ParaRev: Building a dataset for Scientific Paragraph Revision annotated with revision instruction

Léane Jourdan and Nicolas Hernandez and Richard Dufour
Nantes Université, École Centrale Nantes,
CNRS, LS2N, UMR 6004, F-44000 Nantes, France
firstname.lastname@univ-nantes.fr

Florian Boudin
JFLI, CNRS, Nantes University, France
florian.boudin@univ-nantes.fr

Akiko Aizawa
National Institute of Informatics, Japan
aizawa@nii.ac.jp

Abstract

Revision is a crucial step in scientific writing, where authors refine their work to improve clarity, structure, and academic quality. Existing approaches to automated writing assistance often focus on sentence-level revisions, which fail to capture the broader context needed for effective modification. In this paper, we explore the impact of shifting from sentence-level to paragraph-level scope for the task of scientific text revision. The paragraph level definition of the task allows for more meaningful changes, and is guided by detailed revision instructions rather than general ones. To support this task, we introduce ParaRev, the first dataset of revised scientific paragraphs with an evaluation subset manually annotated with revision instructions. Our experiments demonstrate that using detailed instructions significantly improves the quality of automated revisions compared to general approaches, no matter the model or the metric considered.

1 Introduction

In the scientific domain, writing assistance is crucial as researchers share their findings through articles published in conferences or journals. However, writing articles is challenging and time-consuming, notably for non-native English speakers or young researchers (Amano et al., 2023).

The field of writing assistance has grown rapidly to address these challenges leading to the development of various tools (Grammarly, Trinka AI[1], ...) and specialized workshops (In2Writing, WRAICOGS[2]).

[1]https://www.grammarly.com/, https://www.trinka.ai/
[2]https://in2writing.glitch.me/, https://sites.google.com/view/wraicogs1

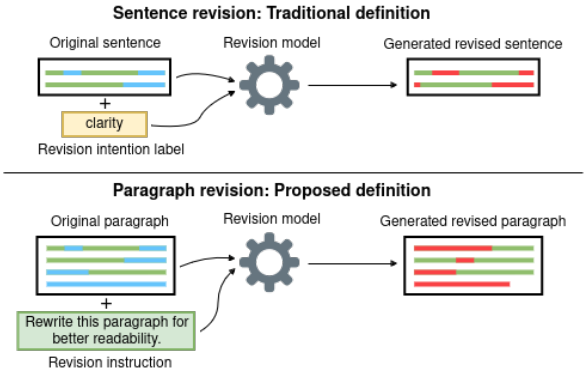

Figure 1: Definitions of the traditional sentence revision task and the proposed paragraph revision task.

The goal of writing assistance is to support researchers throughout the writing process, which includes four steps: Prewriting, Drafting, Revising, and Editing (Jourdan et al., 2023). This paper focuses on the revision task where an input text is substantially modified for clarity, simplicity, style, and other aspects (Du et al., 2022a; Li et al., 2022). Since poor writing quality undermines the communication of research findings and often leads to paper rejection (Amano et al., 2023), effective revision is a critical step in scientific writing.

Due to past limitations in processing long texts, prior research has focused on the sentence revision task (see Figure 1). In this task, a sentence is given to a seq2seq model or a Large Language Model (LLM) along with a general revision prompt, which could take the form of a label (e.g., Coherence, Style) (Du et al., 2022b; Jiang et al., 2022) or a general instruction (Raheja et al., 2023). In this definition of the task, labels are assigned to specific modifications within a sentence, targeting particular spans of text to revise.

Proceedings of the First Workshop on Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WRAICOGS 2025), pages 35–44
January 20, 2025.

