Themis: A Reference-free NLG Evaluation Language Model with Flexibility and Interpretability

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, Xiaojun Wan
Wangxuan Institute of Computer Technology, Peking University
{huxinyu, gaomingqi, xjyin, wanxiaojun}@pku.edu.cn
efsotr_l@stu.pku.edu.cn

Abstract

The evaluation of natural language generation (NLG) tasks is a significant and longstanding research area. With the recent emergence of powerful large language models (LLMs), some studies have turned to LLM-based automatic evaluation methods, which demonstrate great potential to become a new evaluation paradigm following traditional string-based and model-based metrics. However, despite the improved performance of existing methods, they still possess some deficiencies, such as dependency on references and limited evaluation flexibility. Therefore, in this paper, we meticulously construct a large-scale NLG evaluation corpus **NLG-Eval** with annotations from both human and GPT-4 to alleviate the lack of relevant data in this field. Furthermore, we propose **Themis**, an LLM dedicated to NLG evaluation, which has been trained with our designed multi-perspective consistency verification and rating-oriented preference alignment methods. Themis can conduct flexible and interpretable evaluations without references, and it exhibits superior evaluation performance on various NLG tasks, simultaneously generalizing well to unseen tasks and surpassing other evaluation models, including GPT-4.

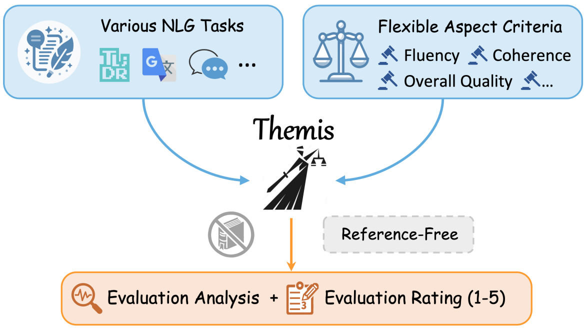

Figure 1: **Themis** is capable of evaluating various NLG tasks based on flexible evaluation aspects and criteria without references and providing corresponding analyses along with the evaluation ratings.

1 Introduction

Automated evaluation is crucial for natural language generation tasks, as it measures the performance of related models and consequently promotes the development of NLG research. In the early years, traditional string-based evaluation metrics, such as BLEU (Papineni et al., 2002), were commonly used. Despite their convenience, surface-level matching cannot reliably evaluate texts as they are easily affected by perturbations (He et al., 2023), and previous work (Sulem et al., 2018a) has indicated their low correlation with human evaluations. With the development of pre-trained language models and related corpora, pre-trained model-based evaluation metrics such as BARTScore (Yuan et al., 2021) and COMET (Rei et al., 2020) have then been proposed. But their performance remains unsatisfactory compared with human evaluation, and they cannot conduct evaluations on customized aspects, like coherence.

Recently, large language models (LLMs) such as ChatGPT and LLaMA (Touvron et al., 2023) have emerged and demonstrated unprecedented performance in instruction following and open-domain generation, which shows great potential for LLM-based automated evaluation to become a new paradigm (Wang et al., 2023a; Gao et al., 2024). Existing related studies can be divided into two major categories: directly prompting LLMs with different optimization methods (Liu et al., 2023b; Chiang and Lee, 2023b; Liu et al., 2024a; Kocmi and Federmann, 2023b) and fine-tuning LLMs with annotated data (Li et al., 2023a; Jiang et al., 2023; Xu et al., 2023; Kim et al., 2023). However, these approaches still have some limitations: the first category often relies on proprietary LLMs such as GPT-4, which are high-cost and possibly irreproducible, while the second category tends to have weaknesses of dependence on references, lacking flexibility, or poor interpretability.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15924–15951
November 12-16, 2024 ©2024 Association for Computational Linguistics

In this paper, we propose **Themis**, an 8B-parameter LLM specifically designed and trained for NLG evaluation with more comprehensive capabilities. Our Themis can evaluate various NLG tasks, including uncommon ones like question-answering evaluation (**Versatility**), in a reference-free manner (**Independence**). Moreover, it allows for specific and customized evaluation aspects and criteria, including overall quality and more fine-grained aspects (**Flexibility**), and its evaluation contains corresponding analysis and explanation together with the rating (**Interpretability**), as shown in Figure 1. We believe that an ideal evaluator should be convenient to use and possess these characteristics. The comparison between related methods and Themis is shown in Table 1.

To obtain high-quality training data for NLG evaluation, we conducted a comprehensive survey of relevant studies and collected corresponding resources, finally selecting 58 evaluation datasets with human annotations across 9 common NLG tasks. Given the importance of evaluation criteria, we meticulously proofread each dataset and manually supplemented the missing descriptions. In addition, GPT-4 is treated as an additional annotator to supplement and validate the collected data while providing evaluation analyses. Ultimately, we constructed a large-scale evaluation corpus, NLG-Eval, which contains about 0.5 million samples with meta information. It aims to alleviate the issue of scattered and scarce data in the area of NLG evaluation and facilitate relevant research.

Furthermore, we proposed a multi-perspective consistency verification method to select relatively more reliable data from the constructed NLG-Eval corpus. Empirical methods are also employed to ensure sufficient diversity and balanced distribution of the data as much as possible, resulting in approximately 67K training samples. Moreover, we designed specific preference alignment, guiding the construction and utilization of preference data through evaluation ratings, to improve the evaluation capabilities of the fine-tuned model. Experimental results show that our Themis achieves better overall evaluation performance over previous evaluation models on common NLG tasks, including summarization, story generation, and so on. We also conducted more in-depth analyses, including generalization tests on unseen tasks like the instruction-following evaluation as well as aspect-targeted perturbation tests to verify the reliability.

Overall, our main contributions are as follows:

Method	Vers.	Inde.	Flex.	Inte.	Open.
UniEval	✗	✗	✓	✗	✓
G-Eval	✓	✓	✓	✓	✗
X-Eval	✓	✗	✓	✗	✗
Prometheus	✓	✗	✓	✓	✓
Auto-J	✓	✓	✗	✓	✓
InstructScore	✓	✗	✗	✓	✓
TIGERScore	✓	✓	✗	✓	✓
Themis (Ours)	✓	✓	✓	✓	✓

Table 1: Comparisons of our Themis with currently common evaluation models, including UniEval (Zhong et al., 2022), G-Eval (Liu et al., 2023b), X-Eval (Liu et al., 2023a), Prometheus (Kim et al., 2023), Auto-J (Li et al., 2023a), InstructScore (Xu et al., 2023) and TIGERScore (Jiang et al., 2023). Vers., Inde., Flex., Inte., and Open. represent versatility, independence, flexibility, interpretability, and open-source, respectively.

- We construct a large-scale NLG evaluation corpus, including about 0.5 million samples and 58 datasets across 9 NLG tasks, with detailed meta information, aspect criteria, and evaluations from both humans and GPT-4.
- We propose Themis, an LLM dedicated to NLG evaluation, which has been trained through our specific consistency and alignment methods and possesses versatility, independence, flexibility, and interpretability.
- Extensive experiments demonstrate the superior evaluation performance of Themis on common NLG tasks, as well as good generalization and robustness. Our model and relevant resource have been released in https://github.com/PKU-ONELab/Themis to facilitate related research.

2 NLG-Eval Corpus

Despite abundant data on NLG tasks, the corresponding high-quality evaluation data remains scarce and scattered due to the high cost of professional human annotations. Previous related methods either used a small amount of human annotations to train regression models, such as UniEval (Zhong et al., 2022) and X-Eval (Liu et al., 2023a), resulting in limited task coverage and a lack of evaluation analyses, or they entirely relied on LLMs to generate synthetic data, which raised reliability concerns (Hu et al., 2024). To address this challenge, we formally defined the NLG evaluation task and clarified the involved elements. Sub-

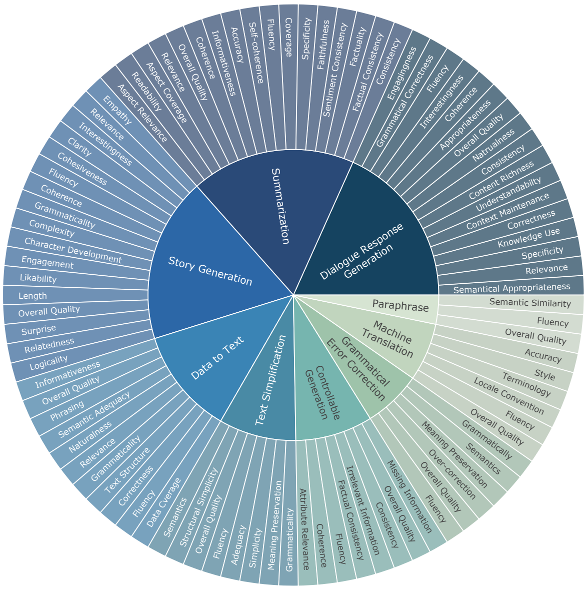

Figure 2: The NLG evaluation tasks and corresponding evaluation aspects in our NLG-Eval corpus.

sequently, we surveyed a large number of existing related studies and compiled 58 evaluation datasets with human annotations across 9 NLG tasks, totaling 0.5 million samples. They have undergone our meticulous proofreading, and the missing but critical content has been supplemented, such as evaluation criteria. Additionally, we utilize the competent GPT-4 for supplementary evaluations, including analyses and ratings. The corpus has also been equipped with meta information, aiming to promote the development of related research.

