Characterizing revision events in students' writing processes using LLMs

Léo Nebel, François Bouchet, Vanda Luengo

Characterizing revision events in students' writing processes using LLMs

Léo Nebel
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
EvidenceB, Paris, France
leo.nebel@lip6.fr

François Bouchet
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
francois.bouchet@lip6.fr

Vanda Luengo
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
vanda.luengo@lip6.fr

ABSTRACT

Revision is a key part of the writing process. Some important studies proposed taxonomies of revision and analysed different corpora through the lens of these taxonomies. These analyses have often been made through manual annotation after collecting the data, even if, more recently, some classifiers were trained to do this task. Based on an annotated public dataset available (ArgRewrite V.2), we go further by exploring an LLM-based classifier of revision events, which paves the way for an automatic online feedback system on the revision process students follow when writing. We compare our results to those obtained from the trained classifier of this dataset, as well as another LLM-based classifier applied in another context and discuss them. We made our code publicly available on a GitHub repository[1] for replication.

Keywords

Writing, Keystroke Logging, Revision

1. INTRODUCTION

Revision is one of the three main parts of the whole writing process, as introduced in the work of Flower and Hayes [6]. Keystroke logging techniques are now commonly used to study this process. It allows a non-intrusive approach to analyse all the changes the writer made and when they made them. In an educational context, it would be interesting to be able to give feedback to students on their revision process when writing texts. Indeed, a fully automatic characterization of revisions would allow providing strategy-specific feedback by aligning a student's revision approach to the main weaknesses of their text (recommending structural revisions for unclear reasoning, surface revisions for grammar or spelling issues). This extends traditional product-focused feedback by addressing task processing and self-regulation—— two levels identified by Hattie et al. [8] as crucial for positively impacting learning. To be able to characterize revision (to build the feedback), a first mandatory step consists of automatically extracting the beginnings and ends of revision events [13]. This step, not detailed here, can be performed automatically using a rule-based approach. It has been shown that this approach gives a precision of 0.79 and an average error of 5 characters away from the truth. The next step, which is the focus of this paper, is to characterize automatically what type of revision occurred. We will first present the different existing revision taxonomies and how they were used in previous research works. Then, after introducing our research question, we will introduce the datasets and methods used before showing our results and discussing them in the last part.

[1] https://anonymous.4open.science/r/argRewrite-v2-llms-7503

2. RELATED WORKS

2.1 Taxonomies

In their cognitive process theory of writing, Flower and Hayes [6] described reviewing as a process that could occur at any time during writing. This reviewing process implies an evaluation (when the writer decides to read their written text, for instance) and/or a revision (when a change is made in the written text), which is the only external and, therefore, easily observable part of the reviewing process.

Faigley and Witte [5] provide more details in their characterization of revision by defining two main categories of revision: (1) meaning changes, which affect the semantics, and (2) surface changes, which do not. Their taxonomy is based on "whether new information is brought to the text or whether old information is removed in such a way that it cannot be recovered through drawing inferences". Within surface changes, the authors separated what they called (a) formal changes (changes on spelling, tenses, abbreviations, punctuation, paragraph, other format) and (b) meaning-preserving changes (changes that paraphrase the concepts in the text but do not alter them) which they characterize by the action type (additions, deletions, substitutions, permutations, distributions, and consolidations). Distribution consists of dividing the content of a single unit into multiple ones. In contrast, consolidation does the opposite (splitting a sentence or reuniting two sentences are respectively distributions and consolidations revisions). Within meaning changes, they separated (a) macrostructure changes and (b) microstructure changes (depending on whether it would or would not affect the summary of a text). These two categories had the same subdivisions as the meaning-preserving changes, which are the different types of actions possible.

More recently, Lindgren and Sullivan [12] developed another taxonomy of revision based on these previous works. They start by distinguishing internal and external revisions. Under the scope of internal revisions, they describe (a) prelinguistic revisions, which are mental revisions of plans or ideas, and (b) pre-text revisions, which are revisions of linguistic formulations that have not been written yet. External revisions are also divided into two categories: (a) precontextual revisions are revisions that happen at the point of inscription (i.e., the leading edge), inspired by Van Gelderen and Oostdam [18], and (b) contextual revisions, which happen within the previously written text. Pre-contextual and contextual revisions are both subdivided into three categories: conceptual revisions, form revisions, and typos. We can interpret these categories as another revision dimension, focused on the effect of the revision on the product.