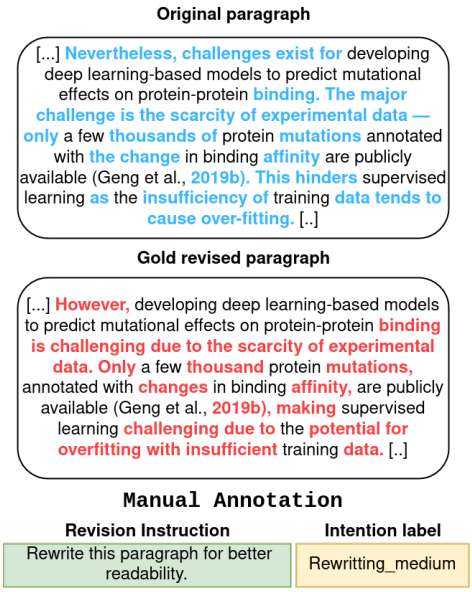

Figure 2: Example of a revised paragraph with its associated revision instruction and label.

Thanks to the recent advances in NLP in the past years, we propose to expand the traditional scope of this sentence-level paradigm to detailed personalised instructions guiding the model on revisions to conduct at the paragraph level, as illustrated in Figure 1.

We argue that this new paradigm aligns better with how human writers revise the text and how LLMs are used today, allowing more comprehensive changes such as merging, splitting, or reorganizing sentences. Additionally, personalised instructions enable more nuanced control over the degree of revision, specifying whether minor edits or major restructuring is required. They can also target specific areas within a paragraph, while other sentences provide essential context.

To support this task, we introduce `ParaRev`, a corpus of paragraphs revised by their authors annotated with human revision intention labels and instructions (e.g. in Figure 2). Our contributions are as follows:

1. We proposed a definition of the text revision task at paragraph-level, with personalised revision instructions.

2. We release a high-quality corpus of 48k revised paragraphs with an evaluation subset of 641 manually annotated paragraphs, facilitating future research in this area [3].

[3] https://huggingface.co/datasets/taln-ls2n/pararev

2 Related work

Existing corpora for scientific text revision provide aligned versions of revised texts, with varying scope. Some datasets focus only on the abstract and introduction sections of scientific papers (Du et al., 2022b; Mita et al., 2024; Ito et al., 2019), while others include full-length articles (Kuznetsov et al., 2022; Jiang et al., 2022; D'Arcy et al., 2023; Jourdan et al., 2024). Most of these resources align revisions at the sentence level, though paragraph-level reconstruction is possible to capture broader, more substantial revisions.

However, not all datasets include revision annotations with explicit intention labels. Some, such as those designed for tasks related to peer-review (Kuznetsov et al., 2022; D'Arcy et al., 2023), focus on tracking changes without offering structured guidance for the revision process. In revision tasks, having an explicit revision intention is crucial for guiding models in performing meaningful modifications. In sentence-level revision datasets, individual modifications (i.e. spans of text) are commonly associated with a label indicating the revision intention. The taxonomies for these labels can vary across corpora (Jiang et al., 2022; Du et al., 2022b). However, labels associated with short spans of text often lack the contextual information needed for more substantial, long-range revisions. They also do not provide the specificity that detailed instructions could offer to guide more precise edits.

Recent efforts have attempted to bridge this gap by converting labels into general instructions to better align with how LLMs are utilized for revision (Raheja et al., 2023). Our work aims to extend this approach by introducing detailed, personalized paragraph-level instructions that provide richer contextual and precise guidance for revisions.

3 Dataset construction

Figure 3 summarizes the overall data pipeline described in this section.

3.1 Paragraph Selection and Extraction

Our dataset consists of pairs of revised paragraphs extracted from the CASIMIR corpus (Jourdan et al., 2024), a large resource containing revised scientific articles aligned at sentence level. This corpus provides paragraph-level IDs for each sentence, which allows us to treat paragraphs as coherent

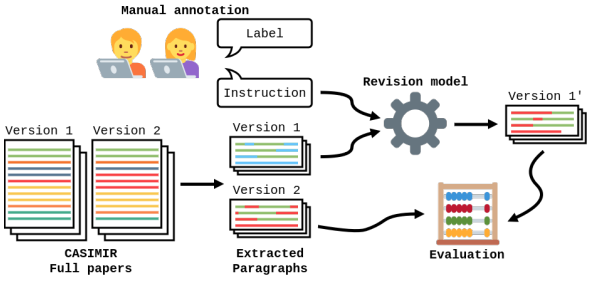

Figure 3: The data pipeline: annotation, paragraph revision and evaluation

units marked by changes in paragraph IDs across both versions of the text.

