diverse forms of evaluation tasks during instruction tuning. Table 5 shows the averaged Spearman correlation on each meta-evaluation dataset. In general, X-EVAL trained on the combination of all forms of evaluation tasks, including *scoring, comparison, ranking*, achieves the highest averaged correlation for nearly all tasks.

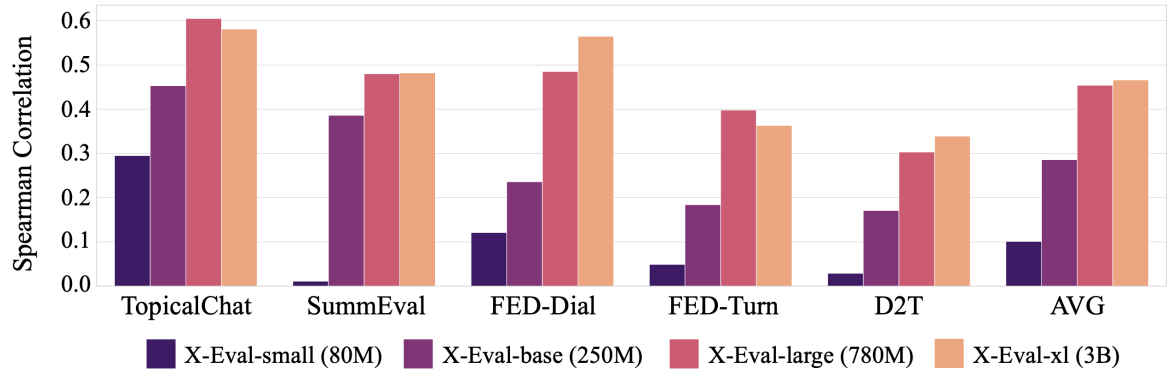

Figure 3: Effect of the scale of language model backbones. For each meta-evaluation benchmark, we report the average Spearman correlation on all the aspects. X-EVAL-large (780M) is the default backbone language model throughout all the experiments if there is no specification.

Metrics	NAT	COH	ENG	GRO	AVG
X-EVAL	0.478	0.622	0.593	0.728	0.605
- Inference w/o Auxiliary Aspects	0.462	0.641	0.577	0.723	0.600
- w/ GT RAA (Upperbound)	**0.552**	**0.651**	**0.703**	**0.751**	**0.664**
- w/ Random RAA (Lowerbound)	0.468	0.601	0.561	0.628	0.564

Table 6: Analysis of error propagation in auxiliary aspects on Topical-Chat in terms of Spearman correlation. We highlight the best results in **bold** and the best results without using ground truths with underline. "RAA" denotes the evaluation Results on Auxiliary Aspects.

Effect of the Scale of Language Model Backbones We adopt the same training and inference pipelines for the backbones with different scales to show the effect of the models' size and justify the use of Flan-T5-large. Specifically, we additionally experiment with Flan-T5-small (80M), Flan-T5-base (250M), and Flan-T5-xl (3B) as the backbone models, and term our X-EVAL respectively. The results are shown in Figure 3. From Figure 3, the evaluators' performance consistently increases as the model size increases in general. However, when we upgrade the backbone model from Flan-T5-large to Flan-T5-xl, the performance improvement becomes less significant. Given the trade-off between efficiency and performance, we select Flan-T5-large as the default backbone model of X-EVAL in our experiments. We include a more detailed performance analysis of the effect of language model backbones in Appendix C.

Error Propagation from Auxiliary Aspects during Inference During inference, X-EVAL may predict inaccurate evaluations for auxiliary aspects. To investigate their impact, we tailor several baselines: (1) directly applying the model after two-stage tuning to evaluate without auxiliary aspects; (2) using the ground truth ("GT") evaluation results instead of predicted results for auxiliary aspects (upperbound), and; (3) using random evaluation results for auxiliary aspects (lowerbound). From Table 6, removing auxiliary aspects makes the overall performance drop. The variant with GT results gains improvement in all aspects, which indicates the error in the evaluation of auxiliary aspects does impact the performance of target aspects, but not to a large degree. Using random results, on the other hand, deteriorates the performance significantly.

