From Model-centered to Human-Centered: Revision Distance as a Metric for Text Evaluation in LLMs-based Applications

Yongqiang Ma[1,2,*], Lizhi Qing[2], Jiawei Liu[1], Yangyang Kang[2†],
Yue Zhang[2], Wei Lu[1†], Xiaozhong Liu[3], Qikai Cheng[1]
[1]School of Information Management, Wuhan University, China
[2]Institute for Intelligent Computing, Alibaba Group, China
[3]Worcester Polytechnic Institute, USA
{mayongqiang,laujames2017,weilu}@whu.edu.cn
{yekai.qlz,shiyu.zy,yangyang.kangyy}@alibaba-inc.com
xliu14@wpi.edu, chengqikai0806@163.com

Abstract

Evaluating large language models (LLMs) is fundamental, particularly in the context of practical applications. Conventional evaluation methods, typically designed primarily for LLM development, yield numerical scores that ignore the user experience. Therefore, our study shifts the focus from model-centered to human-centered evaluation in the context of AI-powered writing assistance applications. Our proposed metric, termed "Revision Distance," utilizes LLMs to suggest revision edits that mimic the human writing process. It is determined by counting the revision edits generated by LLMs. Benefiting from the generated revision edit details, our metric can provide a self-explained text evaluation result in a human-understandable manner beyond the context-independent score. Our results show that for the easy-writing task, "Revision Distance" is consistent with established metrics (ROUGE, Bert-score, and GPT-score), but offers more insightful, detailed feedback and better distinguishes between texts. Moreover, in the context of challenging academic writing tasks, our metric still delivers reliable evaluations where other metrics tend to struggle. Furthermore, our metric also holds significant potential for scenarios lacking reference texts.

1 Introduction

> You can't manage what you can't measure well.—Cruz-Cázares et al. 2013

With the continuous development of large language models (LLMs) such as ChatGPT[1], GPT-4(OpenAI), and Llama(Touvron et al., 2023), a plethora of application research and development work based on LLMs has emerged.

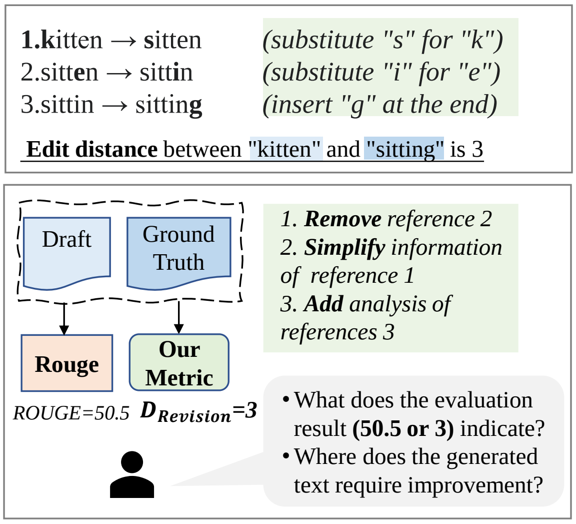

1.kitten → sitten
(substitute "s" for "k")
2.sitten → sittin
(substitute "i" for "e")
3.sittin → sitting
(insert "g" at the end)
Edit distance between "kitten" and "sitting" is 3
Draft
Ground Truth
Rouge
Our Metric
ROUGE=50.5
D_Revision=3
1. Remove reference 2
2. Simplify information of reference 1
3. Add analysis of references 3
• What does the evaluation result (50.5 or 3) indicate?
• Where does the generated text require improvement?

Figure 1: Inspired by the classical edit distance metric, our "Revision Distance" $\mathbf{D}_{Revision}$ can offer a more human-centered and nuanced metric for text evaluation. As illustrated, the $\mathbf{D}_{Revision}(Draft, GroudTruth)$ can provide a more transparent evaluation result, benefiting from the generated revision edit details.

During the model training phase, the main focus is optimizing the model's loss in an isolated environment. However, LLM-based applications should be human-centered, prioritizing user experience and utility. This raises a key question: How do we evaluate LLM-based applications from a human-centric perspective?

Imagining the scenario where developers employ automatic evaluation metrics (Lin, 2004; Papineni et al., 2002; Zhang et al., 2020; Zhao et al., 2019) like Rouge(Lin, 2004) to evaluate LLM-generated text for writing assistance debugging. Rouge only provides a high-level evaluation score to measure textual surface similarity. Since these metrics disregard end-users, the evaluation result is inadequate and misaligns with user needs and preferences. To address this gap, we explore alternative human-centered evaluation metrics, putting the user at the forefront of our evaluation.