However, many articles in CASIMIR contain identical or minimally revised content, which is not suitable for our purpose. We aim to build a high-quality dataset by selecting paragraphs with substantial revisions (beyond minor grammatical fixes) while preserving the original idea of the text.

To achieve this, we developed hand-crafted heuristics through empirical observations of a subset of the corpus, to retain only the sufficiently revised paragraphs (see Appendix A). From the original 1 889 810 paragraph pairs with at least one modification, we kept after this selection process 48 203 paragraphs. Extraction code is openly available [4].

3.2 Paragraph revision taxonomy

To align with prior research and facilitate analysis or example selection for few-shot tasks, we chose to assign revision intention labels to each paragraph pair. Motivated by the works of Du et al. (2022b) and Jiang et al. (2022), we propose a new paragraph-level taxonomy based on their existing sentence-level ones and observations done on a subset of our dataset.

In this taxonomy, we identified nine revision intentions, defined in Appendix B: *Rewriting (light, medium, heavy), Concision, Development, Content (addition, substitution, deletion)* and *Unusable*. These labels are not associated with individual edits: they instead represent the overall revision intention for the paragraph. Each paragraph can receive up to two labels, as multiple revisions with different intentions may occur within a single paragraph.

3.3 Instructions

An instruction is provided only when no new information is introduced in the revised paragraph, as revision models are only supposed to improve existing text and not make up new content. Labels are used to identify the paragraphs that do not require an annotation, i.e. the paragraphs annotated with *Development*, *Content Addition*, or *Content Substitution*.

Annotators are asked to write concise, simple instructions as they would when guiding an LLM to revise the first version of the paragraph into the second. Detailed lists of changes are not allowed. They must also indicate the position and intensity of revisions when necessary, especially when only part of the paragraph requires revision while the rest provides context.

Some examples of instructions and their associated pair of paragraphs are available in Appendix C.

3.4 Annotation

The annotation process involved 10 annotators (2 professors, 3 PhD students, and 5 master's students), all not native from English and specialized in the NLP domain and experienced in reading and writing academic papers. Most paragraphs (73.32%) were double annotated.

Since annotators could assign up to two labels, with 1.2 labels on average per paragraph per annotator, we used Krippendorffs alpha for agreement. It often occurs that some revisions are on the line of two categories, e.g., *Rewriting light* and *medium*. Given this ambiguity, we computed two scores: one for the strict taxonomy (agreement of 0.499) and another for broader super-labels, i.e. merging similar categories (agreement of 0.693), see Appendix D. Agreement with super-labels exceeds the 0.67 threshold for tentative conclusions about the consistency of the annotations (Krippendorff, 2018).

Additionally, 75.32% of paragraphs share at least one label between annotators with strict taxonomy, rising to 95.11% using super-labels.

Those results reflect the inherent complexity of the annotation task.

4 Dataset Statistics

The dataset contains 48 203 paragraph pairs from 16 664 pairs of revised articles. From this total 48K paragraphs, 641 were manually annotated (470 were double annotated). This subset was chosen to represent the overall corpus based on paper revision extent: 218 paragraphs are from heavily revised pa-