Effect of Hyperparameter k We examine the choice of k in selecting top-k auxiliary aspects during inference. Table 7 shows that inference with the top-1 auxiliary aspect generally achieves better correlation. We speculate that this may stem from the error propagation during inference on auxiliary aspects, where using more auxiliary aspects potentially introduces more inaccuracies, offsetting their potential performance benefits.

Qualitative Correlation Analysis on Instruction Tuning To further investigate the effect of instruction tuning, in Figure 4, we visualize the correlation of our X-EVAL and Flan-T5 (i.e., "X-EVAL w/o Training") based on `naturalness` on Topical-Chat and `consistency` on SummEval. The red lines are linear regression fits to show how well the predicted scores correlate to human judgments linearly. Before instruction tuning, the predicted scores are more uniformly distributed regardless of ground truth scores, which results in poor correlation. On the contrary, our X-EVAL can predict scores that not only achieve better correlation but also are more distinctive (either close to 1 or 0), showing the effectiveness of our instruction tuning.

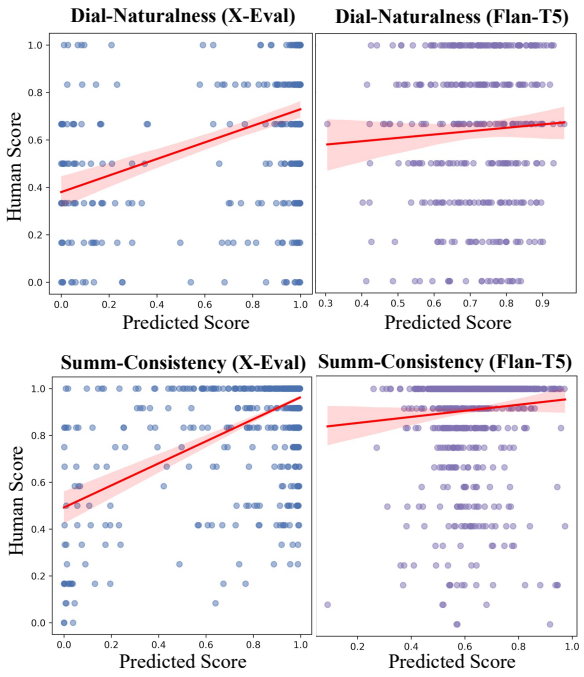

Figure 4: The scatter plots of correlation between human scores and predicted scores of X-EVAL and Flan-T5, respectively.

Selection	Topic.	FED	Summ.	D2T	AVG
Top-1	**0.605**	0.434	**0.480**	**0.303**	**0.456**
Top-3	0.602	0.414	0.466	0.278	0.440
Top-5	0.598	**0.435**	0.463	0.275	0.443

Table 7: Effect of different k in selecting auxiliary aspect in terms of averaged Spearman correlation. The best results are highlighted in **bold**.

Visualization of Auxiliary Aspect Selection In Figure 5, we also report the cosine similarity between the sentence embeddings of the aspect definitions used in turn-level dialogue evaluation as the qualitative analysis of our aspect selection strategy. In general, our strategy can select semantically related aspects for target-aspect evaluation.

Analysis of Auxiliary Aspect Selection Strategy We also experimented to compare the performance of selecting auxiliary aspects based on seen, unseen, or all aspects, as well as randomly selecting aspects regardless of the definitions. We set the number of auxiliary aspects to 1 in this experiment. From Table 8, selecting the auxiliary aspect based on all the aspects achieves the best overall performance. Also, we observe a substantial performance degradation when the auxiliary aspect is randomly selected, which shows the effectiveness of our aspect selection strategy.

