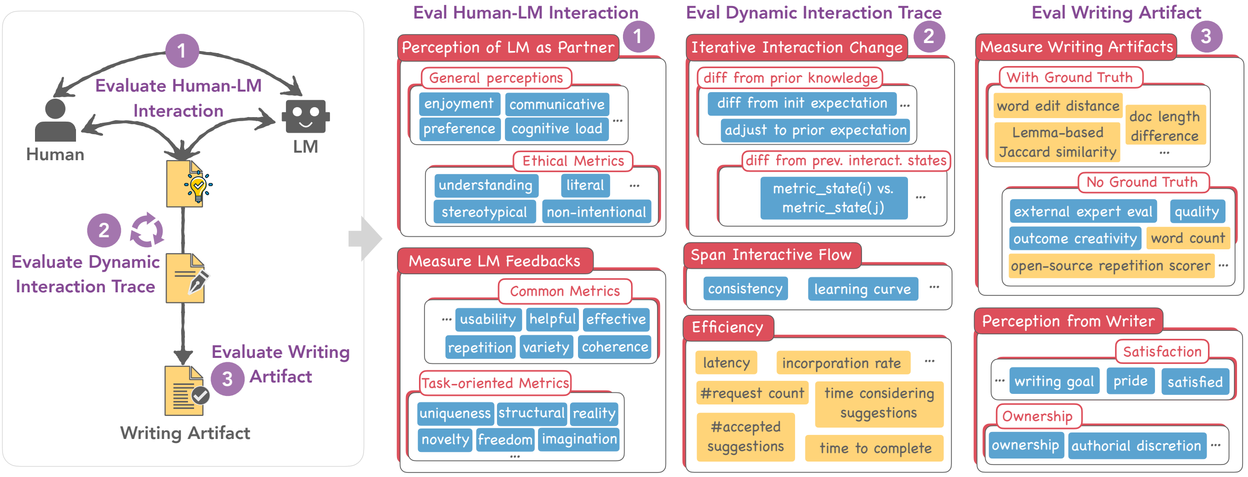
Parachute: Evaluating Interactive Human-LM Co-writing Systems

Hua Shen
huashen218@psu.edu
Pennsylvania State University, USA

Tongshuang Wu
sherryw@cs.cmu.edu
Carnegie Mellon University, USA



(a) Parachute Overview

(b) Unfolding Parachute — The collections of subgroup metrics for each evaluation aspect

Figure 1: Parachute: human-centered integrative evaluations on interactive co-writing systems. (a) Parachute overview. (b) The subgroup metrics for each evaluation aspect. (yellow and blue indicates objective and subjective metrics, respectively.)

ABSTRACT

A surge of advances in language models (LMs) has led to significant interest in using LMs to build co-writing systems, in which humans and LMs interactively contribute to a shared writing artifact. However, there is a lack of studies assessing co-writing systems in interactive settings. We propose a human-centered evaluation framework, Parachute, for interactive co-writing systems. Parachute showcases an integrative view of interaction evaluation, where each evaluation aspect consists of categorized practical metrics. Furthermore, we present Parachute with a use case to demonstrate how to evaluate and compare co-writing systems using Parachute.

1 INTRODUCTION

Language models (LMs) have advanced significantly, showcasing previously unheard-of capabilities in solving a wide spectrum of generation and language understanding tasks [3, 7, 8]. This has spurred great academic and public interest in using LMs to build writing assistants, in which humans collaborate with LMs to paraphrase sentences (*e.g.*, QuillBot), autocomplete sentences [2], write stories [1], etc. Despite the interactive co-writing process, the co-writing systems are at present primarily tested in non-interactive settings [4, 11]. Specifically, current studies commonly conduct one-time evaluations on writing article quality [6], or prior- and post-human assessment on perceiving LMs [10, 11], etc. As a result, these evaluations fall short of capturing the dynamic shift in human-LM interactions and even fair assessments. For instance, suppose a scientific paper writing task involves two users, a freshman and a university professor, to participate in user study evaluations. With the help of LMs, the freshman makes much improvement in writing papers by benefitting from LMs' suggestions (*e.g.*, understanding the paper structure, grammar correction, etc.) While the professor is skilled at writing in general, the system is considerably less helpful for him. However, when making the one-time evaluation of the final paper's quality, the professor's performance will likely outperform the freshman's significantly. This result can not genuinely reflect the co-writing system's influence on users. Therefore, we should assess users' dynamic interaction improvements to indicate system capability, such as measuring relative quality change between two iterated articles from the same user.

