January 20th 2025

Building a dataset for Scientific Paragraph Revision annotated with revision instruction

Léane Jourdan , Florian Boudin , Akiko Aizawa , Nicolas Hernandez , Richard Dufour

Contact: leane.jourdan@univ-nantes.fr

Writing Aids at the Crossroads of AI, Cognitive Science and NLP

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

NII National Institute of Informatics

LS2N Laboratoire des Sciences du Numérique de Nantes

Nantes Université

TALN

Context

Domain

- Scientific writing assistance

Motivations

- Writing an article is challenging
- Strong writing skills are essential
- Especially difficult for junior researchers and non-native English speakers

(Amano et al., 2023)

Context

Domain

- Scientific writing assistance

Motivations

- Writing an article is challenging
- Strong writing skills are essential
- Especially difficult for junior researchers and non-native English speakers

Tools

DeepL Write BETA

LinggleWrite

TRINKA by enago

Workshops

Workshop on Innovative Use of NLP for Building Educational Applications (BEA)

Acronym: BEA
Venue ID: bea

2024 • Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024) 59 papers

2023 • Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) 66 papers

shop on Innovative Use of NLP for Building Educational Applications (BEA 2022) 32 papers

shop on Innovative Use of NLP for Building Educ... 24 papers

Writing Aids at the Crossroads of AI, Cognitive Science and NLP

Abu-Dhabi, UAE, January 20, 2025

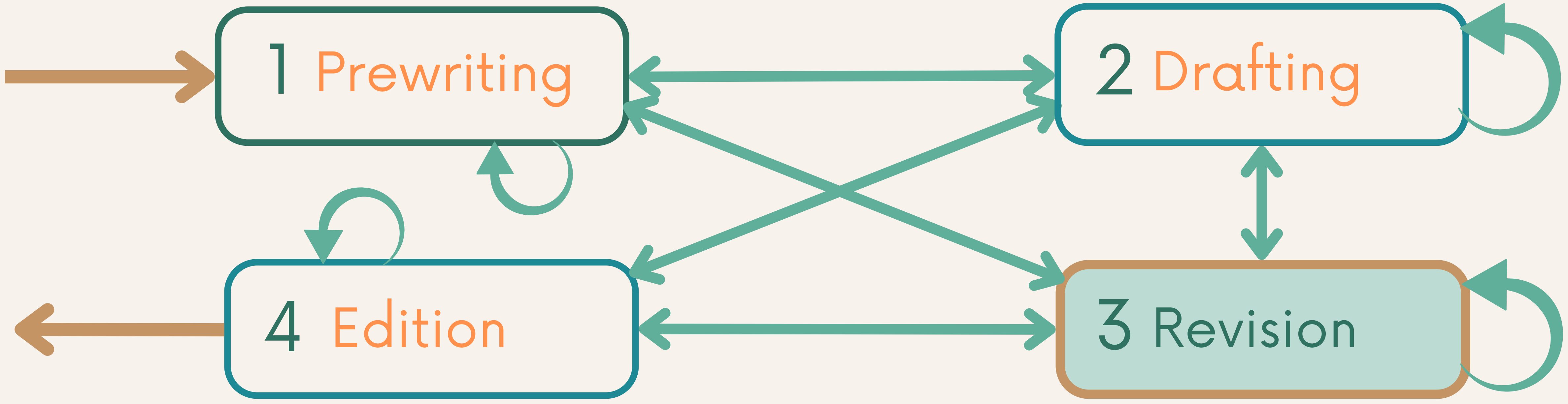
Organizers: Zock, M., Inui, K., & Yuan, Z.

Workshop co-located with COLING
https://coling2025.org

In2Writing

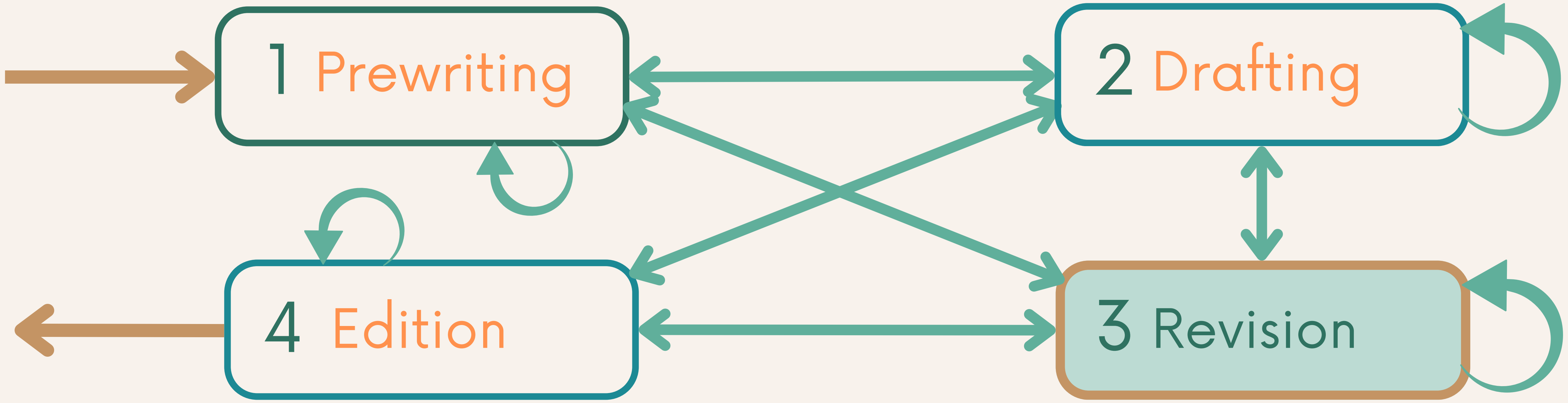
(Amano et al., 2023)   grammarly.com , deepl.com/fr/write , f.linggle.com , trinka.ai , aclanthology.org/venues/bea , sites.google.com/view/wraicogs1 , in2writing.glitch.me

Revision task



(Amano et al., 2023; Jourdan et al., 2023; Du et al., 2022a; Li et al., 2022)

Revision task



Definition

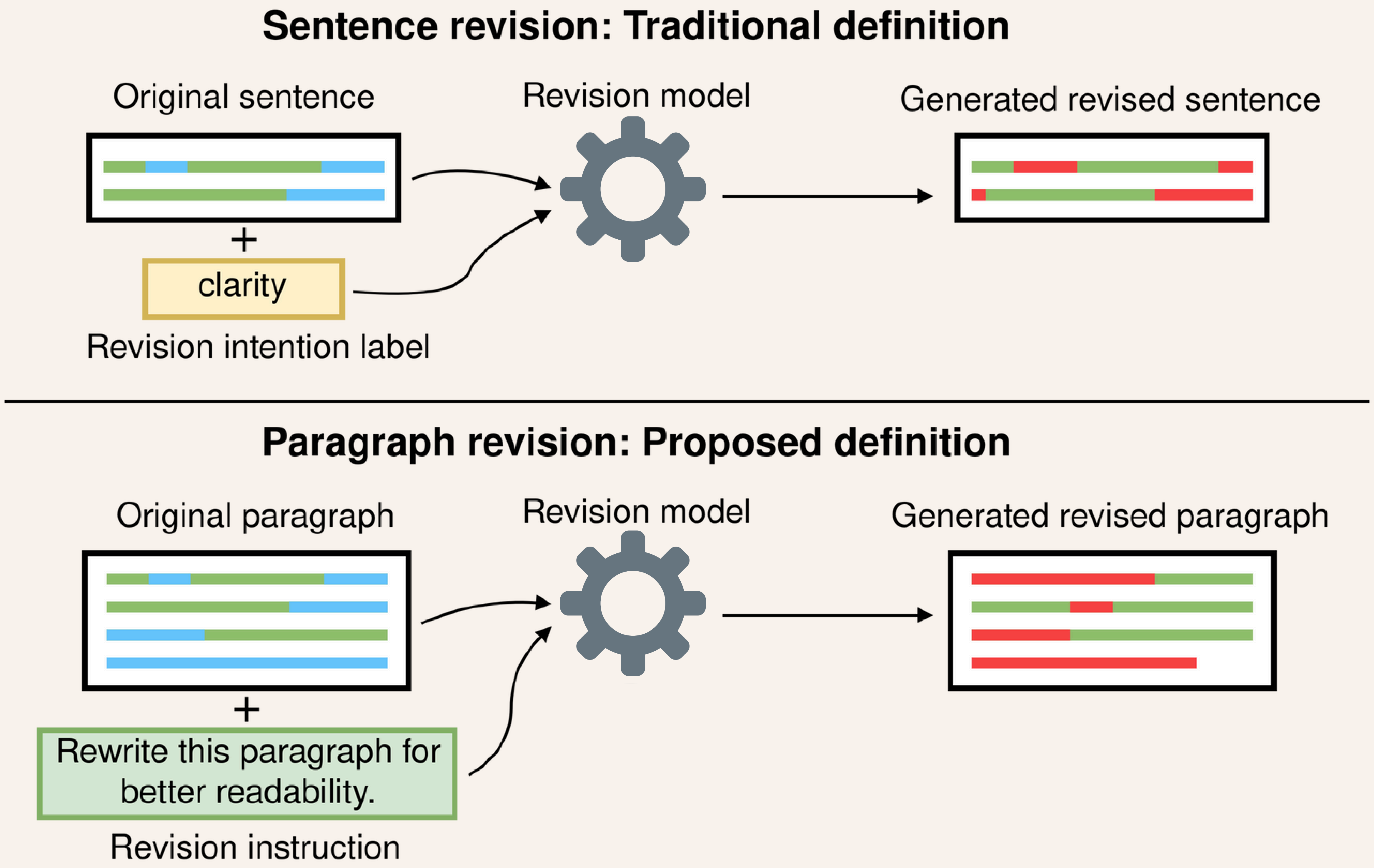
Text revision is the transformation of an input text into an improved version fitting a desired attribute (formality, clarity, etc.), closer to the intended text

Example

The model has good results. → Our model shows good results in this task. → Our model shows excellent performance in this task.

