

Article

AI-Based Assessment of Revision-Associated Research Writing Development in a Peer-Feedback-Supported

Hamed Hilal AlYahmadi 

Faculty of Education, Arab Open University, Muscat 130, Oman; hamed.y@aou.edu.om

Abstract

Despite its ubiquity as a writerly practice in academic writing, there is little publicly available evidence that connects feedback-based revision practices with quantitative changes in research writing. This study proposes a framework based on artificial intelligence (AI) that aids in the assessment of revision-related development in student research exposés based on the Exposía academic writing and peer-feedback corpus (Ec). The raw files analyzed included 3247 records, with 16,068 feedback and comment records, 55 matched draft–final pairs of exposés, and 28 matched draft–final pairs of cases by score. For every text, features of NLP were computed, such as: word count, type–token ratio, moving-average type–token ratio over 100-word windows (MATTR-100), lexical density, readability, and adjacent-sentence cohesion. The differences between the draft and the final text were analyzed with paired statistical tests, and the human assessment scores were estimated using draft-Ridge Regression and draft-Random Forest models. Human assessment scores increased significantly from the draft ($M = 5.26$) to the final version ($M = 7.55$), $t(27) = 9.04$, $p < 0.001$, Cohen’s $d_z = 1.709$. There was also a significant increase in word count, MATTR-100, and lexical density. Random Forest gave a better R^2 (0.426) than Ridge Regression ($R^2 = 0.248$). The exploratory feedback-improvement model resulted in a negative R^2 due to a limited number of automatically merged feedback features, a lack of alignment between comment and revision, and a small number of matched scores. Therefore, AI-supported analysis identified writing development associated with revision and captured human judgment to some extent, but did not provide causal evidence of the independent contribution of peer feedback to writing improvement due to the observational design. The framework offers a repeatable benchmark and calls for enhanced detail in providing feedback–revision alignment.

Keywords: AI-based writing evaluation; peer feedback; research writing; natural language processing; automated assessment; Exposía corpus; machine learning; writing development



Received: 13 June 2026

Revised: 26 July 2026

Accepted: 29 July 2026

Published: 31 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Research writing is one of the key competencies in higher education as it involves identifying, stating, and solving a research problem, summarizing the published literature, developing an evidence-based argument, and presenting ideas in a coherent and scholarly manner [1,2]. Instructors have historically assessed these skills using a rubric-based process, written feedback, and ongoing supervision. While human evaluation is pedagogically significant, it is expensive, time-consuming, difficult to scale, and may be subject to inter-rater variability. Artificial intelligence (AI) and natural language processing (NLP) can then be used to aid, not replace, human assessment by identifying measurable signs of writing progression in draft and completed writing [3,4].

Peer feedback is a common pedagogical practice in which students provide evaluative and constructive feedback on other students' work. Research has shown that peer feedback can encourage metacognitive reflection, awareness of audience, and revision behavior both in providing and receiving [5,6]. However, the literature offers conflicting findings on whether peer feedback results in an improvement in the quality of writing beyond the superficial changes that it promotes [7,8]. Previous studies have used summary scores or qualitative content analysis, and whether computational text analytics can identify the type of linguistic and structural changes that accompany significant progress in writing remains unknown [9,10]. With the development of large-scale academic writing corpora came the opportunity to investigate writing and revision processes by using data. One such resource is the Exposía academic writing and peer-feedback corpus (Ec), a publicly available dataset containing student research exposés in draft and final forms, comments and assessment scores provided by students, peers, and teachers, inline annotations, and structured assessment scores [11]. Ec supports reproducible and publicly accessible research on writing development in a peer-feedback-supported environment [12].

Although they offer these opportunities, most studies have not integrated and applied AI-based text analytics and machine-learning-based scoring within a single framework that not only quantifies writing changes but also investigates the predictive power of linguistic features for human assessment scores and whether feedback-related features can be leveraged as indicators of writing improvement. In the past, the majority of the automated writing evaluation (AWE) systems have been limited to providing either scores or diagnostic feedback, and only a few systems have combined both aspects in one workflow [13,14].

This study would fill this gap by proposing an assessment framework that integrates AI tools to examine revision-associated research writing development in a peer feedback-supported corpus. It combines NLP feature extraction, statistical analysis, and supervised machine learning. It first measures differences between a draft and a final version of an expository text using AI metrics. It then tests whether linguistic features can serve as a proxy for human assessment scores. Lastly, it explores whether automatically derived feedback features can serve as useful signals to facilitate writing improvement and explicitly acknowledges that the available data structure does not lend itself to causal estimates of peer feedback effects.