2.1 Definitions

NLG tasks typically require language models to generate the corresponding output t based on the given input i and task requirements. The input in specific tasks may consist of multiple components; for example, in controlled generation, it is composed of the generation requirements and the constraint labels. NLG evaluation involves evaluating t in conjunction with i through the evaluation model or method E, using evaluation criteria c and references r if available. When they are absent, the evaluation defaults to general and reference-free evaluation, respectively. Evaluation results come in various forms, primarily focusing on rating s and additional information a such as analyses. The formalization of the evaluation process is as follows, with square brackets indicating optional elements:

$$s, [a] = E(t, i, [r], [c])$$

In the corpus construction, we collect and annotate the aforementioned elements for each sample.

2.2 Data Collection

To improve the coverage and diversity of the corpus, we conducted a comprehensive survey of related work on NLG tasks, including summarization, dialogue response generation, data-to-text, paraphrase, and so on. We collected datasets with human evaluation ratings and performed meticulous manual checking and filtering, ultimately obtaining 58 high-quality datasets across 9 different NLG tasks, which contain about 0.5 million samples. However, we discovered that many datasets, despite including evaluation aspects, lacked specific criteria or definitions, which could lead to confusion and ambiguity in evaluations. Such concerns have also been pointed out by Zhou et al. (2022); Hu et al. (2024). So we analyzed all these datasets, taking samples, human labels, and information from related papers into account, and manually supplementing the criteria. The different tasks included in our NLG-Eval corpus and their respective involved evaluation aspects are gathered and shown in Figure 2, with semantically similar aspects being merged. And more detailed statistics and information are presented in Appendix A.

2.3 GPT-4 Annotations

Although human evaluation is often considered the most reliable, many datasets have sparse or inconsistent human annotations, leading to questions about their accuracy. Therefore, we utilized the superior language capabilities of GPT-4 to evaluate each sample in the corpus, as it has been shown to perform well on evaluation (Zheng et al., 2023). GPT-4 can serve as an additional annotator to cross-verify with human evaluations, thereby improving reliability. On the other hand, it can provide evaluation analyses beyond just ratings and increase interpretability, which is lacking in human evaluation. Following conclusions and suggestions from Chiang and Lee (2023b); Bsharat et al. (2023), we carefully designed instructions, ensuring that a concise and accurate analysis is provided before the rating. Each sample is evaluated with a temperature setting of 1, and 10 diverse evaluation results are obtained through multiple samplings, with more details included in Appendix B. Additionally, our annotation and subsequent training did not use references because many datasets lack references or only have automatically constructed references that are unsatisfactory (Pu et al., 2023). And the references are actually difficult to obtain in practice.

3 Methodology

Existing studies have highlighted that data quality is more crucial than quantity in LLM fine-tuning (Li et al., 2023c; Zhou et al., 2023). Therefore, we adopted empirical methods to improve the diversity and balance of data distribution and designed a multi-perspective consistency verification approach to obtain high-quality samples from the constructed corpus. Then they would be used for supervised fine-tuning based on open-source LLMs. To further enhance the evaluation capability and alignment of the model, we proposed a preference data construction and training method guided by evaluation ratings to improve the fine-tuned model.

3.1 Diversity and Balance

Similar to many studies on LLM training, we initially employed empirical approaches to sample data from the original corpus, ensuring adequate diversity and balanced distribution, which was generally considered beneficial for model training and generalization. To be specific, we treated the NLG task, evaluation aspect, and rating triplets as identifiers of data categories to cover as comprehensive a distribution as possible. And we sampled 100 pieces of data from each category while also considering the semantic diversity of the input texts involved. Given the insufficiency of data in certain categories and the requirement to introduce diversity in evaluation analyses and aspect criteria, we conducted multiple samplings of analyses differently for each category to further balance the data distribution, simultaneously rephrasing the evaluation aspect criteria. In the end, we obtained a dataset of approximately 67K for supervised fine-tuning. The detailed sampling method for both samples and evaluations will be introduced in the next subsection.

3.2 Multi-perspective Consistency Verification

During the sampling process, most categories actually include abundant samples, and each sample also involves multiple evaluations annotated by GPT-4. To take full advantage of such sufficient data resources and enhance the reliability of the training data, we verified the consistency of each sample and their evaluations from three different perspectives, ultimately obtaining the potentially high-quality data through screening.

Self-Consistency When dealing with complex tasks like mathematical problems and logical reasoning, the chain of thought (CoT) (Wei et al., 2022) method is always employed on LLMs to improve their performance. It involves generating intermediate steps before providing the final result, which generally leads to higher-quality responses. Furthermore, by performing multiple diverse samplings and selecting the most frequently occurring final result from the responses, the output can be more stable and accurate (Wang et al., 2023c). This process is referred to as self-consistency, which can also be regarded as a metric and calculated as the proportion of the most frequently occurring result to the total number of samplings. A higher self-consistency indicates greater certainty of the model itself and, thus, higher reliability of the responses to the corresponding samples.

In our evaluation annotation, GPT-4 was required to generate an analysis before assigning a rating, similar to a CoT process, which allows using self-consistency for data filtering. The formal definitions of the most frequently occurring rating $\hat{r}$ (called consistent rating) and self-consistency are as follows, where n denotes the number of evaluations for a sample, and r_i represents the rating of the i-th evaluation:

$$\hat{r} = \arg\max_{r} \sum_{i=1}^{n} \mathbb{1}\left(r_i = r\right)$$

$$\text{self-consistency} = \frac{\sum_{i=1}^{n} \mathbb{1}\left(r_i = \hat{r}\right)}{n}$$

And we prioritized samples with high self-consistency and only retained their evaluations that include the consistent rating $\hat{r}$ to collect both potentially reliable samples and evaluations for the following process of verification.

Cross-Validation The evaluation ratings serve as the most critical supervision signal in evaluation, making their accuracy paramount. Unlike previous related work that relied entirely on human labels or LLM-generated labels, we comprehensively considered the evaluations from both sources to conduct cross-validation. As for each sample, the corresponding consistent rating $\hat{r}$ is regarded as its rating from GPT-4 and scaled to the same range as the human rating. We prioritized samples where the two evaluation ratings were close, which reflected the high consistency between evaluations from humans and GPT-4 and also indicated the potential strong reliability of the samples.

Evaluation Inspection Hu et al. (2024) indicated that although current LLMs like GPT-4 possess exciting evaluation capabilities, they may encounter issues of confusing aspect criteria, affecting the accuracy of their analyses and ratings. To address this, we propose two specific criteria for inspecting the evaluations themselves: consistency between the evaluation analysis and rating, and consistency between the evaluation analysis and aspect. We designed additional instructions to prompt GPT-4 to re-evaluate the candidate evaluations, whose details could be found in Appendix B, and prioritized those being assessed as of good quality from both perspectives.

3.3 Preference Alignment

With the rapid rise of InstructGPT (Ouyang et al., 2022), reinforcement learning from human feedback (RLHF), as one of the key technologies, has garnered wide attention and been applied in many subsequent LLMs with great success. So we have specifically modified DPO (Rafailov et al., 2023), a commonly-used implementation of RLHF, based on our NLG evaluation scenarios, using evaluation ratings to guide the construction and training of preference data. The preference alignment was further conducted after the supervised fine-tuning of the model.

3.3.1 Preference Construction

Although the DPO method does not require training a separate reward model, it still necessitates preference pairs to convey preference information. Benefiting from multiple evaluations of each sample, we can directly regard the evaluation that matches the consistent rating $\hat{r}$ introduced in Section 3.2 as the chosen response and that does not as the rejected response to construct preference pairs. The quality gap between the chosen and rejected evaluations can be reflected by the difference in their evaluation ratings, as the closer the rating to $\hat{r}$, the more reliable the evaluation can be deemed. And our preference data is constructed based on the previously obtained training data without involving additional samples.

3.3.2 Rating-guided DPO

During vanilla DPO, we reparameterize the reward function r using the policy as follows:

$$r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$$

where π_θ is the policy model, π_{ref} is the reference model, typically a model undergone supervised fine-tuning, β is the parameter controlling the degree of deviation between them, and $Z(x)$ is the partition function. Moreover, the Bradley-Terry (BT) (Bradley and Terry, 1952) model is employed to model preferences, yielding:

$$p(y_1 \succ y_2|x) = \sigma(r(x, y_1) - r(x, y_2))$$

For a pair of evaluation responses (y_1, y_2), if we prefer y_1 to y_2, then the larger the evaluation rating difference between y_1 and y_2, the greater the preference difference $r(x, y_1) - r(x, y_2)$ should be. Therefore, to treat each preference pair more equally, we modify the BT model by subtracting a value proportional to the rating difference between y_1 and y_2 from $r(x, y_1) - r(x, y_2)$, thereby compensating for the prior preference difference caused by the rating difference, and then obtain:

$$\begin{aligned} p^*(y_1 \succ y_2|x) = \sigma(&r(x, y_1) - r(x, y_2) \\ &- \alpha|R(y_1) - R(y_2)|) \end{aligned}$$

where $R(y)$ denotes the evaluation rating included in response y. Finally, the maximum likelihood of the new preference model serves as the optimization objective for our rating-guided DPO:

$$\begin{aligned} &\mathcal{L}^*_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) \\ &= -\mathbb{E}_{(x,y_c,y_r)\sim\mathcal{D}}\big[\log \sigma\big(\beta \log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} \\ &\quad -\beta \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} - \alpha|R(y_c) - R(y_r)|\big)\big] \end{aligned}$$

where (x, y_c, y_r) contains the input content, the chosen evaluation, and the rejected evaluation, respectively.

