# Proceedings of the 2nd Educational Data Mining in Writing and Literacy Instruction Workshop(WLIEDM)

Co-located with the 18th Educational Data Mining Conference (EDM 2025)

Edited by
Collin Lynch[1], Zhikai Gao[1], Damilola Babalola[1],
Piotr Mitros[2], Paul Deane[2]

Organized by
[1] ArgLab, North Carolina State University
[2] Educational Testing Service

July 2025, Palermo, Italy.

# Contents

**Preface** 3

**Organizing Committee and Program Committee List** 4

**Research Papers** 5

Learning to Grade Efficiently: A Bandit-Driven Prompt-Selection Framework for Low-Cost LLM Essay Scoring (Work-in-Progress) . . . . . 5

Characterizing revision events in students' writing processes using LLMs . . . . . 12

# Preface

This volume contains the proceedings of the selected papers of the 2nd Educational Data Mining in Writing and Literacy Instruction Workshop (WLIEDM), held on July 20, 2025, at Palermo, Italy.

The objective of this workshop is to facilitate discussion among the research community around Educational Data Mining (EDM) and AI in Writing and Literacy Education. Moreover, during a tutorial session, a prototype platform developed by the organizers was introduced to the participants. This platform is currently being developed to support ethical students' writing and learning data management.

We accepted two research papers. Each paper was peer reviewed by our program committee in a double-blinded way, and decisions were made based on these reviews, as well as discussions by the workshop organizers.

We would like to thank all the authors for their contributions and the reviewers for their valuable feedback. Special thanks to all the participants for making this workshop a success.

**WLIEDM Editors**

# Organizing Committee

**Collin Lynch** is an Associate Professor in the Department of Computer Science at North Carolina State University. His primary research is focused on developing robust ITS and adaptive educational systems for Ill-Defined domains such as scientific writing, law, and software development. His current research includes work on argument mining and natural language processing, real-time support for classroom orchestration and writing to learn tasks, advances in student modeling, the development of embodied cognitive agents for collaborative learning, and scaffolding for CS education.

**Paul Deane** is a principal research scientist in the Research & Development division at ETS. He is the author of Grammar in Mind and Brain, a study of the interaction of cognitive structures in syntax and semantics, and the second author of Vocabulary Assessment to Support Instruction. His current research interests include formative assessment design in the English language arts, cognitive models of writing skills, automated essay scoring, and vocabulary assessment. During his career at ETS, he has worked on a variety of natural language processing (NLP) and assessment projects, including automated item generation, tools to support verbal test development, scoring of collocation errors, reading and vocabulary assessment, and automated essay scoring.

**Piotr Mitros** is a Senior Research Scientist at ETS. He is also the original author of the popular Open edX learning platform and the original founder as well as the Chief Scientist for more than five years. He has spent the past few years exploring issues around why educational initiatives go south, and evidence-based practices aren't adopted and converged on issues around governance, transparency, and incentive structures. His current work focuses on how we develop educational measurements that incentivize and support rich classroom instruction supporting diverse (rather than standardized) students.

**Zhikai Gao** is a recent Ph.D. graduate at North Carolina State University. He is currently an Assistant Professor in Western Carolina University. His current research focuses on understanding students' learning behaviors through traceable log data from ITS, CS education, help-seeking behavior, and LLM usage in education across disciplines.

**Damilola Babalola** is a third-year Computer Science Ph.D. student at North Carolina State University with a research focus on using Artificial Intelligence (Educational Data Mining and Natural Language Processing) to improve Education. His current work involves research, software development, data mining, and data visualizations aimed at assisting middle-school and high-school students in improving their essay-writing skills. The core of his research centers around the extraction and classification of student essay revisions based on their edit intention, followed by the visualization of student clusters exhibiting similar revision patterns.

# Program Committee

**Effat Farhana:** Auburn University
**Yuqi Wu:** North Carolina State University
**Caleb Scott:** North Carolina State University

# Learning to Grade Efficiently: A Bandit-Driven Prompt-Selection Framework for Low-Cost LLM Essay Scoring (Work-in-Progress)

Olga Manakina
Carleton University
olgamanakina@cmail.carleton.ca

Igor Bogdanov
Carleton University
igorbogdanov@cmail.carleton.ca

## ABSTRACT

Large Language Models (LLMs) demonstrate strong capabilities in automated essay scoring (AES), but contemporary approaches typically employ fixed prompt selection, failing to address operational cost concerns and evolving optimal configurations. We propose a novel cost-aware approach that treats each prompt type as an arm in a multi-armed bandit (MAB) controller, enabling adaptive selection of optimal prompting strategies during inference. Our experiments on IELTS Writing Task 2 essays show that the MAB framework achieves comparable scoring accuracy to exhaustive grid search while reducing LLM calls by 78.4% to find the best grading approach (Table 1). We implemented four distinct grading recipes (multi-step vs. single-step assessment, with vs. without calibration examples) and found that the multi-step approach with examples achieves the highest accuracy. By tracking token usage and latency alongside agreement metrics, we produce the first cost-reliability learning curves for essay scoring, providing actionable insights for educational technology platforms that must balance operational costs against assessment validity. This work represents the first application of online control mechanisms to adaptively select prompting strategies in AES, transforming prompt selection from an offline hyperparameter optimization problem into an efficient online learning task.