This study sees its contribution as an AI-supported evaluation of revision-related writing development, rather than as causal evidence of the effects of peer feedback alone. The difference is important because there are a number of factors that can account for changes in writing in the revision cycle, such as instructor guidance, self-directed revision, and maturation effects [15]. With clear outcomes of both successful and unsuccessful modeling experiments, the research provides a repeatable methodological model for future research into feedback and writing analytics. To clarify the distinction between confirmatory and exploratory analyses, this study is guided by the following research questions:

- RQ1:** To what extent do AI-derived writing metrics change between draft and final research exposés in the matched text-pair sample?
- RQ2:** Do human assessment scores significantly improve from draft to final submissions in the score-matched subset?
- RQ3:** How accurately can traditional machine learning models approximate human assessment scores using AI-derived writing features?
- RQ4:** Can automatically extracted feedback features predict draft-to-final writing improvement, and what limitations affect this exploratory prediction?

RQ1, RQ2, and RQ3 are considered Major Analytical Questions. The dataset does not provide a complete alignment of comment spans to revisions, and the subset of feedback-improvements matched to score is small, so RQ4 is considered exploratory.

This study is not randomized or controlled, nor does it include a no-feedback comparison group or a counterfactual design to isolate the causal effect of peer feedback. The observed improvements might result from a combination of peer comments, instructor feedback, self-directed revision, additional reading, and time spent. The term “peer-feedback-supported” is therefore employed to describe the instructional context, and the empirical claims pertain to writing development during revision and to predicting improvements with exploratory feedback.

The rest of this paper is structured as follows: In Section 2, related studies in automated writing evaluation, peer feedback, NLP-based writing analysis, and machine-learning-based assessment are reviewed. The dataset, proposed framework, preprocessing, feature extraction, statistical analysis, and machine learning experiments are described in Section 3. The results are given in Section 4. The findings and implications are discussed in Section 5. The limitations, conclusion, and future research directions are given in Sections 6 and 7.

2. Related Work

2.1. Automated Writing Evaluation

In recent years, transformer- and LLM-based automated writing evaluation has attracted increasing research attention because it can better capture semantic, discourse-level, and contextual information than models built on features. These models can enhance scoring accuracy, particularly when trained and/or prompted with criteria aligned with the rubric. They also present computational problems, lack transparency, are sensitive to prompt design, and cannot explain why certain properties of the text are important to the score. Conversely, models created using the NLP feature-based methodology offer interpretable features such as lexical density, readability, cohesion, and syntactic complexity. The field of automated writing evaluation has undergone quite a transformation in the last 20 years. Some of the early systems, like e-rater [16] and the Intelligent Essay Assessor [17], used regression-based techniques that were trained on hand-crafted linguistic features, such as length of essay, vocabulary sophistication, and syntactic complexity. Recent systems have also adopted deep-learning architectures and pre-trained language models to enhance the scoring accuracy [18,19]. However, most AWE studies have been carried out in the context of standardized testing or second-language writing, and relatively little research has been conducted on research writing for higher education purposes [20].

2.2. Peer Feedback in Academic Writing

Peer feedback has been researched extensively as a formative assessment practice. Topping [5] offered a seminal review, which concluded that, with proper structuring, the peer assessment process could be as reliable as the instructor’s assessment process. Moderate positive effects of peer feedback on writing quality were later confirmed by subsequent meta-analyses [6,21]. The mechanisms by which feedback leads to revision, however, have not been well understood, and the quality of the feedback in different contexts is also widely different [7,22].

2.3. NLP-Based Writing Analytics

In particular, the introduction of NLP-driven writing analytics tools like Coh-Metrix [23], TAALES [24], and TAASSC [25] has given researchers more granular measures of textual complexity, cohesion, and sophistication. These tools have been applied to investigating writing development in longitudinal and cross-sectional designs [26]. Re-

search undertaken with these tools has shown that they can differentiate between learners' levels and between a draft and a revised text about lexical diversity, syntactic complexity, and cohesion features [27,28]. This ongoing research builds on this work by using a unique NLP pipeline for research exposé texts and combining statistical analysis with machine-learning-based prediction.