4 Experiments

4.1 Benchmarks

We conducted extensive experiments to evaluate our Themis across six commonly used evaluation datasets for different NLG tasks, including SummEval (Fabbri et al., 2021) for summarization, Topical-Chat (Gopalakrishnan et al., 2019) for dialogue response generation, SFRES&SFHOT (Wen et al., 2015) for data-to-text, QAGS (Wang et al., 2020) for factuality, MANS (Guan et al., 2021) for story generation, and WMT23 (Freitag et al., 2023) for machine translation (zh-en). These datasets were collected into our NLG-Eval corpus but were excluded during the construction of our training data to ensure fairness.

Method	SummEval		Topical-Chat		SFHOT&RES		QAGS		MANS		WMT23		Average
	ρ	τ	r	ρ	ρ	τ	ρ	τ	ρ	τ	r	ρ	ρ
Traditional Metrics													
BLEU†	0.075	0.057	0.356	0.388	0.024	0.018	-	-	0.032	0.009	-0.130	0.021	-
ROUGE†	0.152	0.120	0.393	0.412	0.101	0.076	-	-	-0.002	0.156	0.081	0.151	-
BARTScore	0.329	0.261	0.067	0.086	0.208	0.156	0.425	0.347	0.350	0.260	0.091	0.118	0.253
BERTScore†	0.231	0.182	0.388	0.394	0.139	0.105	-	-	0.285	0.163	0.123	0.219	-
BLEURT†	0.152	0.118	0.384	0.388	0.244	0.184	-	-	0.138	0.221	0.163	0.263	-
CometKiwi	0.228	0.180	0.353	0.340	0.251	0.186	0.094	0.074	0.251	0.176	0.413	0.343	0.251
UniEval†	0.474	0.377	0.533	0.577	0.282	0.211	-	-	-	-	-	-	-
Prompting LLM													
G-Eval (GPT-3.5)	0.409	0.323	0.574	0.585	-	-	0.461	0.337	-	-	-	-	-
G-Eval (GPT-4)	<u>0.523</u>	0.423	0.575	0.588	-	-	0.611	0.532	-	-	-	-	-
GPT-3.5	0.416	0.340	0.592	0.578	0.306	0.239	0.431	0.356	0.328	<u>0.295</u>	0.388	0.347	0.401
GPT-4	0.511	<u>0.423</u>	**0.770**	**0.746**	<u>0.320</u>	<u>0.260</u>	<u>0.637</u>	<u>0.532</u>	<u>0.473</u>	0.260	**0.496**	**0.437**	<u>0.521</u>
Fine-tuned LLM													
X-Eval†	0.480	0.362	0.539	0.605	0.303	-	0.578	-	-	-	-	-	-
Prometheus-13B†	0.163	0.142	0.435	0.434	0.173	0.142	-	-	0.007	0.146	0.144	0.129	-
Auto-J-13B	0.198	0.172	0.427	0.425	0.141	0.120	0.226	0.209	0.380	0.284	0.128	0.104	0.246
TIGERScore-13B	0.384	0.334	0.334	0.346	0.200	0.175	0.504	0.446	0.231	0.207	0.277	0.248	0.319
InstructScore-7B†	0.258	0.226	0.269	0.241	0.247	0.210	-	-	0.298	0.168	0.213	0.219	-
Themis-8B (ours)	**0.553**	**0.499**	<u>0.733</u>	<u>0.725</u>	**0.333**	**0.284**	**0.684**	**0.613**	**0.551**	**0.501**	<u>0.431</u>	<u>0.405</u>	**0.542**

Table 2: The results of our Themis compared with different evaluation metrics and models on six different NLG tasks. † represents reference-based methods, while bold and underline indicate the first and second best results.

4.2 Experimental Settings

We chose Llama-3-8B (Meta, 2024) for supervised fine-tuning and preference alignment in our main experiments, and more training details are described in Appendix C. When assessing the evaluation capability of different models, we calculated the correlation between evaluation ratings from the model and humans using Pearson (r), Kendall (τ), and Spearman (ρ) correlation coefficients. The specific calculations and the selection of correlation coefficients vary across different datasets, and we follow the common setups used in previous work. Furthermore, while some studies (Chiang and Lee, 2023b) employed a temperature setting of T = 1 to perform multiple sampling and aggregate the evaluation results, which may enhance performance, we believe it is costly and instable in practice and actually leads to the loss of evaluation analyses. Therefore, we tested our Themis with T = 0 and single sampling, while using their original settings for other methods. Given that most metrics and models do not support specific evaluation aspects, we calculate the results of different aspects using their overall scores, as in previous work.

4.3 Baselines

We experimented with existing representative and commonly-used evaluation metrics and models from three categories. Traditional metrics include BLEU, ROUGE (Lin, 2004), BARTScore, BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), CometKiwi (Rei et al., 2022), and UniEval. Methods of direct prompting LLMs include G-Eval and the evaluations with GPT-3.5 and GPT-4, implemented following Chiang and Lee (2023b). Fine-tuned evaluation LLMs include X-Eval, Prometheus, Auto-J, TIGERScore, and InstructScore. Many of them are reference-based and therefore cannot be tested on the reference-free QAGS dataset.

4.4 Main Results

The main experimental results of our Themis compared with other baselines are shown in Table 2. We present the average correlation coefficients for each task, with the complete results demonstrated in Appendix D. Our Themis achieves the best overall evaluation performance and surpasses all larger LLMs. Furthermore, Themis also behaves

Method	Avg ρ	Avg τ	Avg r
BARTScore	0.253	0.197	0.262
CometKiwi	0.251	0.195	0.272
GPT-4	0.521	0.417	0.564
TIGERScore-13B	0.319	0.280	0.333
Auto-J-13B	0.246	0.211	0.262
Themis-8B (ours)	**0.542**	**0.486**	**0.569**
Data Sampling			
100% raw data	0.493	0.445	0.515
50% raw data	0.478	0.432	0.505
67K raw data	0.476	0.428	0.507
Preference Alignment			
vanilla DPO	0.528	0.474	0.556
without DPO	0.508	0.456	0.534
Foundational Model			
Mistral-7B	0.489	0.440	0.515
Llama-2-7B	0.478	0.430	0.508
Llama-2-13B	0.500	0.449	0.533

Table 3: The results with different ablation settings.

better on each of six NLG tasks than other baselines, except GPT-4. It only outperforms Themis in dialogue response and translation evaluations, likely because GPT-4 has considerably stronger dialogue and multilingual capabilities than Llama-3-8B, which are almost impossible to improve during our limited training. And our Themis avoids the limitations of high costs and instability in proprietary LLMs and can be conveniently used offline. Moreover, although models such as BLEURT, X-Eval, and Prometheus conduct evaluation based on references, they still lag behind our reference-free Themis, showing its superiority.

4.5 Ablation Study

We conducted ablation studies on our two main methods: consistency-based data sampling for supervised fine-tuning and rating-guided preference alignment. Additionally, considering the superior performance of Llama-3 among LLMs of the same size, we also experimented with other LLMs as the foundational model to verify the effectiveness of our approach. The average results for the six task datasets with different settings are shown in Table 3.

Data Sampling We experimented with approximately 67K raw data that has the same scale as the constructed data for supervised fine-tuning, and further trained the model using more randomly sampled data, namely half and all of the corpus data. The results show that although more data bring some improvements, the training cost increases significantly, and there is still a substantial performance drop compared to our current model.

Preference Alignment The model trained with our rating-guided DPO method exhibits progressive improvements over the model trained with vanilla DPO, and then the model that has only been fine-tuned. It indicates that preference alignment has certain effects on the training of NLG evaluation tasks, and our specific method is better suited for evaluation scenarios.

Foundational Model Although the evaluation models trained based on other LLMs do not perform as well as our current model, the models fine-tuned on Llama-2-7B still significantly outperformed TIGERScore and Auto-J, which use Llama-2-13B as the foundation model. These results demonstrate the effectiveness and necessity of the data sampling and alignment training methods we introduced.

5 Reliability Analyses

To analyze and investigate the reliability of our Themis, we consider two aspects: generalizability, which means the model can effectively evaluate unseen NLG tasks during training, and robustness, which means the model can accurately evaluate texts containing noises and perturbations.

5.1 Unseen Tasks

We tested our model on two common but unseen evaluation tasks during training: long-form question answering and instruction-following evaluation (Li et al., 2023b). The former task involves five evaluation aspects: clarity (CLA), completeness (COM), correctness (COR), politeness (POL), and relevance (REL), while the latter involves example usage (EU), factual accuracy (FA), logical coherence (LC), overall usefulness (OU), and simple language (SL). As shown in Table 4, Themis outperforms both proprietary and open-source LLMs on the overall two tasks, demonstrating good generalization. And its performance is relatively well-balanced without particular weaknesses, unlike Auto-J, which performs well on instruction-following but poorly on long-form question-answering. Furthermore, some evaluation aspects are relatively open, which traditional reference-based metrics that do not support specific

Method	Instruction Following					Long form Question Answering					Average ρ
	CLA	COM	COR	POL	REL	EU	FA	LC	OU	SL	
BARTScore	-0.053	-0.040	0.063	-0.038	0.025	0.098	0.174	0.360	0.296	0.464	0.135
CometKiwi	0.345	0.464	0.349	0.223	0.473	0.155	0.128	0.210	0.138	0.134	0.262
GPT-3.5	0.257	0.488	0.433	0.239	0.401	0.612	0.318	0.443	0.212	0.562	0.396
GPT-4	0.257	**0.653**	0.573	0.338	0.455	0.767	0.431	0.237	-0.004	**0.711**	0.442
Auto-J	**0.422**	0.562	0.571	**0.386**	**0.548**	0.095	0.255	0.109	0.191	0.033	0.317
TIGERScore	0.262	0.247	0.285	0.198	0.239	-0.037	0.075	0.009	0.002	-0.217	0.106
Themis (ours)	0.414	0.381	**0.685**	0.349	0.500	**0.835**	**0.538**	**0.700**	**0.365**	0.684	**0.545**

Table 4: The spearman correlation between evaluations from humans and different models on instruction-following and long-form question-answering tasks.

