Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision

Wanyu Du[1]*, **Zae Myung Kim**[2]* , **Vipul Raheja**[3] , **Dhruv Kumar**[3] , **Dongyeop Kang**[2]
[1]University of Virginia, [2]University of Minnesota, [3]Grammarly
wd5jq@virginia.edu, {kim01756,dongyeop}@umn.edu
{vipul.raheja,dhruv.kumar}@grammarly.com

Abstract

Revision is an essential part of the human writing process. It tends to be strategic, adaptive, and, more importantly, *iterative* in nature. Despite the success of large language models on text revision tasks, they are limited to non-iterative, one-shot revisions. Examining and evaluating the capability of large language models for making continuous revisions and collaborating with human writers is a critical step towards building effective writing assistants. In this work, we present a human-in-the-loop iterative text revision system, $\mathcal{R}$ead, $\mathcal{R}$evise, $\mathcal{R}$epeat ($\mathcal{R}3$), which aims at achieving high quality text revisions with minimal human efforts by reading model-generated revisions and user feedbacks, revising documents, and repeating human-machine interactions. In $\mathcal{R}3$, a text revision model provides text editing suggestions for human writers, who can accept or reject the suggested edits. The accepted edits are then incorporated into the model for the next iteration of document revision. Writers can therefore revise documents iteratively by interacting with the system and simply accepting/rejecting its suggested edits until the text revision model stops making further revisions or reaches a predefined maximum number of revisions. Empirical experiments show that $\mathcal{R}3$ can generate revisions with comparable acceptance rate to human writers at early revision depths, and the human-machine interaction can get higher quality revisions with fewer iterations and edits. The collected human-model interaction dataset and system code are available at https://github.com/vipulraheja/IteraTeR. Our system demonstration is available at https://youtu.be/lK08tIpEoaE.

1 Introduction

Text revision is a crucial part of writing. Specifically, text revision involves identifying discrepancies between intended and instantiated text, deciding what edits to make, and how to make those desired edits (Flower and Hayes, 1981; Faigley and Witte, 1981; Fitzgerald, 1987). It enables writers to deliberate over and organize their thoughts, find a better line of argument, learn afresh, and discover what was not known before (Sommers, 1980; Scardamalia, 1986). Previous studies (Flower, 1980; Collins and Gentner, 1980; Vaughan and McDonald, 1986) have shown that text revision is an *iterative* process since human writers are unable to simultaneously comprehend multiple demands and constraints of the task when producing well-written texts – for instance, covering the content, following linguistic norms and discourse conventions of written prose, etc. Therefore, writers resort to performing text revisions on their drafts iteratively to

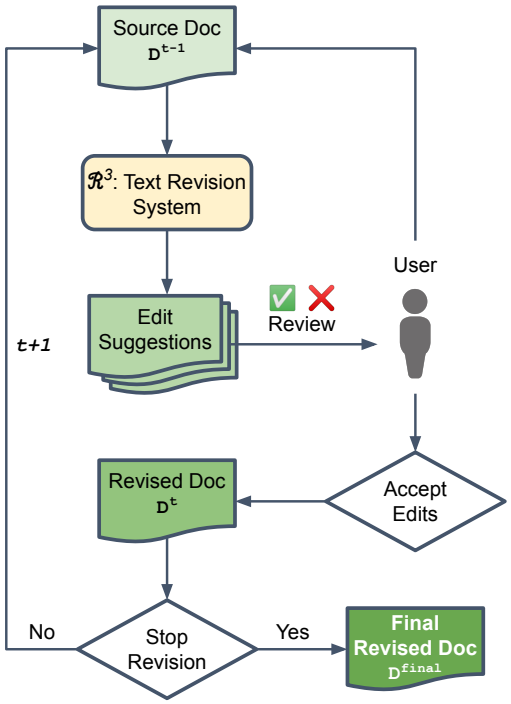


Figure 1: System overview for $\mathcal{R}3$ human-in-the-loop iterative text revision.

*Equal contributions.

Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022), pages 96 - 108

reduce the number of considerations at each time.

Computational modeling of the iterative text revision process is essential for building intelligent and interactive writing assistants. Most prior works on the development of neural text revision systems (Faruqui et al., 2018; Botha et al., 2018; Ito et al., 2019; Faltings et al., 2021) do not take the iterative nature of text revision and human feedback on suggested revisions into consideration. The direct application of such revision systems in an iterative way, however, could generate some "noisy" edits and require much burden on human writers to fix the noise. Therefore, we propose to collect human feedback at each iteration of revision to filter out those harmful noisy edits and produce revised documents of higher quality.

In this work, we present a novel human-in-the-loop iterative text revision system, $\mathcal{R}$ead, $\mathcal{R}$evise, $\mathcal{R}$epeat ($\mathcal{R}3$), which reads model-generated revisions and user feedbacks, revises documents, and repeats human-machine interactions in an iterative way, as depicted in Figure 1. First, users write a document as input to the system or choose one from a candidate document set to edit. Then, the text revision system provides multiple editing suggestions with their edits and intents. Users can accept or reject the editing suggestions in an iterative way and stop revision when no editing suggestions are provided or the model reaches the maximum revision limit. The overall model performance can be estimated by calculating the acceptance rate throughout all editing suggestions.