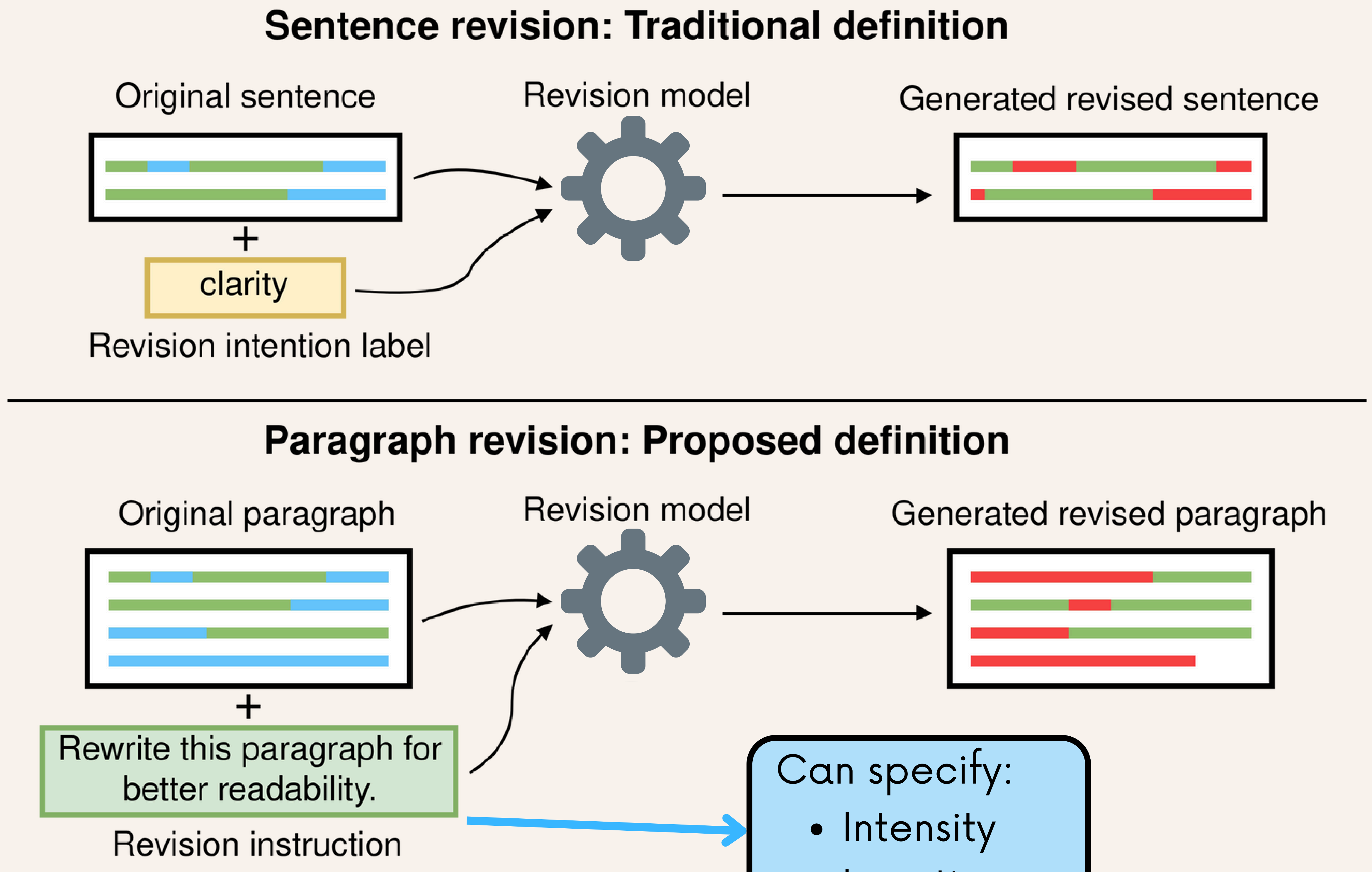
(Amano et al., 2023; Jourdan et al., 2023; Du et al., 2022a; Li et al., 2022; Ito et al. 2019)

Sentence vs paragraph revision tasks



(Du et al., 2022b; Jiang et al., 2022; Raheja et al., 2023)

Sentence vs paragraph revision tasks



Can specify:
- Intensity
- Location

(Du et al., 2022b; Jiang et al., 2022; Raheja et al., 2023)

Contributions

1. Definition of the text revision task at **paragraph-level**, with personalised **revision instructions**
2. **Pararev**, a corpus of **48k revised paragraphs** with an evaluation subset of **641 manually annotated** paragraphs

Original paragraph

[...] Nevertheless, challenges exist for developing deep learning-based models to predict mutational effects on protein-protein binding. The major challenge is the scarcity of experimental data — only a few thousands of protein mutations annotated with the change in binding affinity are publicly available (Geng et al., 2019b). This hinders supervised learning as the insufficiency of training data tends to cause over-fitting. [..]

Gold revised paragraph

[...] However, developing deep learning-based models to predict mutational effects on protein-protein binding is challenging due to the scarcity of experimental data. Only a few thousand protein mutations, annotated with changes in binding affinity, are publicly available (Geng et al., 2019b), making supervised learning challenging due to the potential for overfitting with insufficient training data. [..]

Manual Annotation

Revision Instruction	Intention label
Rewrite this paragraph for better readability.	Rewritting_medium

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓
Include revision intentions		✓	✓		✓	?	✓

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓
Include revision intentions		✓	✓		✓	?	✓

Focus on peer review

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓
Include revision intentions		✓	✓		✓	?	✓

Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D’Arcy et al. 2023; [7] Jourdan et al. 2024

Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓
Include revision intentions		✓	✓		✓	?	✓
Label scope		Span of text	Span of text		Span of text	Multi-sentences?	Span of text

Table - Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

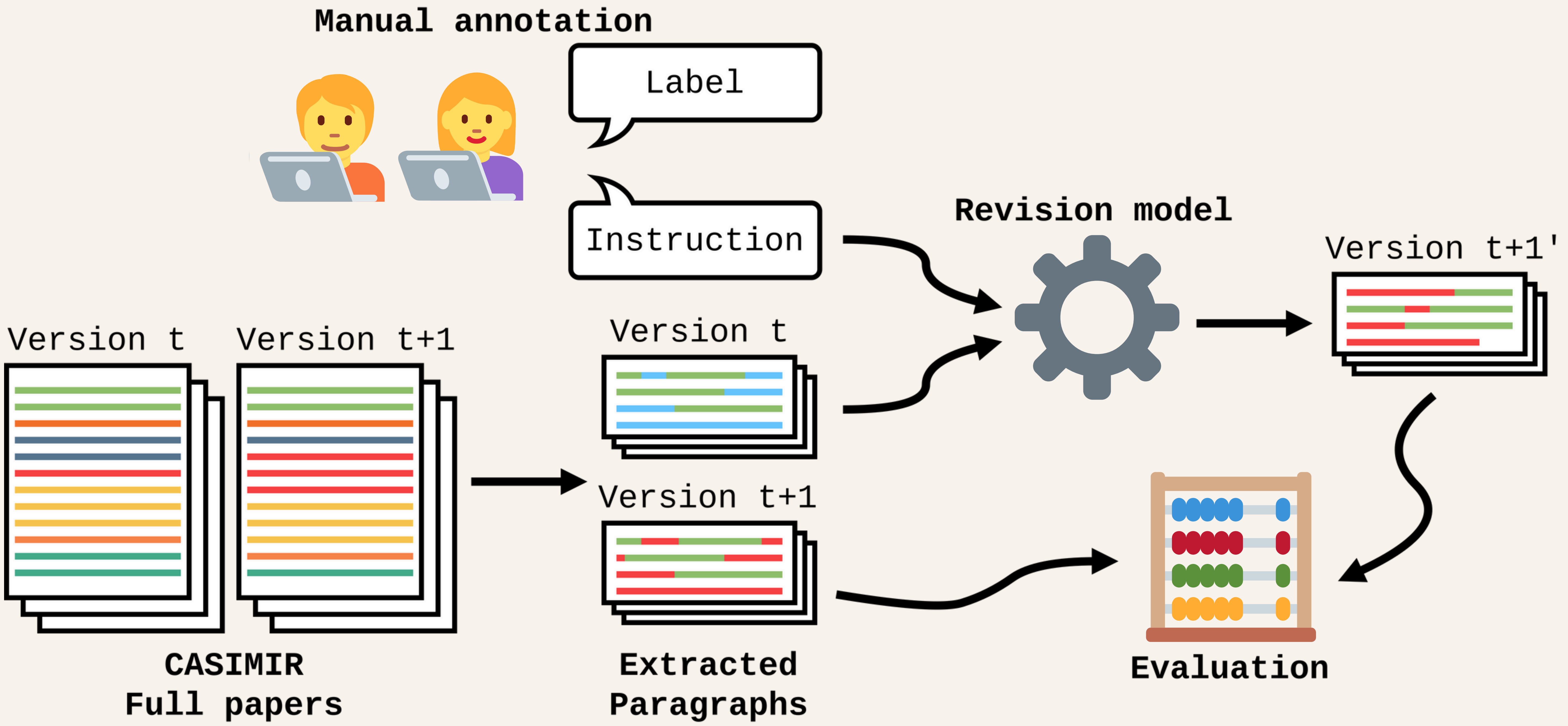
Comparison to existing corpora

	SMITH [1] 10/2019	IteraTeR [2] 03/2022	TETRA [3] 05/2022	F1000RD [4] 07/2022	arXivEdits [5] 10/2022	ARIES [6] 06/2023	CASIMIR [7] 10/2023
Full-length articles				✓	✓	✓	✓
Possible paragraph reconstruction		✓	✓	✓	✓	✓	✓
Include revision intentions		✓	✓		✓	?	✓
Label scope		Span of text	Span of text		Span of text	Multi-sentences?	Span of text

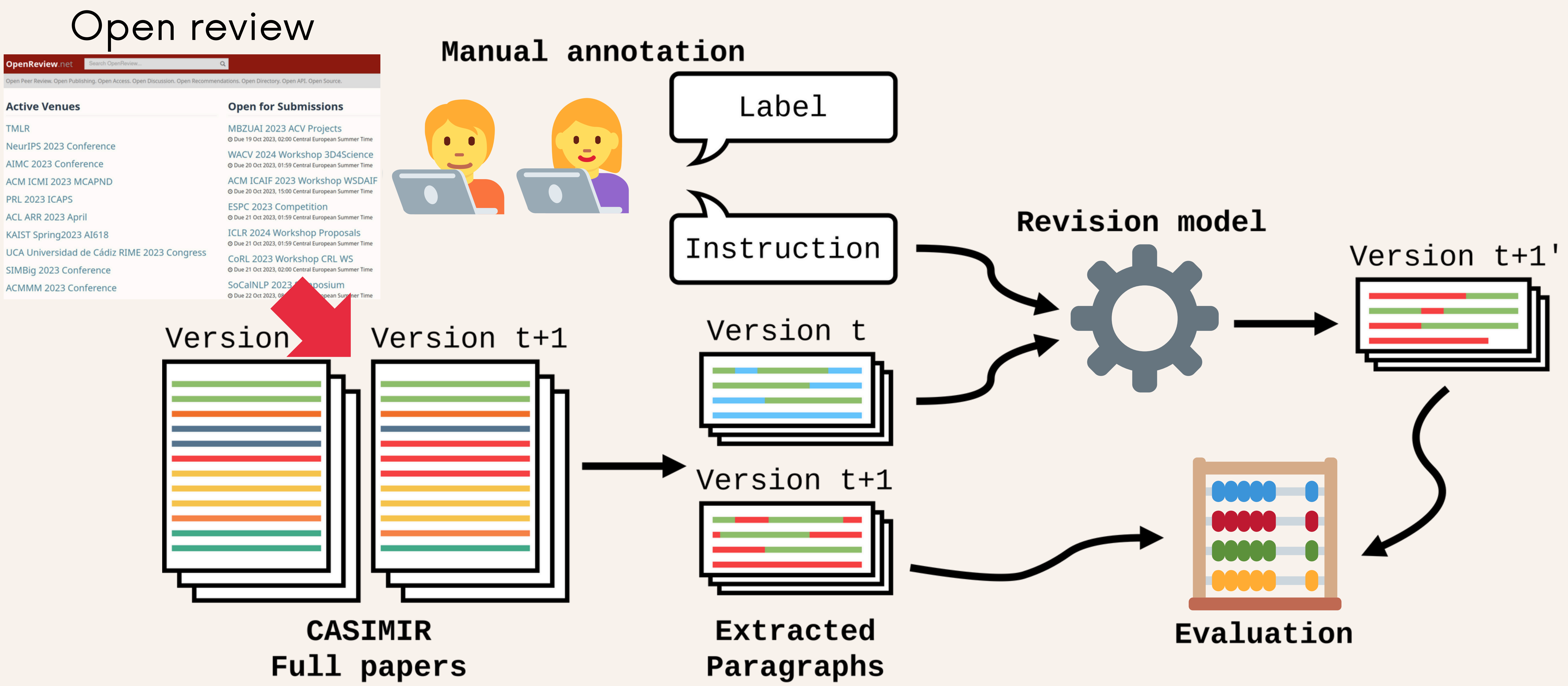
Table – Characteristics of previous datasets for scientific text revision