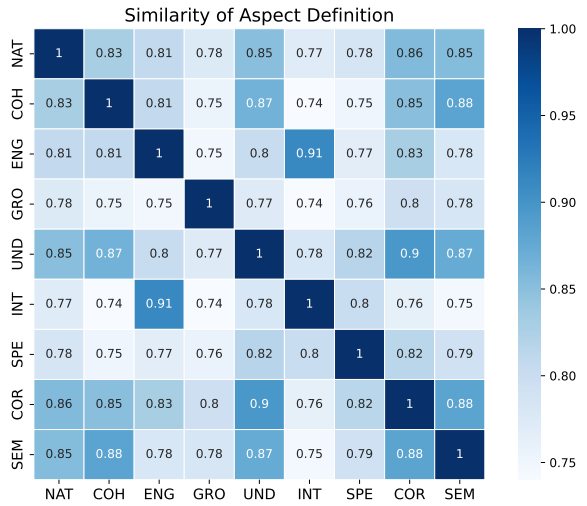

Figure 5: Cosine similarity scores of the sentence embeddings of aspect definition in turn-level dialogue evaluation. `Naturalness` (NAT), `coherence` (COH), `engagingness` (ENG), and `groundedness` (GRO) are seen aspects, while the rest are unseen aspects.

Selection	Topic-Chat	FED-Turn	AVG
All	0.605	0.398	**0.502**
Seen	0.602	**0.399**	0.489
Unseen	**0.608**	0.379	0.481
Random	0.592	0.381	0.475

Table 8: Comparison of different pools of candidate auxiliary aspects in terms of averaged Spearman correlation for turn-level dialogue evaluation. The best results are highlighted in **bold**.

8 Conclusion

In this work, we present X-EVAL, a novel two-stage instruction-tuning framework for text evaluation across both seen and unseen aspects. To facilitate training, we collect ASPECTINSTRUCT, the first instruction-tuning dataset for multi-aspect evaluation. Extensive experiments on meta-evaluation benchmarks demonstrate that with significantly fewer parameters, X-EVAL achieves a comparable if not higher correlation with human judgments compared to the state-of-the-art NLG evaluators.

9 Limitations

Limitation of Data Collection In this work, we mainly target evaluation tasks in English. Future work can explore evaluation tasks in a more diverse language setting and augment our ASPECTINSTRUCT dataset. In addition, our dataset focuses on a limited subset of NLG tasks including dialogue, summarization, and data2text. More NLG tasks

can be considered in the future.

Inference Efficiency Our algorithm may require multiple rounds of predictions to generate evaluation results from auxiliary aspects in the inference time. While this process imposes additional computational costs, given that the backbone we used is lightweight (with 780M parameters) and efficient, our approach is still significantly more efficient than the evaluators that are much larger, e.g., GPT-4. We leave exploring more efficient inference strategies for future work.

Error Propagation During inference, the evaluation results of auxiliary aspects may contain some errors. The errors may affect the final evaluation of the target aspect. We leave developing more robust inference algorithms to address the error propagation problem for future works.

Acknowledgments

This research is partially supported by a research award from the Amazon-Virginia Tech Initiative and award No. 2330940 from the Secure and Trustworthy Cyberspace program of the National Science Foundation (NSF). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Intell. Res.*, 77:103–166.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations.

R. Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David R. Traum, Maxine Eskénazi, Ahmad Beirami, Eunjoon Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar, and Rajen Subba. 2020. Overview of the ninth dialog system technology challenge: DSTC9. *CoRR*, abs/2011.06486.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9230–9240. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *CoRR*, abs/2310.00752.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Minqian Liu, Shiyu Chang, and Lifu Huang. 2022. Incremental prompting: Episodic memory prompt for lifelong event detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2157–2165, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Minqian Liu and Lifu Huang. 2023. Teamwork is not always good: An empirical study of classifier drift in class-incremental information extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2241–2257, Toronto, Canada. 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. *CoRR*, abs/2303.16634.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, Pascale Fung, and Zhiguang Wang. 2021. Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. pages 681–707.

Shuhaib Mehri and Vered Shwartz. 2023. Automatic evaluation of generative models with instruction tuning. *CoRR*, abs/2310.20072.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of SOCRATIC QUESTIONING: zero-shot multimodal reasoning with recursive thinking and self-questioning. *CoRR*, abs/2305.14999.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5008–5020. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020b. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023a. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. *CoRR*, abs/2305.14282.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023b. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

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*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A More Details on ASPECTINSTRUCT

A.1 Annotation Protocol of Instructions

We depict the annotation process for the instructions in ASPECTINSTRUCT as follows. To curate the definition for each aspect, we first refer to the definition of the aspect in the original annotation guideline. When a definition is absent from the guideline, three human annotators (graduate students studying in computational linguistics or natural language processing areas) construct and revise the definition until they reach an agreement. The task descriptions and evaluation protocols are also written by three human annotators in similar annotation protocols.