$\mathcal{R}3$ provides numerous benefits over existing writing assistants for text revision. First, $\mathcal{R}3$ improves the overall writing experience for writers by making it more interpretable, controllable, and productive: on the one hand, writers don't have to (re-)read the parts of the text that are already high quality, and this, in turn, helps them focus on larger writing goals (§4.2); on the other hand, by showing edit intentions for every suggested edit, which users can further decide to accept or reject, $\mathcal{R}3$ provides them with more fine-grained control over the text revision process compared to other one-shot based text revision systems (Lee et al., 2022), and are limited in both interpretability and controllability. Second, $\mathcal{R}3$ improves the revision efficiency. The human-machine interaction can help the system produce higher quality revisions with fewer iterations and edits, and the empirical experiments in §4.2 validate this claim. To the best of our knowledge, $\mathcal{R}3$ is the first text revision system in literature that can perform *iterative* text revision in collaboration by human writers and revision models.

In this paper, we make three major contributions:

- We present a novel human-in-the-loop text revision system $\mathcal{R}3$ to make text revision models more accessible; and to make the process of iterative text revision efficient, productive, and cognitively less challenging.
- From an HCI perspective, we conduct experiments to measure the effectiveness of the proposed system for the iterative text revision task. Empirical experiments show that $\mathcal{R}3$ can generate edits with comparable acceptance rate to human writers at early revision depths.
- We analyze the data collected from human-model interactions for text revision and provide insights and future directions for building high-quality and efficient human-in-the-loop text revision systems. We release our code, revision interface, and collected human-model interaction dataset to promote future research on collaborative text revision.

2 Related Work

Previous works on modeling text revision (Faruqui et al., 2018; Botha et al., 2018; Ito et al., 2019; Faltings et al., 2021) have ignored the iterative nature of the task, and simplified it into a one-shot "original-to-final" sentence-to-sentence generation task. However, in practice, at every revision step, multiple edits happen at the document-level which also play an important role in text revision. For instance, reordering and deleting sentences to improve the coherence.

More importantly, performing multiple high-quality edits at once is very challenging. Continuing the previous example, document readability can degrade after reordering sentences, and further adding transitional phrases is often required to make the document more coherent and readable. Therefore, one-shot sentence-to-sentence text revision formulation is not sufficient to deal with real-world challenges in text revision tasks.

While some prior works on text revision (Coenen et al., 2021; Padmakumar and He, 2021; Gero et al., 2021; Lee et al., 2022) have proposed human-machine collaborative writing interfaces, they are

mostly focused on collecting human-machine interaction data for training better neural models, rather than understanding the iterative nature of the text revision process, or the model's ability to adjust editing suggestions according to human feedback.

Another line of work by Sun et al. (2021); Singh et al. (2022) on creative writing designed human-machine interaction interfaces to encourage new content generation. However, text revision focuses on improving the quality of existing writing and keeping the original content as much as possible. In this work, we provide a human-in-the-loop text revision system to make helpful editing suggestions by interacting with users in an iterative way.

3 System Overview

Figure 1 shows the general pipeline of $\mathcal{R}3$ human-in-the-loop iterative text revision system. In this section, we will describe the development details of the text revision models and demonstrate our user interfaces.

We first formulate an iterative text revision process: given a source document[1] $\mathcal{D}^{t-1}$, at each revision depth t, a text revision system will apply a set of edits to get the revised document $\mathcal{D}^t$. The system will continue iterating revision until the revised document $\mathcal{D}^t$ satisfies a set of predefined stopping criteria, such as reaching a predefined maximum revision depth t_{max}, or making no edits between $\mathcal{D}^{t-1}$ and $\mathcal{D}^t$.

3.1 Text Revision System

We follow the prior work of Du et al. (2022) to build our text revision system. The system is composed of edit intention identification models and a text revision generation model. We follow the same data collection procedure in Du et al. (2022) to collect the iterative revision data.[2] Then, we train the three models on the collected revision dataset.

Edit Intention Identification Models. Following Du et al. (2022), our edit intentions have four categories: FLUENCY, COHERENCE, CLARITY, and STYLE. We build our edit intention identification models at each sentence of the source document $\mathcal{D}^{t-1}$ to capture the more fine-grained edits. Specifically, given a source sentence, the system will make two-step predictions: (1) whether or not to edit, and (2) which edit intention to apply. The decision whether or not to edit is taken by an edit-prediction classifier that predicts a binary label of whether to edit a sentence or not. The second model, called the edit-intention classifier, predicts which edit intention to apply to the sentence. If the edit-prediction model predicts "not to edit" in the first step, the source sentence will be kept unchanged at the current revision depth.

Text Revision Generation Model. We fine-tune a large pre-trained language model like PEGASUS (Zhang et al., 2020) on our collected revision dataset to build the text revision generation model. Given a source sentence and its predicted edit intention, the model will generate a revised sentence, conditioned on the predicted edit intention. Then, we concatenate all un-revised and revised sentences to get the model-revised document $\mathcal{D}^t$, and extract all its edits using *latexdiff*[3] and *difflib*.[4]

In summary, at each revision depth t, given a source document $\mathcal{D}^{t-1}$, the text revision system first predicts the need for revising a sentence, and for the ones that need revision, it predicts the corresponding fine-grained edit intentions – thus, generating the revised document $\mathcal{D}^t$ based on the source document and the predicted edit decisions and intentions.