Nevertheless, as far as we know, there is at present a lack of studies in examining *how to assess the interaction shift along the iterative co-writing process?*, and furthermore, *how to depict an integrative view for evaluating interactive co-writing systems?* To close this gap, we propose Parachute: a human-centered evaluation framework for human-LM interactive co-writing systems. we identify the key components and interaction aspects for evaluating the co-writing systems. We further collect the comprehensive types of evaluation metrics under each aspect, including the novel measurements designed for dynamic interaction assessments, and the other two conventional yet important aspects (*i.e.,* human-LM interaction and writing artifact evaluation). Additionally, we present a use case study to exemplify how to leverage Parachute as a thinking tool to comprehensively evaluate and compare the co-writing systems.

CHI In2Writing Workshop '23, Apr 23, 2023, Hamburg, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PARACHUTE: A THINKING TOOL TO GUIDE INTEGRATIVE INTERACTION EVALUATION

A human-LM co-writing system involves both human and LM collaborating on a shared writing artifact (*e.g.,* essay, story, scientific paper, etc.) as partners [9]. Grounded on the Co-Creative Framework for Interaction Design (COFI) model [9], we identify three key components in co-writing systems, including *human*, *LM*, and *writing artifact*. They act as the joint driving force of the interactive co-writing process. As humans are the essential decision-makers to interact with LMs and write artifacts, in this work, we propose a human-centered evaluation framework[1], Parachute, to comprehensively evaluate the human-LM interactive co-writing systems.

As described in Figure 1, Parachute further recognizes three crucial aspects for evaluating the components' interactions, including ① *evaluate human-LM interaction*, ② *evaluate dynamic interaction trace*, and ③ *evaluate writing artifact*. Compared with existing evaluation approaches, which might capture insufficient evaluation aspects to reflect the system capability holistically [4, 6, 11] or focus much on one-time metrics that neglect dynamic interaction changes [5, 10], Parachute seeks to contribute in two folds. First, Parachute presents a more integrative view of interaction evaluation supported by practical metrics. Besides, Parachute explicitly extracts a set of metrics to assess the dynamic metric change along iterative interactions. Overall, we propose Parachute as a thinking tool, assisting researchers to practically evaluate and analyze interactive co-writing systems more comprehensively and fairly.

3 UNFOLD PARACHUTE: WHICH EVALUATION METRICS CAN WE USE IN PRACTICE?

To consolidate Parachute to be more useful in practice, we **unfold the three evaluation aspects into subgroups of practical evaluation metrics**. To achieve this, we use an inductive approach to collect the practical metrics adopted in state-of-the-art co-writing systems (*e.g., Wordcraft* [11], *Integrative Leaps* [10], *Beyond Text Generation* [4], *Dramatron* [6],etc.) Then we categorize them into subgroups and fit into Parachute framework. Figure 1(b) depicts the categorized evaluation metrics for each Parachute aspect. We next clearly define the aspects and metric subgroups in Parachute, and briefly explain the underlying motivation. Please see Table 1 in Appendix A.1 for more elaborated metrics and details.[2].

① **Evaluating Human-LM interaction.** These metrics measure interactions between the co-writers (*i.e.,* human & LM). Suppose that human perceives LM as a co-author in co-writing systems. Then her evaluations primarily derive from two dimensions: *i) how does the human feel to collaborate with LM?* (*i.e.,* "perception of LM as partner", like *enjoyment*, *stereotypical*, etc.), and *ii) how credible are the LM's feedbacks?* (*i.e.,* "measure LM feedbacks", such as *usability*, *imagination*, etc.).

② **Evaluating dynamic interaction trace.** These metrics focus on evaluating the dynamic change of interactions along iterative writing process. We identify three dimensions for evaluations. First, when human iteratively updates the artifacts, "Iterative Interaction Change" subgroup aims to compare metrics of current and previous states, such as measuring human understanding on LM before start writing and when almost finish the article). Also, we cover "Span Interactive Flow" subgroup to assess metrics that need to observe spanning multiple artifact versions (*e.g.,* consistency). Besides, the responding time matters in the interactive systems. We thus include "Efficiency" metrics (*e.g.,* latency, incorporation rate) to assess the interactive process.