## Keywords

Automated essay scoring, Large language models, Prompt optimization, Multi-armed bandit (adaptive selection), Cost-efficient assessment

## 1. INTRODUCTION

Recent research has shown that Large Language Models (LLMs) have demonstrated significant capabilities in automated essay scoring (AES) when using well-designed prompts. Various prompting techniques for AES have been investigated, including rubric decomposition [8], few-shot approaches [16], chain-of-thought reasoning [1], and comparative judgment methods [7]. Despite these advances, contemporary approaches typically employ fixed prompt selection; researchers either predetermine a template or conduct exhaustive evaluations on a restricted set of options before adopting a standardized solution [15]. This static approach fails to address two key practical limitations in real-world assessment scenarios:

1. Inference costs are heavily influenced by prompt length and call frequency [11]. Rich prompting strategies that incorporate few-shot examples or detailed rationales can consume more tokens than minimalist instructions, which can significantly impact operational costs.
2. The optimal prompt configuration is not static; it evolves as models are updated, pricing structures change, or essay characteristics shift over time. In high-stakes educational settings, such as international language testing programs or large-scale MOOC platforms, applying every essay to every possible prompt configuration is financially unsustainable.

We propose a novel, cost-aware framework that treats each prompt-model combination as an arm in a multi-armed bandit (MAB) controller. During the scoring process, the agent observes a reward signal based on the agreement between the LLM's predicted score and the examiner's ground-truth score. This enables the system to incrementally concentrate calls on the most reliable and cost-effective prompt configurations. This approach transforms prompt selection from an offline hyperparameter optimization problem into an online learning task, similar to recent MAB optimizers for prompt engineering [13] and retrieval-augmented generation [3].

Our preliminary experiments on the IELTS Writing Task 2 corpus from the IELTS Writing Scored Essays Dataset [6] demonstrate the effectiveness of the framework. The bandit-based approach achieves comparable Quadratic Weighted Kappa (QWK) scores to exhaustive grid search while requiring approximately one-tenth of the LLM calls, showing significant potential for reducing operational costs without compromising scoring quality.

This work makes two key contributions. First, to our knowledge, it represents the first application of online control mechanisms, whether bandit-based, reinforcement learning, or otherwise, to select prompting strategies in AES in an adaptive manner. All prior published work has relied on

**Table 1: Detailed Resource Consumption: MAB vs. Grid Search**

| Method | Calculation | LLM Calls | Tokens |
|---|---|---|---|
| Grid Search | 787 essays × 4 recipes<br>Multi-Step: 787 × 4 calls × 2 recipes = 6,296<br>Single-Step: 787 × 1 call × 2 recipes = 1,574<br>Total: 6,296 + 1,574 = 7,870 | 7,870 | 10,915,411 |
| MAB | 500 total essays (each assigned to one of 4 recipes)<br>Multi-Step+Ex: 332 × 4 calls = 1,328<br>Multi-Step-NoEx: 67 × 4 calls = 268<br>Single-Step+Ex: 56 × 1 call = 56<br>Single-Step-NoEx: 45 × 1 call = 45 | 1,697 | 2,964,444 |
| Reduction | | 78.4% | 72.8% |

fixed prompt templates. Second, by tracking token usage and latency alongside agreement metrics, we produce the first cost-reliability learning curves for essay scoring. These findings offer actionable insights for test providers and educational technology platforms that must strike a balance between cost considerations and psychometric validity.

This paper reports work in progress. To date, we have implemented the bandit controller with four prompting approaches (single-step with examples, single-step without examples, multi-step with examples, and multi-step without examples). The four approaches are described in Section 3. For this initial stage of our study, we have used Google Gemini Flash 2.5. In our future steps, we will extend the study to additional models (GPT-4, Llama-3 70B) and the ASAP dataset [5] to assess generalizability. By releasing our code and policy logs, we aim to stimulate further research into adaptive, cost-efficient LLM-based grading in educational assessment.

## 2. RELATED WORK

The related work review is organized into three parts. We begin with LLM-based essay-scoring studies, summarizing research on zero-shot prompting, rubric-aligned, few-shot, and chain-of-thought methods, and noting their reliance on static, cost-blind grid searches. Next, we move to the writing domain to survey adaptive prompt and model-selection techniques in NLP, focusing on multi-armed bandit and other online controllers that optimize accuracy–cost trade-offs for retrieval and question-answering tasks. Finally, we summarize the public essay datasets that underpin most evaluations and pinpoint the still-unfilled gap: no prior work combines these adaptive controllers with essay scoring. This structure clarifies how our study builds directly on advances in prompt engineering and presents the first cost-aware adaptive framework for automated essay scoring.