3.2 Human-in-the-loop Revision

In practice, not all model-generated edits are equally impactful towards improving the document quality (Du et al., 2022). Therefore, we enable user interaction in the iterative text revision process to achieve high quality of text revisions along with a productive writing experience. At each revision depth t, our system provides the user with suggested edits, and their corresponding edit intentions. The user can interact with the system by choosing to accept or reject the suggested edits.

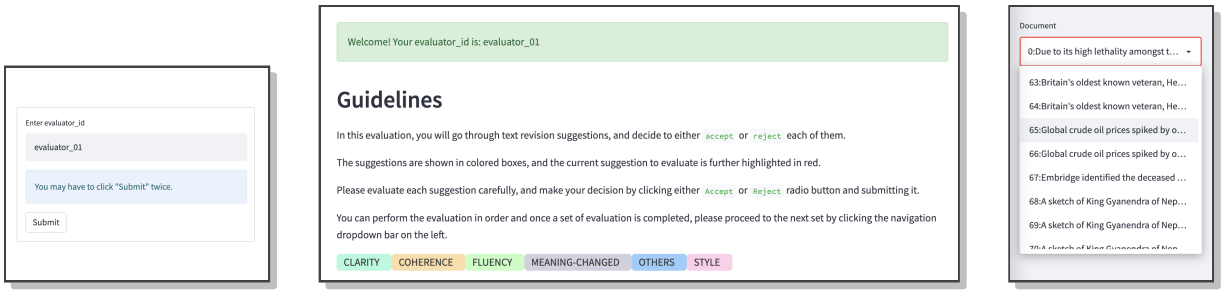
Figure 2 illustrates the details of $\mathcal{R}3$'s user interface. First, a user enters their id to login to the web interface as shown in Figure 2a. Then, the user is instructed with a few guidelines on how to operate the revision as demonstrated in Figure 2b. After getting familiar with the interface, the user can select a source document from the left drop-down menu in Figure 2c. By clicking the source document, all the edits predicted by the text re-

[1] The source document can be chosen by a user in the candidate set of documents or written from scratch by a user.

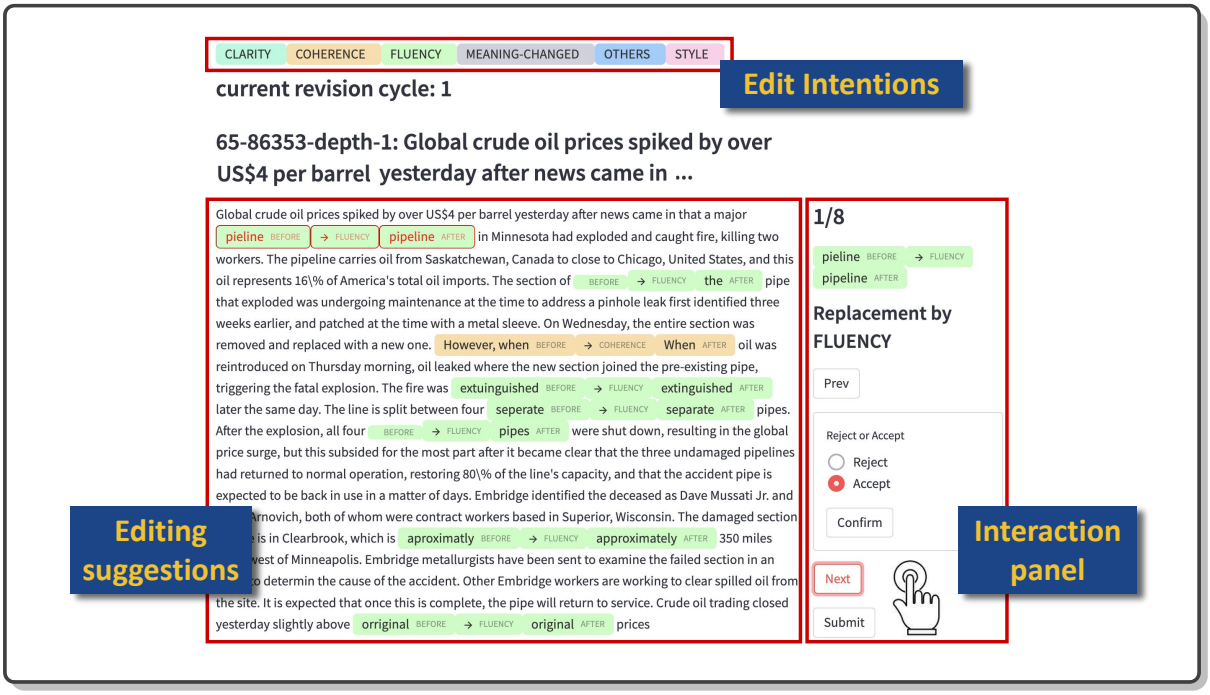
[2] See §4.1 for the detailed data collection.

[3] https://ctan.org/pkg/latexdiff

[4] https://docs.python.org/3/library/difflib.html



(a) Login (b) Read guidelines (c) Select doc



(d) Editing suggestions and interaction panel

Figure 2: User interface demonstration for $\mathcal{R}3$. Anonymized version available at https://youtu.be/lK08tIpEoaE.

vision model, as well as their corresponding edit intentions will show up in the main page as illustrated in Figure 2d (left panel). The user is guided to go through each suggested edits, and choose to accept or reject the current edit by clicking the *Confirm* button in Figure 2d (right panel). After going through all the suggested edits, the user is guided to click the *Submit* button to save their decisions on the edits. Then, the user is guided to click the *Next Iteration!* button to proceed to the next revision depth and check the next round of edits suggested by the system. This interactive process continues until the system does not generate further edits or reaches the maximum revision depth t_{max}.