[1] Ito et al. 2019; [2] Du et al. 2022; [3] Mita et al. 2022; [4] Kuznetsov et al. 2022; [5] Jiang et al. 2022; [6] D'Arcy et al. 2023; [7] Jourdan et al. 2024

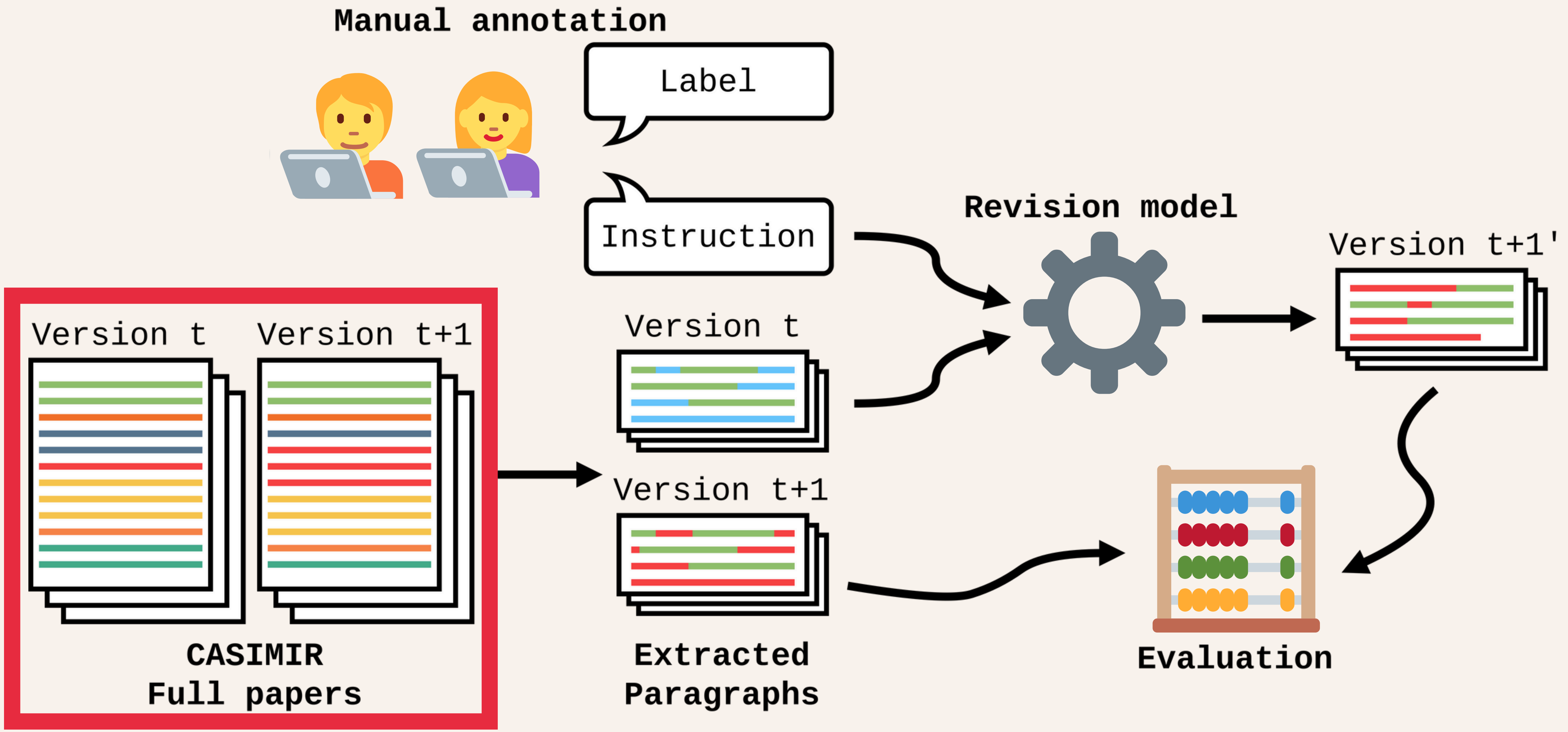
Data pipeline



Data pipeline



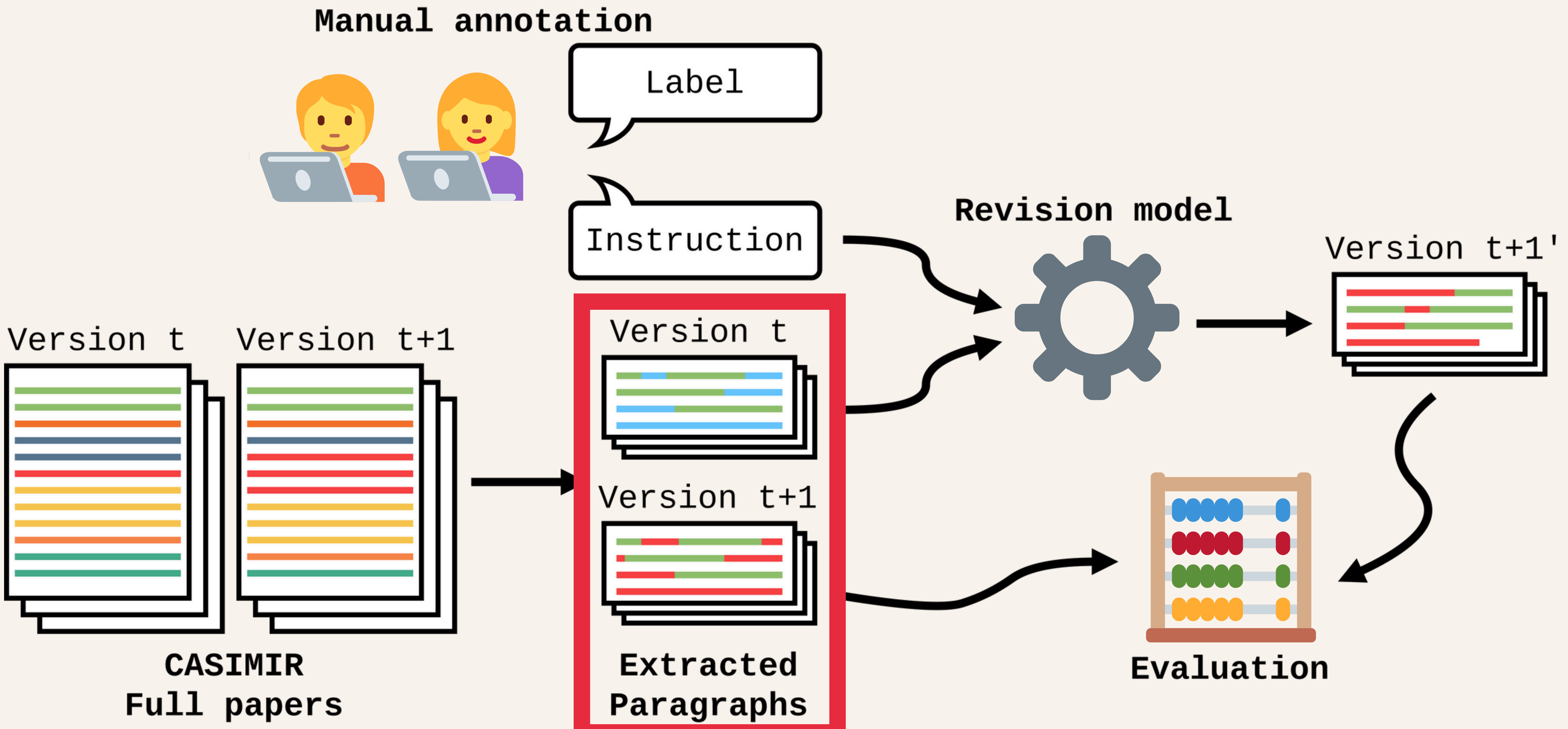
Data pipeline - Data source



1 889 810 modified paragraphs

(Jourdan et al., 2024)