### 2.1 LLM-Based Automated Essay Scoring and Prompting Strategies

Large language models (LLMs) have recently been shown to be effective in grading student writing, eliminating the need for task-specific fine-tuning. Lee et al.'s Multi-Trait Specialization (MTS) framework shows that a zero-shot, rubric-decomposed prompt can lift GPT-3.5 and Llama-2-13B to state-of-the-art Quadratic-Weighted-Kappa (QWK) on the ASAP and TOEFL11 benchmarks, outperforming a straightforward prompting instruction by up to 0.35 QWK [8].

Stahl et al. [15] compare zero-shot, one-shot, and few-shot prompts (with and without chain-of-thought) for joint scoring and feedback generation, finding that AES accuracy improves modestly when the model is asked to explain its scores, but at the cost of much longer prompts.

Alternative prompting paradigms include comparative judgement: Kim & Jo [7] report that GPT-4, asked repeatedly to pick the better of two essays, surpasses a rubric-based direct-scoring prompt on the IELTS and ASAP sets [5].

For multi-trait scoring, Chu et al. [1] generate trait-wise rationales with GPT-4 and feed them to a smaller student model; the rationale-augmented scorer beats strong baselines on ASAP++ [10]and Feedback-Prize [12] datasets while offering transparent explanations.

### 2.2 Cost-efficiency

Most LLM-AES studies evaluate a small, fixed set of prompts offline and then deploy the single best template. Token budgets are rarely reported, even though few-shot prompts, including rubric and rationale, can exceed the context window of mid-tier LLMs. A recent study on rubric granularity by Yoshida [16] shows that a simplified rubric maintains accuracy for three of four LLMs while cutting prompt length by more than half, underscoring the need for cost-sensitive experimentation. Nevertheless, the dominant evaluation paradigm remains static grid search: every prompt or configuration of prompts is tried on every essay, and the winner is chosen post-hoc.

### 2.3 Adaptive Prompt and Model Selection in NLP

Outside essay scoring, prompt and model choice have been framed as online decision problems. A framework TRIPLE, proposed by Shi et al. [13] connects prompt optimization to fixed-budget best-arm identification and shows that a multi-armed-bandit (MAB) policy can identify a high-performing prompt on several NLP benchmarks while using only 50–80% of the LLM calls required by exhaustive search.

In retrieval-augmented generation, AutoRAG-HP [3] formulates hyperparameter tuning (e.g., k retrieved documents, prompt-compression ratio) as a hierarchical MAB; it matches grid-search recall with roughly 20% of the API queries.

Somerstep et al.[14] introduced CARROT (Cost AwaRe Rate

Optimal Router), which applies a contextual bandit router to select, for each query, the cheapest LLM that still meets a target quality level, yielding substantial cost savings without compromising quality.

These successes demonstrate that adaptive controllers can maintain task performance while reducing token budgets; yet, none of these techniques has been incorporated into automated essay scoring. Our work addresses that gap by embedding an MAB inside an AES pipeline to select among four prompting techniques on the fly, delivering human-level reliability with a significant reduction in tokens.

### 2.4 Public Essay-Scoring Datasets and the Remaining Gap

LLM-AES research is typically benchmarked on the ASAP corpus (eight prompts, $\approx$ 12k essays) [5], the TOEFL11 corpus of 12,100 non-native essays [2], and the newer IELTS Writing Band-Score set ($\approx$ 1200 essays) [6].

All published LLM studies apply fixed prompt templates to these datasets; none employ bandits or any online search to allocate prompt-model calls. Consequently, our study is the first to embed a bandit controller inside an LLM-based AES system, adaptively routing each essay to the most cost-effective prompting strategy and thereby uniting efficiency research in NLP with high-stakes writing assessment.

## 3. METHODOLOGY

This study employs a novel approach to automated IELTS essay grading using LLMs with MAB optimization. We divided the IELTS dataset [6] into Task 1 and Task 2 subsets, focusing exclusively on Task 2 essays for this analysis. We developed a modular system with four distinct grading "recipes": multi-step criteria-based assessment with and without essay examples, and single-step direct scoring with and without examples. The multi-step approach separately evaluates Task Response, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy (official IELTS key assessment criteria [4]) before calculating an overall score, while single-step methods directly predict the overall band score. For optimization, we implemented an epsilon-greedy MAB algorithm that balances exploration and exploitation to identify the most effective grading strategy. The reward function incorporates both prediction accuracy (measured by the negative absolute error against human scores) and optionally penalizes token usage for efficiency. We also conducted exhaustive grid search evaluations as a baseline comparison. All experiments utilized Google's Gemini 2.5 model, selected for its balance of performance accuracy, low latency, and cost-effective API access. Experiments were performed on the Task 2 essay subset with human-assigned scores, employing concurrent processing to maximize throughput. Performance metrics included Mean Absolute Error (MAE) against human scores, token utilization, latency, and estimated API costs, allowing for a comprehensive assessment of each recipe's effectiveness and efficiency. The complete implementation code is publicly accessible on GitHub [9].