A.2 Augmenting Instruction-tuning Tasks

We show the seen aspects, their corresponding source datasets where we collect the training data,

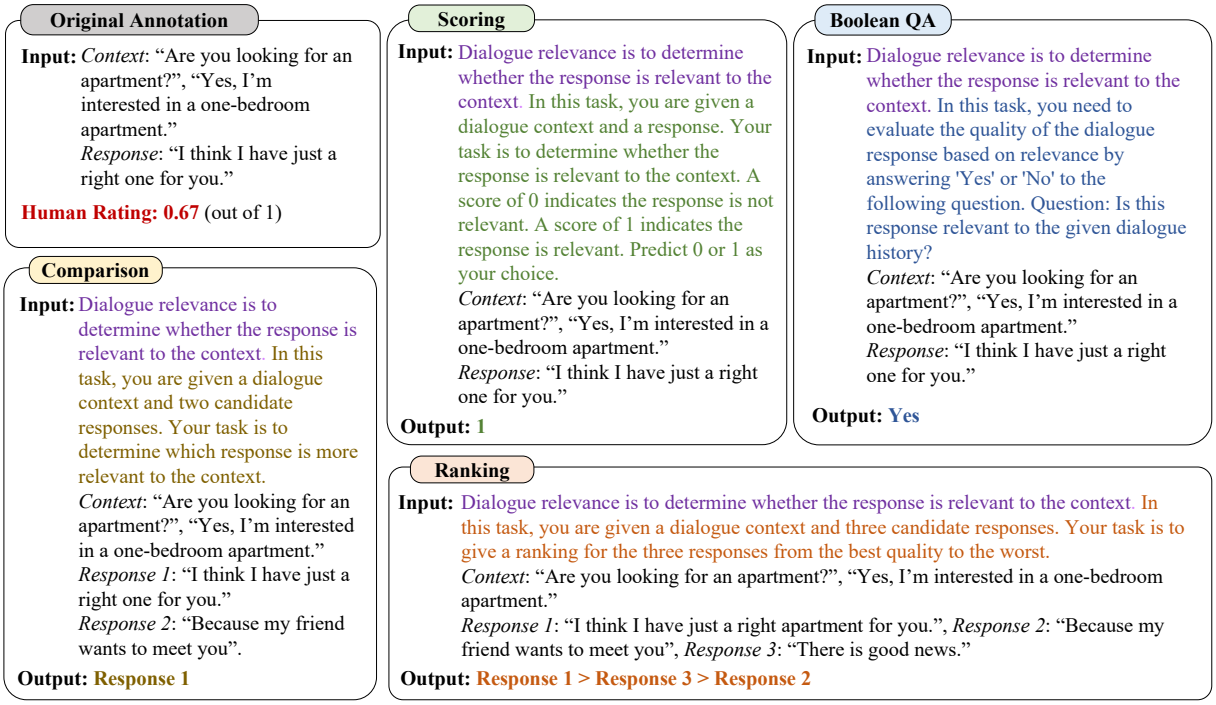

Figure 6: An illustrative example of augmented instruction-tuning tasks from the original annotation. The definition of the aspect is highlighted in purple. The annotated task instructions and the constructed output labels are highlighted in the corresponding colors for each task.

constructed tasks, and the number of training instances for each task in Table 11 and Table 12. For the way we count the number of aspects, we treat the aspects with the same name but in different NLG tasks as different aspects. For example, the *naturalness* aspect in dialogue evaluation and data2text evaluation are considered different under these two settings, although they have the same aspect name. More specifically, in our ASPECTINSTRUCT dataset, *understandability* is counted twice for dialogue-level and turn-level dialogue evaluation; *naturalness* is counted twice for turn-level dialogue evaluation and data-to-text evaluation; *informativeness* is counted twice for dialogue-level dialogue evaluation and data-to-text evaluation. We also include an example of how we augment instruction-tuning tasks from the original annotation in Figure 6.