[4] https://github.com/JourdanL/pararev

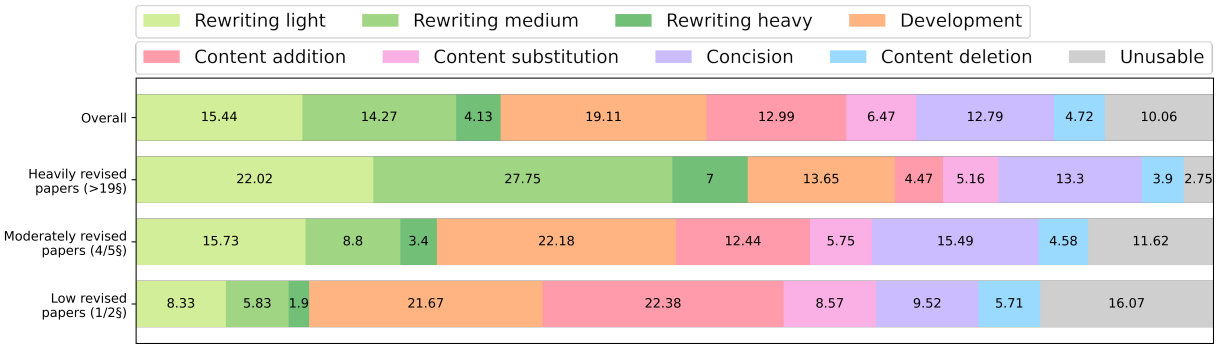

Figure 4: Distribution of labels across the dataset overall and degree of modification of the articles.

pers (where over 19 paragraphs are revised), 213 from moderately revised papers (4-5 revised paragraphs), 210 from low revised papers (1-2 revised paragraphs).

Figure 4 shows the label distribution across the dataset. For fairness in the analysis, when annotators picked two labels, they were weighted 0.5 each. Additionally, paragraphs with only one annotation are counted twice.

The figure distinguishes between paragraphs from articles with different degrees of revision. Heavily revised papers tend to mainly feature *Rewriting* revisions, suggesting that the entire document was evenly reworked. In contrast, low-revised papers are more likely to involve small content modifications, such as adding or removing forgotten information.

Finally, we report the instructions' distribution as follows: of the 641 annotated paragraphs, 328 have no instruction, 55 have one, and 258 have two. These 258 paragraphs form our evaluation set in Section 5.

5 Impact of task definition on revision

To verify our hypothesis that using detailed instructions better guides the revision process compared to generic instruction labels, we conducted a comparative experiment. For this, we evaluated how different models performed when given either a general prompt mapped from an intention label or a personalised instruction tailored to the specific changes needed (see Appendix E).

We experimented with multiple models to ensure the results were robust across various architectures: **CoEdit**[5], a T5-based model fine-tuned on sentence revision task (Raheja et al., 2023), as well as **Llama3**[6]**, Mistral**[7]**, and GPT-4o,** state-of-the-art foundation models with strong language understanding and generation capabilities. All models are used in zero-shot, the prompt used is given in Appendix E.

Additionally, as a control baseline, we included a **CopyInput** method, which does not apply any edits to the input paragraph.

To assess the quality of revisions, we employed traditional sentence revision metrics, ROUGE-L (Lin, 2004) and SARI (Xu et al., 2016), alongside Bertscore (Zhang et al., 2020) to measure similarity between the generated and gold revised paragraphs. The results are summarized in Table 1.

Across all models, we observed consistent improvements when using detailed instructions over general prompts. They are even statistically significant for Mistral, Llama3, and GPT-4o, with p-values below 0.05 (paired Student's t-test).

The experiment confirms our hypothesis: instructions that provide specific revision guidance allow the models to produce more accurate revisions compared to relying solely on general labels.

However, when examining the performances of the models, we observe that the CopyInput and Coedit achieve the best results. A manual overview of a subset of outputs reveals that Co-edit only suggests minor changes, such as grammar corrections, while other models propose more substantial modifications.