4 Experiments

We conduct experiments to answer the following research questions:

- RQ1 How likely are users to accept the editing suggestions predicted by our text revision system? This question is designed to evaluate whether our text revision system can generate high quality edits.
- RQ2 Which types of edit intentions are more likely to be accepted by users? This question is aimed to identify which types of edits are more favored by users.
- RQ3 Does user feedback in $\mathcal{R}3$ help produce higher quality of revised documents? This question is proposed to validate the effectiveness of human-in-the-loop component in $\mathcal{R}3$.

4.1 Experimental Setups

Iterative Revision Systems. We prepare three types of iterative revision systems to answer the above questions:

1. HUMAN-HUMAN: We ask users to accept or reject text revisions made by human writers, which are directly sampled from our collected iterative revision dataset. This serves as the baseline to measure the gap between our text revision system and human writers.
2. SYSTEM-HUMAN: We ask users to accept or reject text revisions made by our system. Then, we incorporate user accepted edits to the system to generate the next iteration of revision. This is the standard human-in-the-loop process of $\mathcal{R}3$.
3. SYSTEM-ONLY: We conduct an ablation study by removing user interaction in reviewing the model-generated edits. Then, we compare the overall quality of final revised documents with and without the human-in-the-loop component.

In both HUMAN-HUMAN and SYSTEM-HUMAN setups where users interacted with the system, they were not informed whether the revisions were sampled from our collected iterative revision dataset, or generated by the underlying text revision models.

User Study Design. We hired three linguistic experts (English L1, bachelor's or higher degree in Linguistics) to interact with our text revision system. Each user was presented with a text revision (as shown in Figure 2d) and asked to accept or reject each edit in the current revision (users were informed which revision depth they were looking at). For a fair comparison, users were not informed about the source of the edits (human-written vs. model-generated), and the experiments were conducted separately one after the other. Note that the users were only asked to accept or reject edits, and they had control neither over the number of iterations, nor over the stopping criteria. The stopping criteria for the experiment were set by us and designed as: (1) no new edits were made at the following revision depth, or (2) the maximum revision depth $t_{max} = 3$ was reached.

Data Details. We followed the prior work (Du et al., 2022) to collect the text revision data across three domains: ArXiv, Wikipedia and Wikinews. This data was then used to train both the edit intention identification models and the text revision generation model. We split the data into training, validation and test set according to their document ids with a ratio of 8:1:1. The detailed data statistics are included in Table 1. Note that our newly collected revision dataset is larger than the previously proposed dataset in Du et al. (2022) with around 24K more unique documents and 170K more edits (sentence pairs).

	# Docs	Avg. Depths	# Edits
Training	44,270	6.63	292,929
Validation	5,152	6.60	34,026
Test	6,226	6.34	39,511

Table 1: Statistics for our collected revision data which has been used to train the edit intention identification model and the text revision generation model. **# Docs** means the total number of unique documents, **Avg. Depths** indicates the average revision depth per document (for the human-generated training data), and **# Edits** stands for the total number of edits (sentence pairs) across the corpus.

For the human evaluation data, we randomly sampled 10 documents with a maximum revision depth of 3 from each domain in the test set in Table 1. For the evaluation of text revisions made by human writers (HUMAN-HUMAN), we presented the existing ground-truth references from our collected dataset to users. Since we do not hire additional human writers to perform continuous revisions, we just presented the static human revisions from the original test set to users at each revision depth, and collected the user acceptance statistics as a baseline for our system.

For the evaluation of text revisions made by our system (SYSTEM-HUMAN), we only presented the original source document at the initial revision depth ($\mathcal{D}^0$) to our system, and let the system generate edits in the following revision depths, while incorporating the accept/reject decisions on model-generated edit suggestions by the users. Note that at each revision depth, the system will only incorporate the edits accepted by users and pass them to the next revision iteration.

For text revisions made by our system without human-in-the-loop (SYSTEM-ONLY), we let the system generate edits in an iterative way and accepted all model-generated edits at each revision depth.

Model Details. For both edit intention identification models, we fine-tuned the RoBERTa-large (Liu et al., 2020) pre-trained checkpoint from HuggingFace (Wolf et al., 2020) for 2 epochs with a learning rate of 1×10^{-5} and batch size of 16. The edit-

	HUMAN-HUMAN				SYSTEM-HUMAN (ours)			
t	**# Docs**	**Avg. Edits**	**Avg. Accepts**	**% Accepts**	**# Docs**	**Avg. Edits**	**Avg. Accepts**	**% Accepts**
1	30	5.37	2.77	51.66	30	**5.90**	**2.90**	49.15
2	30	4.83	3.00	62.06	24	3.83	2.57	**67.02**
3	20	3.80	2.67	70.39	20	3.43	1.94	56.71

Table 2: Human-in-the-loop iterative text revision evaluation results. t stands for the revision depth, **# Docs** shows the total number of revised documents at the current revision depth, **Avg. Edits** indicates the average number of applied edits per document, **Avg. Accepts** means the average number of edits accepted by users per document, and **% Accepts** is calculated by dividing the total accepted edits with the total applied edits.

prediction classifier is binary classification model that predicts whether to edit a given sentence or not. It achieves an F1 score of 67.33 for the edit label and 79.67 for the not-edit label. The edit-intention classifier predicts the specific intent for a sentence that requires editing. It achieves F1 scores of 67.14, 70.27, 57.0, and 3.21[5] for CLARITY, FLUENCY, COHERENCE and STYLE intent labels respectively.