2.4. Machine Learning for Writing Assessment

Various machine learning methods for writing assessment include linear models, ensemble models, and neural networks. Some of the most common applications of educational data mining involve predicting student performance from textual features using Ridge Regression and Random Forest [29,30]. Tree models, such as Random Forests, can be used not only to predict human scores but also to interpret the model, since their feature importance scores indicate which linguistic features are most important for predicting human scores [31,32]. In addition, recent educational AI research indicates that AI tools can aid in personalized learning, data-informed decision-making, and the implementation of blended learning [33]. In parallel, research on postgraduate students has emphasized that scientific research competencies are essential for success in higher education and should be strengthened through structured institutional support [34]. These findings, together with recent research on AI-supported educational evaluation [33], reinforce the need for scalable, transparent, and data-driven frameworks for assessing research writing development.

2.5. The Ec Dataset

The dataset Ec contains academic writing and peer feedback data of student research exposés written in a university seminar [11,12]. It includes LaTeX files for the preliminary exposé and the final version, peer-reviewed versions of the texts, rubric-based assessment scores, and inline comments. This can be used to analyze writing development, feedback characteristics, and relationships between feedback and revision outcomes. The main dataset for all analyses in the present study is the Ec dataset.

3. Materials and Methods

3.1. Research Design

The study employed a quantitative research design using data from the Ec [35] that is publicly available. A first draft and final versions of exposés were compared to measure the development of writing using AI-derived text metrics. Second, peer feedback records were used in this study to summarize and describe the nature and depth of feedback in the dataset using NLP. Third, supervised machine learning models were trained to assess the ability of automatically extracted writing features to approximate the human assessment scores.

3.2. Proposed AI-Based Framework

The proposed framework is depicted in Figure 1 and presented as a layered AI-based pipeline for research writing development in a peer feedback-supported environment. Draft exposés, final exposés, peer feedback records, and human assessment scores from the Ec can be fed into the input layer. These inputs serve as the cues for determining writing improvement, feedback attributes, and human-evaluated writing quality. The preprocessing layer cleans and makes the raw corpus files analyzable. It includes archive extraction, file scanning and classification, conversion of LaTeX/PDF to text, text cleaning, tokenization, sentence splitting, and detection of the draft/final stage. This step normalizes heterogeneous corpus files into normalized text and metadata. The feature extraction layer calculates various writing and feedback indicators based on AI and NLP. The word

count, TTR, MATTR 100, lexical density, average sentence length, Flesch reading ease, and adjacent-sentence cohesion are all features included in writing. The feedback includes word count, lexical density, action word count, question count, and positive and negative indicators. These features are then used to create new composite indicators, such as quality of writing, quality of feedback, and draft final gain vectors. The analytical layer conducts paired statistical tests, machine learning score prediction, and exploratory feedback impact prediction. Draft-to-final changes are examined with paired *t*-tests, Wilcoxon signed-rank tests, confidence intervals, and effect sizes. Human assessment scores are predicted using Ridge Regression and Random Forest models, and the feedback impact model is interpreted with caution because of the limitations of aligning feedback with revision. Finally, the output layer delivers writing development indicators, human score improvement results, automated scoring performance, feature importance rankings, feedback analytics, and reproducible research outputs. The framework is generally designed to be linked to raw corpus data, AI-based feature engineering, statistical analysis, machine learning prediction, and interpretable analysis of writing development.