A.3 Aspect Definition

We present the annotated definitions in ASPECTINSTRUCT in the following. We show the definitions of seen aspects on dialogue evaluation on Table 13, unseen aspects on dialogue evaluation on Table 14, and the aspects on summarization on Table 15.

Metrics	CNN	XSUM	AVG
ROUGE-L (Lin, 2004)	0.324	-0.011	0.156
BERTScore (Zhang* et al., 2020)	0.505	0.008	0.256
MOVERScore (Zhao et al., 2019)	0.347	0.044	0.195
BARTScore (Yuan et al., 2021)	0.680	0.159	0.420
UniEval (Zhong et al., 2022)	0.662	0.488	0.575
GPTScore (Fu et al., 2023)	0.649	0.238	0.443
G-Eval (GPT-3.5) (Liu et al., 2023)	0.516	0.406	0.461
G-Eval (GPT-4) (Liu et al., 2023)	**0.685**	**0.537**	**0.611**
X-EVAL (Ours)	0.656	0.500	0.578

Table 9: Spearman correlation on the summarization task based on the `consistency` aspect on QAGS. The best results are highlighted in **bold**. We also highlight the best results among lightweight (with <7B parameters) and open-source metrics with underline.

A.4 Source Datasets for Meta Evaluation

SummEval (Fabbri et al., 2021) is an evaluation benchmark for summarization which contains human ratings of 100 summaries along four evaluation dimensions: `fluency`, `coherence`, `consistency`, and `relevance`.

QAGS (Wang et al., 2020b) is a benchmark for identifying and evaluating hallucinations in the summarization task. It aims to measure the factual `inconsistencies` of generated summaries.

Topical-Chat (Gopalakrishnan et al., 2019) is a knowledge-grounded human-human conversation dataset. Following (Zhong et al., 2022), we utilize human ratings collected by (Mehri and Eskenazi, 2020b) for Topical-Chat as the benchmark for evaluating dialog response generation. The assessment consider five aspects: `naturalness`, `coherence`, `engagingness`, `groundedness`, and `understandability`.

FED (Mehri and Eskenazi, 2020a) is an evaluation benchmark for fine-grained dialog evaluation. It comprises human annotations evaluated across eighteen dialog aspects at both the turn-level and the dialog-level.

SFHOT & SFRES (Wen et al., 2015) are evaluation benchmarks for data-to-text task. They provide information about restaurants and hotels in San Francisco. The generated text is evaluated based on two aspects: `informativeness` and `naturalness`.

B More Details on X-EVAL

Pseudo-code of Inference Pipeline We provide the pseudo-code of our proposed inference pipeline for X-EVAL in Algorithm 1.

More Details on Verbalizer v and its Templates We design a template-based verbalizer to convert the evaluation results of auxiliary aspects into natural language evaluation that can be integrated into the instructions. More formally, the inputs of the verbalizer v contain aspect a and evaluation score s (in the range of 0-1). We first adopt a threshold δ (we set $\delta = 0.5$ throughout all experiments) to get a *binary* label that indicates the quality is *"positive"* (if $s > \delta$) or *"negative"* (if $s \leq \delta$). Given this label and the aspect a, we map the results into a template in natural language accordingly. The verbalized results will then be integrated into the instructions. We construct the templates for each aspect by deriving from aspect definition. We apply the annotation protocol that three human annotators revise the templates together until they reach a consensus. We show the verbalized templates in Table 16 for dialogue evaluation and Table 17 for summarization evaluation.

C More Details on Experiments

More Implementation Details We use the checkpoint released on HuggingFace for `Flan-T5-large`[2]. In the first training stage, we set the number of epochs to 2, the learning rate to 5e-05, and the maximum source length to 1024. The second training stage shares the same setup except the number of epochs set to 1. We set the maximum source length during inference to 2048 and pick the top-1 aspect during inference, i.e., $k = 1$. We use `sentence-T5-large`[3] to compute the embeddings for aspect definition for auxiliary aspect selection. All the experiments are conducted on NVIDIA A40 GPUs including both training and inference.