[*] This work was done when Yongqiang Ma interned at Alibaba.

[†] Corresponding Authors.

[1] https://openai.com/blog/chatgpt

Findings of the Association for Computational Linguistics: ACL 2024, pages 2127–2137
August 11-16, 2024 ©2024 Association for Computational Linguistics

This paper focuses on the prevalent application scenario for LLMs, specifically, the LLM-powered writing assistant in scenarios from email and letter writing to academic writing.[2] During the AI-human collaborative writing process, AI-generated text often requires extended revisions. Additionally, recent studies suggest that LLMs can produce human-like behavior, such as providing human preferences feedback (Bai et al., 2022; Lee et al., 2023), conducting text quality evaluation (Chiang and Lee, 2023; Fu et al., 2023). Therefore, we assume that the LLM can be a proxy user for generating revision edits that align with actual human editing behaviors.

Drawing from these insights, our proposed metric, $RevisionDistance$, incorporates the iterative process of user-driven text revision. It quantifies the number of edits a user must take to an LLM-generated text to achieve a predefined quality threshold. Compared to typical metrics where higher scores often indicate better quality, our proposed metric operates on the principle that a smaller distance signifies superior text quality. This design is premised on the notion that high-quality texts require fewer revisions to align with a predefined standard of excellence.

In the reference-based evaluation setting, we compared our metric with Rouge, BERT-Score, and GPT-Score across two writing tasks: the easy-writing task and the challenge-writing task. For each task, we sample texts from two models to form a comparison group. Then, we apply text evaluation metrics to assess the text quality. (1) For the easy-writing task, our metric consistently aligns with baseline metrics, supporting the intuition that a stronger model should produce texts with superior evaluation scores. (2) For more challenging tasks, our metrics can still provide stable and reliable evaluation results even if most of the baseline indicators encounter different issues.

In reference-free scenarios, the "Revision Distance" metric aligns closely with human judgment in approximately 76% of cases in the dataset from the ultrafeedback dataset (Bartolome et al., 2023). Furthermore, by categorizing the types of edits made, our metric provides a more fine-grained analysis than those metrics that only yield scores. Additionally, there is no significant difference in format between texts written by humans and AI-generated texts. Our metric is naturally extendable to evaluating human-authored texts in scenarios involving human writing. For example, our metric can be used in educational settings.

The contributions are listed as follows: 1) We highlight the significance of the end-user's perspective in the text evaluation in the context of LLM-power writing assistant. 2) Aligning with real-world human editing behaviors, we propose a human-centered text evaluation metric, which provides a self-explain and fine-grained insight for developers and end-users. 3) Based on broad and various test tasks, we conduct an experiment to demonstrate the utility of our proposed human-centered metrics.

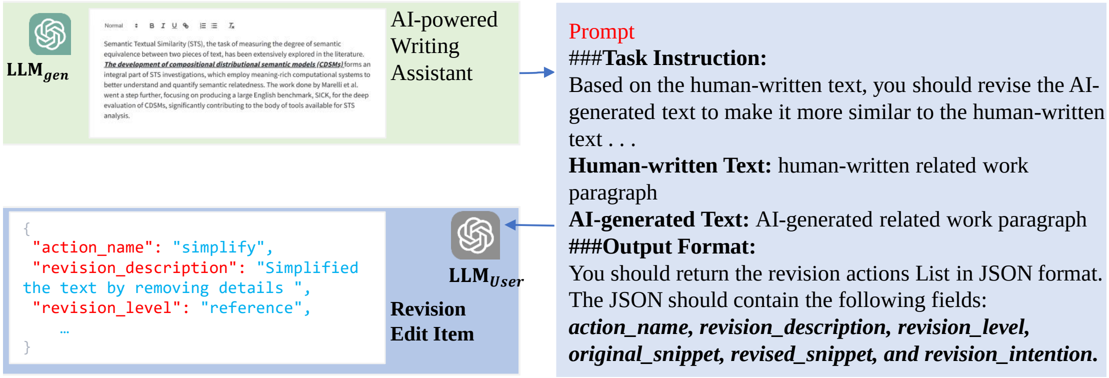
2 Related Work

The text evaluation methods can be categorized into human evaluation and automated approaches. Human evaluation is widely recognized as the most natural way to evaluate the quality of a given text, which often involves human annotators and qualitative analyses (Clark et al., 2021; Belz et al., 2023). This method is often expensive and time-consuming work and requires extensive domain expertise for domain-specific scenarios. On the other hand, current automated evaluation methods tend to generate a comprehensive score that is facilitated in comparing new models with established state-of-the-art approaches. These include metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), BARTScore (Yuan et al., 2021), and DiscoScore (Zhao et al., 2023a) typically compute a similarity (or dissimilarity) score between a model-generated text and a reference text.