Method	Dial	News	Para	Table	Average↓
GPT-4	1.158	0.896	0.929	0.923	0.976
Prometheus	1.095	1.433	1.051	1.485	1.266
Themis (ours)	0.841	0.748	0.698	0.860	0.787

Table 5: The results of perturbation tests, showing the changes in evaluation ratings from different models before and after perturbations.

evaluation aspects cannot handle, highlighting the superiority of our Themis.

5.2 Perturbation Tests

Hu et al. (2024) has pointed out that both proprietary and open-source LLMs have certain issues with confusing evaluation aspects during NLG evaluation. Therefore, we conducted perturbation tests based on their methods to investigate the robustness of different evaluation models that support customized aspects. Specifically, we applied aspect-targeted perturbations on references in four tasks—news summarization, dialogue summarization, paraphrase generation, and data-to-text—focusing on fluency, coherence, informativeness, and consistency. The perturbations were designed to impact only the quality of the target aspect while leaving other aspects unaffected. We tested the models by comparing ratings on the perturbed texts with those on the original references, expecting little change and decreases in unaffected aspects. As shown in Table 5, Themis is closest to what is expected, with the smallest average decreases across the four tasks. And slight decreases may be caused by the fact that the designed pertubations do not entirely avoid affecting other aspects. On the other hand, the evaluation of perturbed texts may indeed be challenging, necessitating further exploration to enhance the model's capabilities.

6 Related Works

6.1 Prompting LLMs for NLG Evaluation

Previous research has evaluated text quality in various tasks by directly prompting proprietary LLMs (e.g. GPT-4) for scoring (Liu et al., 2023b; Chiang and Lee, 2023a; Kocmi and Federmann, 2023b), comparison (Liusie et al., 2024; Wang et al., 2023d), ranking (Ji et al., 2023; Liu et al., 2023d), and error analysis (Kocmi and Federmann, 2023a; Lu et al., 2023), surpassing traditional evaluation metrics in performance and providing significant flexibility. Leiter et al. (2023) have experimented with various prompt formats and incontextual example settings. Moreover, to better assess specific aspects, Liu et al. (2024a) have used LLMs to generate or improve the definitions of evaluation criteria for use in prompting, and Gong and Mao (2023a) have had LLMs first evaluate other relevant aspects. Additionally, Wu et al. (2023); Bai et al. (2023) have enhanced evaluative capabilities through interactions and role-playing among LLMs. Although these studies are compelling, they rely on proprietary models, which are costly and pose reproducibility issues.

6.2 Fine-tuned LLM Evaluators

In response to the issues with prompting LLMs, subsequent studies have shifted to fine-tuning relatively smaller open-source large models to create specialized evaluators (Wang et al., 2023e; Xu et al., 2023; Kim et al., 2023; Wang et al., 2023b; Jiang et al., 2023; Li et al., 2023a; Ke et al., 2023; Zhu et al., 2023; Liu et al., 2023a). Their approaches are generally similar: they use existing human evaluation data or synthetic data created by proprietary

LLMs, such as GPT-4, as training data to fine-tune open-source LLMs like Llama (Touvron et al., 2023). Their training modes vary in details, such as whether references are required and whether they support customized evaluation criteria. However, without specific measures, such fine-tuned evaluators seemed to contain issues similar to those in proprietary LLMs (Hu et al., 2024).

7 Conclusions

In this paper, we propose a large-scale and comprehensive NLG evaluation corpus and an LLM dedicated to NLG evaluation, Themis. The NLG-Eval corpus encompasses 9 common NLG tasks across 58 datasets, comprising about 0.5 million samples with evaluations from both human and GPT-4. Based on the constructed corpus, we propose a specialized multi-perspective consistency verification method to select high-quality supervised fine-tuning data, together with preference alignment guided by evaluation ratings, to train our Themis. It can be applied to various NLG tasks with flexible and interpretable evaluation in a reference-free manner. Extensive experiments demonstrate that our Themis performs well on different NLG tasks and can be generalized well to unseen tasks. Our model and resources have been released, and we hope they will promote further research in the field of NLG evaluation.

Limitations

During the construction of our NLG-Eval corpus, we relied on annotations from GPT-4, resulting in considerable costs. However, it is worthwhile because our Themis, built upon this corpus and proposed training methods, has achieved great NLG evaluation capabilities. We have released the corresponding corpus and model to reduce the cost of future research and promote the development of related studies. Additionally, due to limited computational resources, we did not use larger and more powerful LLMs than Llama-3-8B as the foundational model for training, such as Llama-3-70B. Although larger models might offer stronger evaluation capabilities, their practical usage becomes inconvenient. Balancing model performance and usability requires further exploration in the future.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011), National Science Foundation of China (No. 62161160339), Ant Group Research Fund and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments and all the people who have contributed to this work. Xiaojun Wan is the corresponding author.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4668–4679. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Comput. Linguistics*, 47(4):861–889.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, GPT-3.5/4. *CoRR*, abs/2312.16171.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5794–5836. International Committee on Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15607–15631. Association for Computational Linguistics.

David Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8928–8942. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George F. Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 578–628. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based NLG evaluation: Current status and challenges. *CoRR*, abs/2402.01383.

Mingqi Gao and Xiaojun Wan. 2022. Dialsummeval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5693–5709. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 179–188. Association for Computational Linguistics.

Peiyuan Gong and Jiaxin Mao. 2023a. Coascore: Chain-of-aspects prompting for NLG evaluation. *CoRR*, abs/2312.10355.

Peiyuan Gong and Jiaxin Mao. 2023b. Coascore: Chain-of-aspects prompting for NLG evaluation. *CoRR*, abs/2312.10355.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 1891–1895. ISCA.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6394–6407. Association for Computational Linguistics.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 379–391. Association for Computational Linguistics.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James R. Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12067–12097. Association for Computational Linguistics.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. PARABANK: monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6521–6528. AAAI Press.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing NLG quality criteria? *CoRR*, abs/2402.12055.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020a. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 446–469. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020b. 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.

Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences. *CoRR*, abs/2303.07610.

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.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation. *CoRR*, abs/2311.18702.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. Ctrleval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2306–2319. Association for Computational Linguistics.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *CoRR*, abs/2310.08491.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 768–775. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Naomi Kong-Vega, Mingxin Shen, Mo Wang, and Luis Fernando D'Haro. 2018. Subjective annotation and evaluation of three different chatbots WOCHAT: shared task report. In *9th International Workshop on Spoken Dialogue System Technology, IWSDS 2018, Singapore, April 18-20, 2018*, volume 579 of *Lecture Notes in Electrical Engineering*, pages 371–378. Springer.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1777–1788. Association for Computational Linguistics.

Seolhwa Lee, Heuiseok Lim, and João Sedoc. 2020. An evaluation protocol for generative conversational systems. *CoRR*, abs/2010.12741.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2023, Bali, Indonesia, November 1, 2023*, pages 117–138. Association for Computational Linguistics.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *CoRR*, abs/2310.05470.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023b. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *CoRR*, abs/2310.19740.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2023c. One shot learning as instruction data prospector for large language models. *CoRR*, abs/2312.10302.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2023a. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. *CoRR*, abs/2311.08788.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yixin Liu, Alexander R. Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023c. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *CoRR*, abs/2311.09184.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023d. On learning to summarize with large language models as references. *CoRR*, abs/2305.14239.

Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. 2023e. Evaluate what you can't evaluate: Unassessable generated responses quality. *arXiv preprint arXiv:2305.14658*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024a. Calibrating llm-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2638–2656. ELRA and ICCL.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024b. Calibrating llm-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 2638–2656. ELRA and ICCL.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024c. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. *CoRR*, abs/2402.15754.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 139–151. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16383–16408. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235. Association for Computational Linguistics.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. `https://ai.meta.com/blog/meta-llama-3/`.

Courtney Napoles, Maria Nadejde, and Joel R. Tetreault. 2019. Enabling robust grammatical error correction in new domains: Datasets, metrics, and analyses. *Trans. Assoc. Comput. Linguistics*, 7:551–566.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 72–78. Association for Computational Linguistics.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *INLG 2016 - Proceedings of the Ninth International Natural Language Generation Conference, September 5-8, 2016, Edinburgh, UK*, pages 265–273. The Association for Computer Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural*

Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4812–4829. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 3619–3629. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *CoRR*, abs/2309.09558.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Max Schwarzer, Teerapaun Tanprasert, and David Kauchak. 2021. Improving human text simplification with sentence fusion. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 106–114, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Scialom, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021. Rethinking automatic evaluation in sentence simplification. *CoRR*, abs/2104.07560.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle H. Ungar, and Chris Callison-Burch. 2019. Chateval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 60–65. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3178–3190. Association for Computational Linguistics.

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization. *CoRR*, abs/2310.18122.

Tejpalsingh Siledar, Swaroop Nath, Sankara Sri Raghava Ravindra Muddu, Rupasai Rangaraju, Swaprava Nath, Pushpak Bhattacharyya, Suman Banerjee, Amey Patil, Sudhanshu Shekhar Singh, Muthusamy Chelliah, and Nikesh Garera. 2024. One prompt to rule them all: Llms for opinion summary evaluation. *CoRR*, abs/2402.11683.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 8776–8788. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4,* *2018*, pages 738–744. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 685–696. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018c. Simple and effective text simplification using semantic and neural methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 162–173. Association for Computational Linguistics.