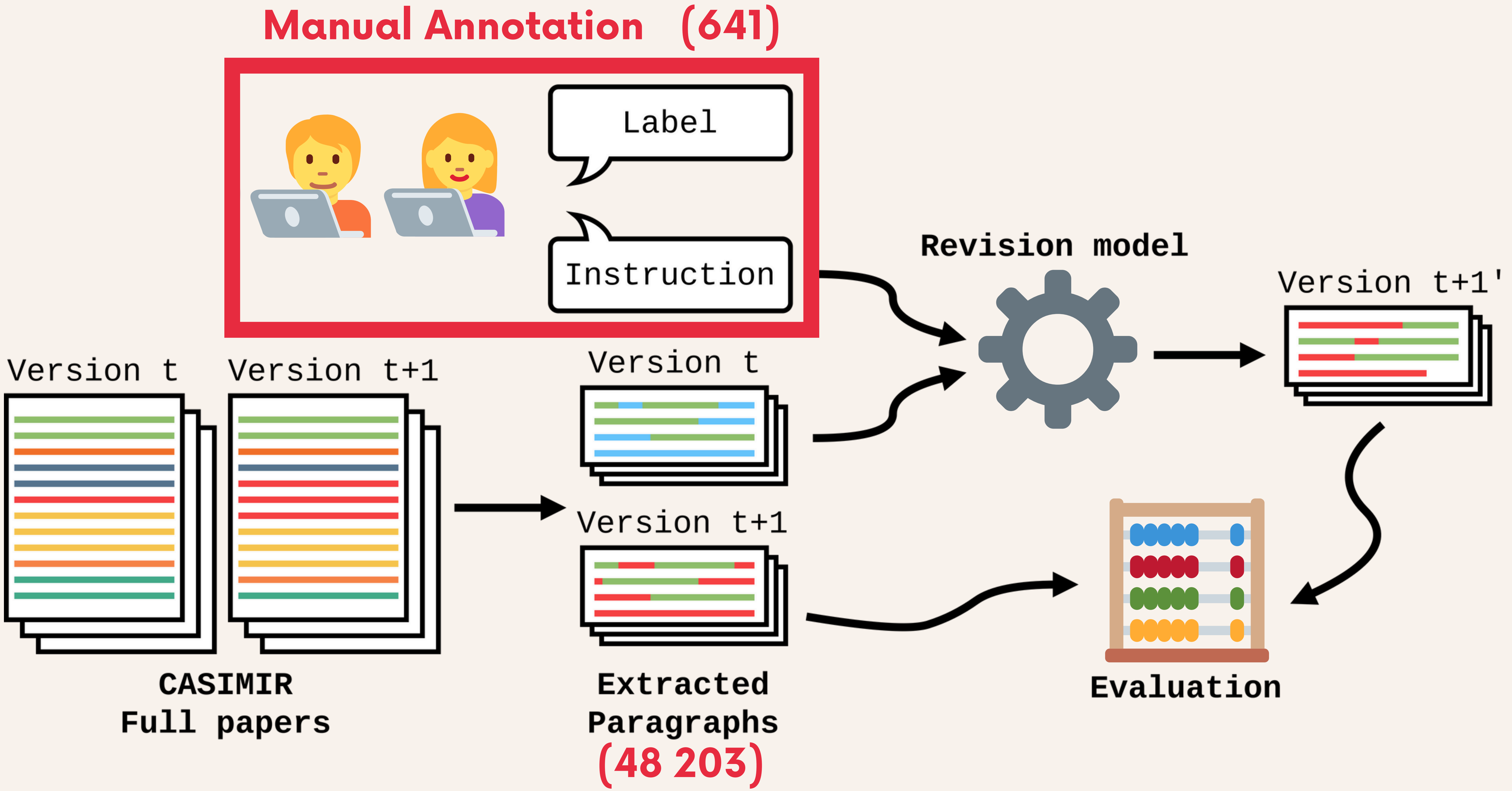
Data pipeline - Extraction



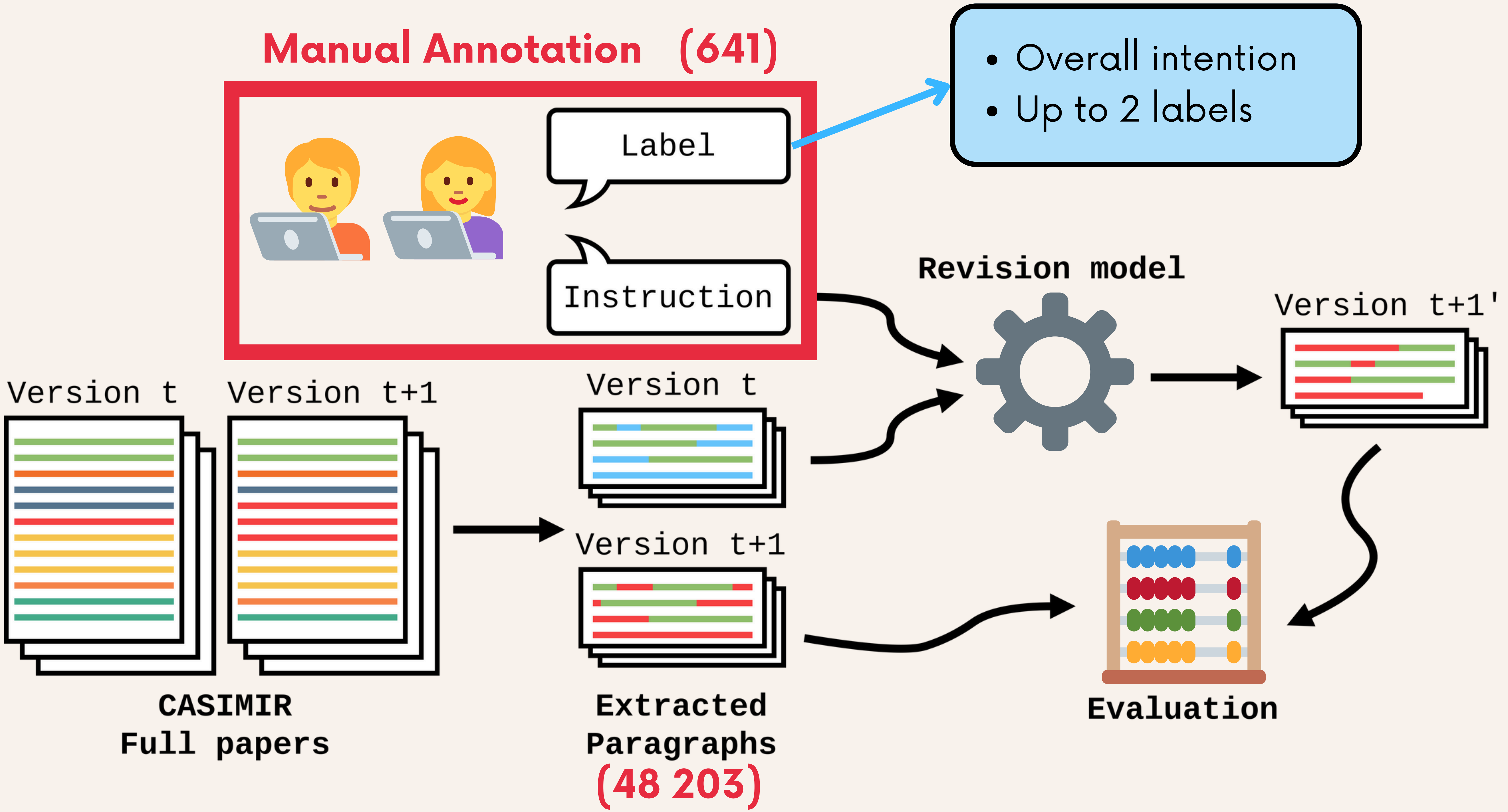
Interesting paragraphs exctraction (48 203)

Extraction code: https://github.com/JourdanL/pararev

Data pipeline – annotation



Data pipeline – annotation



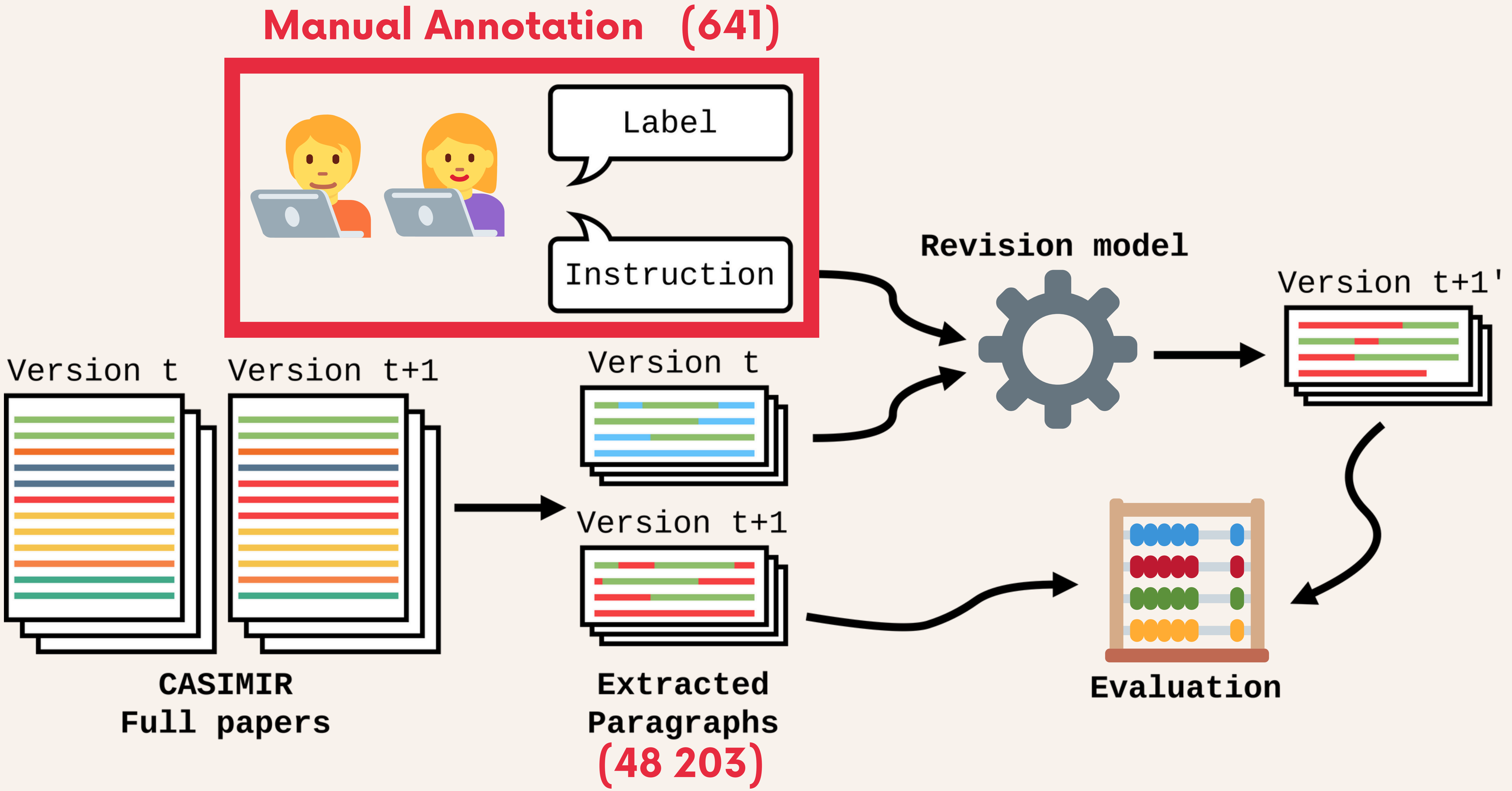
(Du et al., 2022b; Jiang et al., 2022)

Paragraph Revision taxonomy

Rewritting	**Light**	Changes in the choice of words
	Medium	Complete rephrasing of sentences
	Heavy	Complete rephrasing of the paragraph
Concision		Same idea, stated more briefly. Details are deleted
Development		Same idea, stated out at greater length by adding details or definitions of the terms used
Content	**Addition**	Modification on the content — Addition of a new idea
	Substitution	Modification on the content — Substitution of an idea or a fact by an other
	Deletion	Modification on the content — Deletion of an idea
Unusable	Segmentation problems (Footnote mixed with text), misalignment (paragraphs that have nothing to do with each other) and others problems coming from document processing	

(Du et al., 2022b; Jiang et al., 2022)

Data pipeline – annotation



Instruction

When is an instruction provided?

A paragraph have an associated instruction only when "Development", "Content addition" and "Content substitution" are not part of the list of intentions.

Concision & Rewritting Heavy

Rewritting light & Development

Instruction

When is an instruction provided?

A paragraph have an associated instruction only when "Development", "Content addition" and "Content substitution" are not part of the list of intentions.

How is an instruction written?

Instructions are simple and concise.

- Fluidify this paragraph.
- Edit this paragraph by making more formal choices of wording.
- Remove unnecessary details.

- Replace "A" with "B", change "C" to "D" and "E" to "F".
- 1. Remove "X" in sentence 1.
  2. Replace "Y" by "Z" in sentence 2.
  3. Make sentence 2 shorter.

Instruction

How is an instruction written?

Instructions can be used to direct the model on the location of the modifications.

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. [. . .]

Instruction

When is an instruction provided?

A paragraph have an associated instruction only when "Development", "Content addition" and "Content substitution" are not part of the list of intentions.

How is an instruction written?

Instructions are simple and concise.

Instructions can be used to direct the model on the location of the modifications.

In real world usage, a paragraph can be revised on a specific portion and the rest serve as context.