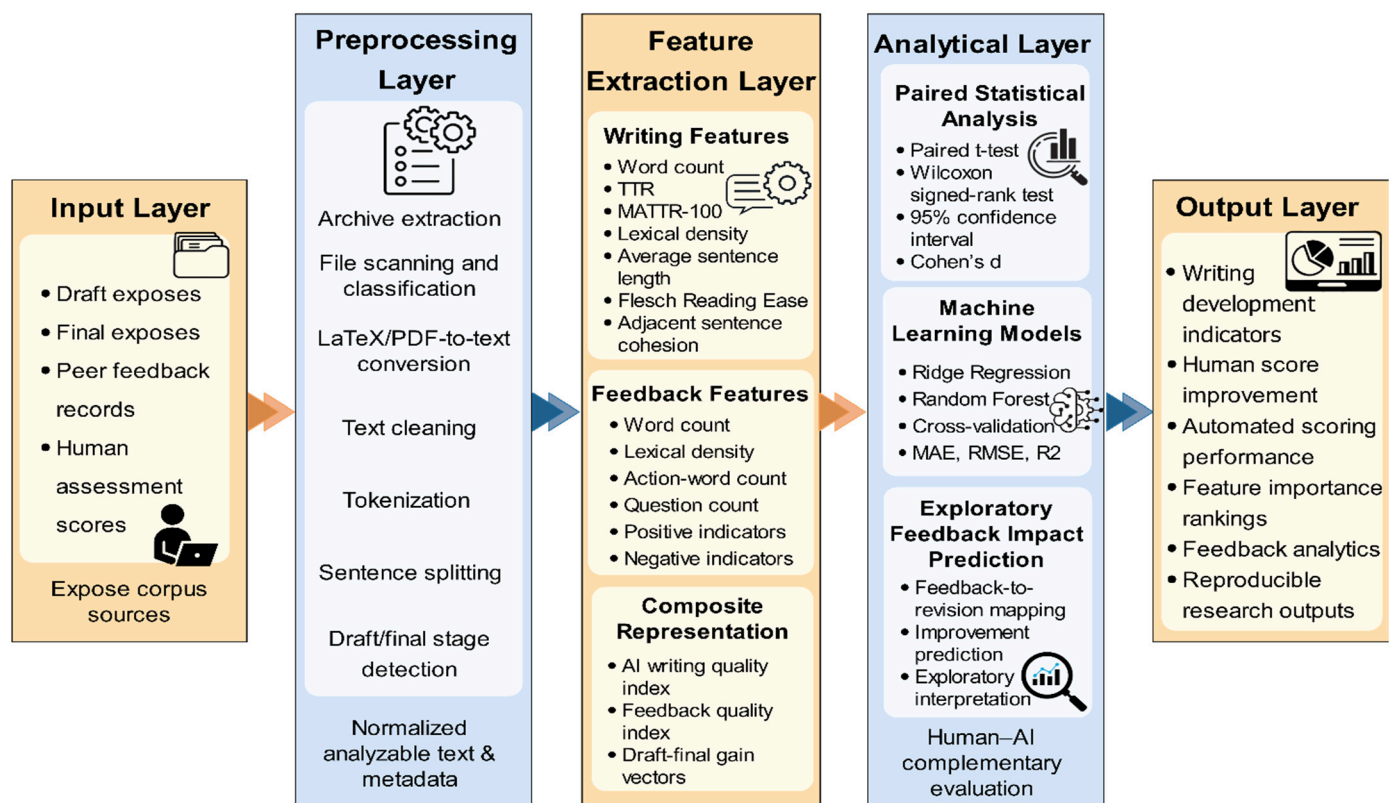


Figure 1. Proposed AI-based framework architecture for peer feedback and research writing evaluation.

The complete workflow of the proposed framework is presented in Algorithm 1, where data are extracted, text is processed, features are extracted using artificial intelligence, the drafted results are compared with the actual results, statistical tests are performed, machine learning is used to predict the results, and the impact of the feedback on the drafted results is carefully considered.

3.3. Dataset and Analytical Sample

This dataset was processed through a computational workflow created for this work. The pipeline retrieved the Ec archive, checked all available files, categorized structured tables, text files, LaTeX exposé files, PDF files, feedback and comment records, and score files. An automatic scanner strategy was deliberately designed to be broad to avoid

excluding useful material, resulting in a large number of intermediate records. Defensible analytical subsets reported in Table 1 were used in this study.

Algorithm 1: Proposed AI-Based Peer Feedback and Research Writing Evaluation Framework

Input: Draft exposés D , final exposés F , peer feedback records P , assessment scores S

Output: Writing development metrics, score predictions, feature importance, and feedback analytics

1. Load Ec and extract archive
 2. Scan and classify all files (LaTeX, PDF, JSON, CSV, text)
 3. Identify draft and final exposé versions from file paths and labels
 4. Clean and normalize text (remove LaTeX commands, braces, citations, whitespace)
 5. Extract AI/NLP writing features for each document (lexical, syntactic, readability, cohesion)
 6. Extract feedback features for each peer feedback record (length, density, action words, questions)
 7. Match draft–final pairs by author/exposé identifier
 8. Compute draft-to-final gains (Δ = final value – draft value) for all metrics
 9. Conduct paired statistical tests (t -test, Wilcoxon, Cohen’s d_z , 95% CI)
 10. Train Ridge Regression and Random Forest on writing features $\rightarrow$ human scores
 11. Evaluate models using MAE, RMSE, and R^2 via cross-validation
 12. Compute feature importance from Random Forest
 13. Interpret feedback impact prediction cautiously (exploratory only)
 14. Generate tables, figures, and conclusions.
-

Table 1. Dataset Processing and Analytical Subsets.