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. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. 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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good NLG evaluator? A preliminary study. *CoRR*, abs/2303.04048.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023b. Shepherd: A critic for language model generation. *CoRR*, abs/2308.04592.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. 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.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023d. Automated evaluation of personalized text generation using large language models. *CoRR*, abs/2310.11593.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023e. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. *CoRR*, abs/2306.05087.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Peihao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721. The Association for Computational Linguistics.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I*, volume 14302 of *Lecture Notes in Computer Science*, pages 695–707. Springer.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference, INLG 2023, Prague, Czechia, September 11 - 15, 2023*, pages 323–351. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3511–3535. Association for Computational Linguistics.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6516–6522. International Committee on Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael E. Banchs, and Alexander Rudnicky. 2021. Automatic evaluation and moderation of open-domain dialogue systems. *CoRR*, abs/2111.02110.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 26–33. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2023–2038. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 314–324. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *CoRR*, abs/2310.17631.

A Details for NLG-Eval Corpus

Our constructed NLG-Eval corpus includes evaluation datasets on nine NLG tasks, including Controllable Generation, Data to Text, Dialogue Response Generation, Grammatical Error Correction, Machine Translation, Paraphrase Generation, Story Generation, Summarization, and Text Simplification. An example from the hsplit dataset for the text simplification task is presented in Figure 3 with simplified illustrations, whose situation is similar for other datasets in NLG-Eval corpus. The statistics are shown in Table 6 and more detailed meta information is described in Table 7 to Table 17. Through such data and information, we hope to facilitate related research and advance the field of NLG evaluation.

B Settings and Prompts for GPT-4 Annotations

We follow the suggestions of Bsharat et al. (2023) to design the instructions and prompts for GPT-4 to annotate NLG evaluation tasks, as shown in Table 18. Those for re-evaluating evaluations are shown in Table 19 and Table 20, where the task descriptions are appropriately modified to prevent confusion for GPT-4.

C Training Details

The sizes of training data used in our supervised fine-tuning and preference alignment are 67,180 and 10,000, respectively. In the construction of the latter, we prioritize selecting preference pairs with larger differences in evaluation ratings, as intuitively, their preference relations are more reliable. In addition, following the construction method of the training data, we built an additional validation set based on the NLG Eval corpus, comprising a total of 2,479 samples, to find the most suitable training hyperparameters. During supervised fine-tuning, the learning rate is 1e-5, the batch size is 128, the training epoch is 3, and the optimizer is AdamW with the zero weight decay. And in preference alignment, the learning rate is 3e-6, the batch size is 128, the training epoch is 1, α is 1.0, and the optimizer is AdamW. Our experiments utilize 8 A100 GPUs during training and inference.

Task	#Datasets	#Aspects	#Samples
Controllable Generation	4	8	11299
Data to Text	6	11	36548
Dialogue Response Generation	17	17	91111
Grammatical Error Correction	3	6	41058
Machine Translation	2	6	347504
Paraphrase Generation	2	3	18299
Story Generation	6	17	12636
Summarization	10	17	61977
Text Simplification	8	8	27026

Table 6: The statistics of NLG-Eval corpus.

Dataset	Size	Aspect
Chiang-LLM-Evaluation (Chiang and Lee, 2023a)	1600	Cohesiveness: How well do the sentences in the story fragment fit together? Grammaticality: How grammatically correct is the text of the story fragment? Likability: How enjoyable do you find the story fragment? Relevance: How relevant is the story fragment to the story prompt?
CoEval (Li et al., 2023b)	1400	Character Development: The characters in the generated story should be well-developed. Clarity: The generated story should be clear and easy to understand, with no confusing or ambiguous elements. Coherence: The generated story should have a logical flow and provide closure. Engagement: The generated story should be engaging from beginning to end. Grammaticality: The generated story should be grammatically correct. Length: The generated story should have an appropriate length according to the story requirement. Relevance: The generated story should be relevant to the story requirement (beginning or topic) and the daily events it aims to capture.
Hanna (Chhun et al., 2022)	6336	Coherence: How much does the generated story make sense? Complexity: How elaborate is the generated story, involving complex concepts, realistic characters, an intricate plot, an underlying history or circumstances, or precise descriptions? Empathy: How well do you understand the character's emotions in the generated story (regardless of whether you agree with them)? Engagement: How much do you engage with the generated story? Relevance: How well does the generated story match the story prompt? Surprise: How surprising is the end of the generated story, with enough clues for a reasonable prediction?
Mans_roc (Guan et al., 2021)	1000	Overall Quality: The overall quality of the generated story, considering whether it has global errors like chaotic scenes (difficult to understand as a whole) and local errors, including repetitive plots (repeating similar texts), unrelated events (to the story beginning or within its own context), and conflicting logic (against common sense or with wrong causal or temporal relationship).
Mans_wp (Guan et al., 2021)	1000	Overall Quality: The overall quality of the generated story, considering whether it has global errors like chaotic scenes (difficult to understand as a whole) and local errors, including repetitive plots (repeating similar texts), unrelated events (to the story prompt or within its own context), and conflicting logic (against common sense or with wrong causal or temporal relationship).
nextchapter (Xie et al., 2023)	1300	Coherence: How well do the sentences in the generated story fit together? Fluency: How grammatically correct is the text of the generated story? Interestingness: How enjoyable do you find the generated story? Logicality: How much does the generated story obey commonsense? Relatedness: How relevant is the generated story to the story prompt?

Table 7: Datasets for Story Generation task.

Dataset	Size	Aspect
CTRLEval (Ke et al., 2022)	3960	Attribute Relevance: Measure whether the generated text satisfies the attribute label. Coherence: Measure whether the sentences in the generated text are semantically relevant to compose a coherent body, which reflects the quality of the generated text itself. Consistency: Evaluate whether the generated text is consistent to the content prefix.
FUDGE (Yang and Klein, 2021)	2088	Fluency: Is the generated text fluent, i.e., well-written and grammatical?
PPLM (Dathathri et al., 2020)	3251	Fluency: Whether the generated text has no grammatical errors, formatting problems, or obviously ungrammatical issues (e.g., fragments, missing components) that make the text difficult to read?
InstruSum (Liu et al., 2023c)	2000	Factual Consistency: Is the summary consistent with the facts presented in the article, without contradicting or misrepresenting any information? Irrelevant Information: Does the summary include any information that is not relevant to the summary requirement? Missing Information: Does the summary omit any crucial information from the article concerning the summary requirement? Overall Quality: Assess the overall quality of the summary in relation to the summary requirement.

Table 8: Datasets for Controllable Generation task.

Dataset	Size	Aspect
E2E NLG (Dusek et al., 2020)	6300	Naturalness: Could the utterance have been produced by a native speaker? Overall Quality: How is the overall quality of the utterance in terms of its grammatical correctness, fluency, adequacy and other important factors?
INLG16 (Novikova et al., 2016)	3726	Informativeness: Is this utterance informative? (i.e. do you think it provides enough useful information from the data?) Naturalness: Is this utterance natural? (e.g. could it have been produced by a native speaker?) Phrasing: Is this utterance well phrased? (i.e. do you like how it is expressed?)
RankMe (Novikova et al., 2018)	900	Informativeness (= adequacy): Does the utterance provide all the useful information from the meaning representation? Naturalness (= fluency): Could the utterance have been produced by a native speaker? Overall Quality: How do you judge the overall quality of the utterance in terms of its grammatical correctness, fluency, adequacy and other important factors?
SFRES_SFHOT (Wen et al., 2015)	6168	Informativeness: Does the utterance provide all the useful information from the meaning representation? Naturalness: Could the utterance have been produced by a native speaker? Overall Quality: How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?
webnlg_2017 (Gardent et al., 2017)	5214	Fluency: Does the text sound fluent and natural? Grammaticality: Is the text grammatical (no spelling or grammatical errors)? Semantic Adequacy: Does the text correctly represent the meaning in the data?
webnlg_2020 (Castro Ferreira et al., 2020)	14240	Correctness: When describing predicates which are found in the data, does the text mention the correct objects and adequately introduce the subject for this specific predicate? Data Coverage: Does the text include descriptions of all predicates presented in the data? Fluency: Is it possible to say that the text progresses naturally, forms a coherent whole and it is easy to understand the text? Relevance: Does the text describe only such predicates (with related subjects and objects), which are found in the data? Text Structure: Is the text grammatical, well-structured, written in acceptable English language?

Table 9: Datasets for Data to Text task.

