Universal Multi-Dimensional Text Evaluation Enhanced with Auxiliary Evaluation Aspects

Minqian Liu Ying Shen Zhiyang Xu Lifu Huang
Department of Computer Science, Virginia Tech
`{minqianliu, yings, zhiyangx, lifuh}@vt.edu`

Abstract

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

1 Introduction

Recent advancements of pre-training (Chung et al., 2022; Touvron et al., 2023a,b), prompting (Brown et al., 2020; Wei et al., 2022b; Wang et al., 2023; Qi et al., 2023), and instruction tuning (Wei et al., 2022a) have improved the quality of machine generated texts by a significant degree. Nevertheless, the evaluation of various Natural Language Generation (NLG) tasks still lags far behind compared

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

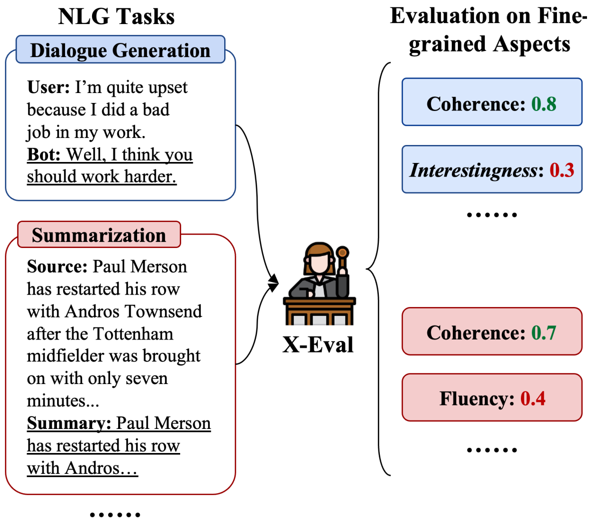

Figure 1: Illustration of X-EVAL for multiple seen and unseen fine-grained evaluation aspects across various NLG tasks. The unseen aspect (i.e., `Interestingness`) is highlighted in *italics*. The text to be evaluated is highlighted with underline. In this example, each evaluation score is from 0 to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

with the rapid progress of large language models (LLMs). Previous similarity-based metrics such as ROUGE (Lin, 2004), BLUE (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020) predominantly measure the similarity between the generated and reference text, failing to accurately reflect the quality of generated text (Gehrmann et al., 2023), especially for open-ended generation tasks.

To obtain a more comprehensive assessment of text quality, multi-aspect evaluation (Fabbri et al., 2021) has been proposed to evaluate the generated text from multiple fine-grained evaluation *aspects*, e.g., `fluency`. While most existing studies (Mehri and Eskenazi, 2020; Yuan et al., 2021; Zhong et al., 2022) consider a closed set of aspects, in many realistic scenarios, the users may need to evaluate the text with their customized aspects, calling for building an evaluator that can be flexibly extended to any *unseen* aspects without the need of training data.

Metrics	Dialogue-level							Turn-level					
	DEP	LIK	UND	FLE	INF	INQ	AVG	INT	SPE	COR	SEM	UND	AVG
BARTScore (Yuan et al., 2021)	0.082	0.099	-0.115	0.093	0.092	0.062	0.052	0.159	0.083	0.076	0.100	0.120	0.128
DynaEval (Zhang et al., 2021)	0.498	0.416	0.365	0.383	0.426	0.410	0.416	0.327	0.346	0.242	0.202	0.200	0.263
UniEval (Zhong et al., 2022)	0.046	0.009	-0.024	-0.003	-0.070	0.085	0.030	0.435	**0.381**	0.125	0.051	0.082	0.215
GPTScore (GPT-3-d03) (Fu et al., 2023)	0.341	0.184	0.196	0.072	0.317	-0.101	0.168	0.224	0.151	0.428	0.405	0.311	0.304
G-Eval (GPT-3.5)† (Liu et al., 2023)	0.339	0.392	0.123	0.344	0.232	0.101	0.259	0.30	0.280	0.430	0.390	0.274	0.335
G-Eval (GPT-4)† (Liu et al., 2023)	**0.583**	**0.614**	**0.602**	**0.587**	**0.510**	**0.551**	**0.573**	**0.506**	0.368	**0.522**	**0.443**	**0.438**	**0.455**
X-EVAL (Ours)	**0.583**	0.436	0.588	0.324	0.480	0.497	0.485	0.421	0.370	0.492	0.376	0.332	0.398

Table 1: **Meta-evaluation on dialogue** based on *unseen* aspects in terms of dialogue-level and turn-level Spearman (ρ) correlations on FED. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with underline.