More Details on Baselines We include more details for the following baselines that are omitted in the main paper due to page limit: **(4) ROUGE-L** (Lin, 2004) counts the overlap (i.e., longest common subsequence) between the text to be evaluated and reference to indicate text quality; **(5) DynaEval** (Zhang et al., 2021) adopts a graph convolutional network to model dialogue's structure to facilitate evaluation; **(6) BERTScore** (Zhang* et al., 2020) is a similarity-based evaluator. It uses the contextualized representation from BERT (Devlin et al., 2019) to compute the similarity between the generated text and reference; **(7) MoverScore** (Zhao et al., 2019) goes beyond BERTScore by utilizing soft alignments and new aggregation methods on the layer-wise information; **(8) USR** (Mehri and Eskenazi, 2020b) is an unsupervised and reference-free evaluation metric to measure multiple desirable qualities of dialog; **(9) BARTScore** (Yuan et al., 2021) is a unified evaluator based on BART (Lewis et al., 2019), which uses the average likelihood of the model output as the metric. Note that for all single-aspect metrics, we compute the correlation between the single predicted evaluation and the human rating of each fine-grained aspect, respectively.

More Results on the Effect of the Scale of Language Model Backbones We further conducted an experiment on using another language model backbone. Specifically, we adopt LLaMA-7B-chat (Touvron et al., 2023a) as the backbone model and adopt LoRA parameter-efficient tuning (Hu et al., 2022) during the two-stage instruction tuning. We report the performance in Table 10.

[2] https://huggingface.co/google/flan-t5-large

[3] https://huggingface.co/sentence-transformers/sentence-t5-large

Model	# Parameters	TopicalChat	SummEval	FED-Dialog	FED-Turn	Data2Text	AVG
X-EVAL-large (Default Ver.)	780M	0.605	0.480	0.485	0.398	0.303	0.454
X-EVAL-LLaMA-LoRA	7B	0.519	0.448	0.427	0.351	0.337	0.416

Table 10: Effect of the scale of language model backbones. For each meta-evaluation benchmark, we report the average Spearman correlation on all the aspects.

Aspect	Datasets	Task	# Instances
Accuracy	TL;DR (Völske et al., 2017)	Scoring Boolean QA Comparison Ranking	5,000 5,000 898 599
Coherence	TL;DR (Völske et al., 2017), UniEval (Zhong et al., 2022)	Scoring Boolean QA Comparison Ranking	5,000 5,000 734 425
Coverage	TL;DR (Völske et al., 2017)	Scoring Boolean QA Comparison Ranking	5,000 4,354 1,028 964
Consistency	UniEval (Zhong et al., 2022)	Boolean QA	15,000
Fluency	UniEval (Zhong et al., 2022)	Boolean QA	15,000
Relevance	UniEval (Zhong et al., 2022)	Boolean QA	15,000

Table 11: The full list of apects, the corresponding datasets and tasks on summarization evaluation collected in the training split of ASPECTINSTRUCT.

Aspect	Datasets	Task	# Instances
Relevance	DailyDialog++ (Sai et al., 2020)	Scoring Boolean QA Comparison Comparison (w/ NOTA)	2,000 2,000 2,000 2,000
Coherence	HolisticDial (Pang et al., 2020); DSTC9 (Gunasekara et al., 2020); UniEval (Zhong et al., 2022)	Scoring Boolean QA	2,400 17,200
Consistency	DSTC9 (Gunasekara et al., 2020)	Scoring Boolean QA	2,200 2,200
Diversity	DSTC9 (Gunasekara et al., 2020)	Scoring Boolean QA	2,200 2,200
Engagingness	UniEval (Zhong et al., 2022)	Boolean QA	15,000
Groundedness	UniEval (Zhong et al., 2022)	Boolean QA	15,000
Naturalness	UniEval (Zhong et al., 2022)	Boolean QA	15,000
Fluency	HolisticDial (Pang et al., 2020)	Scoring	200

Table 12: The full list of apects, the corresponding datasets and tasks on dialogue evaluation collected in the training split of ASPECTINSTRUCT. "NOTA" indicates the comparison task consists of the case of "None Of The Above", where the quality of two candidates is tied.