Large language models have been adeptly utilized for roles such as aiding in data annotation (Li et al., 2023) and delivering feedback that mirrors human preferences (Bai et al., 2022; Lee et al., 2023; Pang et al., 2023). For the evaluation stage, Chiang and Lee (2023) found that the LLM evaluation is consistent with the human evaluation results. The GPTScore (Fu et al., 2023) has been proposed to score the model-generated text. Similarly, Jain et al. (2023) also studied the efficacy of LLMs as multi-dimensional evaluators.

In conclusion, current metrics tend to yield a comprehensive score that detaches the task context

[2] We use the "Related Work" section Generation (RWG) (Liu et al., 2023; Chen et al., 2021) as the testbed for academic writing, which requires heavy knowledge reasoning work and complex concept understanding ability.



Semantic Textual Similarity (STS), the task of measuring the degree of semantic equivalence between two pieces of text, has been extensively explored in the literature. ***The development of compositional distributional semantic models (CDSMs)*** forms an integral part of STS investigations, which employ meaning-rich computational systems to better understand and quantify semantic relatedness. The work done by Marelli et al. went a step further, focusing on producing a large English benchmark, SICK, for the deep evaluation of CDSMs, significantly contributing to the body of tools available for STS analysis.

```
{
 "action_name": "simplify",
 "revision_description": "Simplified
the text by removing details ",
 "revision_level": "reference",
    …
}
```

Prompt
###Task Instruction:
Based on the human-written text, you should revise the AI-generated text to make it more similar to the human-written text . . .
Human-written Text: human-written related work paragraph
AI-generated Text: AI-generated related work paragraph
###Output Format:
You should return the revision actions List in JSON format. The JSON should contain the following fields: ***action_name, revision_description, revision_level, original_snippet, revised_snippet, and revision_intention.***

Figure 2: The evaluation flow of “Revision Distance”. We require the LLM_{User} to produce results in JSON format with detailed information, In this work, we primarily use the **action_name** to analyze the revisions.

for model development and optimization. However, the ultimate application of LLMs is human-centered, prioritizing the user experience and utility. Consequently, a context-independent numerical score is insufficient in LLM application scenarios. Our proposed metric shifts the text evaluation to a human-centered perspective, which incorporates the iterative process of user-driven text revision.

3 Revision Distance

As depicted in Figure 2, we frame the context of AI-powered writing assistance, with LLMs serving dual functions: as the proxy of the user (LLM_{User}) and as the generator (LLM_{gen}). The LLM_{gen} is the pivotal component of the writing assistant application. Given the user input content, the LLM_{gen} generates a Y_{draft}, such as emails, letters, articles, and “Related Work” sections. The $\mathbf{D}_{Revision}$ quantifies the number of edits from $LLM_{User}(Y_{draft}, Y)$.

To avoid the generated revision focusing excessively on surface-level textual features, we have explicitly required LLM_{User} to focus on coherence and logical flow, as shown in Appendix D. Additionally, we have calibrated the revision generation process by setting the temperature parameter to 0, aiming to reduce the likelihood of generating unstable outputs. We also separated complex revisions that contained multiple revisions, such as “simplify and reorganize,” into individual actions to ensure consistency in the granularity of revisions. By setting a consistent scope and clear objectives for revisions, our metric promotes a uniform revision process that is less prone to variance, thus enhancing reproducibility and reliability.

For the reference-based evaluation setting, we utilize the human-written text or ChatGPT output as the ground truth. The LLM_{User} is designed to produce structured revision edits, improving the consistency of the Y_{Draft} to the ground-truth text Y. In scenarios where no ground truth text is available, we require the LLM_{User} to refine the given text towards an ideal form, as envisioned by the LLM_{User} itself.[3] These revision edits are produced to improve Y_{draft} to closer align with the ideal version, which mimics the revision process of human writers.