Recent studies (Fu et al., 2023; Liu et al., 2023) propose to leverage LLMs such as GPT-4 (OpenAI, 2023) as NLG evaluators, yielding promising performance on unseen aspects. However, such evaluations, especially with proprietary LLMs, are cost-intensive, time-consuming, and pose concerns about data privacy and reproducibility.

2 Approach

In this work, we propose X-EVAL, an automatic evaluation framework that can conduct fine-grained evaluation on both seen and unseen aspects across various NLG tasks with a single model, as illustrated in Figure 1. X-EVAL follows a two-stage training paradigm: we first instruction-finetune an open-source language model to equip it with the instruction-following capability for evaluation. Then, motivated by the observation that evaluation aspects usually exhibit inter-connections (Fu et al., 2023) and thus their evaluations can benefit each other, we introduce an additional training stage to finetune the model on the instruction-tuning tasks enriched with the evaluations of a set of *auxiliary aspects*, which are expected to provide clues for evaluating the target aspect and encourage consistent evaluations across multiple aspects. To support our proposed two-stage training of X-EVAL, we construct ASPECTINSTRUCT, the first multi-aspect evaluation instruction tuning dataset spanning 27 diverse aspects over 65 tasks. This dataset is anchored around three core categories of NLG tasks: dialogue, summarization, and data-to-text. We present the illustration of our X-EVAL framework in the Figure 2 in the Appendix.

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

3 Experiments

Experiment Setup We evaluate our X-EVAL on the test split of ASPECTINSTRUCT with 13 unseen aspects. We adopt Flan-T5-large as our base language model for two-stage instruction tuning.

Main Results To assess X-EVAL's ability to generalize to *unseen* aspects, we present the Spearman correlation dialogue evaluation on FED in Table 1. X-EVAL surpasses the traditional metrics and evaluators based on lightweight language models in the top section. Also, X-EVAL matches the performance of GPT-based baselines with much fewer parameters. The bottom section of the table highlights the improvement achieved by two-stage tuning, incorporating instructions, and integrating auxiliary aspects. We report more evaluation results of data-to-text in Table 2, dialog in Table 3, and summarization in Table 4 in Appendix.

4 Conclusion

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

References

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

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

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

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

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

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. *CoRR*, abs/2303.16634.

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

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

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

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. The art of SOCRATIC QUESTIONING: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. 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. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

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

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

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and

Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online. Association for Computational Linguistics.

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

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

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

A More Details on ASPECTINSTRUCT

We define a unified instructions format for tasks included in ASPECTINSTRUCT. Each instruction consists of three parts: (1) *task description* that briefly introduces the evaluation task, (2) *aspect definition*, and (3) *evaluation protocol* that details what the model should output to perform the evaluation. In total, we construct 65 tasks in ASPECTINSTRUCT, where we split 32 tasks and 14 seen aspects for instruction tuning and 33 tasks and 13 unseen aspects for meta-evaluation. We collect 72,637 instances in total with 55,602 instances for training and 17,035 instances for inference.

B More Details on X-EVAL

We present the illustration of the training and inference processes in Figure 2.

Metrics	**SFRES**		**SFHOT**		**AVG**
	NAT	**INFO**	**NAT**	**INFO**	
ROUGE-L	0.169	0.103	0.186	0.110	0.142
BERTScore	0.219	0.156	0.178	0.135	0.172
MOVERScore	0.190	0.153	0.242	0.172	0.189
BARTScore	0.289	0.238	0.288	0.235	0.263
UniEval (Summ)	0.333	0.225	0.320	0.249	0.282
GPTScore	0.190	0.232	0.036	0.184	0.161
G-Eval (GPT-3.5)†	0.144	0.118	0.072	0.102	0.109
G-Eval (GPT-4)†	**0.351**	0.189	**0.338**	0.198	0.269
X-EVAL (Ours)	0.316	**0.265**	0.322	**0.310**	**0.303**

Table 2: Spearman correlation on the data-to-text NLG task. NAT and INFO indicate `Naturalness` and `Informativeness`, respectively. The best results are highlighted in **bold**. †: our re-implementation.