### 3.1 Dataset

For our experiments, we utilized the IELTS Writing Scored Essays Dataset, a publicly available resource on Kaggle. This collection comprises 787 Academic Task 2 compositions, each accompanied by an official band score ranging from 1 to 9, assigned by qualified IELTS examiners. The compositions respond to conventional Task 2 instructions that challenge candidates to develop arguments or present perspectives on various societal and academic subjects. This corpus serves as an ideal testing ground for evaluating automated grading systems within the context of authentic, consequential language assessment.

### 3.2 Dynamic Prompt Assembly

Our system employs a sophisticated two-tiered approach to prompt engineering. The core architectural design separates prompt templates from their dynamic instantiation, enabling systematic experimentation with different prompting strategies. The foundation of our approach is the prompt library module, which serves as a centralized repository of templated prompts. It contains detailed system prompts that establish the LLM's role as an IELTS examiner, along with specific instructions for each assessment criterion. For calibration purposes, some templates incorporate annotated sample essays (high and low-scoring) to provide reference points for the model's evaluation. The library maintains distinct template variants for each experimental condition (multi-step vs. single-step assessment, with vs. without examples). At runtime, the essay grader agent dynamically assembles the final prompts through the following process:

1. Selection of appropriate system and user templates based on the current grading recipe
2. Injection of the specific essay question and text into the templates via string formatting
3. Construction of properly structured message dictionaries with distinct "system" and "user" roles
4. Assembly of these messages into the final prompt sequence transmitted to the LLM

This modular design allows us to systematically compare different prompt structures while maintaining consistent content across experimental conditions. For multi-step assessment recipes, the system sequentially generates criterion-specific prompts for Task Response, Coherence & Cohesion, Lexical Resource, and Grammatical Range & Accuracy, before programmatically calculating the overall score. In contrast, single-step recipes construct comprehensive prompts requesting direct overall assessment, with or without calibrating examples.

## 4. RESULTS

### 4.1 Learning Behavior and Approach Selection

The Multi-Armed Bandit (MAB) algorithm demonstrated clear preferences among the four grading recipes throughout our experiments. Figure 1 illustrates the cumulative average shaped reward for each approach over 500 steps. After an initial exploration phase with considerable fluctuation,

the algorithm's assessment stabilized around step 100. The multi-step approach with calibration examples (Multi-step Ex) consistently achieved the highest reward values (approximately -0.75), followed by the single-step approach with examples (Single-step Ex) at around -1.0.