4 Results and Discussion

4.1 Evaluation for Reference-based Setting

4.1.1 Task and Dataset

Task Level	Weak 😐	Strong 😎
Easy	Mistral-7B	Mixtral-8x7B
Challenge	vanilla GPT-4	CoT-based GPT-4

Table 1: The employed models for both writing tasks.

To validate the utility of our proposed metric, we have constructed two distinct datasets to address both the easy-writing task and the challenge-writing task. The challenge-writing task refers to the scenario that requires heavy knowledge reasoning and complex concept understanding. For the easy-writing task, we use emails, letters, and article generation as a testbed. For the challenge-writing task, we employ academic writing as the testbed. The test dataset details in this evaluation setting are described in Appendix A.

[3] This ideal version is not explicitly generated but rather serves as an implicit standard within the revision edits generation prompt.

	Easy-writing Task			Challenge-writing Task		
Metrics	😐 Mistral-7b	😎Mixtral-8x7b		😐 Vanilla GPT-4	😎CoT GPT-4	
Rouge 1 ↑	50.53	51.65	▲ 2.2%	31.62	29.78	▼ -6.2%
Rouge 2 ↑	19.52	22.06	▲ 11.5%	7.64	6.86	▼ -11.4%
Rouge L ↑	26.74	29.21	▲ 8.5%	15.09	15.59	▲ 3.2%
Bert-Score ↑	–	–	–	57.36	56.36	▼ -1.8%
GPT-Score ↑	90.56	88.63	▼ -2.2%	87.67	87.47	▼ -0.2%
$\mathbf{D}_{Revision}$ ↓	3.20	2.79	▲ **14.7%**	3.94	3.73	▲ **5.3%**

Table 2: The symbols ▲ and ▼ indicate directional changes in performance as delivered by evaluation metrics. Specifically, ▲ and ▼ denote performance improvement and decline, respectively, from weaker to stronger models. For the easy-writing task, our $\mathbf{D}_{Revision}$ aligns well with other evaluation measures, showing our metric's utility. For the more challenging writing task, it offers stable evaluations and better distinguishes model quality, whereas other metrics struggle. The limited input length of Bert-Score, capped at 512 tokens, precluded its use in the easy-writing task where many texts exceeded this limit.

4.1.2 Text Generation Models

To assess the discriminative capacity of our revision distance metric, we designed strong and weak writing applications. The terms "strong" and "weak" refer to the generation ability of utilized LLM, as detailed in Table 1. (1) For the easy-writing task, we employ two Mistral-series models (Jiang et al., 2023); (2) For the challenge-writing task, we employ GPT-4 and its variant.[4]

We have validated the text generated by both strong and weak models. For most of the text pairs, the strong model consistently produced higher-quality text. Aligning with human evaluations, a good metric should yield higher evaluation scores for texts generated by stronger models compared to those from weaker models. Here, we analyze the evaluation capability of different metrics by assessing texts generated by strong and weak models.

4.1.3 Result Analysis

As shown in Table 2, our metric shows utility for easy and challenging writing tasks. Different from other metrics, smaller $\mathbf{D}_{Revision}$ indicate better text quality. To assess the metric's ability to differentiate between models, we calculate the relative change rate from the "Weak" model to the "Strong" model. Notably, existing metrics have reached saturation for the easy-writing tasks, exhibiting a limited relative change rate regarding the performance of distinct models. For example, the GPT score tends to assign higher scores to texts, primarily clustering between 85 to 95, resulting in smaller relative variations.

Conversely, our metric demonstrates better efficacy in discerning the nuanced capabilities of diverse models. Our metric operates within a range of 0 to n, which, by its nature, may exhibit large relative changes for small numeric variations. It's observed that $\mathbf{D}_{Revision}$ yields a larger change rate, highlighting the enhanced discriminative capacity of our metric.

Additionally, for the complex academic writing task, we conducted a human evaluation, as shown in Appendix G. Based on the evaluation results, we categorized texts as "Chosen" or "Rejected." As shown in Table 3, $\mathbf{D}_{Revision}$ metric aligns with human preferences, indicating superior text quality with fewer revisions for "Chosen Texts." In contrast, the Rouge metric often misaligns with human judgments, erroneously assigning higher scores to "Rejected Texts."