Analytical Component	n	Description	Role in Study
Raw files scanned	3247	All files detected after extracting the Exposía ZIP archive	Corpus auditing and file classification
Peer feedback/comment records	16,068	Feedback-like records extracted from reviews, annotations, and comment files	Descriptive feedback analytics using NLP features
Matched draft–final exposé text pairs	55	Exposé.tex files matched by author between draft and final stages	Draft-to-final writing metric comparison
Score-matched draft–final cases	28	Subset with both draft and final human assessment scores	Paired statistical testing of score improvement
Supervised automated scoring records	2437	Score-bearing records with valid human assessment scores and NLP features	Cross-validated machine learning scoring

3.4. Preprocessing Procedure

Five steps were used in the preprocessing. All the files were recursively extracted first. Second, file metadata were cataloged for a replicable record of sources. Third, files in the formats CSV, JSON, JSONL, Excel, Parquet, Plain Text, Markdown, LaTeX, and PDF were converted to data frames or plain text if possible. Fourth, all LaTeX commands, all citation commands, references, braces, markup symbols, and excessive whitespace were stripped from the document to create a plain text document that can be analyzed. Fifth, file names

and labels including the words “draft,” “initial,” “final,” and “revised” were used to make inferences about the document stage.

3.5. AI-Based Writing Feature Extraction

Lexical, syntactic, readability, and cohesion-related features were extracted using the NLP feature extraction module for every writing sample. The tokens were extracted with regular-expression-based word extraction, and sentences were segmented with punctuation-based sentence splitting. The final value minus draft value yielded draft-to-final gains for each metric. The features analyzed are summarized in Table 2.

Table 2. AI/NLP Writing and Feedback Features.

Category	Feature	Operational Definition	Formula/Method	Interpretation
Lexical	Word count	Total tokens in the exposé	Count of tokenized words	Text development and expansion
Lexical	TTR	Unique words/total words	V/N	Raw vocabulary variation; length-sensitive
Lexical	MATTR-100	Average TTR across 100-word moving windows	Mean of sliding-window TTR	Length-normalized lexical richness
Lexical	Lexical density	Content words/total words	Content words/N	Academic information density
Syntactic	Avg. sentence length	Words/sentences	N/S	Syntactic density and control
Readability	Flesch reading ease	Readability from sentence and syllable length	$206.835 - 1.015(N/S) - 84.6(\text{syl}/N)$	Complexity of academic prose
Cohesion	Adjacent cohesion	TF-IDF cosine similarity between neighboring sentences	Mean pairwise cosine($S_i, S_i + 1$)	Local semantic continuity
Composite	AI writing quality index	Standardized composite of selected text features	Weighted combination	Overall AI-derived quality indicator
Feedback	Action-word count	Revision-oriented verbs in feedback	Count of suggest, revise, clarify, etc.	Feedback actionability
Feedback	Question count	Interrogative sentences in feedback	Count of sentences ending in “?”	Inquiry depth

The AI writing quality index did not constitute a trained scoring model. The set of documents it was standardized against was the matched text-pair sample, which consisted of 110 documents: 55 draft exposés and 55 final exposés. To assess both draft and final texts against the same reference distribution, the texts were compiled and standardized before the draft and final texts were collected. Then, the mean μ_j and the standard deviation σ_j were calculated for each feature j , across the 110 documents, and the value for document i was transformed as $z_{ij} = (x_{ij} - \mu_j) / \sigma_j$. Word Count, Average Sentence Length, MATTR-100, Lexical Density, Flesch reading ease, Adjacent-sentence cohesion, and Citation count (where available) were considered standardized features. To handle missing values, the feature median was calculated from the same 110-document reference set prior to standardization and used to impute the missing values. Finally, a composite index was created using the same weights for each standardized feature. The index was not optimized for score prediction, so equal weighting was used to make it comparable across groups. As such, the

index does not measure the quality of writing on a standardized scale and should not be used to judge the quality of writing.

3.6. Rationale for Feature Selection

The features were chosen to serve as a basis for understanding one's research writing development. Word count reflects text elaboration and expansion. The average sentence length indicates syntactic density and the level of control they have. TTR and MATTR-100 measure lexical variation, and MATTR-100 was used because raw TTR is affected by text length. The proportion of content-bearing words is called lexical density and applies to academic information density. Flesch reading ease is a readability score that reflects sentence and word complexity. The adjacent-sentence cohesion is computed via local semantic continuity calculated by the TF-IDF cosine similarity between adjacent sentences. Those features were chosen because they can be reproduced, are easily computed, and are easy to understand for educational assessment.