Aspect	Definition
Naturalness	Naturalness in dialogue evaluation refers to the degree to which a response in a conversational context mirrors the characteristics, language use, and structure typical of a human conversational partner.
Coherence	Coherence refers to the logical and consistent interconnection of utterances and exchanges throughout a conversation. It represents the extent to which a dialogue system maintains relevance, consistency, and meaningful progression within the discourse, ensuring that the flow and structure of the conversation align with expected conversational norms and the ongoing context.
Engagingness	Engagingness in the context of dialogue evaluation refers to the degree to which a response fosters continued interaction, maintains or elevates interest, and stimulates a compelling exchange of ideas, emotions, or information between participants.
Groundedness	Dialogue groundedness measures how well does the response use the given fact. A response with weak groundedness means the response does not mention or refer to the fact at all. A response with good groundedness means the response uses the fact well.
Relevance	Relevance in dialogue evaluation refers to the measure of applicability, pertinence, or connection of a given response to the preceding conversational context and/or the explicitly posed question or statement.
Fluency	Fluency in dialogue evaluation refers to the degree of fluidity, coherence, and linguistic correctness in a generated response. It encompasses not only the grammatical and syntactic accuracy but also the seamless flow of ideas, the smooth transition between topics, and the naturalness of the language used, echoing human-like conversation patterns.

Table 13: The full list and definitions of *seen* aspects on dialogue evaluation collected in ASPECTINSTRUCT.

Aspect	Definition
Topic Depth	Topic depth refers to the ability of a dialogue system to engage in extensive, detailed, and multi-turn discussions on a particular subject.
Likeability	Likeability refers to the degree to which an interactive system presents a pleasant, engaging, and affable conversational style that resonates positively with the user.
Understandability	Understandability reflects the ability of a conversational system to correctly parse and interpret user inputs, reflect an appropriate comprehension of the context, and generate contextually relevant responses.
Flexibility	Flexibility measures the system's capacity to understand and react appropriately to a wide range of conversational scenarios, and not merely those for which it was explicitly programmed or trained. It implies the capacity to engage in a diverse array of topics, offer meaningful responses in unexpected situations, and adjust conversational strategies based on the evolving context or user input.
Informativeness	Informativeness refers to the quality and relevance of the information that a dialogue system provides in response to user inputs. It captures the system's ability to offer novel, detailed, accurate, and appropriate information that aligns with the user's requests or needs.
Inquisitiveness	Inquisitiveness pertains to the consistent exhibition of the capacity to ask meaningful, contextually appropriate, and well-timed questions within a conversation by a dialogue system. This behavior is exhibited in the pursuit of greater comprehension, clarifying ambiguities, furthering the dialogue, or driving deeper engagement with the conversation partner.
Interestingness	Interestingness refers to the degree to which a response stimulates engagement, thought, or emotional reaction in the average user, fostering a desire to continue the conversation or explore the topic further. It is a measure of the response's capacity to capture the user's attention and maintain their engagement over time.
Specificity	Specificity measures to what degree the response is unique, personalized, or pertinent to the specific details of the preceding user inputs or dialogue context, as opposed to being generic, universally applicable, or independent of the conversational specifics.
Correctness	Correctness in dialogue evaluation measures to the extent to which a generated response correctly reflects, comprehends, and addresses the salient elements, inferences, and implications in the preceding conversation context.
Semantic Appropriateness	Semantic appropriateness is the measure of the extent to which a response in a dialogue maintains logical, meaningful, and contextually fitting alignment with the preceding discourse elements, while adhering to the rules and principles of the language used in the conversation.

Table 14: The full list and definitions of *unseen* aspects on dialogue evaluation collected in ASPECTINSTRUCT.

Aspect	Definition
Accuracy	The accuracy aspect measures how the factual information in the summary accurately matches the post. A summary is accurate if it doesn't say things that aren't in the article, it doesn't mix up people, and generally is not misleading. If the summary says anything at all that is not mentioned in the post or contradicts something in the post, it should be considered as an inaccurate summary.
Coherence	The coherence aspect measures how coherent is the summary on its own. A summary is coherent if, when read by itself, it's easy to understand and free of English errors. A summary is not coherent if it's difficult to understand what the summary is trying to say. Generally, it's more important that the summary is understandable than it being free of grammar errors.
Coverage	The coverage aspect measures how well does the summary cover the important information in the post?" A summary has good coverage if it mentions the main information from the post that's important to understand the situation described in the post. A summary has poor coverage if someone reading only the summary would be missing several important pieces of information about the situation in the post. A summary with good coverage should also match the purpose of the original post (e.g. to ask for advice).
Consistency	The consistency aspect measures the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. You also need to penalize summaries that contained hallucinated facts.
Fluency	Fluency measures the quality of individual sentences. A fluent summary should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
Relevance	Relevance measures the selection of important content from the source. The summary should include only important information from the source document. You should penalize summaries which contain redundancies and excess information.