Metric	**Chosen text**	**Rejected text**
$\mathbf{D}_{Revision}$ ↓	3.4	3.9
Rouge 1 ↑	30.92	32.50
Rouge 2 ↑	6.58	7.29
Rouge L ↑	14.71	15.04
Bert-score ↑	58.01	58.54
GPT-score ↑	87.4	87.7

Table 3: Comparison of Evaluation Metrics for Chosen and Rejected Texts

Our metric shows a negative Spearman correlation (R = -0.38) with human evaluation scores, unlike the near-zero or marginally positive coefficients of Rouge, BERTScore, and GPTScore. This indicates our metric's unique approach, where

[4] The models employed in both tasks are detailed in Appendix B and Appendix C, respectively.

lower scores mean better text quality.

4.2 Evaluation for Reference-free Setting

To demonstrate the performance of our evaluation method in scenarios, where ground truth is unavailable, we extracted 41 cases related to writing tasks from the UltraFeedback dataset(Bartolome et al., 2023). Each case contains a chosen response and a rejected response.

When applied to the selected cases, our "Revision Distance" metric aligns with human judgments in 76% of instances, indicating that chosen responses typically necessitated fewer revisions.

4.3 Qualitative Analysis

Based on the analysis of revision edit details from the Challenge-writing Task (writing the "Related Work" section), we classify the revision actions into three categories: (1) Reference Order Revision, (2) Reference Comparison Revision, and (3) Reference Description Revision. The description of three categories is shown in Appendix F.

For complex writing tasks, the challenge often lies in knowledge reasoning of concepts. CoT prompting can dramatically improve the multi-step reasoning abilities of LLMs (Wang et al., 2023). As shown in Table 4, for texts generated by Vanilla GPT-4, the average revisions required are 0.80 for Order, 0.84 for Comparison, and 2.29 for Description. For the texts generated by CoT GPT-4, the average revisions required are 0.67 for Order, 0.71 for Comparison, and 2.36 for Description.

Model Type	Order	Comp	Desc	Total
Vanilla GPT-4	0.80	0.84	2.29	3.93
CoT GPT-4	0.67	0.71	2.36	3.73

Table 4: The result of revision edits analysis. To calculate the average number of revisions across different dimensions, we divided the total number of revisions (in Order, Comparison, and Description) by the total number of cases.

Based on the fine-grained analysis of revision edits, we find that CoT-based GPT-4 can provide text with fewer revisions related to Order and Comparison issues in "Related work" writing tasks. A slight decline in the reference description dimension exists compared with Vanilla GPT-4. In conclusion, the fine-grained analysis revision edits can provide insightful feedback for future model improvement.

5 Conclusion

With the rapid advancement of LLM-based applications, the pivotal question arises: "how can we evaluate LLM-based applications from a human-centered perspective?" Existing evaluation metrics, typically used for model development, merely yield a context-independent numerical score, lacking user relevance. Our research shifts text evaluation from a predominantly model-centered perspective to a human-centered one.

Using the LLM-powered writing assistant as a test scenario, we take a comprehensive experiment on diverse writing tasks to validate the effectiveness and reliability of our "Revision Distance" metric. This metric converts text evaluation into contextualized text revisions, highlighting textual discrepancies and offering users a detailed, transparent rationale for the scores. Our findings demonstrate the metric's applicability and dependability in both reference-based and reference-free contexts.

Limitations

This paper introduces a metric that leverages GPT-4, specifically applied to evaluating LLM-powered writing assistants. In our initial attempts, we utilized the GPT-3.5 API to generate revisions. However, GPT-3.5 struggled to follow the instructions for generating revisions that improve logic flow and coherence. However, the computational and financial costs of using GPT-4 are considerable. Exploring the use of a smaller, specialized model to generate initial edits could reduce costs and improve efficiency.

LLMs have a wide array of applications, and for this study, we have chosen the "Related Work" section generation task as a testbed for challenging writing scenarios. As a knowledge-intensive cognitive task, writing the "Related Work" section requires writers to integrate multi-source knowledge into the manuscript. Therefore, writing a comprehensive "Related Work" section is a labor-intensive and time-consuming endeavor. Future studies could explore the application of our metric in longer text generation tasks, such as code generation and scientific reports, to validate its effectiveness and applicability across different domains.

In this study, each generated revision item is assigned equal weight. Future research should focus on developing a dynamic revision edit weighting method to evaluate textual differences more finely.