3.7. Peer Feedback Feature Extraction

Peer feedback records were analyzed separately from the texts of exposé. The word count, average sentence length, TTR, lexical density, action-word count, question count, positive-word count, and negative-word count were used to summarize each feedback item. Furthermore, the feedback items were assigned to three descriptive categories using rule-based indicators. Question marks and interrogative forms in feedback were considered as interrogative feedback. Revision-oriented verbs (e.g., suggest, revise, clarify, expand, improve, restructure, explain) in feedback were considered to be modification-instruction feedback. Evaluative feedback consisted primarily of evaluative language, including positive or negative appraisal words. This classification was employed to represent the feedback environment and to separate the amount of feedback from its quality. Peer feedback was collected separately from the exposé texts. Word count, Average Sentence length, TTR, Lexical density, Action-word count, Question count, Positive-word count, and Negative-word count were shown as a summary of each feedback item. Words that were used as action words were verbs that were revision-oriented: explain, restructure, improve, clarify, expand, suggest, and revise. These feedback indicators were used to characterize the feedback environment and to develop a preliminary feedback-quality indicator. The first automatic merge of feedback to writing is not a good source of variation for the prediction of the impact of the feedback, and, therefore, feedback features are not considered as causal evidence in the current results.

3.8. Machine Learning Experiments

Two machine learning experiments were carried out. In the first experiment, human assessment scores were predicted from AI-derived writing features using Ridge Regression and a Random Forest Regressor (RFR), with cross-validation. Five-fold cross-validation with shuffled folds and a fixed random seed for reproducible performance was performed on Ridge Regression and Random Forest Regressor. Standard scaling was performed, followed by Ridge Regression with $\alpha = 1.0$. Random Forest: 300 trees, random_state = 42, minimum samples per leaf = 2. The performance of the models was assessed using MAE, RMSE, and R^2 . These settings are better considered a stable baseline for a small-to-moderate educational corpus, rather than hyper-optimized settings. No tuning was performed on the test folds.

To reduce the risk of leakage, cross-validation was performed at the record level for the 2437 score-bearing records used in automated scoring, as well as for the 55 matched text pairs and 28 score-matched cases (draft–final improvement analysis). The 2437 score-

bearing records were not used to interpret improvements from draft to final, but rather to measure the ability of AI-based linguistic features to match human assessment scores.

In the exploratory feedback-improvement experiment, the impact target was the difference between the draft and final human assessment score (for cases where a human assessment score was provided for both draft and final products):

$$\text{impact}_{\text{target}} = \text{final human assessment score} - \text{draft human assessment score} \quad (1)$$

In the absence of human score gains, the AI writing quality index gains were included only in the diagnostic analysis and not as the primary reported result. This experiment was conducted as a first exploratory experiment, as only 28 cases were scored and the automatic feedback-to-writing merge was insufficiently reliable in aligning comments with the writing to serve as evidence of the impact of causal feedback. The performance was measured by mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). The second experiment was designed to predict improvements in writing based on features and draft-level writing indicators. The second experiment is reported as exploratory, as the mapping from feedback to writing should be further developed before reliable predictions can be made.

3.9. Statistical Analysis

Paired-sample tests were used to evaluate draft–final differences. For the same human score subset ($n = 28$), paired-sample t -tests and Wilcoxon signed-rank tests were performed to determine if scores at the conclusion of the tournament were significantly higher than scores at the beginning of the tournament. The effect size was reported as Cohen’s d_z , which is the average of the paired gains divided by the standard deviation of the paired gains. Paired mean changes, 95% confidence intervals, p -values, and standardized paired effect sizes were computed for writing metrics obtained from AI ($n = 55$).

3.10. Experimental Setup

A single-cell computational pipeline, implemented in Python 3.11, was used on the Kaggle platform for all analyses. For data manipulation, pandas were used; for numerical computation, NumPy; for text processing and LaTeX removal, regular expressions; and for machine learning model training and evaluation, scikit-learn. TF-IDF vectorization was used to extract sentence vectors and calculate cosine similarity between adjacent sentences. Five-fold cross-validation was used for training the Ridge Regression and Random Forest Regressor with default hyperparameters. Feature extraction was not performed using external tools or pre-trained language models; the results are fully reproducible. The experimental setup comprised the following matched text pairs ($n = 55$) for an analysis of the draft–final NLP, score-matched cases ($n = 28$) for human assessment analysis, and 2437 scored text passages for training machine learning models. Data manipulation and numerical computation were performed using pandas and NumPy; text cleaning and removal of LaTeX commands were executed with the help of regular expressions; feature processing, model training, and evaluation were executed using scikit-learn and all implemented in a single-cell Python 3.11 pipeline on Kaggle. Adjacent-sentence cohesion was measured by converting sentences into TF-IDF vectors and calculating the cosine similarity between neighboring sentences. For automated scoring, the five-fold cross-validation with shuffled folds and a fixed random seed of 42 was tested. After standard scaling, $\alpha = 1.0$ was used for Ridge Regression. The Random Forest Regressor had the following parameters: 300 trees, `random_state = 42`, and the minimum samples per leaf was 2. These parameters were not optimized hyperparameters for validation but rather baseline parameters deemed repeatable. Analytical samples included 55 pairs of draft–final text for the AI-derived analysis

of writing metrics, 28 pairs of cases that were drafted and finalized with the same scores for the human assessment analysis, and 2437 texts with scores for the automated score prediction analysis. In the experiments reported here, no external service that performs text analysis, nor any pre-trained language model, was used.