Evaluation remains a significant challenge in the text revision domain, as widely used metrics compare the proposed revision to a single reference version. This approach penalizes revisions that deviate from the gold standard, even if they result in valid improvements. Consequently, unless the

[5]https://huggingface.co/grammarly/coedit-xl

[6]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[7]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Metric	ROUGE-L		SARI		Bertscore	
Approach	Label	Instruction	Label	Instruction	Label	Instruction
CopyInput- no edits	78.49		60.69		95.98	
coedit-xl	67.50	67.70	39.56	39.68	93.88	93.93
Mistral-7B-Instruct-v0.2	45.70	48.23†	28.47	30.43†	91.38	91.78†
Meta-Llama-3-8B-Instruct	50.37	55.73†	30.59	35.07†	91.84	92.68†
GPT4o	57.99	66.17†	33.33	41.39†	92.89	94.11†
Average gain	+4.07		+3.66		+0.75	

Table 1: Results on the paragraph revision task. Symbol † marks a significative improvement.

model's modifications exactly replicate those made by the original author, the score will be lower than proposing no modifications (CopyInput). This limitation need to be address in future work to develop more robust and reliable evaluation methods for this task.

6 Conclusion

We proposed a definition of the scientific text revision task at paragraph-level, enabling more context-aware revisions using full-length instruction. Additionally, we presented `ParaRev`, a dataset of revised paragraphs, with an evaluation split annotated with revision instructions. Our experiments demonstrate that providing detailed personalised instructions leads to more effective revisions than general ones, across multiple models.

In future work, as manual annotation is costly and time-consuming, we aim to annotate the remaining non-annotated wide split of the dataset automatically. This silver dataset will then be used to fine-tune an open-source model specifically for paragraph-level revision tasks.

7 Limitations

The primary limitation of this work is the size of the evaluation subset, as it was manually annotated by volunteer researchers whose availability constrained the number of annotations. A larger annotated subset would enhance the reliability of our evaluation, allowing us to determine if smaller improvements in revision scores are statistically significant.

While the core focus of this study is on introducing personalized annotated instructions, we also labelled paragraphs with revision intention labels. Labelling revisions is a challenging task since multiple modifications can occur within a single paragraph, and annotators may interpret boundaries between similar categories differently. However, this limitation can be mitigated in practice by using super-labels or considering the union of the two annotations.

8 Ethical Considerations

Data availability All the data are extracted from the CASIMIR corpus, collected from OpenReview where all articles fall under different "non-exclusive, perpetual, and royalty-free license" [8].

Computational resources Our experiments with revision models ran CoEdit on a local GPU for approximately two hours, while Mistral and Llama ran for nine hours on the supercomputer Jean Zay, emitting less than 0.001 tons of CO_2, with an additional 3.16$ spent on GPT API credits.

Use of revision models We release this dataset to support future research on writing assistance for researchers. We believe that revision models based on LLMs should be used as tools to enhance clarity and structure, not to generate the primary content and analysis.

Acknowledgments

We thank Jiahao Huang, Xanh Ho, Juan Junqueras, Ken Kim, Jonas Luhrs, Julian Schnitzler and Tomás Vergara Browne for their participation in annotating the dataset.

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013901R1 made by GENCI.

References

Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, et al. 2023. The manifold

[8] https://openreview.net/legal/terms

costs of being a non-native english speaker in science. *PLoS Biology*, 21(7):e3002184.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *Preprint*, arXiv:2306.12587.

Wanyu Du, Zae Myung Kim, Vipul Runderstandaheja, Dhruv Kumar, and Dongyeop Kang. 2022a. Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 96–108, Dublin, Ireland. Association for Computational Linguistics.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022b. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

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.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arxivedits: Understanding the human revision process in scientific writing. *In Proceedings of EMNLP 2022*.

Léane Jourdan, Florian Boudin, Nicolas Hernandez, and Richard Dufour. 2024. CASIMIR: A corpus of scientific articles enhanced with multiple author-integrated revisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2883–2892, Torino, Italia. ELRA and ICCL.

Léane Jourdan, Florian Boudin, Richard Dufour, and Nicolas Hernandez. 2023. Text revision in scientific writing assistance: A review. In *13th International Workshop on Bibliometric-enhanced Information Retrieval (BIR)*, number 3617 in CEUR Workshop Proceedings, pages 22–36, Aachen.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986.

Jingjing Li, Zichao Li, Tao Ge, Irwin King, and Michael Lyu. 2022. Text revision by on-the-fly representation optimization. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, pages 58–59, Dublin, Ireland. 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.

Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2024. Towards automated document revision: Grammatical error correction, fluency edits, and beyond. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 251–265, Mexico City, Mexico. Association for Computational Linguistics.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. 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.

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*.

A Paragraph selection criteria

We keep only paragraphs that met the following requirements: Criteria for selection (threshold obtained empirically):

- **Size:** The longer version must at least be 250 characters
- **Percentage of modification:**
  - The most edited sentence should be at least modified at 25%
  - The whole paragraph should be at least edited at 10%
  - In a paragraph, the set of sentences modified at more than 90% should not represent more than 40% or 200 characters in the whole paragraph
  - If a paragraph does not contain sentences revised at more than 50%: The set of modified sentences should be modified at least by 20%
- **Quantity of transcribed equations:** The quantity of transcribed equations captured by regular expression should not represent more than 9% of the set of modified sentences in the paragraph.

- If the paragraph starts with a modification: We check that it is not a segmentation mistake
  - Is the beginning of the sentences correctly formed.
  - If only one sentence was completely added or deleted: Accepted if it is only tags
  - If the sentence is revised at more than 50%
    - Refused if the shorter version is equal to the end of the longer one
    - Refused if the longer version is more than 3 times the length of the shorter one
  - If the sentence is revised at less than 50%
    - If the modification is at the beginning on both sides: Refused if the shorter version is equal to the end of the longer one
    - If the modification is at the beginning on one side: Refused if the modification is longer than 10 characters (without spaces and tags)
- If the paragraph ends with a modification: We check that it is not a segmentation mistake
  - Is the end of the sentences correctly formed
  - If only one sentence was completely added or deleted: Always rejected. A second version of the function exists to include cases where a full correctly formed sentence is deleted/added, resulting in 11k additional paragraphs in the corpus.
  - If the sentence is revised at more than 50%
    - Refused if the shorter version is equal to the beginning of the longer one
    - Refused if the longer version is more than 3 times the length of the shorter one
  - If the sentence is revised at less than 50%: Always accepted
- Check if a part of the text has not been transformed into a tag during PDF conversion

B Paragraph revision taxonomy

See Table 2

C Examples of instructions

See Table 3.

D Super-labels mapping

In our taxonomy, boundaries between categories may be ambiguous, allowing for interpretation and discussion. Given this ambiguity, we defined super-labels that encompass categories of revision where similar actions are taken in Table 4. For example, the limit between *Rewriting light* and *Rewriting medium* or *Content addition* and *Development* can be blurry, and they totalise 59.43% of complete disagreements (disagreement where there is no overlap between the two sets of labels). However, both opinions from annotators can be justified in discussions, as some paragraphs can be on the line of the two definitions.

E Prompting

To work with the different models for revision, we use the following prompt (**Bold blue text** correspond to the input data, the instruction and the paragraph to revise):

```
You are a writing assistant specialised in academic writing. Your task is to revise the paragraph from a research paper draft that will be given according to the user's instructions. Please answer only by "Revised paragraph: <revised_version_of_the_paragraph>"
instruction : original_paragraph
```

For the comparative evaluation, based on the work of (Raheja et al., 2023), the labels are mapped to general instructions, given in Table 5.

Type		Description
Rewriting	Light	Minor changes in word choice or phrasing.
	Medium	Complete rephrasing of sentences within the paragraph.
	Heavy	Significant rephrasing, affecting at least half of the paragraph.
Concision		Same idea, stated more briefly by removing unnecessary details.
Development		Same idea, expanded with additional details or definitions.
Content	Addition	Modification of content through the addition of a new idea.
	Substitution	Modification of content through the replacement of an idea or fact.
	Deletion	Modification of content through the deletion of an idea.
Unusable		Issues due to document processing errors (e.g., segmentation problems, misaligned paragraphs, or footnotes mixed with the text).