Conijn et al. [2] go further in that way and proposed a tagset rather than a strict taxonomy, arguing that many revision events warrant multiple tags instead of a single label drawn from a large, overlapping set. Their tagset includes tags reflecting not only the effect of the revision on the final product but also on its temporal or spatial occurrence, which are considered process-related attributes. As demonstrated in their study, these process-related attributes can often be automatically extracted from keystroke data.

2.2 Automatic annotation

All these studies proposed manual annotations of the proposed taxonomies, except Conijn et al., who proposed automatic feature extraction (slightly different from really determining automatically the revision types). Regarding automatic processing to identify these types of revisions, we need to focus on more recent works. While many studies analyze revisions at the keystroke level to characterize individual events, others, such as Du et al. [4], compare drafts at specific stages. Rather than focusing on revision effects, they propose an 'edit intention' taxonomy, a distinction we find particularly relevant in educational contexts, where intended revisions may not yield the expected outcomes (e.g., a student correcting a spelling error introduces another). As in other frameworks, they distinguish between (1) meaning-changing and (2) non-meaning-changing edits, with the latter subclassified into (a) fluency, (b) clarity, (c) coherence, and (d) style. Their main contribution is the Iterater dataset, which is manually and then automatically annotated. A similar approach was taken by Kashefi et al. [9], who introduced the ArgRewrite V.2 corpus, comprising annotated argumentative revisions from two revision cycles. Both studies used these corpora to train revision purpose classifiers — RoBERTa-based in Du et al.'s case, and XGBoost in Kashefi et al.'s, the latter optimized via randomized parameter search with cross-validation, using textual, syntactic, semantic, and discourse-level features.

These two last works brought novelty by exploring the use of machine learning techniques to train prediction models and automatically classify revision types (that were only manually annotated in the other cited works). More recent ones even try to classify edit intention using Llama2-70b [14, 15]. In [14], the data are aligned scientific papers, collaborative revisions that were manually labeled, and the authors are doing In-Context Learning with Llama2-70b, including Chain-Of-Thought and a dynamic selection of examples through a computation of embedding distances based on RoBERTa, to predict the labels. In [15], a subset of the same authors are pushing this work further by fine-tuning models. The results are promising, but the question remains open on whether their approach is reproducible in other contexts. Here, we are especially interested in educational ones which may substantially differ: students are still learning writing strategies and therefore the produced text may be less fluent and coherent, and the revision intents, not always clear.

To sum up, Ruan et al. explored the use of LLMs to predict revision types, Kashefi et al. explored the automatic prediction of revision types in an educational context. We now aim to close this research gap by exploring the following questions:

- To what extent can existing LLM-based revision-type classification methods be effectively applied within the context of student argumentative writing?
- How do large language models perform compared to other techniques for revision-type classification existing within the context of student argumentative writing?

Specifically, we evaluate GPT-4o and o3-mini using In-Context Learning techniques—including Chain-of-Thought prompting on the ArgRewrite V.2 dataset.

3. DATASET

We used the ArgRewrite V.2 annotated dataset to predict revision labels based on the taxonomy from the original study. This dataset was selected for several reasons: it was previously used to train a revision purpose classifier, enabling direct comparison with our results; its manual annotation process is well-documented, supporting more explainable outcomes; and its educational context and focus on argumentative writing align closely with our research objectives, offering a distinct perspective from prior LLM-based studies on revision type prediction.