Dataset	Size	Aspect
convai2-grade (Huang et al., 2020b)	600	Coherence: The response should be coherent with the dialogue context, maintaining a good logical flow.
dailydialog-grade (Huang et al., 2020b)	300	Coherence: The response should be coherent with the dialogue context, maintaining a good logical flow.
dailydialog-gupta (Gupta et al., 2019)	500	Appropriateness: Whether the response is appropriate given the dialogue context in grammar, topic, and logic?
dailydialog-zhao (Zhao et al., 2020)	3600	Appropriateness: The response should be appropriate given the preceding dialogue. Content Richness: The response should be informative, with long sentences including multiple entities and conceptual or emotional words. Grammatical Correctness: The response should be free of grammatical and semantic errors. Relevance: The response should be on-topic with the immediate dialogue history.
DialogADV (Liu et al., 2023e)	16416	Coherence: The logical and semantic coherence between the response and dialogue history (previous context). Consistency: The logical and factual consistency between the response and dialogue history (previous context), facts also include external commonsense knowledge. Fluency: The fluency and grammatical correctness of the response. Relevance: The degree to response is connected or relevant to a particular topic, question, or situation of the dialogue history (previous context).
dstc10-persona_clean (Zhang et al., 2021)	19316	Appropriateness: The response should be appropriate given the preceding dialogue. Content Richness: The response should be informative, with long sentences including multiple entities and conceptual or emotional words. Grammatical Correctness: The response should be free of grammatical and semantic errors. Relevance: The response should be on-topic with the immediate dialogue history.
dstc10-topical_clean (Zhang et al., 2021)	18000	Appropriateness: The response should be appropriate given the preceding dialogue. Content Richness: The response should be informative, with long sentences including multiple entities and conceptual or emotional words. Grammatical Correctness: The response should be free of grammatical and semantic errors. Relevance: The response should be on-topic with the immediate dialogue history.
empathetic-grade (Huang et al., 2020b)	300	Coherence: The response should be coherent with the dialogue context, maintaining a good logical flow.
esl (Lee et al., 2020)	1242	Appropriateness: The response should be appropriate given the dialogue context, in grammar, topic, and logic.
fed-turn (Mehri and Eskénazi, 2020)	3375	Correctness: Is the response correct or was there a misunderstanding of the conversation? Engagingness: Is the response engaging to user and fulfill the particular conversational goals implied by the user? Fluency: Is the response fluently written and free of grammatical and semantic errors? Interestingness: To the average person, is the response interesting? Overall Quality: What is the overall impression of the response, and quality of and satisfaction with the conversation? Relevance: Is the response relevant to and on-topic with the conversation history? Semantical Appropriateness: Is the response semantically appropriate given the conversation history? Specificity: Does the response produce unique and non-generic information that is specific to the conversation history? Understandability: Is the response understandable?
holistic dialogue (Pang et al., 2020)	400	Coherence: Measures the meaningfulness of the response within the context of prior query. Fluency: Measures the quality of phrasing of the response relative to a human native speaker.
humod (Merdivan et al., 2020)	19000	Fluency: The response should be written naturally and free of grammatical and semantic errors. Relevance: The Response should be on-topic with the immediate dialogue history.
jsalt (Kong-Vega et al., 2018)	741	Appropriateness: Measure how well the response is semantically and pragmatically valid given the previous recent dialogue history.

Table 10: Datasets for Dialogue Response Generation task. (Part 1)

Dataset	Size	Aspect
ncm (Sedoc et al., 2019)	2461	Appropriateness: Whether the response is appropriate given the dialogue history, in grammar, topic, and logic?
persona-usr (Sedoc et al., 2019)	1800	Context Maintenance: Does the response serve as a valid continuation of the dialogue context (conversation history)? Interestingness: Is the response dull or interesting? Knowledge Use: Given the fact that the response is conditioned on, how well does the response use that fact? Naturalness: Does the response seem to be something that a person would naturally say? Overall Quality: What is the overall impression of this utterance? Please consider whether the response is understandable and natural, and how well it maintains context and uses knowledge. Understandability: Is the response understandable given the previous dialogue context? (Not if its on topic, but for example if it uses pronouns they should make sense)
persona-zhao (Zhao et al., 2020)	900	Appropriateness: The response should be appropriate given the preceding dialogue.
topical-usr (Sedoc et al., 2019)	2160	Context Maintenance: Does the response serve as a valid continuation of the dialogue context (conversation history)? Interestingness: Is the response dull or interesting? Knowledge Use: Given the fact that the response is conditioned on, how well does the response use that fact? Naturalness: Does the response seem to be something that a person would naturally say? Overall Quality: What is the overall impression of this utterance? Please consider whether the response is understandable and natural, and how well it maintains context and uses knowledge. Understandability: Is the response understandable given the previous dialogue context? (Not if its on topic, but for example if it uses pronouns they should make sense)

Table 11: Datasets for Dialogue Response Generation task. (Part 2)

Dataset	Size	Aspect
GMEG (Napoles et al., 2019)	27195	Overall Quality: Whether the corrected text is perfect, namely grammatical and not garbled?
protagolabs (Sottana et al., 2023)	1200	Grammaticality: The quality of the correction and the extent to which errors are left in the corrected text, regardless of whether they are present in the original text or they are newly introduced errors in the supposed corrected version. Over-correction: Assess whether the correction avoids unnecessary syntax changes or being unnecessarily verbose, since there can be multiple ways to correct a text. The best correction should be done with the minimum number of edits. Semantics: Whether the meaning of the original text is preserved following the grammatical error correction? NOTE: You should penalize corrections which change the meaning unnecessarily.
TMU-GFM (Yoshimura et al., 2020)	12663	Fluency: How natural does the corrected text sound for native speakers? Grammaticality: The grammatical correctness of the corrected text and how comprehensible it is. Meaning Preservation: The extent to which the meaning of the original text is preserved in the corrected text.

Table 12: Datasets for Grammatical Error Correction task.

Dataset	Size	Aspect
parabank (Hu et al., 2019)	11140	Fluency: Whether the paraphrase is meaningful and grammatical? Semantic Similarity: Whether the paraphrase maintains similar semantics to the original text?
twitter para (Shen et al., 2022)	7159	Overall Quality: The paraphrase should not only maintain similar semantics to the original text, but also possess lexical or syntactic differences from the original text, with fluent and coherent content.

Table 13: Datasets for Paraphrase Generation task.

Dataset	Size	Aspect
WMT_zhen (Freitag et al., 2021)	346504	Accuracy: The translation should accurately represent the source text, not including information not present in the source text or missing content from the source text. Fluency: The translation should not have incorrect punctuation and spelling, problems with grammar, internal inconsistency, or garbled characters due to incorrect encoding. Locale Convention: The translation should not have the wrong formats for addresses, currency, dates, names, telephone numbers, and time expressions. Overall Quality: The overall quality of the translation, including accuracy (e.g., mistranslation, omission, addition), fluency (e.g., grammar, spelling, punctuation, inconsistency), locale convention (e.g., format for names, currency, address), terminology (e.g., inappropriate or inconsistent usage), and style (e.g., stylistic problems). Style: Does the translation have no stylistic problems? Terminology: Whether the terminology of the translation is standard, appropriate for the context, and used consistently?
HumanMT (Kreutzer et al., 2018)	1000	Overall Quality: The overall quality of the translation, including accuracy (e.g., mistranslation, omission, addition), fluency (e.g., grammar, spelling, punctuation, inconsistency), locale convention (e.g., format for names, currency, address), terminology (e.g., inappropriate or inconsistent usage), and style (e.g., stylistic problems).

Table 14: Datasets for Machine Translation task.

Dataset	Size	Aspect
DialSummEval (Gao and Wan, 2022)	5600	Coherence: Measure the quality of all sentences of the summary collectively, to fit together and sound naturally. Consider the quality of the summary as a whole. Consistency: Measure whether the facts in the summary are consistent with the facts in the dialogue. Consider whether the summary does reproduce all facts accurately and does not make up untrue information. Fluency: Measure the quality of individual sentences of the summary, whether they are well-written and grammatically correct. Consider the quality of individual sentences. Relevance: Measure how well the summary captures the key points of the dialogue. Consider whether all and only the important aspects are contained in the summary.
frank (Pagnoni et al., 2021)	2246	Factuality: Measure whether the facts in the summary are correct according to the article.
Newsroom (Grusky et al., 2018)	1680	Coherence: Do phrases and sentences of the summary fit together and make sense collectively? Fluency: Are the individual sentences of the summary well-written and grammatical? Informativeness: How well does the summary capture the key points of the article? Relevance: Are the details provided by the summary consistent with details in the article?
OpenAI (Stiennon et al., 2020)	34197	Accuracy: Does the factual information in the summary accurately match the post? A summary is accurate if it doesn't say things that aren't in the post, it doesn't mix up people, and generally is not misleading. Coherence: 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: 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). Overall Quality: How good is the summary overall at representing the post? This can encompass coherence (how coherent is the summary on its own), accuracy (does the factual information in the summary accurately match the post), and coverage (how well does the summary cover the important information in the post) of the summary, as well as other important aspects.
QAGS (Wang et al., 2020)	474	Factual Consistency: Is the summary factually consistent with the article? Non-grammatical sentences should be considered not consistent and copies of article sentences should be considered consistent.

Table 15: Datasets for Summarization task. (Part 1)