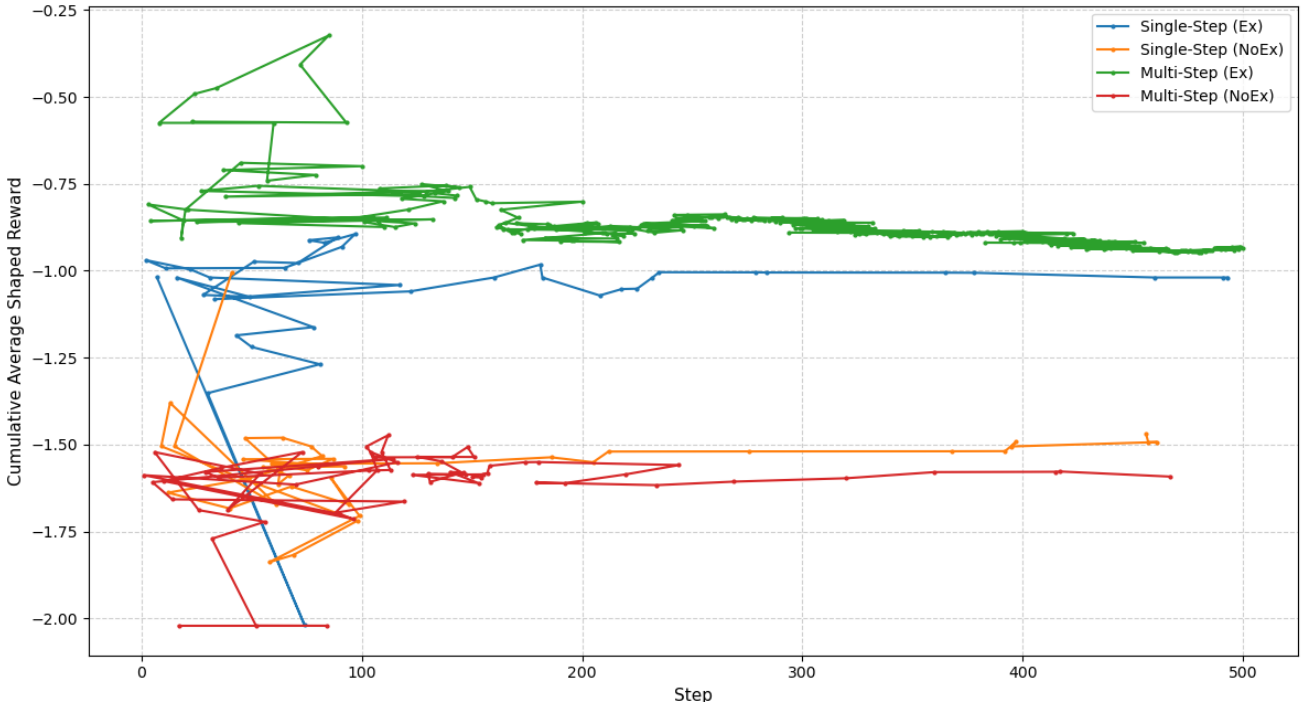


**Figure 1: Cumulative average shaped reward over steps**

The approaches without examples performed notably worse, with Single-step NoEx at approximately -1.5 and Multi-step NoEx showing the lowest rewards at about -1.6. This learning behavior directly influenced arm selection frequency, as depicted in Figure 2. The MAB algorithm favored the Multi-step Ex recipe, allocating approximately 350 pulls to this approach, over 70% of the total experiment. The remaining three approaches received similar, but much lower attention, with each receiving between 40 and 65 pulls, demonstrating the algorithm's strong preference for the most effective recipe.

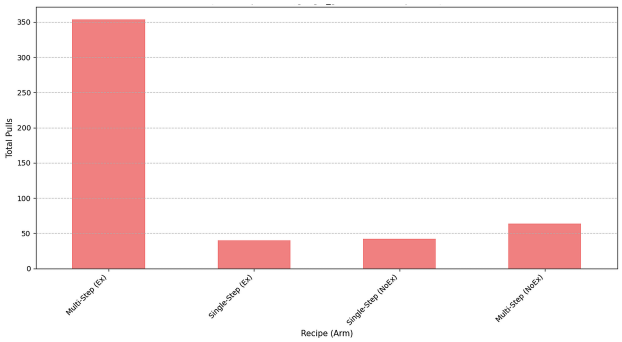


**Figure 2: Distribution of arm pulls across the four grading recipes during the MAB experiment. The histogram shows the total number of times each recipe was selected by the algorithm, reflecting its learned preferences.**

## 4.2 Scoring Accuracy

The MAB's preference for the Multi-step Ex approach is justified by its superior accuracy metrics. Figure 3 displays the Mean Absolute Error (MAE) for each recipe, revealing that Multi-step Ex achieved the lowest error rate (approximately 0.85), followed by Single-step Ex (1.0). The approaches without examples performed substantially worse, with Single-step NoEx showing an MAE of approximately 1.45 and Multi-step NoEx demonstrating the highest error rate at 1.55.

Our Quadratic Weighted Kappa (QWK) analysis in Figure 4 corroborates these findings. QWK is a statistical measure of inter-rater reliability that accounts for both agreement and the magnitude of disagreement between ratings, making it particularly suitable for educational assessment tasks with ordinal scales. Unlike MAE, which treats all disagreements linearly, QWK penalizes larger disagreements more heavily, with values ranging from 0 (no agreement) to 1 (perfect agreement). The analysis shows that Multi-step Ex achieved the highest QWK score (approximately 0.55) across both MAB and Grid Search experiments, indicating the strongest agreement with human graders. Interestingly, Multi-step NoEx performed second-best in QWK (around 0.35), despite its poor MAE, suggesting that it occasionally

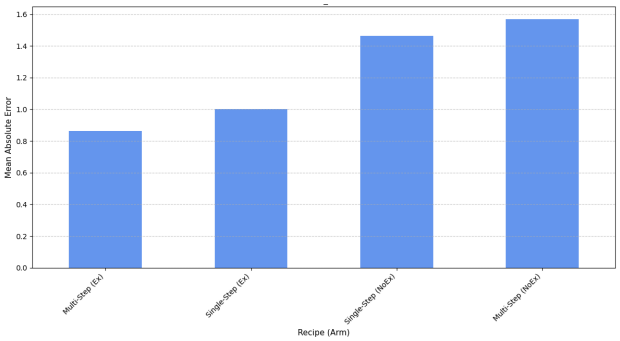


**Figure 3: Mean Absolute Error (MAE) between predicted and human-assigned scores for each grading recipe in the MAB experiment. Lower values indicate greater scoring accuracy**