Annotation

10 annotators

- 👩‍💻 2 professors, 🧑‍🏫 3 PhD students, 🧑‍🎓 and 5 master's students
- not native from English
- specialized in the NLP domain
- experienced in reading and writing academic papers

(Krippendorff, 2018)

Annotation

10 annotators

- 2 professors, 3 PhD students, and 5 master's students
- not native from English
- specialized in the NLP domain
- experienced in reading and writing academic papers

Agreement

73.32% are double annotated
≈ 1.2 labels/paragraph

Krippendorff's alpha
0.499 (strict), 0.693(super-labels)

Paragraphs sharing at least one label
75.32% (strict) 95.11% (super-labels)

Mapping between super-labels and labels

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

(Krippendorff, 2018)

Statistiscs

48 203 paragraphs in total from 16 664 pairs of revised articles

641 annotated paragraphs (470 with cross annotation)

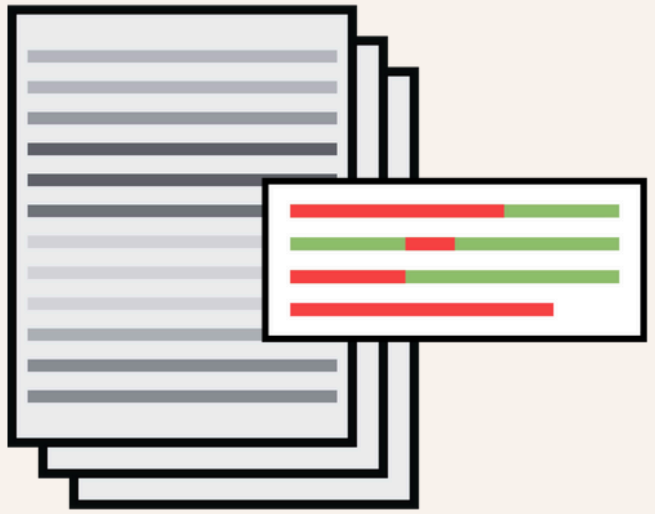
Heavily revised papers
>19 paragraphs revised