For the text revision generation model, we fine-tuned the PEGASUS-LARGE (Zhang et al., 2020) pre-trained checkpoint from HuggingFace. We set the edit intentions as new special tokens (e.g., <STYLE>, <FLUENCY>), and concatenated the edit intention and source sentence together as the input to the model. The output of the model is the revised sentence, and we trained the model with cross-entropy loss. We fine-tuned the model for 5 epochs with a learning rate of 3×10^{-5} and batch size of 4. Finally, our text revision generation model achieves 41.78 SARI score (Xu et al., 2016), 81.11 BLEU score (Papineni et al., 2002) and 89.08 ROUGE-L score (Lin, 2004) on the test set.

4.2 Result Analysis

Iterativeness. The human-in-the-loop iterative text revision evaluation results are reported in Table 2. Each document is evaluated by at least 2 users. **We find that $\mathcal{R}3$ achieves comparable performances with ground-truth human revisions at revision depth 1 and 2, and tends to generate less favorable edits at revision depth 3.** At revision depth 1, $\mathcal{R}3$ is able to generate more edits than ground-truth human edits for each document, and gets more edits accepted by users on average. This shows the potential of $\mathcal{R}3$ in generating appropriate text revisions that are more favorable to users.

At revision depth 2, while $\mathcal{R}3$ generates less edits than human writers on average, it gets a higher acceptance rate than human writers. This result suggests that for the end users, more edits may not necessarily lead to a higher acceptance ratio, and shows that $\mathcal{R}3$ is able to make high-quality edits for effective iterative text revisions. At revision depth 3, $\mathcal{R}3$ generates even less edits compared both to human writers and its previous revision depths. This result can be attributed to the fact that our models are only trained on static human revision data, while at testing time they have to make predictions conditioned on their revisions generated at the previous depth, which may have a very different distribution of edits than the training data. Table 7 shows an example of iterative text revision in ArXiv domain generated by $\mathcal{R}3$. We also provide some other iterative revision examples generated by $\mathcal{R}3$ in Appendix A.

Edit Intentions. Table 3 demonstrates the distribution of different edit intentions, which can help us further analyze the which type of edits are more likely to be accepted by end users. For human-generated revisions, we find that FLUENCY edits are most likely to be accepted since they are mainly fixing grammatical errors.

For system-generated revisions, we observe that CLARITY edits are the most frequent edits but end users only accept 58.73% of them, which suggests that our system needs further improvements in learning CLARITY edits. Another interesting observation is that STYLE edits are rarely generated by human writers (1.2%) and also gets the lowest acceptance rate (33.33%) than other intentions, while they are frequently generated by our system (16.7%) and surprisingly gets the highest acceptance rate (64.6%) than other intentions. This observation indicates that $\mathcal{R}3$ is capable for generating favorable stylistic edits. Table 4 shows some examples of edit suggestions generated by $\mathcal{R}3$.

Role of Human Feedback in Revision Quality. Table 5 illustrates the quality comparison results of

[5]We note that the F1 score for STYLE is low as the number of training samples for that intent is particularly small.

	HUMAN-HUMAN			SYSTEM-HUMAN (ours)		
	# Edits	**# Accepts**	**% Accepts**	**# Edits**	**# Accepts**	**% Accepts**
CLARITY	197	119	60.40	**332**	195	58.73
FLUENCY	178	146	**82.02**	91	41	45.05
COHERENCE	103	41	39.80	141	68	48.22
STYLE	6	2	33.33	113	73	**64.60**

Table 3: The distribution of different edit intentions. **# Edits** indicates the total number of applied edits under the current edit intention, **# Accepts** means the total number of edits accepted by users under the current edit intention, and **% Accepts** is calculated by dividing the total accepted edits with the total applied edits.

Edit Intention	**Edit Suggestion**
CLARITY	Emerging new test procedures ~~, such as antigen or RT-LAMP tests,~~ might enable us to protect nursing home residents.
FLUENCY	For Radar tracking, we show how a model can reduce the tracking errors.
COHERENCE	However, ~~we show that~~ even a small violation can significantly modify the effective noise.
STYLE	There has been ~~numerous~~extensive research focusing on neural coding.

Table 4: Edit suggestion examples generated by $\mathcal{R}3$.

	Avg. Depths	**# Edits**	**Quality**
SYSTEM-HUMAN (ours)	2.5	148	**0.68**
SYSTEM-ONLY	2.8	175	0.28

Table 5: Quality comparison results of final revised documents with and without human-in-the-loop. **Avg. Depths** indicates the average number of iterations conducted by the system, **# Edits** means the total number of accepted edits by the system, and **Quality** represents the human judgements of the overall quality of system-revised final documents.

final revised documents with and without human-in-the-loop for $\mathcal{R}3$. We asked another group of three annotators (English L2, bachelor's or higher degree in Computer Science) to judge whether the overall quality of system-generated final document is better than the ground-truth reference final document. The quality score ranges between 0 and 1. We evaluated 10 unique documents in ArXiv domain, and took the average score from all 3 annotators. As shown in Table 5, **SYSTEM-HUMAN produces better overall quality score for the final system-generated documents with fewer iterations of revision and fewer edits**, which validates the effectiveness of the human-machine interaction proposed in $\mathcal{R}3$.