③ **Evaluating writing artifact.** These metrics gauge the content of the final writing artifact that human and LM accomplish jointly. We broadly divide these metrics into "measure writing artifacts" (*e.g., Jaccard similarity* (with ground truth), or *external expert evaluation* (without ground truth)) and "perception from writer" dimensions (*e.g., satisfaction* or *ownership*).

4 HOW TO USE PARACHUTE TO EVALUATE AND COMPARE CO-WRITING SYSTEMS?

We consider Parachute as a thinking tool for researchers to fairly evaluate and compare co-writing systems. Parachute can be helpful for researchers to: *1)* **identify key interactive evaluation aspects** among human, LM, and writing artifact interactions during the co-writing process, *2)* **select appropriate metrics** to assess and compare the co-writing systems, *3)* **comprehensively analyze and describe** the human-LM interactive performance of the co-writing systems. Next we present a concrete use case of re-evaluating the *Beyond Text Generation* (BTG) [4] system to showcase how to use Parachute for evaluation.

Suppose the researchers have built BTG system and need to evaluate its performance in the interactive setting. The researchers may now apply Parachute to conduct comprehensive evaluation on it and compare with existing baseline systems. For example, by checking the Parachute framework, in the first step, the researchers decide to leverage Parachute to validate: "**the BTG system can help humans write better articles** (*i.e., aspect2: evaluate writing artifact*) **by enabling better human-LM interactions** (*i.e., aspect1: evaluate human-LM interaction*) **in efficient ways** (*i.e., aspect3: evaluate dyanmic interaction trace*). Next, they delve into each aspect to select the appropriate metrics to support this statement. For instance, they assess "writers' perceptions of LM" by choosing *enjoyable*, *preference* metrics, and "how writers think about LM's feedbacks" with *usabiity*, *effective*, *coherence* metrics, etc. Also, they assess the dynamic interaction efficiency by analyzing the objective logging data (*e.g., incorporation rate*, *latency*). Besides, they also evaluate the final writing article with a set of measures (*e.g., external expert review*, *quality*, *satisfaction*, *ownership*, etc.). Thirdly, after selecting these metrics, they design user study and approaches to accomplish these evaluations. Note that the researchers applies all measurements to both proposed BTG system and the baselines they choose to compare co-writing systems.

5 CONCLUSION

This work present Parachute: a human-centered framework to evaluate human-LM interactive co-writing systems. It provides a thinking tool for researchers to design comprehensive interaction evaluations and analyses. We further feature a use case study introducing how to use Parachute step-by-step for fair evaluations.

[1] "human-centered evaluation" means the metrics are designed to help humans achieve their various needs (*e.g.,*, better user experience, higher-quality writing artifact, etc.)

[2] Note that we do not aim to build enumerated lists of evaluation metrics. Instead, we focus on introducing the motivation and process of creating these evaluation aspects and subgroups, which can be generalized in broader use.

6 ACKNOWLEDGEMENT

We especially **thank** the insightful feedback and valuable support from **Dr. Ting-Hao 'Kenneth' Huang**!

REFERENCES

[1] Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6470–6484. https://doi.org/10.18653/v1/2020.emnlp-main.525

[2] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2287–2295.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[4] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[5] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating Human-Language Model Interaction. *arXiv preprint arXiv:2212.09746* (2022).

[6] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958* (2022).

[7] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[8] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[9] Jeba Rezwana and Mary Lou Maher. 2022. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* (2022).

[10] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. 2022. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction* (2022).

[11] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

A APPENDIX

A.1 `Parachute` Metric Details

We list the practical evaluation metrics of `Parachute` with details (*e.g.*, measure questions, references) in Table 1.