Ethics Statement

In this paper, we propose a new automatic evaluation metric, $RevisionDistance$, to evaluate the LLM-generated text in an AI-power writing assistant setting. The positive impact of Revision Distance is that it can provide a more nuanced and self-explain representation of the quality of LLM-generated text. Notably, our metric is human-centered and transparent, which can help demystify the evaluation process for LLM-generated text, making it more accessible to a wider user, including those who are not experts in AI. The negative impact is that over-reliance on $RevisionDistance$ might lead to the overlooking of qualitative aspects of text generation that are harder to quantify, such as creativity. Additionally, if the reference texts within $RevisionDistance$ are biased or of low quality, this could amplify the biases in the LLMs-generated text.

Acknowledgements

This work was supported by the National Science and Technology Major Project (2022ZD0116204), the National Natural Science Foundation of China (72174157), and Alibaba Group through Alibaba Innovation Research Program.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Alvaro Bartolome, Gabriel Martin, and Daniel Vila. 2023. Notus. `https://github.com/argilla-io/notus`.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. 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 6068–6077, Online. Association for Computational Linguistics.

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

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Claudio Cruz-Cázares, Cristina Bayona-Sáez, and Teresa García-Marco. 2013. You can't manage right what you can't measure well: Technological innovation efficiency. *Research Policy*, 42(6):1239–1250.

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

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8487–8495, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

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.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, Liang Hu, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2148–2159, Singapore. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

OpenAI. GPT-4 technical report. Technical report, OpenAI.

Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. 2023. Language model self-improvement by reinforcement learning contemplation. *arXiv preprint arXiv:2305.14483*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The acl ocl corpus: advancing open science in computational linguistics. *arXiv preprint arXiv:2305.14996*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

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

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, Hong Kong, China. Association for Computational Linguistics.

Wei Zhao, Michael Strube, and Steffen Eger. 2023a. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2023b. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

A Dataset for Reference-based Setting

- For the easy-writing task, we use the task of emails, letters, and articles generation as testbeds. Specifically, we extracted 147 relevant instances of emails, letters, and articles written from Wildchat (Zhao et al., 2023b), a real-world user-ChatGPT interactions corpus, as the easy-writing dataset.

- For the challenge-writing task, we employ the "Related Work" generation task (Liu et al., 2023; Chen et al., 2021) as the testbed. The "Related Work" sections in academic manuscripts position the authors' contributions within the existing academic context. They highlight the novelty and advantages of the presented research compared to existing works. Specifically, we randomly selected 90

“Related Work" paragraphs from scholarly articles within the ACL dataset (Rohatgi et al., 2023) as the challenge-writing dataset.

B Text Generation Models for Easy Writing Task

We use the APIs of Mistral-7B[5] and Mistral-8x7B[6], hosted on Huggingface, to generate responses for the prompts within our constructed easy-writing task dataset. The parameters for the inference process are shown in Table 5.

Parameter	Value
temperature	0.9
max_new_tokens	4096
stop_sequences	["</s>"]
top_p	0.95
repetition_penalty	1.0
do_sample	True
seed	41

Table 5: Generation Configuration Parameters

C Text Generation Models for Challenge Writing Task

In academic manuscripts, the “Related Work” sections position the authors’ contributions within the existing academic context. They highlight the novelty and advantages of the presented research in comparison to existing works.

In the “Related Work” generation task, the model utilizes a set of reference papers, denoted as Ref, along with a description of the user’s current research denoted as D, to generate a “Related Work” draft, denoted as Y_{Draft}, for the user. In this work, the input data is sourced from the ACL papers (Rohatgi et al., 2023). We select the related work section paragraph as the test data based on the section title. Notably, the abstracts of both the reference papers and the user’s target paper are utilized to encapsulate the core content of the respective research, thereby assisting in the generation process. For both Vanilla GPT-4 and CoT-based GPT-4, the temperature is set as 1.0 in the generation process.

$$Y_{Draft} = LLM_{gen}(Ref, D) \tag{1}$$

Vanilla GPT-4 We concatenate the task instruction and metadata of reference and source papers to directly prompt the LLM to get the final “Related Work”.

CoT-based GPT-4 We initially prompt the LLM to generate learned relevant knowledge in the training stage(Sun et al., 2023) and then create three segments for different perspectives. Finally, these segments, along with the recalled knowledge and the metadata of the input papers, are integrated to generate the comprehensive “Related Work” paragraph. Based on the intermediate step, the LLM can better capture interrelationships among scientific publications and concepts.