User Feedback. We also collected qualitative feedback about $\mathcal{R}3$ from the linguistic experts through a questionnaire. The first part of our questionnaire asks participants to recall their experience with the system, and evaluate various aspects of the system (in Table 6). They were asked to rate how easy it was to get onboarded and use the system (*convenience*), whether they were satisfied with the system (revision quality and usage experience) (*satisfaction*), whether they felt it improved their productivity for text revision (*productivity*), and whether they would like to use the system again (*retention*) for performing revisions on their documents.

In general, the users gave positive feedback towards the ease of use of the system. However, they were neutral on the potential productivity impact, owing to the lack of domain knowledge of the documents they were evaluating. This issue could be mitigated by asking users to revise their own documents of interest. The retention and satisfaction scores were leaning slightly negative, which was explained as primarily attributed to gaps in the user interface design (eg. improperly aligned diffs, suboptimal presentation of word-level edits, etc.).

We also asked them to provide detailed comments on their experience, and the potential impact of the system on their text revision experience. Specifically, upon asking the users whether using the system to evaluate the model-suggested edits would be more time-efficient compared to actually revising the document themselves, we received many useful insights that help better design better interfaces and features of our system in future work, as some users noted:

> *I think it would be faster using the system, but I would still be checking the text myself in case edits were missed. The system made some edits where there were letters and parts of words being added/re-*

Criterion	Avg. Score	Std. Deviation
Convenience	3.66	0.58
Satisfaction	2.33	0.58
Productivity	3.00	1.00
Retention	2.66	0.58

Table 6: User feedback survey ratings. Ratings are on 5-point Likert scale with 5 being strongly positive experience, 3 being neutral, and 1 being strongly negative. However, we'd like to point out that as the number of users (linguists) who participated in the study is small, the statistical significance of the results should be taken lightly.

> *moved/replaced, which sometimes took some time to figure out. That wouldn't be the case if I were editing a document.*
>
> *Ultimately, I would use the system for grammar/coherence/clarity edits, and then still research (a lot) to ensure that meaning was preserved throughout the document. For topics that I was more familiar with/more general topics, using the system would probably reduce my time by a third or so. For topics that required more in-depth research for me, the time saved by using the system might be minimal.*

5 Discussion and Future Directions

When $\mathcal{R}3$ generates revisions at deeper depths, we observe a decrease in the acceptance ratio by human users. It is crucial to create a text revision system that can learn different revision strategies at each iteration and generate high quality edits at deeper revision levels.

Editing suggestions provided by our text revision generation models could be improved. Particularly, FLUENCY edits show a huge gap between human and system revisions (45.05% and 82.02%). Future work could focus on developing more powerful text revision generation models.

In our human-machine interaction, we restrict the users' role to accept or reject the model's predictions. Even with minimal human interaction, our experiment shows comparable or even better revision quality as compared to human writers at early revision depths. A potential future direction for human-machine collaborative text revision would be to develop advanced human-machine interaction interfaces, such as asking users to re-write the machine-revised text.

Also, a larger-scale user study could be carried out to derive more meaningful statistics (e.g. optimal number of revision depths and edit suggestions) and investigate if there is any intriguing user behavior in the iterative revision process. For example, as mentioned in the users' feedback, it would be interesting to check if users behave differently when they are asked to accept/reject edit suggestions provided for their own texts as opposed to the texts written by a third party.

6 Conclusion

In this work, we develop an interactive iterative text revision system $\mathcal{R}3$ that is able to effectively assist users to make revisions and improve the quality of existing documents. $\mathcal{R}3$ can generate higher quality revisions while minimizing the human efforts. Users are provided with a reviewing interface to accept or reject system suggesting edits. The user-validated edits are then propagated to the next revision depth to get further improved revisions. Empirical results show that $\mathcal{R}3$ can generate iterative text revisions with acceptance rates comparable or even better than human writers at early revision depths.

Acknowledgments

We thank all linguistic expert annotators at Grammarly for participating in the user study and providing us with valuable feedback during the process. We also thank Karin de Langis at University of Minnesota for narrating the video of our system demonstration. We would like to extend our gratitude to the anonymous reviewers for their helpful comments.

References

Jan A. Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from Wikipedia edit history. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 732–737, Brussels, Belgium. Association for Computational Linguistics.

Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a human-ai collaborative editor for story writing. *arXiv preprint arXiv:2107.07430*.