Interaction Aspects	Eval Types	Subgroups	Measure Questions	Reference
❶ Evaluating Human--LM Interaction	**Perception of LM as Partner [General perceptions]**	enjoyment	*I enjoyed writing the story.*	[6, 11]
		effort	*I put lots of effort into getting AI suggestions.*	[10]
		preference	*I prefer the suggestions from the AI agent.*	[6]
		communicative	*I can communicate well with the AI agent.*	[10]
		cognitive load	*Interacting with the AI agent requires much cognitive load.*	[10]
		collaborative	*I feel the AI agent collaborative to work together.*	[6, 10, 11]
		ease	*The AI agent is easy to learn and work with.*	[6, 11]
	Perception of LM as Partner [Ethical Metrics]	understanding	*I can understand the AI agent.*	[10]
		literal	*The AI agent's suggestions are literal*	[6]
		stereotypical	The AI agent's suggestions are stereotypical	[6]
		non-intentional	The AI agent does not have intentions to generate suggstions.	[10]
	Measure LM Feedbacks [Common Metrics]	coherent	*The AI generations are coherent with prompts.*	[6, 10]
		variety	*The AI agent's suggestion are various.*	[6, 10]
		helpful	*I found the AI agent helpful.*	[6, 10, 11]
		effective	*The AI agent is effective at suggesting ideas.*	[11]
		repetition	*The AI agent generates repetitive suggestions.*	[10]
		usability	*The AI agent is useful for my writing.*	[10]
		combinatorial	*I feel the AI agent combines a broad set of information.*	[10]
	Measure LM Feedbacks [Task-oriented Metrics]	uniqueness	*The AI's suggestions are unique.*	[6, 10, 11]
		reality	emphThe AI's suggestion aligns with common sense.	[6, 10]
		novelty	*The AI agent often generates unexpected suggestions.*	[10]
		freedom	*I feel the AI agent can express freely.*	[10]
		structural	*The AI suggestions are structural.*	[6]
		imagination	*I feel the AI agent has much imagination.*	[10]
		unexpected	*The AI suggestions are often unexpected to me.*	[6, 10]
❷ Evaluating Dynamic Interaction Trace	**Iterative Interaction Change [diff from prior knowledge]**	different from initial expectation	*what's the difference before the initial expression.*	[10]
		adjust to prior expectation	*I adjusted my expectation to prior ones.*	[10]
	Iterative Interaction Change [diff from prev. interact. states]	dynamics of suggestion integration	*how does the suggestion integration change dynamically.*	[10]
	Span Interactive Flow	learning curve	*I can learn to use this system quickly.*	[10]
		consistency	*The AI generate consistent suggestions along interaction.*	[6, 10]
		flow and ordering	*the flow and ordering of co-writing are smooth.*	[6, 10]
	Efficiency	latency	*The elapsed time from human request to AI response.*	[4]
		incorporation rate	*The rate of incorporating AI suggestions.*	[4]
		request count	*The count of human requests.*	[4]
		time considering suggestions	*The average time for human to consider AI suggestions.*	[4]
		#accepted suggestions	*The count of accepted AI suggestions.*	[4]
		time to complete	*The elapsed time for human to complete the task.*	[4]
❸ Evaluating Writing Artifact	**Measure Writing Artifacts [With Ground Truth]**	word edit distance	*The word edit distance between prior- and post- articles.*	[6]
		lemma-based Jaccard similarity	*The similarity of ground truth and outcome article.*	[6]
		document length difference	*The difference between prior- and post- articles.*	[6]
	Measure Writing Artifacts [No Ground Truth]	outcome creativity	*The article I wrote with AI is creative.*	[6, 10]
		quality	*The outcome article is high-quality.*	[5]
		external expert evaluation	*The external experts assess the writing artifacts.*	[5]
		word count	*The total words count of the outcome article.*	[4]
		open-source repetition scorer	*Computing repetition score using exisint tools.*	[6]
	Perception from Writer [Satisfaction]	writing goal	*The outcome article reaches my writing goal.*	[4, 6]
		pride	*I'm proud of the final article.*	[4, 6]
		satisfied	*I feel satisfied with the final article.*	[6]
	Perception from Writer [Ownership]	ownership	*I feel ownership over the final article.*	[6, 10, 11]
		authorial discretion	*I can decide what/how to put the AI suggestions into the article.*	[10]

Table 1: The subgroup metrics for each evaluation aspect in `Parachute`.(`yellow` and `blue` indicates objective and subjective metrics, respectively.)