The ArgRewrite V.2 dataset is described and used in the work of Kashefi et al. [9]. It comprises three versions of 86 essays (258 raw text files) written by undergraduate and graduate students, in three different sessions after receiving feedback. The corpus is available at `http://argrewrite.cs.pitt.edu`. Revisions were annotated at two different levels: sentential and subsentential. We chose to focus only on the sentential revisions, knowing that the subsentential revisions are extracted from the same textual data, it is just a more precise level of annotation that we did not consider here (there were also more sentential revision than subsentential ones in the whole dataset, meaning that all subsentential revisions were probably not encoded). As their binary classifier model was trained through 5-fold cross-validation, we created a test set by taking a fifth of the whole dataset while respecting label repartition through stratified sampling across texts (and not revision pairs) to ensure that the context is still available when making predictions. It resulted in a test dataset of 564 revision pairs.

Surface revisions	Content revisions
Word-Usage/Clarity (WRD)	Claim (CLM)
Spelling and Grammar (SPL)	Evidence (EVD)
Organization (ORG)	Reasoning (RSN)
	Rebuttal (RBL)
	General Content Development (GCD)
	Precision (PRN)

Table 1: Available revision types in ArgRewrite V.2

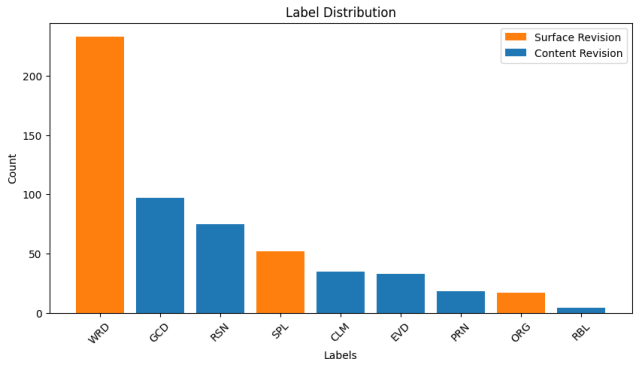

Figure 1: Repartition of the labels in the ArgRewrite test set (N = 564)

To proceed with the annotation of the revision purpose categories, the idea was to align sentences between the different drafts. The annotators then had a choice between different categories, which are shown in Table 1.

The repartition of these labels in the test dataset is shown in Figure 1. Among these, *Word-Usage / Clarity*, *Spelling and Grammar*, and *Organization* correspond to surface revisions, and the others are considered content revisions. The Fleiss κ between the three trained annotators was computed after a preliminary annotation of 5 revised essays to ensure consistency and understanding of the process. It reached 0.65 regarding the categories above and went up to 0.71 when considering a coarser-grained category scheme (composed only of surface vs content revisions). The whole corpus was then split into three equal parts, each of them being attributed to a unique annotator. These results underline the complexity of the task, even when it is done by humans. Therefore, we may not reasonably expect our models to outperform human-level agreement (and a too high accuracy/F1-Score might actually suggest some kind of overfitting).

4. METHOD

4.1 Mapping

To answer our first research question, we wanted to compare our results to those of Ruan et al. [14] as they used an LLM to predict revision types within their corpus. Their annotation grid and dataset are not the same as the one used with ArgRewrite V.2, so the comparison must be done with caution. Their grid has fewer classes, only five, which are Clarity, Grammar, Fact/Evidence, Claim, and finally, Other. Comparing the two annotation grids and the definitions of the categories existing in the articles, we established a mapping between them (see Table 2) to be able to compare our results more reasonably with theirs. Indeed, the *Word-usage/Clarity* category can be directly mapped to the *Clarity* category, which is defined as "altering word choice,e phrase usage, expressions and/or text format to be more formal, concise and understandable without meaning changes". The text format being mentioned here, the *Organization* label can also be mapped to this category. The *Conventions/Grammar/Spelling* category can be mapped to the *Grammar* category as it is defined as the correction of "any errors related to grammar and/or conventions to improve the language", explicitly quoting spelling earlier in the definition. Concerning content revisions, we chose to link the *Evidence* category to their *Fact/Evidence* category, defined as being a modification of "fact and/or evidence from third parties, or the author's factual manipulations and observations". We then mapped the different argumentative categories (*Claim*, *Warrant/Backing/Reasoning* and *Rebuttal*) to their *Claim* category defined as a change linked to "the claim, statement opinion, idea of the authors, or their overall aim of the document". Finally, we mapped the last categories (*Precision* and *General Content Development*) to *Other*, which was created for the revisions that did not match any of the categories defined earlier.