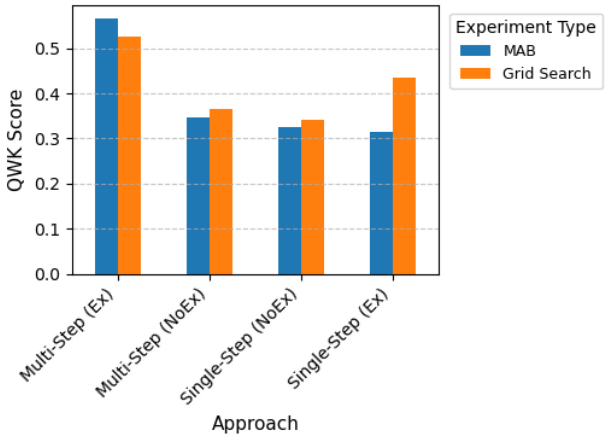


**Figure 4: Quadratic Weighted Kappa (QWK) scores comparing agreement with human graders for each recipe across both MAB and Grid Search experiments. Higher values indicate better agreement with human assessments. The comparison shows consistency between experimental approaches.**

captures assessment patterns that align with human judgment, despite the higher average error.

## 4.3 Cost-Efficiency Tradeoffs

While the Multi-step Ex approach demonstrated superior accuracy, our cost analysis revealed important efficiency implications across recipes. Figure 5 plots MAE against average estimated API cost per grading attempt, highlighting the accuracy-cost tradeoff. Multi-step Ex, while most accurate (MAE of 0.85), incurred a relatively high cost of approximately $0.0011 per essay. Single-step Ex offered a balanced alternative with moderate accuracy (MAE of 1.0) at a lower cost ($0.0003). The Single-step NoEx approach provided the most economical option at $0.0002 but with higher error (MAE of 1.45), while Multi-step NoEx performed poorly on both metrics with the highest error (MAE of 1.55) at a moderate cost ($0.0004).

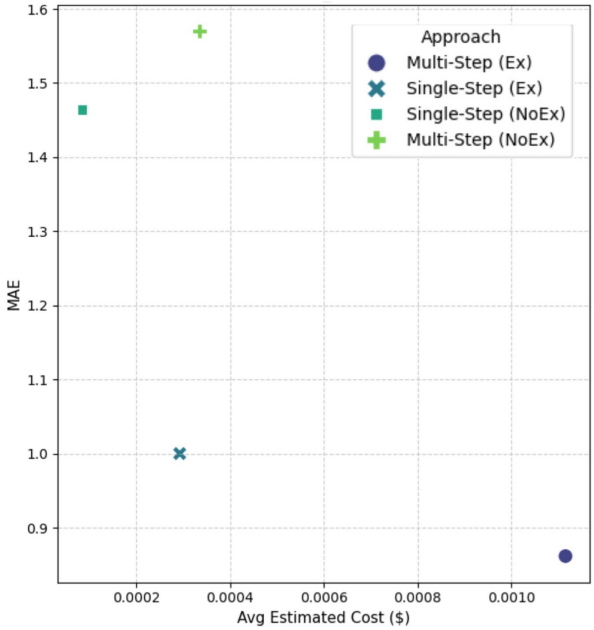


**Figure 5: Accuracy–cost tradeoff for four grading recipes. Multi-Step (Ex) achieves the lowest error (MAE 0.85) at a higher average API cost of about $0.0011 per essay. Single-Step (Ex) provides a balanced option (MAE 1.0 at $0.0003). Single-Step (NoEx) is the most economical ($0.0002) but with higher error (MAE 1.45), while Multi-Step (NoEx) performs worst overall (MAE 1.55 at $0.0004)**

The cumulative token consumption comparison in Figure 6 demonstrates the significant efficiency advantage of MAB over Grid Search.

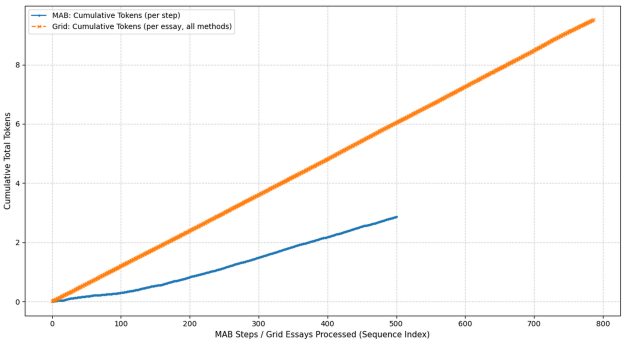


**Figure 6: Cumulative token consumption over time for MAB versus Grid Search approaches. The steeper slope of the Grid Search curve demonstrates the efficiency advantage of the adaptive MAB strategy, which consumes approximately one-fourth the tokens for comparable coverage.**

While Grid Search exhaustively evaluates all approaches for each essay, consuming tokens at approximately five times the rate of MAB, the adaptive MAB strategy selectively allocates resources to promising approaches. At the experi-