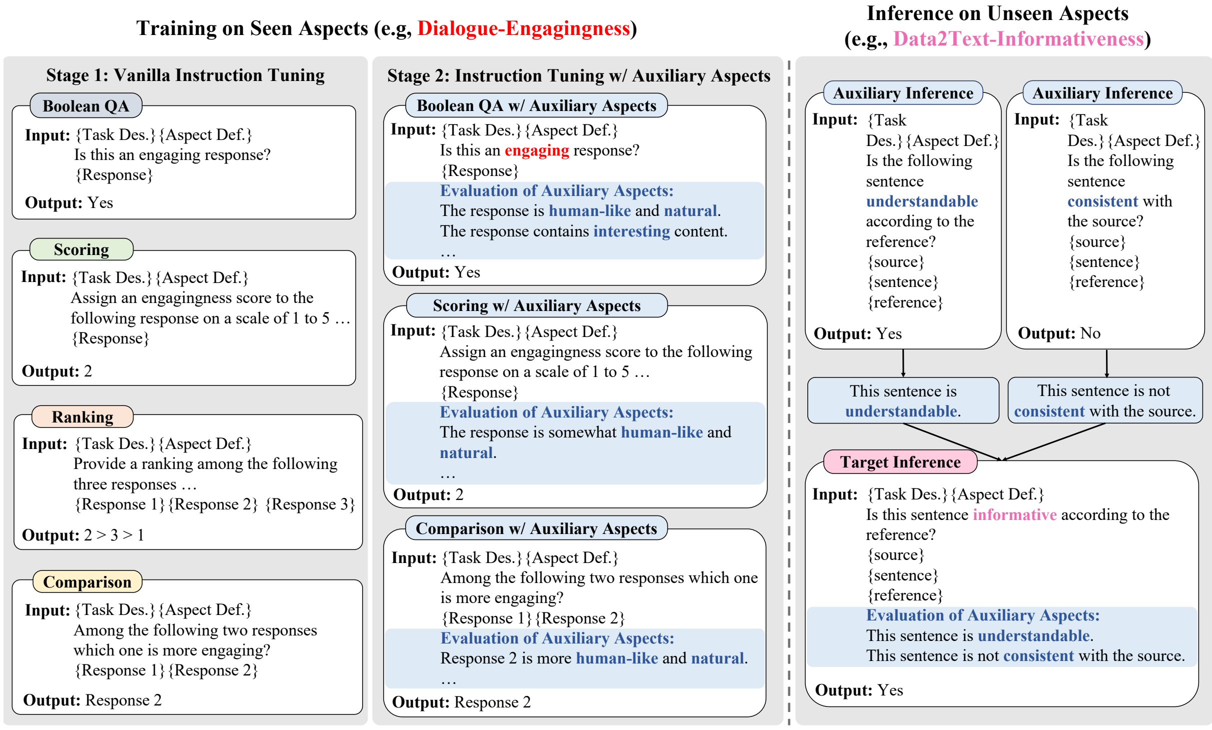

Figure 2: **Illustration of our X-EVAL framework.** The left section depicts our two-stage training approach: vanilla instruction tuning on diverse tasks and subsequent training on instruction tasks enriched with auxiliary aspects. The right section illustrates the inference pipeline with auxiliary aspects.

Metrics	Naturalness		Coherence		Engagingness		Groundedness		AVG	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L (Lin, 2004)	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BERTScore (Zhang* et al., 2020)	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR (Mehri and Eskenazi, 2020)	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval (Zhong et al., 2022)	<u>0.480</u>	<u>0.512</u>	0.518	0.609	<u>0.544</u>	0.563	0.462	0.456	0.501	0.535
G-Eval (GPT-3.5) (Liu et al., 2023)	0.532	0.539	0.519	0.544	**0.660**	**0.691**	0.586	0.567	0.574	0.585
G-Eval (GPT-4) (Liu et al., 2023)	**0.549**	**0.565**	**0.594**	0.605	0.627	0.631	0.531	0.551	**0.575**	0.588
X-EVAL (Ours)	0.417	0.478	<u>0.558</u>	<u>**0.622**</u>	0.449	<u>0.593</u>	<u>**0.734**</u>	<u>**0.728**</u>	<u>0.540</u>	<u>**0.605**</u>

Table 3: Turn-level Pearson (r) and Spearman (ρ) correlations on *seen* aspects on Topical-Chat. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with <u>underline</u>.

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L (Lin, 2004)	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
MOVERSscore (Zhao et al., 2019)	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BERTScore (Zhang* et al., 2020)	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
BARTScore (Yuan et al., 2021)	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval (Zhong et al., 2022)	0.495	0.374	<u>0.435</u>	<u>0.365</u>	0.419	0.346	0.424	0.327	0.443	0.353
GPTScore (Fu et al., 2023)	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-Eval (GPT-3.5) (Liu et al., 2023)	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
G-Eval (GPT-4) (Liu et al., 2023)	**0.582**	**0.457**	**0.507**	**0.425**	0.455	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
X-EVAL (Ours)	<u>0.530</u>	<u>0.382</u>	0.428	0.340	<u>**0.461**</u>	<u>0.365</u>	<u>0.500</u>	<u>0.361</u>	<u>0.480</u>	<u>0.362</u>

Table 4: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval. All aspects are *seen* aspects. The best overall results are highlighted in **bold**. We also highlight the best results excluding GPT-based metrics with <u>underline</u>.