Dataset	Size	Aspect
OpinSummEval (Shen and Wan, 2023)	5600	Aspect Relevance: Measure whether the mainly discussed aspects in the reviews are covered exactly by the summary. It focuses on whether the summary correctly reflects the mainly discussed aspects in the reviews. Readability: Measure whether the summary is fluent and informative. It focuses on whether the summary is well-written and valuable. Self-coherence: Measure whether the summary is consistent within itself in terms of sentiments and aspects. It focuses on whether the summary is coherent and does not reflect conflicting opinions. Sentiment Consistency: Measure whether the summary is consistent with the reviews in terms of sentiments for each aspect. It focuses on whether the summary aspect-wisely captures the main sentiment in the reviews.
PolyTope (Huang et al., 2020a)	1268	Overall Quality: The overall quality of the summary, including accuracy and fluency. Accuracy-related issues refer to the extent to which the summary does not match the article, including unnecessary snippets, missing key points, content unfaithful to the article, content not present in the article and factually incorrect, and statements that contradict the article in attitudes (e.g. from positive to negative). Fluency-related issues refer to the linguistic quality of the summary, which is independent of the relationship between the article and the summary, including unnecessary repetition, problems in the word form, and problems in word order.
protagolabs (Sottana et al., 2023)	1600	Coherence: Measure the quality of all sentences of the summary collectively, to fit together and sound naturally. Consider the quality of the summary as a whole. Consistency: Measure whether the facts in the summary are consistent with the facts in the article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information. Fluency: Measure the quality of individual sentences of the summary, whether they are well-written and grammatically correct. Consider the quality of individual sentences. Relevance: Measure how well the summary captures the key points of the article. Consider whether all and only the important aspects are contained in the summary.
SummEval (Fabbri et al., 2021)	6400	Coherence: Measure the quality of all sentences of the summary collectively, to fit together and sound naturally. Consider the quality of the summary as a whole. Consistency: Measure whether the facts in the summary are consistent with the facts in the article. Consider whether the summary does reproduce all facts accurately and does not make up untrue information. Fluency: Measure the quality of individual sentences of the summary, whether they are well-written and grammatically correct. Consider the quality of individual sentences. Relevance: Measure how well the summary captures the key points of the article. Consider whether all and only the important aspects are contained in the summary.
SummEval-OP (Siledar et al., 2024)	2912	Aspect Coverage: The summary should cover all the aspects that are majorly being discussed in the reviews. The summary should be penalized if it misses out on an aspect that was majorly discussed in the reviews and awarded if it covers all. Coherence: Measure the collective quality of all sentences. The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentences to a coherent body of information. Faithfulness: Every piece of information mentioned in the summary should be verifiable/supported/inferred from the reviews only. The summary should be penalized if any piece of information is not verifiable/supported/inferred from the reviews or if the summary overgeneralizes something. Fluency: Measure the quality of the summary in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure and should contain no errors. The summary should be easy to read, follow, comprehend and should contain no errors. Relevance: The summary should not contain opinions that are either not consensus or important. The summary should include only important opinions from the reviews. The summary should be penalized if it contains redundancies and unimportant information. Sentiment Consistency: All the aspects discussed in the summary should accurately reflect the consensus sentiment of the corresponding aspects from the reviews. The summary should be penalized if it does not cover accurately the sentiment regarding any aspect within the summary. Specificity: The summary should avoid containing generic opinions. All the opinions within the summary should contain detailed and specific information about the consensus opinions. The summary should be penalized for missing out details and should be awarded if they are specific.

Table 16: Datasets for Summarization task. (Part 2)

Dataset	Size	Aspect
ASSET (Alva-Manchego et al., 2020)	300	Adequacy (or Meaning Preservation): The simplified sentence should adequately express the meaning of the original sentence, perhaps omitting the least important information. Fluency (or Grammaticality): The simplified sentence should be fluent, and there are no grammatical errors. Simplicity: The simplified sentence should be easier to understand than the original sentence.
Fusion (Schwarzer et al., 2021)	10490	Adequacy: To which degree does the simplified sentence retain the meaning of the original sentence? Simplicity: To which degree is the simplified sentence simpler than the original sentence?
HSplit (Sulem et al., 2018c)	7560	Grammaticality: Is the simplified text fluent and grammatical? Meaning Preservation: Does the simplified text preserve the meaning of the original text? Simplicity: Is the simplified text simpler than the original text? Structural Simplicity: Is the simplified text simpler than the original text, ignoring the complexity of the words?
Human-likert (Scialom et al., 2021)	336	Fluency: How fluent is the simplified text? Meaning Preservation: How well does the simplified text express the original meaning? Simplicity: To what extent is the simplified text easier to read and understand?
LENS (Maddela et al., 2023)	3840	Overall Quality: The simplified sentence should be fully simplified, entirely fluent, and preserve the core meaning of the original sentence. Adequacy: The simplified sentence should adequately express the meaning of the original sentence, perhaps omitting the least important information. Fluency: The simplified sentence should be fluent, and there are no grammatical errors. Simplicity: The simplified sentence should be easier to understand than the original sentence.
metaeval (Alva-Manchego et al., 2021)	1800	Adequacy (or Meaning Preservation): Judge by looking at both the original and simplified texts, and judge whether or not the changes made preserve the original meaning. Fluency (or Grammaticality): Judge by looking solely at the simplified text. Mainly consider the grammatical and/or spelling errors, but also 'how well' (or natural) the text reads. Do not take capitalization into consideration. Simplicity: Judge by looking at both the original and simplified texts, and judge whether or not the changes made the simplified text easier to understand than the original text.
protagolabs (Sottana et al., 2023)	1200	Fluency (or Grammaticality): Whether the simplified text remains grammatical and understandable? Semantics (or Adequacy): Whether the meaning is preserved further to the simplification? Simplicity: Whether the simplified text is simpler than the original text?
SAMSA (Sulem et al., 2018b)	1500	Grammaticality: Is the simplified text grammatical? Meaning Preservation: Does the simplified text add information or remove important information, compared to the original text? Structural Simplicity: Is the simplified text simpler than the original text, ignoring the complexity of the words?

Table 17: Datasets for Text Simplification task.

Prompts and Instructions
###Instruction### Please act as an impartial and helpful evaluator for natural language generation (NLG), and the audience is an expert in the field. Your task is to evaluate the quality of {task} strictly based on the given evaluation criterion. Begin the evaluation by providing your analysis concisely and accurately, and then on the next line, start with "Rating:" followed by your rating on a Likert scale from 1 to 5 (higher means better). You MUST keep to the strict boundaries of the evaluation criterion and focus solely on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion and example content, carefully. ###Evaluation Criterion### {aspect} ###Example### {source_des}: {source} {target_des}: {target} ###Your Evaluation###

Table 18: Prompts and instructions used for evaluating NLG tasks.

Prompts and Instructions
###Instruction### You are a professional and helpful evaluator for natural language generation (NLG). You will be given an example of Review Generation task, which includes an aspect description and a review based on it. Your task is to evaluate the quality of the review strictly based on the given evaluation criterion. Your evaluation MUST begin with the accurate analysis, followed by 'Rating:' and then include the corresponding evaluation rating. Your rating MUST be an integer ranging from 1 to 5, following a five-point Likert scale. (higher means better) You MUST keep to the strict boundaries of the given evaluation criterion and focus ONLY on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion and example content, carefully. ###Evaluation Criterion### Aspect Alignment: Is the review strictly aligned with and solely based on the corresponding aspect description, without mentioning any other points out of scope? ###Example### Aspect Description: {aspect} Review: {analysis} ###Your Evaluation###

Table 19: Prompts and instructions used for re-evaluating evaluations based on the alignment between evaluation analyses and aspects.

Prompts and Instructions
###Instruction### You are a professional and helpful evaluator for natural language generation (NLG). You will be given an example of Review Generation task, which includes a review of the {target_des} and a polarity. Your task is to evaluate the quality of the review strictly based on the given evaluation criterion. Your evaluation MUST begin with the accurate analysis, followed by 'Rating:' and then include the corresponding evaluation rating. Your rating MUST be an integer ranging from 1 to 5, following a five-point Likert scale. (higher means better) You MUST keep to the strict boundaries of the given evaluation criterion and focus ONLY on the issues and errors involved; otherwise, you will be penalized. Make sure you read and understand these instructions, as well as the following evaluation criterion and example content, carefully. ###Evaluation Criterion### Polarity Consistency: Is the polarity of the review towards the {target_des} consistent with the given polarity (including negative, slightly negative, neutral, slightly positive, and positive)? ###Example### Polarity: {rating_to_polarity} Review: {analysis} ###Your Evaluation###

Table 20: Prompts and instructions used for re-evaluating evaluations based on the alignment between evaluation analyses and ratings.

D Complete Results of Six NLG Tasks

We display the complete results of our Themis and other models on six NLG tasks respectively in Table 21 to Table 26.

```
{
    "task": "Text Simplification",
    "source": "their granddaughter hélène langevin-joliot is a professor of nuclear physics at the university of pari
s, and their grandson pierre joliot, who was named after pierre curie, is a noted biochemist.",
    "target": "their granddaughter hélène langevin-joliot is a professor of physics, and their grandson pierre jolio
t, who was named, is a biochemist.",
    "dataset": "hsplit",
    "aspect": "Meaning Preservation: Does the simplified text preserve the meaning of the original text?",
    "human_score": [3.0, 3.0, 2.0],
    "gpt4_evaluation": [
        {
            "Analysis": "The simplified text omits the detail that Hélène Langevin-Joliot is...",
            "Rating": 3,
        }, ...
    ], ...
}
```


Figure 3: A simplified example from the hsplit dataset for the text simplification task.

Method	**Coherence**		**Consistency**		**Fluency**		**Relevance**		**Average**	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Traditional Metrics										
BLEU	0.062	0.044	0.048	0.040	0.046	0.036	0.145	0.108	0.075	0.057
ROUGE	0.107	0.080	0.145	0.123	0.113	0.093	0.241	0.183	0.152	0.120
BARTScore	0.474	0.367	0.266	0.220	0.258	0.214	0.318	0.243	0.329	0.261
BERTScore	0.285	0.220	0.151	0.122	0.186	0.154	0.302	0.232	0.231	0.182
BLEURT	0.150	0.112	0.089	0.074	0.133	0.107	0.238	0.178	0.152	0.118
CometKiwi	0.353	0.273	0.151	0.124	0.207	0.170	0.203	0.151	0.228	0.180
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
Prompt LLM										
G-Eval (GPT-3.5)	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.409	0.323
G-Eval (GPT-4)	0.582	0.457	0.507	0.425	0.455	0.378	0.548	0.433	0.523	0.423
GPT-3.5	0.459	0.371	0.393	0.331	0.355	0.296	0.455	0.363	0.415	0.340
GPT-4	0.540	0.434	0.531	0.464	0.480	0.409	0.491	0.395	0.511	0.426
AUTOCALIBRATE (Liu et al., 2024b)	0.570	0.493	0.500	0.467	0.487	0.452	0.560	0.483	0.529	0.474
CoAScore$_{(n=10)}$ (Gong and Mao, 2023b)	0.541	0.419	0.339	0.299	0.367	0.308	0.478	0.379	0.431	0.351
HD-EVAL-NN (Liu et al., 2024c)	0.657	-	0.451	-	0.435	-	0.599	-	0.535	-
Fine-tuned LLM										
X-Eval	0.530	0.382	0.428	0.340	0.461	0.365	0.500	0.361	0.480	0.362
Prometheus	0.150	0.126	0.150	0.137	0.189	0.168	0.164	0.138	0.163	0.142
AUTO-J	0.245	0.203	0.131	0.121	0.154	0.141	0.262	0.222	0.198	0.172
TIGERScore	0.381	0.318	0.427	0.387	0.363	0.327	0.366	0.304	0.384	0.334
InstructScore	0.328	0.276	0.232	0.213	0.260	0.237	0.211	0.179	0.258	0.226
Themis (ours)	0.566	0.485	0.600	0.566	0.571	0.533	0.474	0.412	0.553	0.499

Table 21: Complete results on SummEval for Summarization task.