4. Results and Analysis

4.1. Dataset Processing and Analytical Sample

The first phase of the analysis focused on the structure of the extracted Exposía corpus and on identifying the analytical subsets selected for this work. 3247 raw files were identified during the processing pipeline. These files contained structured tables, text records, LaTeX exposé files, feedback-related files, score files, metadata files, and other supporting materials. Of this amount, 16,068 peer feedback/comment records were found for descriptive feedback analysis. This main writing development analysis, however, was limited to 55 matched draft and final exposé cases, as these cases offered comparable versions of the writing. A smaller group of 28 cases had full human assessment scores for both drafts and final versions. Additionally, 2437 score-bearing records were available to be used for supervised automated writing score prediction. The distribution of file types in the Ec is shown in Figure 2. The figure depicts that the most prevalent document types were JSON, TXT, HTML, CSV, Delta, PDF, configuration, and BibTeX. The distribution shows heterogeneity, meaning the corpus was not collected from a single source and required systematic preprocessing before feature extraction and analysis.

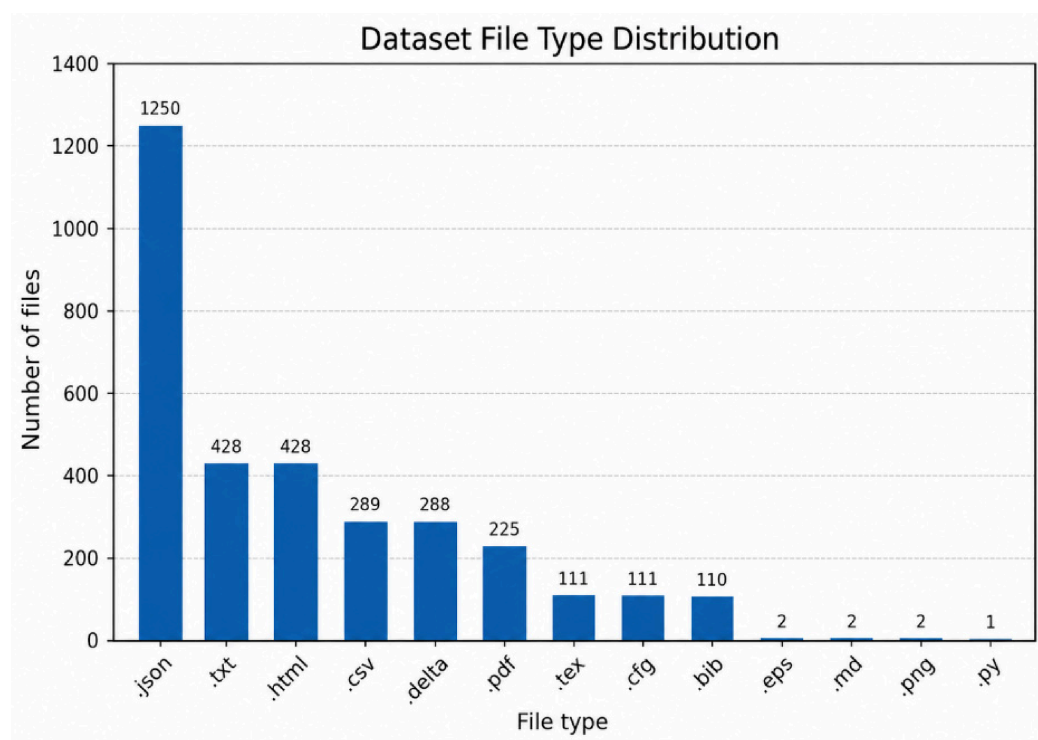


Figure 2. Dataset file type distribution after Exposía corpus extraction.

4.2. Draft-to-Final Development in AI-Derived Writing Metrics

The second phase of the study compared the extent of exposure improvement in final documents with its presentation in drafts, following the definition of the analytical sample. The AI-derived features used to measure writing development and provide feedback are listed in Table 3. These features were used to measure text length, syntactic structure, vocabulary variation, lexical richness, academic information density, readability, semantic

cohesion, and feedback characteristics. These features enabled the study to compare draft and final writing in a way that could be repeated and made transparent to computers.