Allan Collins and Dedre Gentner. 1980. A framework for a cognitive theory of writing. In *Cognitive processes in writing*, pages 51–72. Erlbaum.

t	HUMAN-HUMAN	SYSTEM-HUMAN (ours)
0	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 storm. Emerging new test procedures , such as antigen or RT-LAMP tests, might enable us to protect nursing home residents by means of preventive screening strategies. Here, we develop a novel agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategiesto curb this spread . The model is microscopically calibrated to high-resolution data from actual nursing homes in Austria, including the detailed networks of social contacts of their residents and information on past outbreaks.	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 storm. Emerging new test procedures , such as antigen or RT-LAMP tests, might enable us to protect nursing home residents by means of preventive screening strategies. Here, we develop a novel agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategiesto curb this spread . The model is microscopically calibrated to high-resolution data from actual nursing homes in Austria, including the detailed networks of social contacts of their residents and information on past outbreaks.
1	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 storm. ~~Emerging new~~ With test procedures becoming available at scale , such as antigen or RT-LAMP tests, might enable us to protect nursing home residents by means of preventive screening strategies. Here, we develop a novel agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal ~~preventive testing strategiesto curb this spread~~ prevention strategies . The model is microscopically calibrated to high-resolution data from actual nursing homes in Austria, including the detailed networks of social contacts of their residents and information on past outbreaks.	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 storm. Emerging new test procedures ~~, such as antigen or RT-LAMP tests,~~ might enable us to protect nursing home residents by means of preventive screening strategies. Here, we develop a novel agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategies ~~to curb this spread . The model is microscopically.~~ The model is calibrated to high-resolution data from actual nursing homes in Austria, including the detailed networks of social contacts of their residents and information on past outbreaks.
2	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 ~~storm~~ pandemic . Emerging new test procedures , such as antigen or RT-LAMP tests, might enable us to protect nursing home residents by means of preventive screening strategies. Here, we develop a ~~novel~~ detailed agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategiesto curb this spread . The model is microscopically calibrated to high-resolution data from actual nursing homes in Austria, including ~~the detailed networks of social contacts of their resident~~ detailed social contact networks and information on past outbreaks.	~~Due to its high lethality amongst the elderly, n~~ N ursing homes are in the eye of the COVID-19 storm. Emerging new test procedures might enable us to protect nursing home residents by means of preventive screening ~~strategies~~ . Here, we develop a novel agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategies. The model is calibrated to high-resolution data from actual nursing homes in Austria, including ~~the~~ detailed networks of social contacts of their residents and information on past outbreaks.
3	-	Due to its high lethality amongst the elderly, nursing homes are in the eye of the COVID-19 storm. Emerging new test procedures might enable us to protect nursing home residents by means of preventive screening. Here, we develop a ~~novel~~ n agent-based epidemiological model for the spread of SARS-CoV-2 in nursing homes to identify optimal preventive testing strategies. The model is calibrated to high-resolution data from actual nursing homes in Austria, including detailed networks of social contacts of ~~their~~ residents and information on past outbreaks.

Table 7: A sample snippet of iterative text revisions in ArXiv domain generated by $\mathcal{R}3$, where t is the revision depth and $t = 0$ indicates the original input text. Note that ~~text~~ represents user accepted deletions, text represents user accepted insertions, and text represents user rejected edits.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding Iterative Revision from Human-Written Text. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College composition and communication*, 32(4):400–414.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Jill Fitzgerald. 1987. Research on revision in writing. *Review of Educational Research*, 57(4):481–506.

Linda Flower. 1980. The dynamics of composing: Making plans and juggling constraints. *Cognitive processes in writing*, pages 31–50.

Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Katy Ilonka Gero, Vivian Liu, and Lydia B Chilton. 2021. Sparks: Inspiration for science writing using language models. *arXiv preprint arXiv:2110.07640*.

Takumi Ito, Tatsuki Kuribayashi, Hayato Kobayashi, Ana Brassard, Masato Hagiwara, Jun Suzuki, and Kentaro Inui. 2019. Diamonds in the rough: Generating fluent sentences from early-stage drafts for academic writing assistance. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 40–53, Tokyo, Japan. Association for Computational Linguistics.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. *arXiv preprint arXiv:2201.06796*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach. In *International Conference on Learning Representations*.

Vishakh Padmakumar and He He. 2021. Machine-in-the-loop rewriting for creative image captioning. *arXiv preprint arXiv:2111.04193*.

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.

M. Scardamalia. 1986. Research on written composition. *Handbook of reserch on teaching*.

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 Trans. Comput.-Hum. Interact.*

Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College composition and communication*, 31(4):378–388.

Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and Mohit Iyyer. 2021. IGA: An intent-guided authoring assistant. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5972–5985, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marie M. Vaughan and David D. McDonald. 1986. A model of revision in natural language generation. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 90–96, New York, New York, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

A $\mathcal{R}3$ Iterative Revision Samples

We present more iterative revision examples generated by $\mathcal{R}3$ in Table 8 and Table 9.