The different categories seem to be similar, but even if some disparities still exist, the main point is to be able to have a closer comparison by having the same number of categories between the two studies.

Original	Mapped Category in Ruan et al. [14]
Word-Usage/Clarity	Clarity
Organization	Clarity
Conventions/Gram./Spell.	Grammar
Evidence	Fact/Evidence
Claim	Claim
Warrant/Backing/Reasoning	Claim
Rebuttal	Claim
Precision	Other
General Content Development	Other

Table 2: Mapping between annotation schemes

4.2 In-context learning and prompting techniques

We designed prompts (available in the code) and compared several approaches to classify the revision of the dataset [16], basically doing in-context learning (ICL) while implementing a Chain-of-Thought (CoT) mechanism. As the inter-rater agreement is not very high on the fine-grained annotation task, we used OpenAI's GPT-4o and o3-mini models that have already shown interesting results in other reasoning contexts. Our first attempt, which we will refer to as **4o** in the rest of the article, consisted of a few-shot prompt including CoT, using GPT-4o. The second one will be referred to as **o3** and consists of the same attempts, using the o3-mini model. Then, we tried for both models to separate the task into two steps: the first one being classifying whether each revision concerns a surface or content revision, and then sub-

classifying it into the corresponding categories; these will be referred to as **4o-multiple-steps** and **o3-multiple-steps** further on. Finally, we tried to reverse the label and reasoning order in the model's output. Indeed, we previously asked the model to return the reasoning first, then the label, but Ruan et al. [14] showed that their results were significantly better when reversing those two steps. These last two attempts will be referred to as **4o-reverse** and **o3-reverse**.

4.3 Evaluation

In order to answer our first research question, we will compare our results to those of Ruan et al. using the mapping presented above. In their work [14], they evaluated their LLM predictors through their accuracies, average unweighted F1-score, and precision, recall, and F1-score for each category. Therefore, we used all these metrics, as well as Cohen's κ [1], to compare our model's results to those of Ruan et al. We remind that we can not directly compare Cohen's κ (with two annotators) to Fleiss's κ (for more than two annotators), but we can discuss the interpretation that comes out of these values.

To answer our second research question, we will compare our results to what was obtained on ArgRewrite V.2 by Kashefi et al., keeping the 9 original categories (i.e, not using the mapping presented above). The ArgRewrite V.2 classifiers were evaluated through an average unweighted F1-Score and accuracy (with detailed F1-Score for each label). Weighted scores were also computed using weights derived from the label distribution in the dataset.

We also kept a majority baseline (constantly predicting the most frequent label) to compare our models to a naive approach.

5. RESULTS

Concerning our first research question, Table 3 shows our results when mapping the labels to fit Ruan et al.'s classes, as shown in Table 2. We can see that our best models have performance that matches the previous work of Ruan et al. when looking at the overall accuracy. At the same time, the macro-average F1 scores are significantly higher.

Regarding our second research question, the overall accuracy and unweighted F1-Score, as well as the different F1-Scores for each class, are reported in Table 4. The best classifier from Kashefi et al.[9] (which is an XGBoost trained on semantic, textual, syntactic, and discourse features described earlier, with data augmentation) is better than all of our models that are still significantly better than the baseline consisting of selecting the majority class at each prediction. Weighted results not reported here were significantly higher than these, as low-represented classes have weaker results; however, we do not know them for the classifier from Kashefi et al., so we cannot compare them to previous results. Globally, it seems that the o3-based approaches perform better than their equivalent based on GPT-4o.

Table 5 shows the different Cohen's κ and unweighted F1-score when considering fine granularity (the 9 categories presented above) or coarse granularity (surface vs content revisions). Once again, all of our models are more performant than the majority baseline, but the best F1-score, even for the coarse granularity, is still the classifier from the previous work. We can also see that dividing the task into multiple steps helps to improve results on the coarse granularity, as the model can focus only on the question: is the revision changing or not the meaning of the text? The Cohen's κ are showing a moderate agreement in the case of fine granularity, and a substantial one (or even almost perfect one for o3-multiple-steps and o3-reverse), if we refer to the largely used, even if arbitrary, interpretation of Kappa proposed by Landis et Koch [11].