218 paragraphs

Moderately revised papers
4-5 revised paragraphs

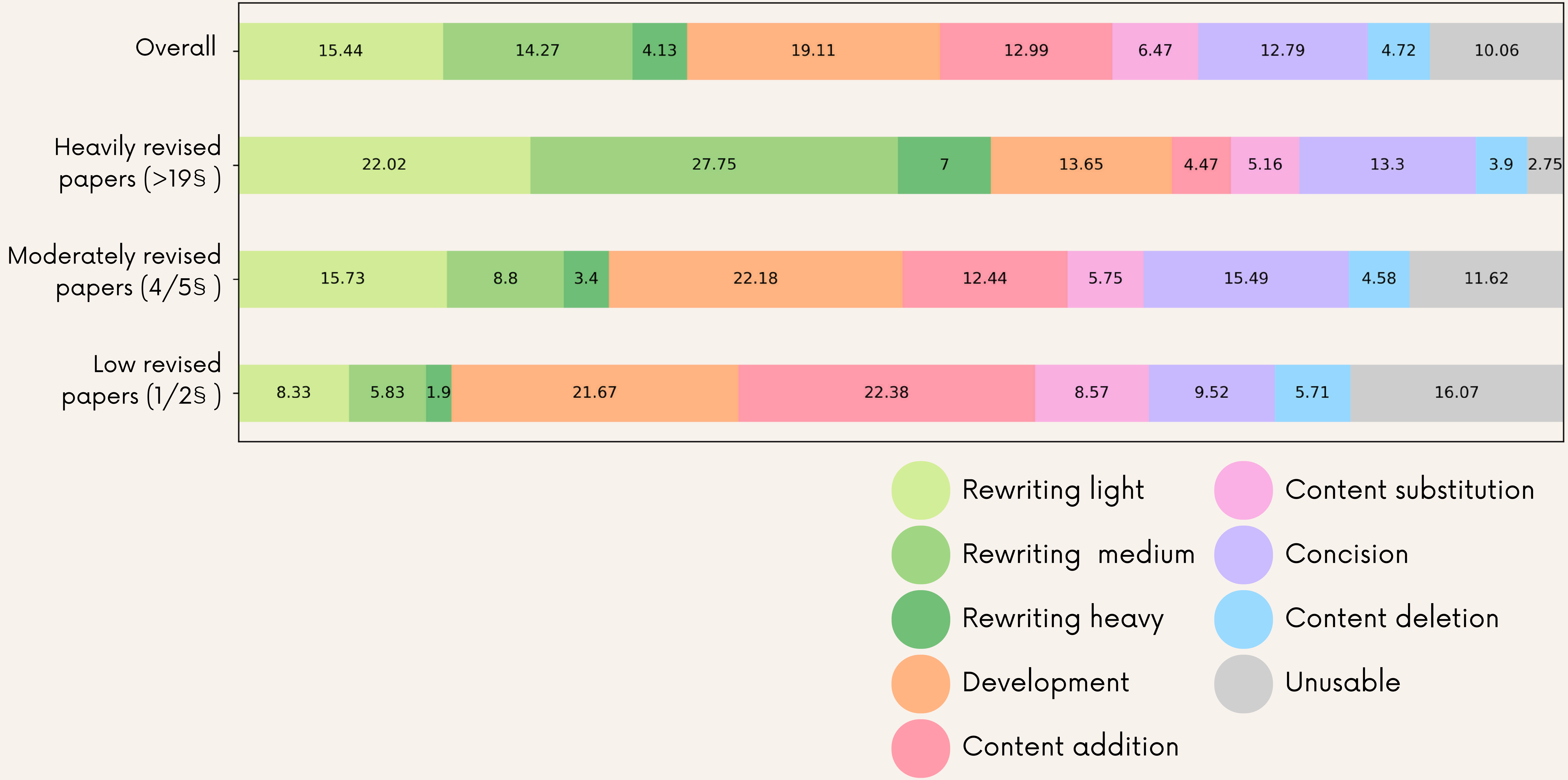
213 paragraphs

Low revised papers
1-2 revised paragraphs

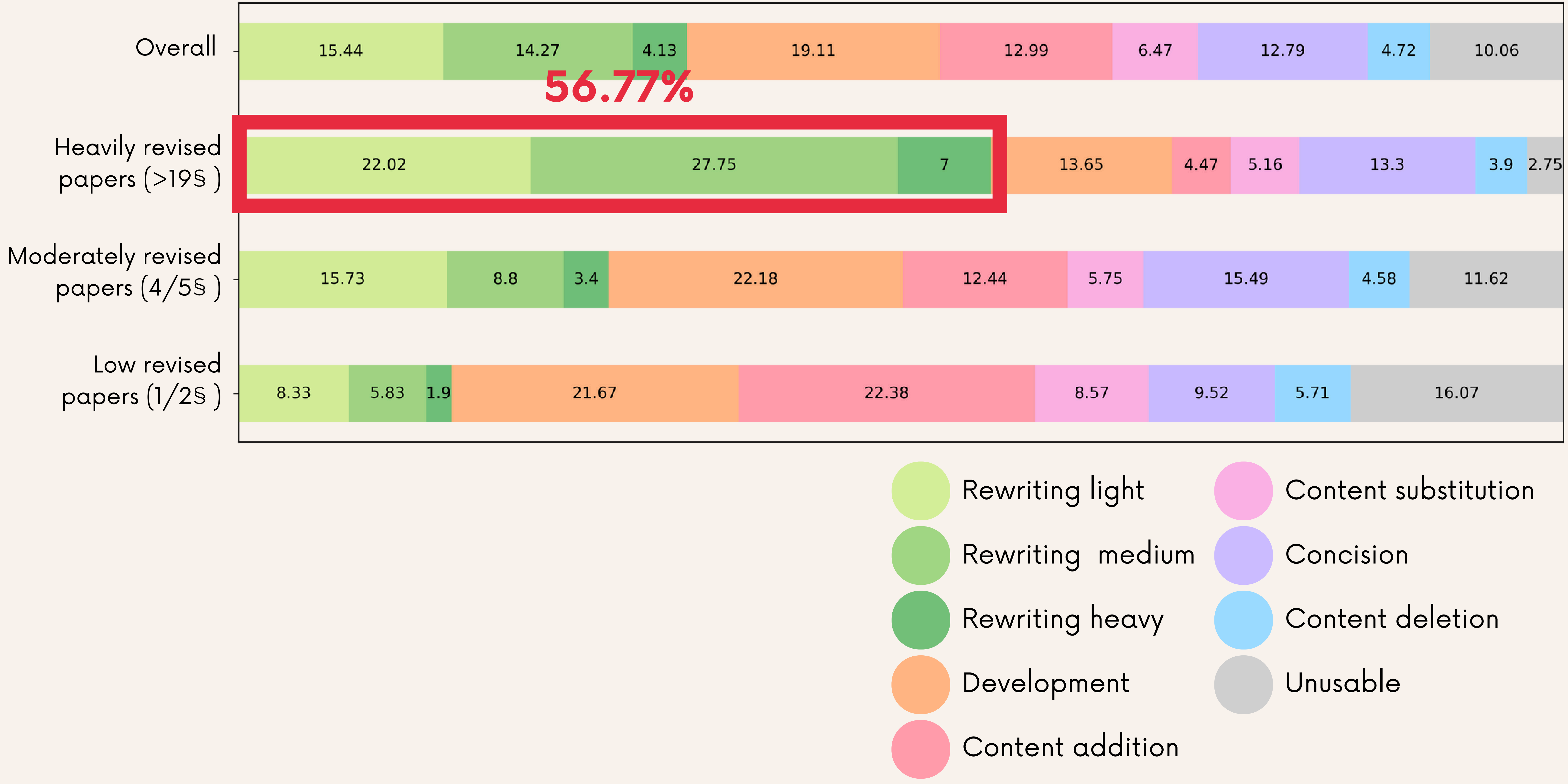


210 paragraphs

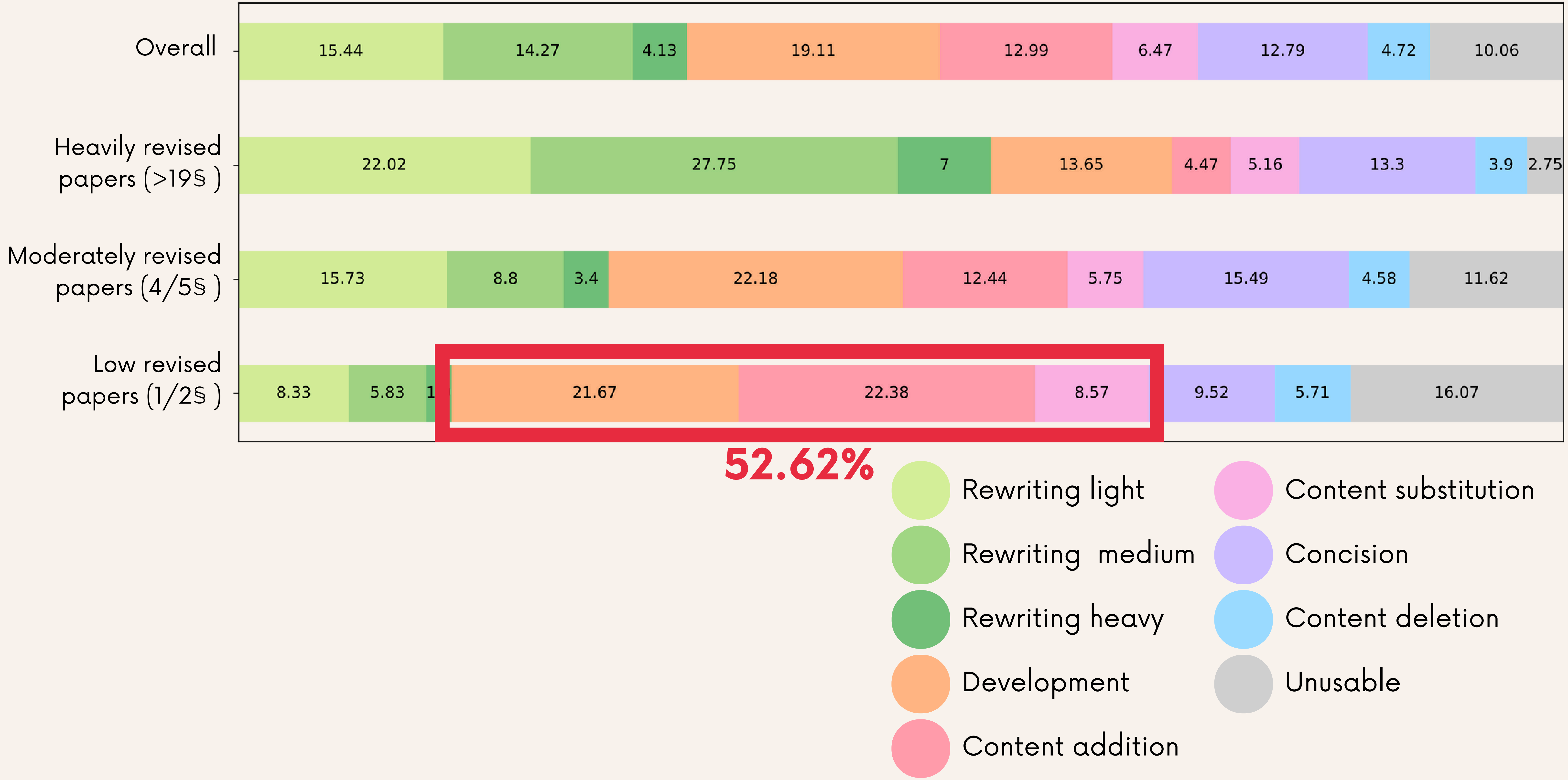
Statistiscs



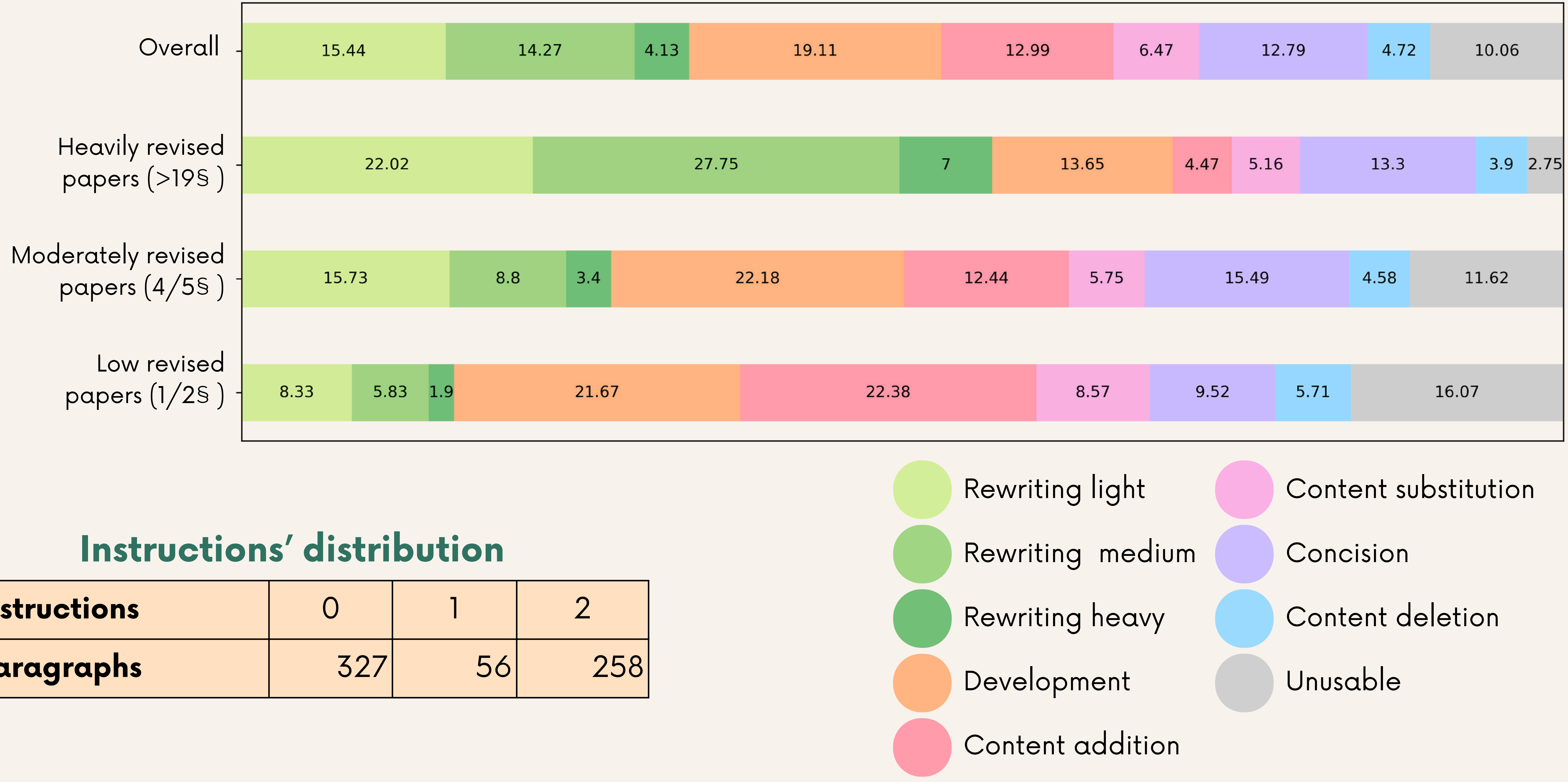
Statistiscs



Statistiscs



Statistiscs



Instructions' distribution

# instructions	0	1	2
# paragraphs	327	56	258

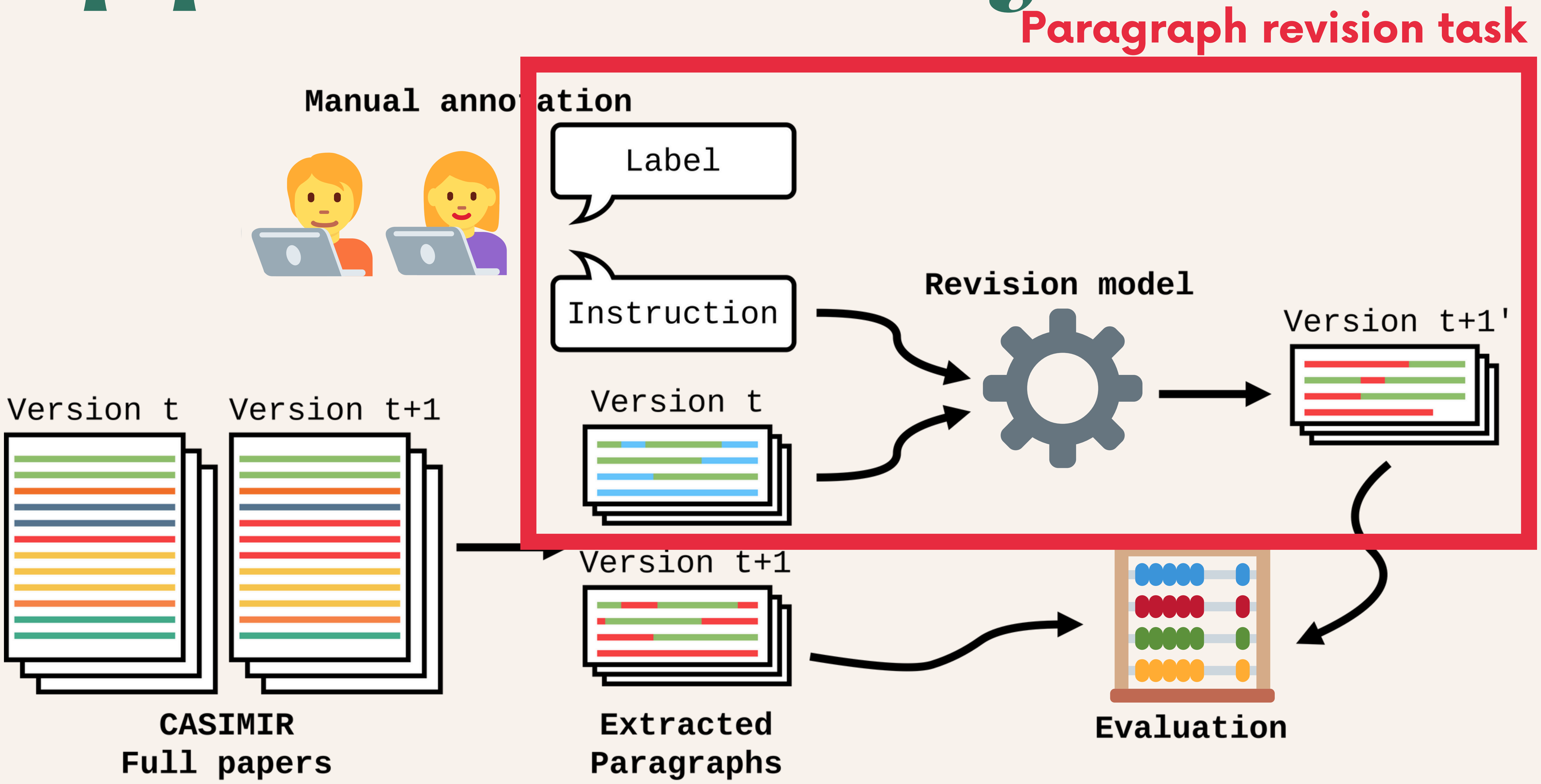
Statistiscs

Instructions' distribution

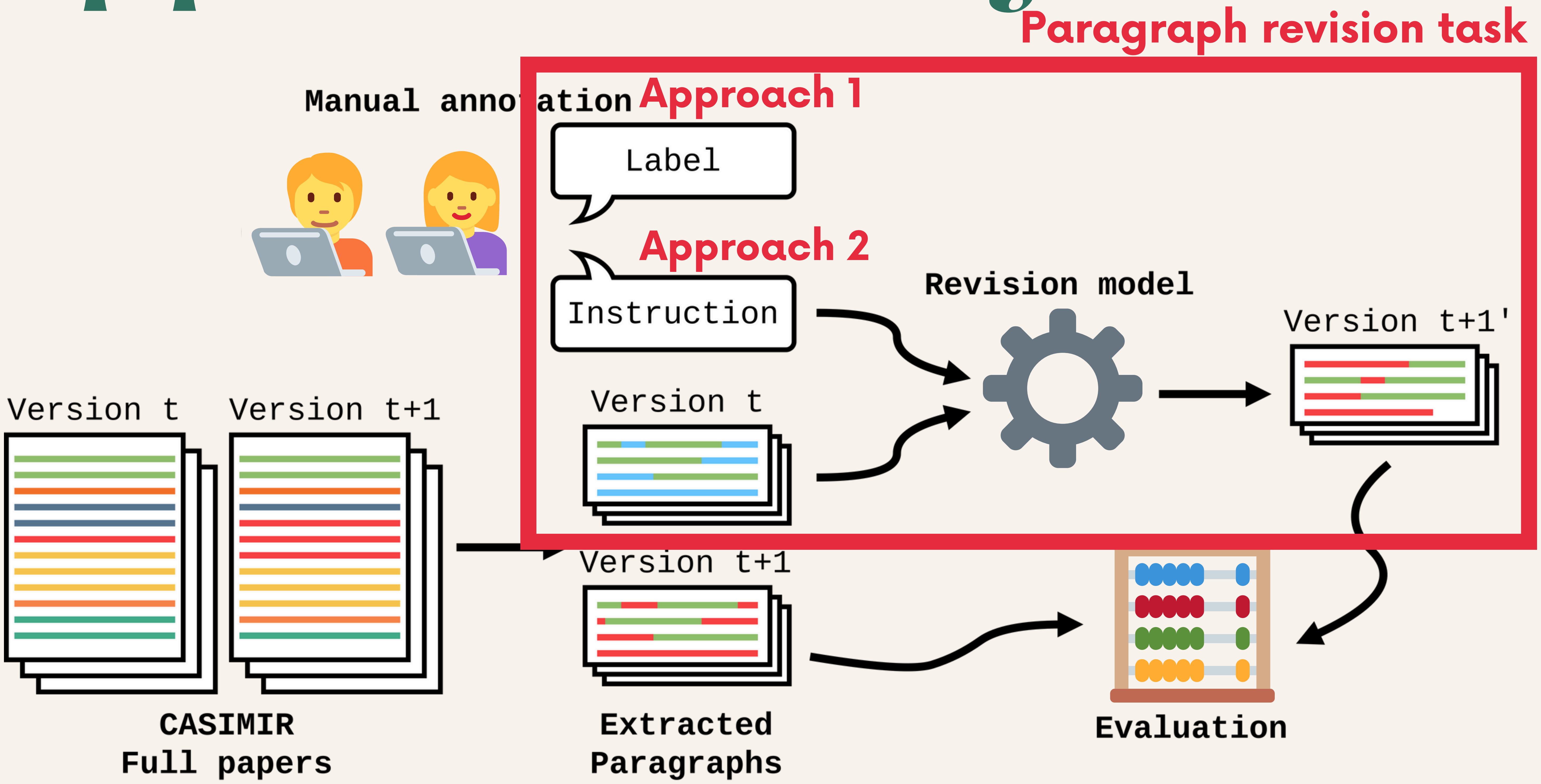
# instructions	0	1	2
# paragraphs	327	56	258