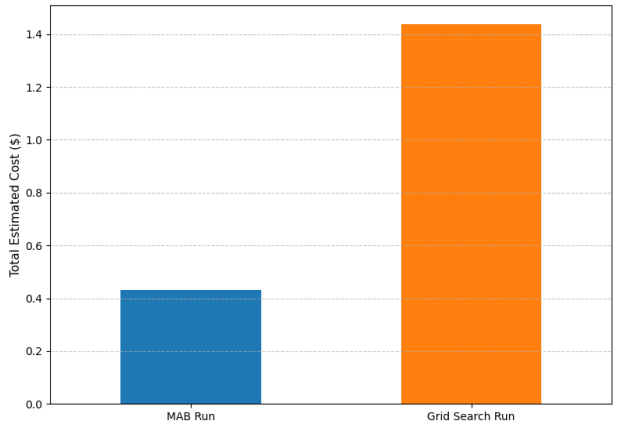


**Figure 7: Total estimated API cost comparison between MAB and Grid Search experiments using Google's Gemini 2.5 model. The MAB approach demonstrates approximately 75% cost reduction while maintaining comparable grading accuracy.**

ment conclusion, Grid Search had consumed approximately 9 million tokens compared to MAB's 1.8 million for similar coverage. This efficiency translated directly to cost savings, as shown in Figure 7, where MAB reduced total experimental costs by approximately 70% compared to Grid Search ($0.4 versus $1.4) while maintaining comparable accuracy outcomes.

## 4.4 Ablation Study

To investigate the impact of prompt design on grading accuracy, we conducted an ablation study focusing on the inclusion of detailed explanations for assessment criteria. In all prompts for our primary experiments, we incorporated comprehensive descriptions of how each IELTS criterion is evaluated, stated in the official IELTS Writing Key Assessment Criteria document [4]. For example, the Task Response assessment guidelines include detailed points such as: 1) How fully the candidate responds to the task; 2) How adequately the main ideas are extended and supported; 3) How relevant the candidate's ideas are to the task; 4) How clearly the candidate opens the discourse, establishes their position, and formulates conclusions; 5) How appropriate is the format of the response to the task.

Similar detailed explanations were provided for Coherence & Cohesion, Lexical Resource, and Grammatical Range & Accuracy criteria in their respective prompts. We then compared these results against a control experiment using simplified prompts without these detailed assessment guidelines. Our ablation study revealed a surprising finding: removing detailed assessment criteria descriptions from prompts actually improved grading accuracy across all approaches. The Multi-Step with Examples recipe without explicit rubrics achieved an MAE of 0.862 and QWK of 0.566, outperforming its rubric-enhanced counterpart (MAE 0.965, QWK 0.485) while consuming fewer tokens (7,402 vs. 7,862) and reducing costs. This suggests that detailed rubric explanations may introduce noise or unnecessary constraints that interfere with the LLM's inherent understanding of essay quality. The MAB controller showed a stronger preference for the simplified approach, allocating 70.8% of pulls to Multi-Step with Examples in the Basic experiment versus 66.4% in the Detailed version. These results challenge the conventional wisdom that more detailed assessment guidelines necessarily lead to better automated scoring performance.

# 5. DISCUSSION & FUTURE WORK

Our findings demonstrate that adaptive prompt selection via MAB can achieve substantial cost savings without compromising scoring quality, a finding that is particularly valuable for large-scale assessment programs. Surprisingly, our ablation study revealed that simplified prompts without detailed rubric explanations outperformed comprehensive ones, suggesting LLMs may have internalized academic writing assessment norms during pre-training. The strong performance of multi-step approaches aligns with human assessment practices and may apply to other LLM assessment applications.

This is a work-in-progress in paper; therefore, it has several limitations that we are planning to address in our next steps. At this initial stage, we used only one LLM model (Gemini 2.5) exclusively on IELTS Task 2 essays, and our epsilon-greedy MAB implementation could be further refined. We maintained a constant exploration rate ($\epsilon = 0.2$) throughout our experiments to ensure sufficient exploration of all arms, although adaptive $\epsilon$ strategies will be investigated in future work. Future work will also explore contextual bandits that incorporate essay-specific features, evaluate performance across diverse models and essay types, develop dynamic cost models that adapt to changing pricing structures, and investigate hybrid approaches that combine single-step efficiency with multi-step accuracy when needed.

# 6. CONCLUSION

This paper introduced a MAB-based approach to automated essay scoring that adaptively selects optimal prompting strategies. Our initial experiments demonstrated comparable accuracy to exhaustive grid search while reducing LLM calls by 78.4% and token consumption by 72.8% (Table 1). Multi-step assessment with calibration examples consistently performed best (MAE 0.86, QWK 0.57), while, surprisingly, simplified prompts outperformed those with detailed rubric criteria. By framing prompt selection as an online learning problem rather than static hyperparameter optimization, our approach enables assessment systems to continuously adapt to changing model capabilities, pricing structures, and essay characteristics. As LLMs integrate into high-stakes assessment, such frameworks that balance accuracy, cost, and adaptability will become increasingly valuable.