6. DISCUSSION

Regarding our first research question, we have achieved results comparable to those of Ruan et al., using our corpus and context with the new GPT-o3 mini model. Even if the overall accuracies are close, we get significantly higher F1-scores, suggesting that the approach may generalize to argumentation-related tasks in educational contexts despite the challenge that it represented. Indeed, revisions made by students could be imperfect (introducing grammatical errors while revising, supporting the wrong claim, or introducing out-of-topic ideas). We observed that the o3-mini model performed better than GPT-4o in most cases. Its "reasoning" nature might explain these slight differences that are still light, as we include Chain-Of-Thought in the process anyway (a comparison of the computation cost would have been interesting to complete the discussion if these models were open-source).

In contrast, for our second research question, our results did not match the performance of the XGBoost-based classifier introduced by Kashefi et al. [9]. However, the ability to capture and retain model reasoning offers a significant advantage over traditional classifiers. It enables interpretability and highlights ambiguous cases where the model's prediction may be justifiable despite diverging from manual annotations—particularly since the annotators' reasoning is not documented. Even when predictions are incorrect, providing learners with the model's reasoning can prompt critical reflection and evaluation, thereby initiating a metacognitive process that supports learning consolidation. For teachers, such explanations—despite occasional errors—can facilitate analysis of writing processes and foster trust in the tool by enabling informed correction of its outputs. Figure 2 illustrates such a case, with four very similar examples appearing consecutively in our test set.

However, our classifiers Cohen's κ values seem to indicate higher disagreement than between the human annotators for finer granularity. When it comes to coarser classification (between surface and content revision), we do have a Cohen's κ that is quite high with our best models, and that we can interpret as indicating an agreement close to the one between human annotators on that task (with a Fleiss's κ of 0.71). Therefore, it is logical that we still observe imperfect performance between our models and the given label on that coarse granularity, due to the inherent uncertainty over this label. While it would be interesting to conduct a new annotation with multiple annotators on this data, such an endeavor would be difficult to justify. The goal would be to determine whether new annotators align more with the LLM's classification—accompanied by its chain-of-thought reasoning—or with the original human annotator's decision.

Type of prompt	Accuracy	Macro-F1	Clarity	Grammar	Fact/Evidence	Claim	Other
Majority Baseline	0.46	0.13	0.58	0.00	0.00	0.00	0.00
Llama2-70b (Ruan et al.)	0.65	0.48	0.66	0.75	0.61	0.01	0.36
o3-multiple-steps	0.63	0.55	0.78	0.73	0.45	0.58	0.23
o3-reverse	0.63	0.60	0.78	0.74	0.54	0.55	0.42

Table 3: Results compared to Ruan et al. on their own dataset. Accuracy and Macro-F1 are reported alongside per-category F1-scores.

Type of prompt	Accuracy	Macro-F1	WRD	SPL	ORG	CLM	RSN	EVD	RBL	PRN	GCD
Majority Baseline	0.41	0.06	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XGBoost from Kashefi et al.	**0.68**	**0.56**	**0.78**	0.38	**0.47**	0.44	**0.60**	0.57	**0.56**	**0.66**	**0.59**
4o	0.44	0.37	0.65	0.73	0.21	0.30	0.28	0.63	0.18	0.11	0.22
o3	0.49	0.40	0.65	**0.74**	0.13	**0.48**	0.35	0.54	0.13	0.17	0.44
4o-multiple-steps	0.51	0.38	0.73	0.71	0.24	0.32	0.41	0.61	0.14	0.10	0.21
o3-multiple-steps	0.54	0.37	**0.78**	0.73	0.20	0.31	0.33	0.45	0.24	0.07	0.25
4o-reverse	0.45	0.37	0.63	**0.74**	0.16	0.29	0.21	**0.64**	0.13	0.21	0.32
o3-reverse	0.56	0.39	**0.78**	**0.74**	0.10	0.35	0.27	0.54	0.14	0.11	0.46

Table 4: Combined results on the ArgRewrite V.2 dataset. Accuracy and Macro-F1 are reported alongside per-category F1-scores. Bold scores are the best of the given column.