Evaluation set

Data pipeline - Revision generation



Data pipeline - Revision generation

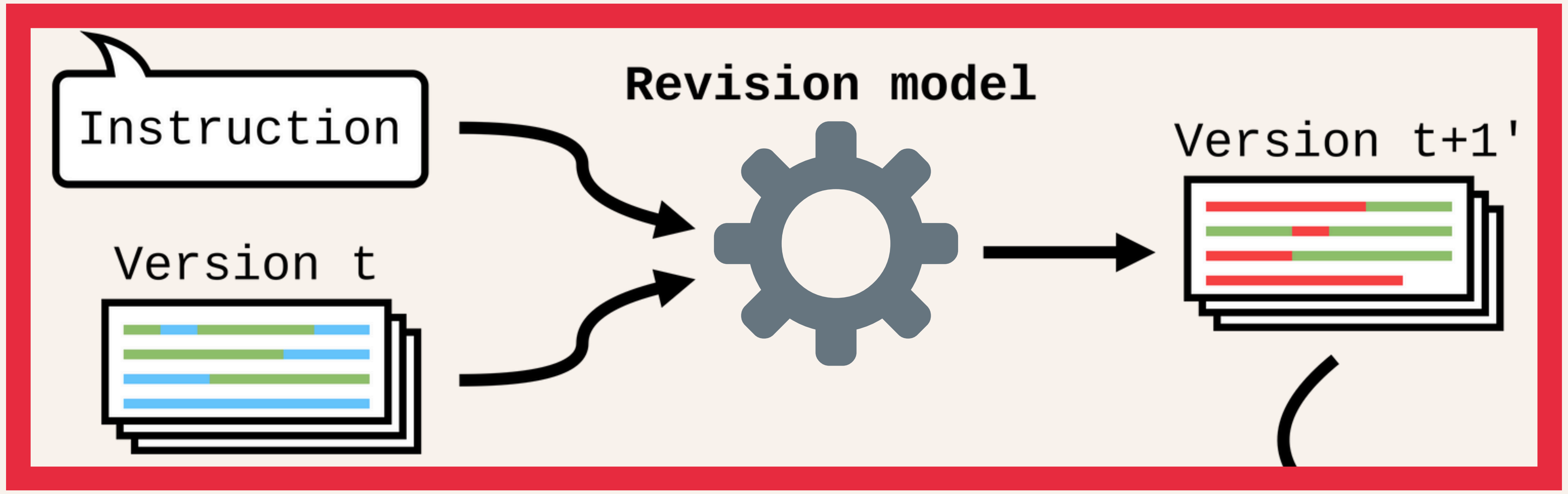


Data pipeline - Revision generation

Models

- CoEdIT (XL) (Grammarly)
- Mistral-7B-Instruct-v0.2 (Mistral AI)
- Llama-3-8B-instruct (Meta)
- GPT4o (OpenAI)

Paragraph revision task



Prompting

Prompt (**Bold blue text** correspond to the input data):

```
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
```

(Raheja et al., 2023)

Prompting

Prompt (**Bold blue text** correspond to the input data):

```
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
```

Approach 1: Label

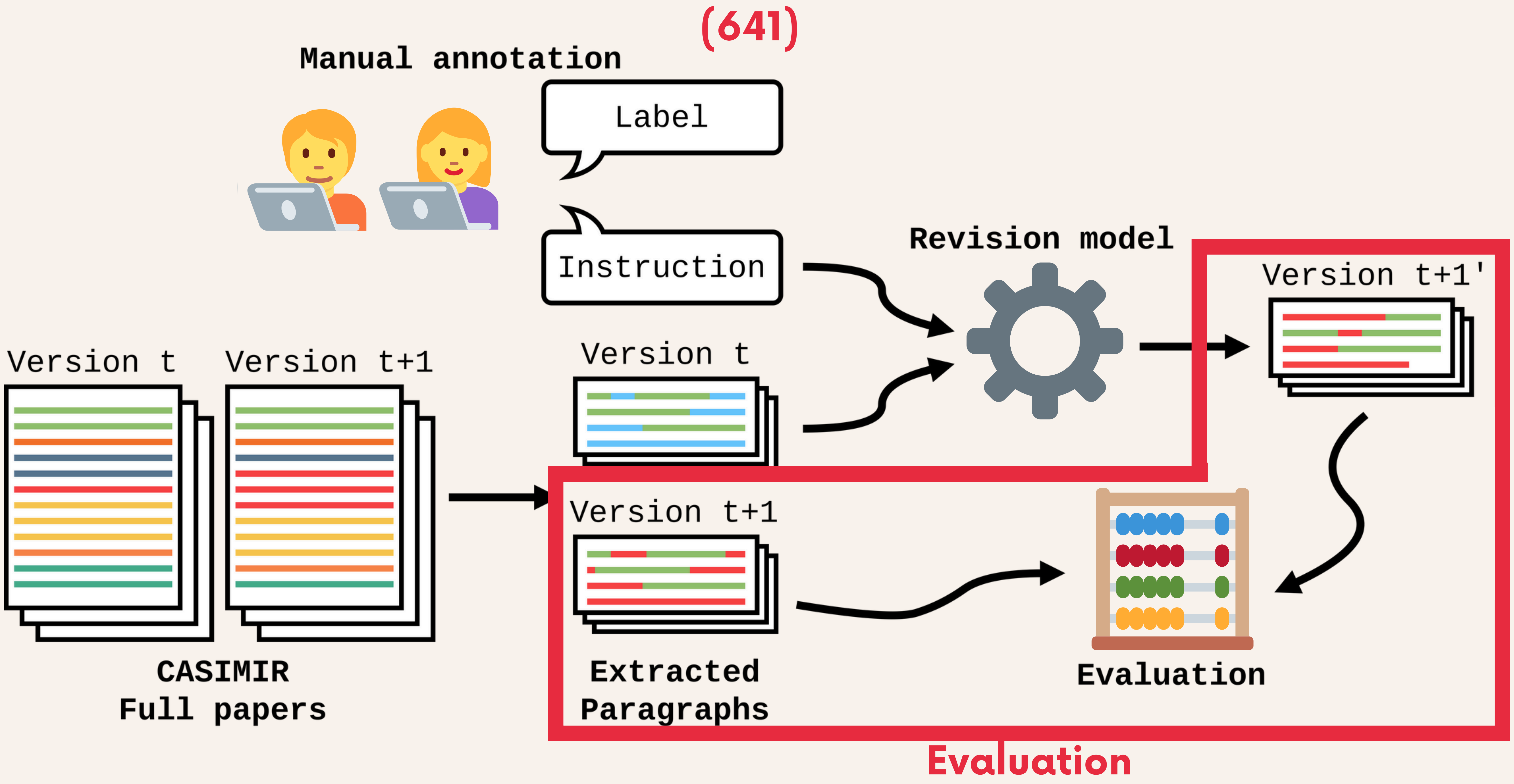
Rewritting	**Light**	Improve the English of this paragraph
	Medium	Rewrite some sentences to make them more clear and easily readable
	Heavy	Rewrite and reorganize the paragraph for better readability
Concision		Make this paragraph shorter
Content	**Deletion**	Remove unnecessary details

Approach 2: Instruction

Control baseline: no edits

(Raheja et al., 2023)

Data pipeline - Evaluation

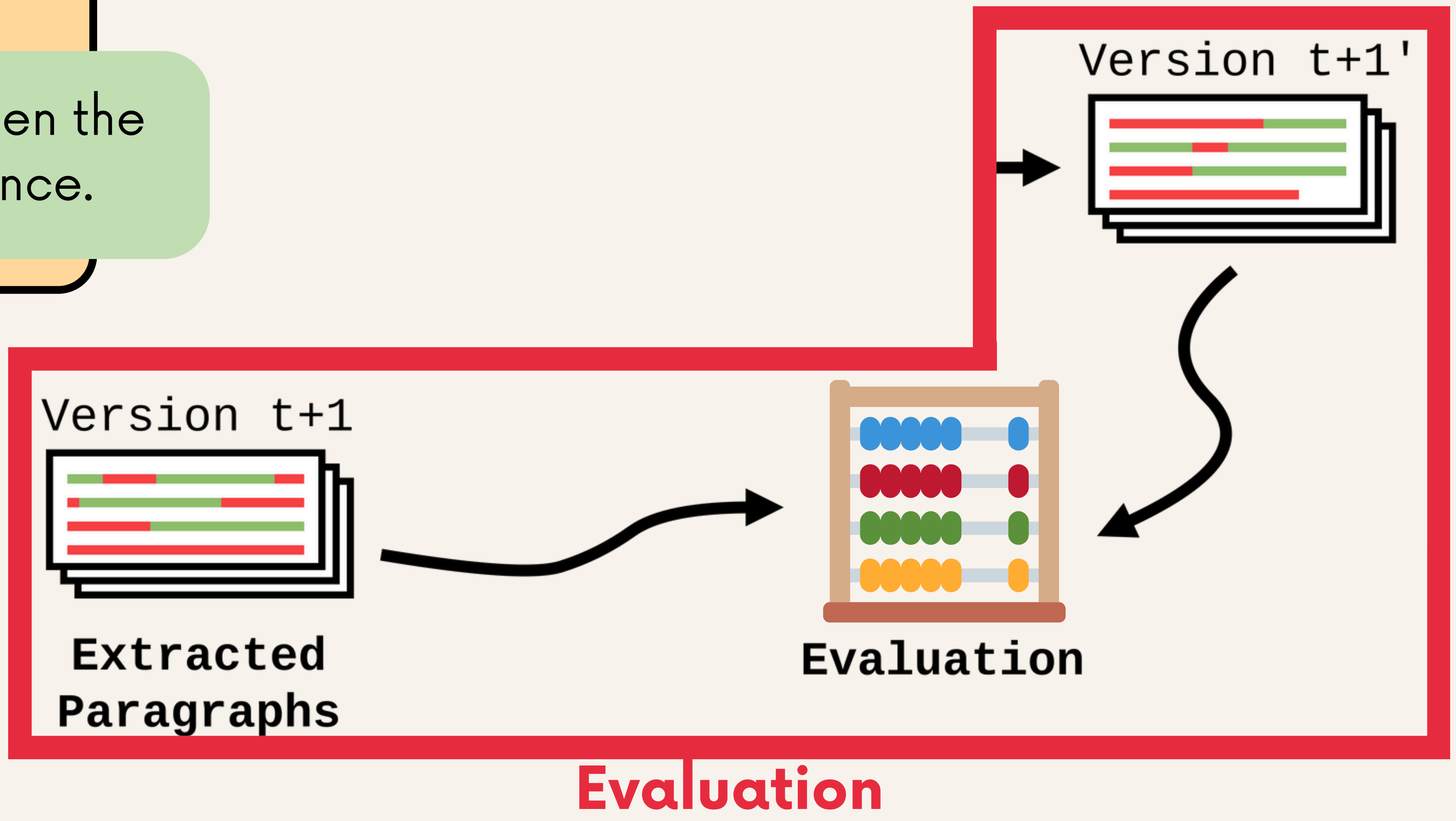


Data pipeline - Evaluation

Metrics

- SARI
- ROUGE-L
- Bert-score

Every metric measure the similarity between the predicted sentence and the gold sentence.