t	HUMAN-HUMAN	SYSTEM-HUMAN(ours)
0	Corporal Nathan Hornburg. A Reserve soldier serving with Canadian Forces in Afghanistanwas killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old Corporal Nathan Hornburg of Calgary, Alberta. A Canadian Forces statement said Cpl. Hornburg was killed during Operation Sadiq Sarbaaz (Honest Soldier) approximately 47 kilometres west of Kandahar City in Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.	Corporal Nathan Hornburg. A Reserve soldier serving with Canadian Forces in Afghanistanwas killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old Corporal Nathan Hornburg of Calgary, Alberta. A Canadian Forces statement said Cpl. Hornburg was killed during Operation Sadiq Sarbaaz (Honest Soldier) approximately 47 kilometres west of Kandahar City in Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.
1	~~Corporal Nathan Hornburg. A Reserve soldier serving with Canadian Forces in Afghanistanwas killed on September 24, 2007~~ On MOnday, a 24-year old Calgary Reservist became the 71st Canadian soldier killed in Afghanistan . Four others were injured in the incident which killed ~~24-year-old~~ Corporal Nathan Hornburg ~~of Calgary, Alberta. A Canadian Forces statement said Cpl. Hornbur was killed during~~ , who was operating as part of Operation Sadiq Sarbaaz (Honest Soldier) approximately 47 kilometres west of Kandahar City in Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time ~~(12:00 UTC)~~ while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.	~~Corporal Nathan Hornburg. A Reserve~~ A Canadian soldier serving with Canadian Forces in Afghanistanwas killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old Corporal Nathan Hornburg of Calgary, Alberta. ~~A Canadian Forces statement said Cpl.~~ Nathan Hornburg was killed during Operation Sadiq Sarbaaz (Honest Soldier) approximately 47 kilometres west of Kandahar City in Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.
2	Corporal Nathan Hornburg. A Reserve soldier serving with Canadian Forces in Afghanistanwas killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old Corporal Nathan Hornburg of Calgary, Alberta. A Canadian Forces statement said Cpl. Hornburg was killed during Operation Sadiq Sarbaaz ~~(Honest Soldier) approximately 47 kilometres west of Kandahar City in Panjwaii District~~ , a joint Afghan-NATO mission designed to "set the conditions for a continuous security presence and the establishment of a new police sub-station in the northern part of (Panjwaii)." . Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.	A Canadian soldier serving with the Canadian Forces in Afghanistanwas killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old Corporal Nathan Hornburg of Calgary, Alberta. Nathan Hornburg was killed during Operation Sadiq Sarbaaz (Honest Soldier) , approximately 47 kilometres west of Kandahar City in Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.
3	-	A Canadian soldier serving with the Canadian Forces in Afghanistan was killed on September 24, 2007. Four others were injured in the incident which killed 24-year-old ~~Corporal~~ Cpl. Nathan Hornburg of Calgary, Alberta. Nathan Hornburg was killed during Operation Sadiq Sarbaaz (Honest Soldier), approximately 47 kilometres west of Kandahar City in the Panjwaii District. Media reports indicated he died from mortar fire at around 4 :30 p.m. local time (12:00 UTC) while he was repairing the track on a Canadian Leopard tank near a cluster of villages known as Zangabad.

Table 8: A sample snippet of iterative text revisions in Wikinews domain generated by $\mathcal{R}3$, where t is the revision depth and $t = 0$ indicates the original input text. Note that ~~text~~ represents user accepted deletions, text represents user accepted insertions, and text represents user rejected edits.

t	HUMAN-HUMAN	SYSTEM-HUMAN(ours)
0	Jecon Gregory is or was a nomadic artist, whose autobiographical fragments and poems, dictated to an acquaintance, were published as the book "History of a Nation of One" (Harcourt Brace, New York, 1969, and Michael Joseph, London, 1971). Jecon apparently did not know his place, date, language or even name of birth, began his wanderings as a child in Malta; walked through many lands, barefoot, tall and thin, pulling all his possessions in a basket on wheels, sleeping on the ground, and making a living by drawing portraits.	Jecon Gregory is or was a nomadic artist, whose autobiographical fragments and poems, dictated to an acquaintance, were published as the book "History of a Nation of One" (Harcourt Brace, New York, 1969, and Michael Joseph, London, 1971). Jecon apparently did not know his place, date, language or even name of birth, began his wanderings as a child in Malta; walked through many lands, barefoot, tall and thin, pulling all his possessions in a basket on wheels, sleeping on the ground, and making a living by drawing portraits.
1	Jecon Gregory is or was a nomadic artist, whose autobiographical fragments and poems, dictated to an acquaintance, were published as the book "History of a Nation of One : An Unlikely Memoir " (Harcourt Brace, New York, 1969, and Michael Joseph, London, 1971). .. Jecon apparently did not know his place, date, language or even name of birth, began his wanderings as a child in Malta; walked through many lands, barefoot, tall and thin, pulling all his possessions in a basket on wheels, sleeping on the ground, and making a living by drawing portraits.	Jecon Gregory is or was a nomadic artist, whose autobiographical fragments and poems, dictated to an acquaintance, were published as the book " History of a Nation of One" (Harcourt Brace, New York, 1969, and Michael Joseph, London, 1971). Jecon apparently did not know his place, date, language or even name of birth, began his wanderings as a child in Malta; walked through many lands, barefoot, tall and thin, pulling all his possessions in a basket on wheels, sleeping on the ground, and ~~making a living by~~ drawing portraits.
2	-	Jecon Gregory is or was a nomadic artist, whose autobiographical fragments and poems, dictated to an acquaintance, were published as the book "History of a Nation of One" (Harcourt Brace, New York, 1969, and Michael Joseph, London, 1971). Jecon apparently did not know his place, date, language or even name of birth, began his wanderings as a child in Malta; walked through many lands, barefoot, tall and thin, pulling all his possessions in a basket on wheels, sleeping on the ground, and ~~making a living by~~ drawing portraits.
3	-	-

Table 9: A sample snippet of iterative text revisions in Wikipedia domain generated by $\mathcal{R}3$, where t is the revision depth and $t = 0$ indicates the original input text. Note that ~~text~~ represents user accepted deletions, text represents user accepted insertions, and text represents user rejected edits.