Text Before Revision:
"As The New York Times suggests, self-driving cars would greatly help elderly and disabled people travel."
Text After Revision:
"Furthermore, as The New York Times suggests, self-driving cars would greatly help elderly and disabled people travel."
Label: Precision
Prediction: Word-Usage/Clarity
Model's Reasoning: The revision adds the transition word *"Furthermore"* at the beginning of the sentence, enhancing the sentence's flow and connection to prior ideas without altering its factual content.

Figure 2: Example of a questionable revision label and the model prediction

Beyond our research questions, we would like to highlight, from a methodological perspective, a limitation frequently observed in studies involving LLMs across various contexts: their predominantly exploratory nature. While guidelines and emerging best practices are available, there remains a notable lack of systematic approaches in the application of these technologies within research settings to date. The framework DSPy [10] appears to be an interesting way to answer this problem by having a more systematic approach. This framework offers optimization algorithms along with other development tools to create and use LLM pipelines in complex use cases. It also allows users to switch easily between different language models. We made a first try using the BootstrapFewShotWithRandomSearch and a train-set representing 20% of the whole dataset (while keeping the same test set) as recommended. The results were not convincing enough to be reported here, and further work is needed to explore the use of this framework in our context.

7. CONCLUSION

This study examined the possibility of using LLM through ICL to automatically predict revision types in students' argumentative writings. While our results match those of other works using LLMs in different contexts [14], they are still weaker than the ones obtained through a trained classifier [9]. However, in our context, the ability to access the model's reasoning represents a significant advantage. Having fewer and clearer categories significantly improves the results. Further applications of this type of prediction could be made on existing taxonomies like the ones of Faigley and Witte [5] or Lindgren and Sullivan [12], which will be made easier thanks to our release source code. We should remind that the ArgRewrite V.2 corpus is based on undergraduate or graduate students; it would be interesting to see if this still works with high-school or middle-school students, as their writing process might be more chaotic and their revision intents less clear. This work could also be strengthened by incorporating the study of subsentential revisions that have not been considered here. Predicting sentential revision types could, for instance, be a multiple-step process where we first determine subsentential revision types and then the sentential revision depending on which subsentential revision is the most important, as described in [9].

These results are promising enough to consider a fully autonomous characterization pipeline of revision within keystroke data, a first step of automatic detection having been made in a previous work [13], inspired by [3]. Providing formative feedback on revision is challenging, as revision is largely an internal process [7]. Writers must interpret instructions, set appropriate goals, and continuously diagnose their text to identify discrepancies between its current state and intended outcomes—yet only the final revision is observable, which still a strong limitation. This makes it difficult to pinpoint the specific cognitive barriers students face. Some studies, however, suggest that informative feedback combined with guided self-reflection can enhance text quality [17], though it remains unclear whether the revision process

Type of prompt	Fine Granularity		Coarse Granularity	
	Cohen's κ	F1-score	Cohen's κ	F1-score
Majority baseline	0.00	0.06	0.00	0.24
XGBoost from Kashefi et al.	/	**0.56**	/	**0.93**
4o	0.43	0.37	0.56	0.77
o3	0.49	0.40	0.54	0.77
4o-multiple-steps	0.46	0.38	0.62	0.80
o3-multiple-steps	0.43	0.37	0.68	0.84
4o-reverse	0.40	0.37	0.59	0.79
o3-reverse	0.46	0.39	0.70	0.85

Table 5: Cohen's κ and unweighted F1-score for fine (9 categories) and coarse (2 categories) granularity for each prompting approach

itself was affected. Building on this work, we aim to leverage our precise automatic characterization of revision to design new feedback models, which we will evaluate in a dedicated experimental setting. Unlike conventional feedback that focuses primarily on the final product, this method emphasizes underlying cognitive and self-regulatory processes, which are critical for effective learning according to Hattie et al. [8].