D Prompt for Revision Generation

> You are a text revision system for revising the text generated by AI.
> You should simulate the revision process of an end-user who is interacting with an AI text generation system. You should produce the revision action from the perspective of the content and the structure of the text that the end-user might be interested in. Given an AI-generated text, you should revise the AI-generated text by comparing it with the human-written text. You should first read the AI-generated text carefully and mark the possible problems in the AI-generated text. Then, you should revise the AI-generated text to improve its quality by fixing the problems you have marked. You can ignore the grammar, and word choice problems in the AI-generated text. You should focus on the concept and the structure of the text.
> **Human-written Text**: human-written related work paragraph
> **AI-generated Text**: AI-generated related work paragraph
>
> Output Format: refered as Figure 6

Figure 3: Prompt for easy-writing task in reference-based setting.

In the prompt, we require the GPT-4 to output in JSON. Because the output format setting is the same across multiple prompts, it is explained separately here, as shown in Figure 6.

[5] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[6] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

You are a text revision system for revising the related work text in scientific papers.
You are given a human-written text and an AI-generated text. Based on the human-written text, you should revise the AI-generated text to make it more similar to the human-written text. You can ignore the reference mark, grammar, and word choice problems in the AI-generated text. You can not copy the sentence directly into human-written text. You should focus on the concept and the structure of the text. The revision actions should be reference level. For the reference level, you can expand, simplify, compare, or group the description of the reference in the AI-generated text. You can also reorganize the order of the reference description in the AI-generated text. The revision description should be clear and concise.
Human-written Text: human-written related work paragraph
AI-generated Text: AI-generated related work paragraph
Output Format: referred to in Figure 6

Figure 4: Prompt for challenge-writing task in reference-based setting.

You are a text revision system for revising the text generated by AI.
You should simulate the revision process of an end-user interacting with an AI text generation system. Given an AI-generated text, you should revise the AI-generated text to improve its quality from the perspective of the content and the structure of the text that the end-user might be interested in. You should first read the AI-generated text carefully and mark the possible problems in the AI-generated text. Then, you should revise the AI-generated text to improve its quality by fixing the issues you have marked. You can ignore the grammar and word choice problems in the AI-generated text. You should focus on the concept and the structure of the text.
AI-generated Text: AI-generated related work paragraph
Output Format: referred to in Figure 6

Figure 5: prompt in reference-free setting.

You should return the revision actions List in JSON format. The JSON should contain the following fields: ***action_name, revision_description, revision_level, original_snippet, revised_snippet, and revision_intention***.

Figure 6: Output format prompt. It is referred to in Figures 3, 4, and 5.

E Example for Revision Action Item

E.1 Revision Regarding the Text Content

###Human-written Text: To deal with the STS task, previous studies have resorted to various features (e.g. word overlap, synonym/antonym), linguistic resources (e.g. WordNet and pre-trained word embeddings) and a wide assortment of learning algorithms (e.g. Support Vector Regression (SVR), regression functions and NNs). Among these works, several techniques extract multiple features of sentences and apply regression functions to estimate these similarity scores (Lai & Hockenmaier, 2014; Zhao et al., 2014; Bjerva et al., 2014; Severyn et al., 2013). Lai & Hockenmaier (2014) analyzed distinctive word relations (e.g. synonyms, antonyms, and hyperonyms) with features based on counts of co-occurences with other words and similarities between captions of images. Zhao et al. (2014) predicted the sentence similarity from syntactic relationship, distinctive content similitudes, length and string features. Bjerva et al. (2014) also utilized a regression algorithm to foresee the STS from different features (WordNet, word overlap, and so forth). Finally, Severyn et al. (2013) combined relational syntactic structures with SVR.