# 7. REFERENCES

[1] S. Chu, J. Kim, B. Wong, and M. Yi. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. *arXiv preprint arXiv:2410.14202*, 2024.

[2] ETS. Toefl11: A corpus of non-native english. `https://www.ets.org/research/policy_research_reports/publications/report/2013/jrkv.html`, 2013. Accessed: 2024.

[3] J. Fu, X. Qin, F. Yang, L. Wang, J. Zhang, Q. Lin, Y. Chen, D. Zhang, S. Rajmohan, and Q. Zhang. Autorag-hp: Automatic online hyper-parameter tuning for retrieval-augmented generation. *arXiv preprint arXiv:2406.19251*, 2024.

[4] IELTS. Ielts writing: Key assessment criteria. `https://assets.ctfassets.net/unrdeg6se4ke/2IRKF0bLwaZSzKUhqgSYn/d3fc8b11109e0da3ec5d65d705c1c3da/ielts-writing-key-assessment-criteria.ashx.pdf`, 2024.

[5] Kaggle. Asap-aes: Automated essay scoring. `https://www.kaggle.com/competitions/asap-aes`, 2024. Accessed: 2024.

[6] Kaggle. Ielts writing band scores dataset. `https://www.kaggle.com/datasets/mazlumi/ielts-writing-scored-essays-dataset`, 2024. Accessed: 2024.

[7] S. Kim and M. Jo. Is gpt-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 315–319, 2024.

[8] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu. Unleashing large language models' proficiency in zero-shot essay scoring. *arXiv preprint arXiv:2404.04941*, 2024.

[9] O. Manakina. Automated essay scoring system. `https://github.com/oemanakina/aes-agent`, 2025.

[10] S. Mathias and P. Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[11] T. McDonald, A. Colosimo, Y. Li, and A. Emami. Can we afford the perfect prompt? balancing cost and accuracy with the economical prompting index. *arXiv preprint arXiv:2412.01690*, 2024.

[12] A. Picou, A. Franklin, Maggie, M. Benner, P. Baffour, P. Culliton, R. Holbrook, S. Crossley, Terry_yutian, and ulrichboser. Feedback Prize – evaluating student writing. `https://www.kaggle.com/competitions/feedback-prize-2021`, 2021. Kaggle competition.

[13] C. Shi, K. Yang, J. Yang, and C. Shen. Best arm identification for prompt learning under a limited budget. *arXiv preprint arXiv:2402.09723*, 2024.

[14] S. Somerstep, F. M. Polo, A. F. M. de Oliveira, P. Mangal, M. Silva, O. Bhardwaj, M. Yurochkin, and S. Maity. Carrot: A cost aware rate optimal router. *arXiv preprint arXiv:2502.03261*, 2025.

[15] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*, 2024.

[16] L. Yoshida. Do we need a detailed rubric for automated essay scoring using large language models? *arXiv preprint arXiv:2505.01035*, 2025.

# 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**

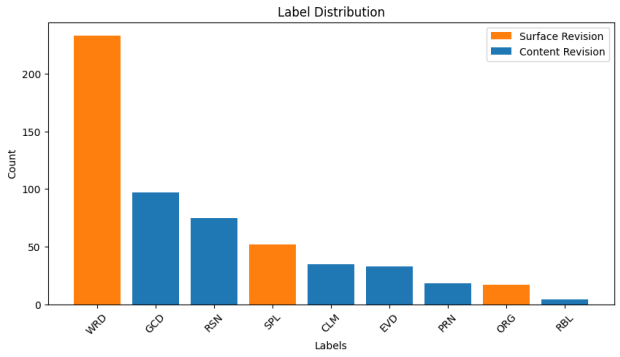


**Figure 1: Repartition of the labels in the ArgRewrite test set (N = 564)**

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 $\kappa$ 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 $\kappa$ [1], to compare our model's results to those of Ruan et al. We remind that we can not directly compare Cohen's $\kappa$ (with two annotators) to Fleiss's $\kappa$ (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 $\kappa$ 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 $\kappa$ 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 $\kappa$ 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 $\kappa$ 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 $\kappa$ 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 of 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 |
| Best Performance in [9] | **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 itself was affected. Building on this work, we aim to leverage

| Type of prompt | Fine Granularity | | Coarse Granularity | |
|---|---|---|---|---|
| | Cohen's $\kappa$ | F1-score | Cohen's $\kappa$ | F1-score |
| Majority baseline | 0.00 | 0.06 | 0.00 | 0.24 |
| Best Performance in [9] | / | **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 $\kappa$ and unweighted F1-score for fine (9 categories) and coarse (2 categories) granularity for each prompting approach**

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.