8. REFERENCES

[1] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[2] R. Conijn and E. D. Speltz, "A Product- and Process-Oriented Tagset for Revisions in Writing," Written Communication.

[3] R. Conijn, E. Dux Speltz, and E. Chukharev-Hudilainen, "Automated extraction of revision events from keystroke data," Read Writ, Nov. 2021, doi: 10.1007/s11145-021-10222-w.

[4] W. Du, V. Raheja, D. Kumar, Z. M. Kim, M. Lopez, and D. Kang, "Understanding Iterative Revision from Human-Written Text," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 3573–3590. doi: 10.18653/v1/2022.acl-long.250.

[5] Faigley, Lester, and Stephen Witte. "Analyzing Revision." College Composition and Communication 32, no. 4 (December 1981): 400. https://doi.org/10.2307/356602.

[6] Flower, Linda, and John R. Hayes. "A Cognitive Process Theory of Writing." College Composition and Communication 32, no. 4 (December 1981): 365. https://doi.org/10.2307/356600.

[7] L. Flower, J.R. Hayes, L. Carey, K. Schriver, and J. Stratman. "Detection, Diagnosis, and the Strategies of Revision." College Composition and Communication 37, no. 1 (February 1986): 16. https://doi.org/10.2307/357381.

[8] J. Hattie and H. Timperley, "The Power of Feedback," Review of Educational Research, vol. 77, no. 1, Art. no. 1, Mar. 2007, doi: 10.3102/003465430298487.

[9] O. Kashefi et al., "ArgRewrite V.2: an Annotated Argumentative Revisions Corpus," Lang Resources and Evaluation, vol. 56, no. 3, pp. 881–915, Sep. 2022, doi: 10.1007/s10579-021-09567-z.

[10] O. Khattab et al., "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," Oct. 05, 2023, arXiv: arXiv:2310.03714. doi: 10.48550/arXiv.2310.03714.

[11] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[12] Lindgren, E., Sullivan, K. P. H. (2006). Analysing on-line revision. In G. Rijlaarsdam (Series Ed.) and K. P. H. Sullivan, E. Lindgren. (Vol. Eds.), Studies in Writing, Vol. 18, Computer Keystroke Logging: Methods and Applications (pp. 157–188). Oxford: Elsevier.

[13] Nebel, L., Bouchet, F., Luengo, V., Couraud, M., "Towards Automated Characterization of Revision Events in Student Writing" in Two Decades of TEL: from Lessons Learnt to Challenges Ahead Newcastle and Durham, UK, 15-19 September 2025, Proceedings (in press)

[14] Q. Ruan, I. Kuznetsov, and I. Gurevych, "Re3: A Holistic Framework and Dataset for Modeling Collaborative Document Revision," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 4635–4655. doi: 10.18653/v1/2024.acl-long.255.

[15] Q. Ruan, I. Kuznetsov, and I. Gurevych, "Are Large Language Models Good Classifiers? A Study on Edit Intent Classification in Scientific Document Revisions," Oct. 17, 2024, arXiv: arXiv:2410.02028. doi: 10.48550/arXiv.2410.02028.

[16] S. Schulhoff et al., "The Prompt Report: A Systematic Survey of Prompting Techniques," Dec. 30, 2024, arXiv: arXiv:2406.06608. doi: 10.48550/arXiv.2406.06608.

[17] N. Vandermeulen, M. Leijten, and L. Van Waes, "Reporting Writing Process Feedback in the Classroom. Using Keystroke Logging Data to Reflect on Writing Processes," Journal of Writing Research vol. 12 issue 1, pp. 109–140, Jun. 2020, doi: 10.17239/jowr-2020.12.01.05.

[18] Van Gelderen, A., & Oostdam, R. (2004). Revisions of form and meaning in learning to write comprehensible text. In: G. Rijlaarsdam (Series Ed.), L. Allal, L. Chanquoy, & P. Largy (Vol. Eds), Studies in writing: Volume 13. Revision: Cognitive and instructional processes (pp. 103–124). Dordrecht: Kluwer Academic Publishers.