Method	Context Maintenance		Interestingness		Knowledge Use		Naturalness		Average	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
Traditional Metrics										
BLEU	0.370	0.374	0.406	0.454	0.281	0.369	0.366	0.356	0.356	0.388
ROUGE	0.400	0.376	0.452	0.488	0.339	0.423	0.381	0.360	0.393	0.412
BARTScore	0.119	0.165	0.069	0.059	0.031	0.053	0.050	0.065	0.067	0.086
BERTScore	0.395	0.383	0.439	0.449	0.330	0.378	0.388	0.366	0.388	0.394
BLEURT	0.401	0.408	0.431	0.427	0.321	0.364	0.383	0.354	0.384	0.388
comet22	0.491	0.496	0.544	0.544	0.407	0.450	0.489	0.492	0.483	0.496
CometKiwi	0.334	0.327	0.369	0.355	0.331	0.309	0.380	0.368	0.353	0.340
UniEval	0.595	0.613	0.557	0.605	0.536	0.575	0.444	0.514	0.533	0.577
Prompt LLM										
G-Eval (GPT-3.5)	0.519	0.544	0.660	0.691	0.586	0.567	0.532	0.539	0.574	0.585
G-Eval (GPT-4)	0.594	0.605	0.627	0.631	0.531	0.551	0.549	0.565	0.575	0.588
GPT-3.5	0.550	0.531	0.651	0.648	0.653	0.581	0.515	0.550	0.592	0.578
GPT-4	0.680	0.680	0.822	0.779	0.810	0.786	0.769	0.739	0.770	0.746
CoAScore$_{(n=20)}$	0.539	0.553	0.578	0.595	-	-	0.558	0.596	-	-
HD-EVAL-NN	0.584	0.607	0.682	0.701	0.549	0.568	0.648	0.674	0.616	0.638
Fine-tuned LLM										
X-Eval	0.558	0.622	0.449	0.593	0.734	0.728	0.417	0.478	0.539	0.605
Prometheus	0.451	0.465	0.495	0.473	0.437	0.412	0.355	0.384	0.435	0.434
AUTO-J	0.452	0.449	0.490	0.459	0.339	0.357	0.425	0.437	0.427	0.425
TIGERScore	0.417	0.438	0.328	0.333	0.137	0.138	0.455	0.477	0.334	0.346
InstructScore	0.299	0.297	0.264	0.233	0.140	0.102	0.374	0.332	0.269	0.241
Themis (ours)	0.639	0.644	0.790	0.766	0.778	0.761	0.727	0.729	0.733	0.725

Table 22: Complete results on Topical-Chat for Dialogue Response Generation task.

Method	SFHOT INF.		SFHOT NAT.		SFRES INFO.		SFRES NAT.		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Traditional Metrics										
BLEU	0.070	0.054	0.055	0.040	-0.023	-0.018	-0.004	-0.004	0.024	0.018
ROUGE	0.107	0.082	0.075	0.055	0.118	0.090	0.105	0.078	0.101	0.076
BARTScore	0.211	0.162	0.130	0.094	0.265	0.201	0.226	0.165	0.208	0.156
BERTScore	0.135	0.104	0.126	0.093	0.157	0.120	0.139	0.102	0.139	0.105
BLEURT	0.219	0.171	0.229	0.171	0.244	0.186	0.282	0.211	0.244	0.184
CometKiwi	0.220	0.169	0.235	0.172	0.203	0.153	0.345	0.252	0.251	0.186
UniEval	0.249	0.191	0.320	0.238	0.225	0.169	0.333	0.247	0.282	0.211
Prompt LLM										
GPT-3.5	0.242	0.196	0.294	0.220	0.304	0.250	0.385	0.291	0.306	0.239
GPT-4	0.302	0.263	0.359	0.283	0.213	0.178	0.405	0.316	0.320	0.260
AUTOCALIBRATE	0.357	0.313	0.440	0.383	0.315	0.272	0.416	0.351	0.382	0.330
Fine-tuned LLM										
Prometheus	0.169	0.141	0.211	0.171	0.161	0.134	0.150	0.122	0.173	0.142
Auto-J	0.176	0.152	0.127	0.106	0.179	0.153	0.084	0.070	0.141	0.120
TIGERScore	0.215	0.191	0.204	0.175	0.160	0.141	0.221	0.191	0.200	0.175
InstructScore	0.222	0.194	0.273	0.231	0.194	0.164	0.300	0.251	0.247	0.210
Themis (ours)	0.259	0.226	0.380	0.321	0.298	0.258	0.395	0.332	0.333	0.284

Table 23: Complete results on SFRES & SFHOT for Data to Text task.

Method	CNN-DM			XSUM			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
Traditional Metrics									
BARTScore	0.732	0.680	0.555	0.175	0.171	0.139	0.454	0.425	0.347
CometKiwi	0.176	0.158	0.123	0.027	0.030	0.025	0.101	0.094	0.074
UniEval	0.682	0.662	0.532	0.461	0.488	0.399	0.572	0.575	0.466
Prompt LLM									
G-Eval (GPT-4)	0.477	0.516	0.410	0.211	0.406	0.343	0.344	0.461	0.377
G-Eval (GPT-3.5)	0.631	0.685	0.591	0.558	0.537	0.472	0.595	0.611	0.532
GPT-3.5	0.454	0.514	0.417	0.279	0.348	0.295	0.366	0.431	0.356
GPT-4	0.735	0.746	0.626	0.541	0.528	0.439	0.638	0.637	0.532
AUTOCALIBRATE	0.740	0.744	0.663	0.662	0.662	0.662	0.701	0.703	0.663
Fine-tuned LLM									
Auto-J	0.291	0.238	0.214	0.225	0.214	0.203	0.258	0.226	0.209
TIGERScore	0.574	0.562	0.479	0.424	0.445	0.412	0.499	0.504	0.446
InstructScore	0.287	0.278	0.233	-0.096	-0.134	-0.119	0.095	0.072	0.057
Themis (ours)	0.747	0.761	0.680	0.599	0.607	0.546	0.673	0.684	0.613

Table 24: Complete results on QAGS for Factuality Evaluation task.

Method	ROC			WP			Average		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
Traditional Metrics									
BLEU	0.034	0.035	0.022	0.009	0.029	0.021	0.021	0.032	0.009
ROUGE	0.012	0.000	0.006	0.003	-0.004	-0.004	0.008	-0.002	0.156
BARTScore	0.344	0.330	0.273	0.403	0.370	0.306	0.373	0.350	0.260
BERTScore	0.288	0.269	0.222	0.293	0.300	0.248	0.291	0.285	0.163
BLEURT	0.189	0.155	0.123	0.144	0.122	0.104	0.166	0.138	0.221
CometKiwi	0.246	0.218	0.176	0.289	0.283	0.238	0.268	0.251	0.176
Prompt LLM									
GPT-3.5	0.372	0.363	0.312	0.312	0.294	0.258	0.342	0.328	0.295
GPT-4	0.590	0.578	0.518	0.382	0.368	0.336	0.486	0.473	0.260
CoAScore$_{(n=20)}$	-	-	-	-	-	-	0.414	0.411	0.315
Fine-tuned LLM									
Prometheus	0.031	0.013	0.014	0.001	0.001	0.003	0.016	0.007	0.146
Auto-J	0.460	0.454	0.410	0.308	0.306	0.278	0.384	0.380	0.284
TIGERScore	0.283	0.271	0.221	0.196	0.191	0.158	0.240	0.231	0.207
InstructScore	0.383	0.368	0.316	0.258	0.228	0.194	0.321	0.298	0.168
Themis (ours)	0.637	0.607	0.551	0.507	0.495	0.452	0.572	0.551	0.501

Table 25: Complete results on MANS for Story Generation task.

Method	WMT23		
	r	ρ	τ
Traditional Metrics			
BLEU	-0.130	0.021	0.018
ROUGE	0.081	0.151	0.117
BARTScore	0.091	0.118	0.093
BERTScore	0.123	0.219	0.170
BLEURT	0.163	0.263	0.208
CometKiwi	0.413	0.343	0.273
Prompt LLM			
GPT-3.5	0.388	0.347	0.278
GPT-4	0.496	0.437	0.361
Fine-tuned LLM			
Prometheus	0.144	0.129	0.107
Auto-J	0.128	0.104	0.087
TIGERScore	0.277	0.248	0.211
InstructScore	0.213	0.219	0.181
Themis (ours)	0.431	0.405	0.357

Table 26: Complete results on WMT23 (zh-en) for Machine Translation task.