(Raheja et al., 2023; Lin, 2004; Xu et al., 2016; Zhang et al., 2020)

Impact of task definition on revision

Do the instructions improve the revision?

Metric	rougeL		sari		bert-score	
Model	Label	Inst	Label	Inst	Label	Inst
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†
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	

Impact of task definition on revision

Do the instructions improve the revision?

Metric	rougeL		sari		bert-score	
Model	Label	Inst	Label	Inst	Label	Inst
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†
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**	

Impact of task definition on revision

Do the instructions improve the revision?

Metric	rougeL		sari		bert-score	
Model	Label	Inst	Label	Inst	Label	Inst
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†
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	

Impact of task definition on revision

Do the instructions improve the revision?

Metric	rougeL		sari		bert-score	
Model	Label	Inst	Label	Inst	Label	Inst
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†
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	

Exemples of revisions with Coedit

Categories	Instruction
Content_deletion - Rewriting_light	Delete the second sentence. Improve the english in the first sentence.
Original paragraph	**Model A** gpt-4o
Here the higher valued θ i,j means the higher probability for the edge from node i to node j to be sampled. More importantly, notice that we use matrix θ ∈ R n × n to parameterize the probabilisticdistribution of n ! discrete feasible solutions. The compact, continuous and differentiable spaceof θ allows us to leverage gradient-based optimization without costly MDP-based construction offeasible solutions, which has been a bottleneck for scaling up in representative DRL solvers so far. Inother words, we also no longer need costly MCMC-based sampling for optimizing our model dueto the chain-rule decomposition. Instead, we use autoregressive factorization for sampling from theauxiliary distribution, which is faster than sampling with MCMC from the distribution defined by theenergy function.	Here, a higher valued θ i,j indicates a greater probability for the edge from node i to node j to be sampled. The compact, continuous, and differentiable space of θ allows us to leverage gradient-based optimization without costly MDP-based construction of feasible solutions, which has been a bottleneck for scaling up in representative DRL solvers so far. In other words, we also no longer need costly MCMC-based sampling for optimizing our model due to the chain-rule decomposition. Instead, we use autoregressive factorization for sampling from the auxiliary distribution, which is faster than sampling with MCMC from the distribution defined by the energy function.
Original paragraph	**Model B** coedit-xl
Here the higher valued θ i,j means the higher probability for the edge from node i to node j to be sampled. More importantly, notice that we use matrix θ ∈ R n × n to parameterize the probabilisticdistribution of n ! discrete feasible solutions. The compact, continuous and differentiable spaceof θ allows us to leverage gradient-based optimization without costly MDP-based construction offeasible solutions, which has been a bottleneck for scaling up in representative DRL solvers so far. Inother words, we also no longer need costly MCMC-based sampling for optimizing our model dueto the chain-rule decomposition. Instead, we use autoregressive factorization for sampling from theauxiliary distribution, which is faster than sampling with MCMC from the distribution defined by theenergy function.	Here the higher valued i,j means the higher probability for the edge from node i to node j to be sampled. More importantly, notice that we use matrix R n n to parameterize the probabilisticdistribution of n! discrete feasible solutions. The compact, continuous and differentiable spaceof allows us to leverage gradient-based optimization without costly MDP-based construction of feasible solutions, which has been a bottleneck for scaling up in representative DRL solvers so far. In other words, we also no longer need costly MCMC-based sampling for optimizing our model dueto the chain-rule decomposition. Instead, we use autoregressive factorization for sampling from theauxiliary distribution, which is faster than sampling with MCMC from the distribution defined

Categories	Instruction
Concision - Rewriting_light	Make first sentence more concise. Rewrite phrases, prefer short formulations and avoid we.
Original paragraph	**Model A** coedit-xl
Listwise RL (CDQN) : To solve the combinatorial action space problem of listwise actions, we follow the Cascaded DQN (CDQN) framework of Chen et al. (2019a). The main challenge is that building the list all at once is not feasible due to the intractably large number of possible lists. Therefore, the key is to build the list incrementally, one action at a time. Thus, each list index can be treated as an individual non-combinatorial action which can be trained with RL. We replace the Q-network of CDQN with AGILE in order to accommodate a varying action space. We share the weights of the cascaded Q-networks. Algorithm 1 provides complete details on listwise AGILE.	Listwise RL (CDQN) : To solve the combinatorial action space problem of listwise actions, we follow the Cascaded DQN (CDQN) framework of Chen et al. (2019a).
Original paragraph	**Model B** Mistral-7B-Instruct-v0.2
Listwise RL (CDQN) : To solve the combinatorial action space problem of listwise actions, we follow the Cascaded DQN (CDQN) framework of Chen et al. (2019a). The main challenge is that building the list all at once is not feasible due to the intractably large number of possible lists. Therefore, the key is to build the list incrementally, one action at a time. Thus, each list index can be treated as an individual non-combinatorial action which can be trained with RL. We replace the Q-network of CDQN with AGILE in order to accommodate a varying action space. We share the weights of the cascaded Q-networks. Algorithm 1 provides complete details on listwise AGILE.	Listwise RL (CDQN): Following the Cascaded DQN (CDQN) framework of Chen et al. (2019a), we address the combinatorial action space problem of listwise actions. Due to the large number of possible lists, constructing them all at once is impractical. Therefore, we build lists incrementally, one action at a time. Each list index is treated as an individual non-combinatorial action, trainable through RL. The Q-network in CDQN is replaced with AGILE to accommodate a varying action space. Weights of the cascaded Q-networks are shared. (Algorithm 1 outlines listwise AGILE in detail.)
Original paragraph	**Model A** Meta-Llama-3-8B-Instruct
Listwise RL (CDQN) : To solve the combinatorial action space problem of listwise actions, we follow the Cascaded DQN (CDQN) framework of Chen et al. (2019a). The main challenge is that building the list all at once is not feasible due to the intractably large number of possible lists. Therefore, the key is to build the list incrementally, one action at a time. Thus, each list index can be treated as an individual non-combinatorial action which can be trained with RL. We replace the Q-network of CDQN with AGILE in order to accommodate a varying action space. We share the weights of the cascaded Q-networks. Algorithm 1 provides complete details on listwise AGILE.	To address the combinatorial action space problem, the Cascaded DQN (CDQN) framework is employed, as proposed by Chen et al. (2019a). The challenge lies in the intractably large number of possible lists, making it necessary to build the list incrementally, one action at a time. Each list index is treated as an individual non-combinatorial action, trained with RL. The Q-network is replaced with AGILE to accommodate a varying action space, with shared weights among cascaded Q-networks. Algorithm 1 provides complete details on listwise AGILE.

Data pipeline – Evaluation

Metrics

- SARI
- ROUGE-L
- Bert-score

Every metric measure the similarity between the predicted sentence and the gold sentence.

- Penalize revisions that deviate from the gold strandard
- Don't reflect the real quality of the revision

(Raheja et al., 2023; Lin, 2004; Xu et al., 2016; Zhang et al., 2020)

Conclusion

Contributions

- Definition of the scientific **text revision task at paragraph-level** using full-length instruction
- **ParaRev**, a **dataset** of revised paragraphs, with an evaluation split annotated with revision instructions
- Providing personalised instructions leads to more effective revisions than general ones

Perspectives

- **Automatically annotate** the train data
- **Fine tune** a specialised small open source model
- Question and explore new **evaluation methods**

COLING 2025 • Abu Dhabi

The 31st International Conference on Computational Linguistics

Writing Aids at the Crossroads of AI, Cognitive Science and NLP

CASIMIR: A Corpus of Scientific Articles enhanced with Multiple Author-Integrated Revisions

Contact: leane.jourdan@univ-nantes.fr

TALN

taln.ls2n.fr

Article:
https://arxiv.org/abs/2501.05222

Corpus:
https://huggingface.co/datasets/taln-ls2n/pararev

Appendix - Annotation environment

ryESgXktV.BJ4dKdWmr.01	
Source paragraph	**Target paragraph**
In our prior work (Chakraborti et al. 2017), we encapsulate such inconsistencies as model differences , while considering the discrepancies between the human and its own model when generating explanations. An explanation then becomes a request to the human to adjust the model differences in his mind so that the robot's behavior would make sense in the updated model, which captures the human's expectation of the robot. The general decision-making process of an agent in the presence of such model differences is termed model reconciliation (Chakraborti et al. 2017; Zhang et al. 2017).	To address this challenge, the agent should consider the discrepancies between the human and its own model while generating explanations. In our prior work [7], we encapsulate such inconsistencies as model differences . An explanation then becomes a request to the human to adjust the model differences in his mind so that the robot's behavior would make sense in the updated model, which is used to produce the human's expectation of the robot. The general decision-making process of an agent in the presence of such model differences is termed model reconciliation [7], [8].

Category main	**Category secondary**	**Instruction**
Rewritting_medium		Revise the opening of this paragraph to make it more compelling.

txe2sPPkO.id6Xr1pUq.00	
Source paragraph	**Target paragraph**
In this section we discuss how SafeNet can be instantiated in practice. There are two aspects the data owners need to agree upon before instantiating SafeNet: i) The MPC framework used for secure training and prediction phase and ii) the parameters in Theorem 6 to achieve poisoning robustness. The MPC framework is agreed upon by choosing the total number of outsourced servers N participating in the MPC, the number of corrupted servers T and the nature of the adversary (semihonest or malicious in the SOC paradigm). The owners then agree upon a filtering threshold ϕ and the number of poisoned owners t that can be tolerated. Once these parameters are chosen the maximum allowed error probability of the local models trained by the honest owners based on Lemma 5 and	In this section we discuss how SafeNet can be instantiated in practice. There are two aspects the data owners need to agree upon before instantiating SafeNet: i) The MPC framework used for secure training and prediction phase and ii) the parameters in Theorem 6 to achieve poisoning robustness. The owners agree upon the number of outsourced servers N participating in the MPC, the number of corrupted servers T along with the role of the adversary (semi-honest or malicious) in the MPC and consequently choose an appropriate training framework that satisfies this criteria. The owners then agree upon a filtering threshold ϕ and the number of poisoned owners t that can be tolerated. Once these parameters are chosen the maximum allowed error probability of the local models trained by the honest owners based on Lemma 5 and

Category main	**Category secondary**	**Instruction**
Rewritting_medium		Rewrite the middle sentence of this paragraph to make it clearer.

Appendix - Additionnal exemples

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_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.

Appendix - Additionnal exemples

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. [...]

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 [...]

Appendix - Additionnal statistics

		Min	Avg	Max	Std
# characters	Source	47	5202	680.16	374.11
	Target	48	5588	715.58	394.20
# words	Source	3	913	109.28	59.55
	Target	3	1037	114.80	62.95
# sentences	Source	1	99	5.38	3.13
	Target	1	81	5.59	3.24

Table 1 - Distribution of the length of the paragraphs

	Min	Avg	Max	Std
% words deleted	0	21.54	96.51	18.19
% words added	0	25.63	97.90	18.15
levenshtein distance	0	194.80	2265	160.10

Table 2 - Amount of edition between version 1 and 2 of the paragraphs