Table 2: Taxonomy of revisions at paragraph level

Type	**Instruction**
Parag source	Parag target
Rewriting_light	**Improve the english in the paragraph, make it slightly more formal.**
[...] Therefore, the generalization rapidly decreases after augmentationinterrupted when training with a single background because the learning direction toward generalization about various backgrounds is not helpful to train. On the other hand, the training can have helpwhen their difculty is solved by augmentation, such as Figure 2(b) and Figure 2(c). [...]	[...] Therefore, the generalization rapidly decreases after augmentation is interrupted during training with a single background because the learning direction toward generalization about various backgrounds is not helpful to train. In contrast, the training can help when their difficulty is solved by augmentation (Figure 2(b), 2(c)).[...]
Rewriting_medium	**Modify the logical flow of ideas to improve the readability of the paragraph.**
Patrick et al. proposed the Mouse Ether technique on finding out that when using multiple displays with different resolutions, a user loses the cursor because of unnatural cursor movement between displays [5]. The results showed that the technique improved [...]	Patrick et al. found out that a user loses the cursor when using multiple displays with different resolutions based on an unnatural cursor movement between displays, and proposed a Mouse Ether technique [5]. The proposed technique improved [...]
Rewriting_heavy	**Rewrite this paragraph to bring the argument through the idea that the goal is to learn a pixel-wise feature for semantic segmentation.**
[...] We consider propagating the labels from an annotated set to an unlabeled set by nearest neighbor search in the featurespace. We assume that semantic clustersemerge during training with sparse supervision, reinforced by aforementioned pixel-to-segment relationships. By propagating labels in the feature space, we reinforce the learning of semantic clusters.	[...] Our goal is to learn a pixel-wise feature that indicates semantic segmentation. It is thus reasonable to assume that pixels and segments of the same semantics form a cluster in the feature space, and we reinforce such clusters with a featural smoothness prior: We find nearest neighbours in the feature space and propagate labels accordingly.
Concision and Rewriting_light	**Combine sentences 3 and 4 into a really short one keeping only the main idea. Improve the choice of wording.**
[...] Our method seeks to best approximate some target distribution that is potentially multivariate, using some chosen set of control distributions. We provide an implementation which gives unique, interpretable weights in a setting of regular probability measures. For general probability measures, we construct our projection by first creating a regular tangent space through applying barycentric projection to optimal transport plans. Our application [...] demonstrates the methods efficiency and the necessity to have a method that is applicable for general proabbility measures. [...]	[...] Our method seeks to best approximate some general target measure using some chosen set of control measures. In particular, it provides a global (and in most cases unique) optimal solution. Our application [...] demonstrates the methods utility in allowing for a method that is applicable for general probability measures. [...]
Content_deletion and Concision	**Heavily remove details from this paragraph to make it more concise.**
[...] They should only contain the name of the medication. Their design should be such that the user can decide whether to add or remove them from the display. [...] On-calendar conflict representation should not be used as the main indication of an error after a rescheduling activity. The user should instead be notified of the impending conflict beforehand. Participants preferred that normal, dismissible error messages be displayed and show the full information regarding the conflicts being introduced by the action. [...]	[...] These summaries should only contain the name of the medication and users should be able to show or hide them. [...] The user should be notified of a newly created conflict upon rescheduling an entry, preferably via dismissible error messages that describe the conflict. [...]

Table 3: Examples of revised paragraph with their associated annotation. Colouration based on difflib output.

Super-label	Label
Rewriting	Rewriting Light
	Rewriting Medium
	Rewriting Heavy
Concision and Content Deletion	Concision
	Deletion
Development and Content Addition	Development
	Content Addition
	Content Substitution
Unusable	Unusable

Table 4: Mapping between super-labels and labels

Type		Description
Rewriting	Light	Improve the English of this paragraph
	Medium	Rewrite some sentences to make them more clear and easily readable
	Heavy	Rewrite and reorganize the paragraph for better readability
Concision		Make this paragraph shorter
Content	Deletion	Remove unnecessary details

Table 5: Mapping of labels with general instructions