###AI-generated Text: Semantic Textual Similarity (STS), the task of measuring the degree of semantic equivalence between two pieces of text, has been extensively explored in the literature. The development of compositional distributional semantic models (CDSMs) forms an integral part of STS investigations, which employ meaning-rich computational systems to better understand and quantify semantic relatedness. The work done by Marelli et al. went a step further, focusing

```
{
    "action_name": "simplify",
    "revision_description": "Simplified the text by removing details regarding
        CDSMs and the inclusion of the SICK benchmark.",
    "revision_level": "reference",
    "revision_intention": "simplify",
    "original_snippet": "Semantic Textual Similarity (STS), the task of measuring
        the degree of semantic equivalence between two pieces of text, has been
        extensively explored in the literature. The development of compositional
        distributional semantic models (CDSMs) forms an integral part of STS
        investigations, which employ meaning-rich computational systems to better
        understand and quantify semantic relatedness. The work done by Marelli et
        al.  went a step further, focusing on producing a large English benchmark,
        SICK, for the deep evaluation of CDSMs, significantly contributing to the
        body of tools available for STS analysis.",
    "revised_snippet": "In the broad field of Semantic Textual Similarity (STS),
        earlier works have explored numerous computational models to comprehend and
        quantify the semantic relatedness between texts."
}
```

Figure 7: An example of content-based revision. The generated revision is about simply the background introduction in the AI-generated text.

on producing a large English benchmark, SICK, for the deep evaluation of CDSMs, significantly contributing to the body of tools available for STS analysis. . . ..

Metric	Rouge-1	Rouge-2	Bert-Score
Value	100.0	100.0	99.0

Table 6: Current typical automated metrics' output.

E.2 Revision Regarding the Text Structure

In this case study, we compare a human-written text and an AI-generated text that superficially appears very similar but exhibits subtle differences in the structure and order of information. The human-written text follows a logical sequence, while the AI-generated text swaps the order of key details.

###Human-written Text: First, **Authors A** proposed a BERT-based method. Second, **Authors B** proposed a GPT-based method.

###AI-generated Text: First, **Authors B** proposed a GPT-based method. Second, **Authors A** proposed a BERT-based method.

This case is used to demonstrate the limitations of current automated metrics in capturing such structural differences and demonstrates the superiority of our novel metric in evaluating text quality.

As shown in Table 6, current commonly used metrics tend to assign near-perfect scores to AI-generated texts, implying a high degree of equivalence with their human-written counterparts. However, this fails to capture the underlying structural differences between the texts.

As depicted in Figure 8, our metric can better capture the text's structural differences. Notably, our metric can offer users a coherent and transparent explanation of the scores assigned, benefiting from the detail of revision actions.

```
{
 "action_name": "Reorder",
 "revision_description": "Reordered the
     sequence of references to match
     the human-written text",
 ...
}
```

Figure 8: Example output of our metric, demonstrating the structural difference between the human-written and AI-generated texts. As shown in " revision_description," the order of the related work statements is adjusted to reflect the original argumentation structure.

F Revision Categories for the LLM-Generated "Related Work"

1. Reference Order Revision: This refers to reorganizing the sequence of references from various viewpoints such as chronological order, methodological approach, or problem context.

2. Reference Comparison Revision: This refers to integrating comparative discussions among a collection of references, thereby stating their congruities or discrepancies.

3. Reference Description Revision: This refers

to modifying the description of a particular reference paper, either by elaborating it or by making it more concise.

G Human Evaluation for Metrics in Challenge-writing Task

We selected 20 paragraphs from both methods for expert analysis. Five AI field specialists assessed the LLM-generated content, focusing on content quality, structural coherence, and argumentative strength. Evaluators assign scores across three dimensions on a scale from 1 to 5. The sum of the scores from the three dimensions is taken as the final score for the text.

The rate of alignment with humans is 50% for our metric. The human alignment rate for Rouge 1, Rouge-2, Rouge-L, and BertScore is lower than 50%, which is 40%. The GPT-Score's human alignment rate is 60%. However, the GPT-Score's evaluation scores for human-chosen and rejected samples are 87.4 and 87.7, respectively. This demonstrates that the GPT-Score overall lacks discriminative power. Additionally, compared to our metric, the GPT-Score also falls short in terms of interpretability.

While their potential optimization revision edits might differ for two texts, they can have the same number of optimization steps, leading to instances where their Revision Distance is equal (accounting for 20% of cases). Therefore, our metric achieved only a 50% human alignment rate. We will refine our metric by implementing a weighting method to address this issue.

H Stability Analysis

The stability of evaluation results when employing LLMs as proxies of evaluators is a concern echoed by other metrics such as the GPT-Score. Based on the "TextAttack" framework (Morris et al., 2020), we utilized word embeddings to make slight alterations to the words in the text (originally with a Revision Distance of 4). This yielded four perturbed samples, with resultant revision edits numbering 4, 4, 5, and 4, illustrating our metric's stability.