Table 15: The full list and definitions of aspects of summarization evaluation collected in ASPECTINSTRUCT.

Aspect	Verbalizer Template
Naturalness	NEG: The response is unnatural. POS: The response is natural.
Coherence	NEG: The response drastically changes topic or ignores the conversation history. POS: The response is on topic and strongly acknowledges the conversation history.
Engagingnes	NEG: The response is generic and dull. POS: The response is interesting or presents an interesting fact.
Groundedness	NEG: Given the interesting fact that the response is conditioned on, the response does not mention or refer to the fact at all. POS: Given the interesting fact that the response is conditioned on, the response uses the fact well.
Relevance	NEG: The response is not relevant to the conversation. POS: The response is relevant to the conversation.
Fluency	NEG: The response is not fluently written. POS: The response is fluently written.
Topic Depth	NEG: The system cannot discuss topics in depth. POS: The system is able to discuss topics in depth.
Likeability	NEG: The system cannot display a likeable personality. POS: The system is able to display a likeable personality.
Understandability	NEG: The response is difficult to understand. You do not know what the person is trying to say. POS: The response is understandable. You know what the person is trying to say.
Flexibility	NEG: The system is not flexible and adaptable to the user and their interests. POS: The system is flexible and adaptable to the user and their interests.
Informativeness	NEG: The system is not informative throughout the conversation. POS: The system is informative throughout the conversation.
Inquisitiveness	NEG: The system is not inquisitive throughout the conversation. POS: The system is inquisitive throughout the conversation.
Interestingness	NEG: To the average person, the response is not interesting. POS: To the average person, the response is interesting.
Specificity	NEG: The response is too generic and not specific to the conversation. POS: The response is specific to the conversation.
Correctness	NEG: There was a misunderstanding of the conversation. POS: The response is correct in the context of the conversation.
Semantic Appropriateness	NEG: The response is not semantically appropriate. POS: The response is semantically appropriate.

Table 16: The full list of verbalizer templates that are used to convert the evaluation results of auxiliary aspects for dialogue evaluation collected in ASPECTINSTRUCT. "POS" and "NEG" indicate *"positive"* and *"negative"*, respectively.

Aspect	Verbalizer Template
Accuracy	NEG: The factual information in the summary cannot accurately match the post. It says things that aren't in the article, it mixes up people, or generally is misleading. POS: The factual information in the summary accurately match the post. It doesn't say things that aren't in the article, it doesn't mix up people, and generally is not misleading.
Coherence	NEG: The summary is not coherent as it lacks a logical flow and has disjointed information, making it difficult to understand the main topic or argument. POS: The summary is well-structured and well-organized and it is built from sentence to sentence to a coherent body of information about a topic.
Coverage	NEG: The summary has poor coverage on the important information in the post, e.g., someone reading only the summary would be missing several important pieces of information about the situation in the post. POS: The summary has good coverage since it mentions the main information from the post that's important to understand the situation described in the post and also match the purpose of the original post.
Consistency	NEG: The summary is not factually consistent with the original post as it introduces factual inaccuracies or hallucinated facts that are not present in or supported by the original source document. POS: The summary has good factual alignment between the summary and the summarized source. It contains only statements that are entailed by the source document.
Fluency	NEG: The summary is not fluent as it contains formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read. POS: This is a fluent summary as it generally does not have formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
Relevance	NEG: This summary is not relevant to the source document as it contains redundancies or excess information. POS: The summary generally includes relevant content, capturing some key points from the source.

Table 17: The full list of verbalizer templates that are used to convert the evaluation results of auxiliary aspects for summarization evaluation collected in ASPECTINSTRUCT. "POS" and "NEG" indicate *"positive"* and *"negative"*, respectively.