Table 3. AI/NLP writing and feedback features used in the analysis.

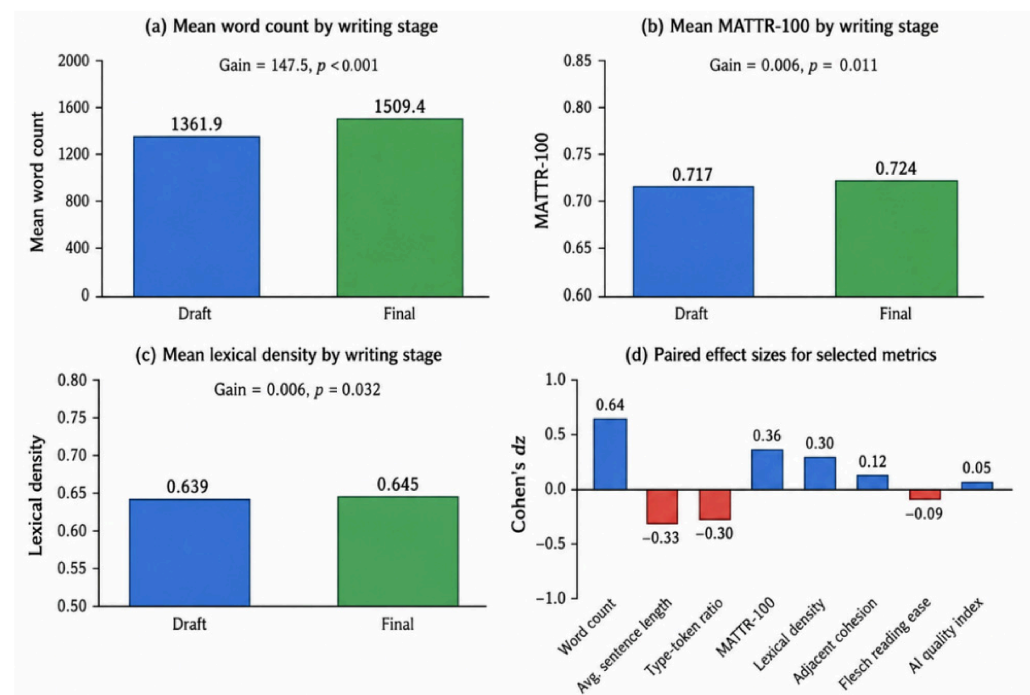
Feature Category	Feature	Operational Definition	Interpretation
Text development	Word count	Total number of tokens in the exposé	Indicates expansion and development of the research text
Syntax	Average sentence length	Number of words divided by number of sentences	Indicates syntactic density and sentence complexity
Vocabulary	Type–token ratio	Unique words divided by total words	Indicates raw vocabulary variation but is sensitive to length
Lexical richness	MATTR 100	Moving-average type–token ratio across 100-word windows	Indicates length-normalized lexical richness
Academic density	Lexical density	Content words divided by total words	Indicates information-rich academic expression
Readability	Flesch reading ease	Readability score based on sentence and syllable length	Indicates readability and complexity
Cohesion	Adjacent-sentence cohesion	Mean TF-IDF cosine similarity between neighboring sentences	Indicates local semantic continuity
Composite quality	AI writing quality index	Standardized composite of selected text features	Indicates overall AI-derived writing quality
Feedback quality	Action-word count	Count of revision-oriented words	Indicates directive and improvement-focused feedback
Feedback interaction	Question count	Count of question marks or question-based feedback	Indicates inquiry-based feedback

The draft and final comparisons are presented in Table 4. The biggest gains, in terms of improvement, were in text expansion. The length of final exposés compared to draft exposés is longer, with an increase in the mean number of words from 1361.909 to 1509.436 words. The difference in mean gain was found to be statistically significant ($p < 0.001$ and Cohen's $d_z = 0.642$). This means that the revision process in students' research writing mainly involved extending and developing their writing. For the final versions, there were small but statistically significant gains in MATTR 100 and lexical density, indicating modest gains in length-normalized lexical richness and academic information density. The raw type–token ratio fell slightly; however, this is not considered a sign of poor vocabulary, as a document tends to have a lower ratio as it gets longer and academic vocabulary is used more consistently.

The draft-to-final development patterns for the matched text-pair sample of an exposé are shown in Figure 3. As indicated in Figure 3a, the final exposés were longer than the draft exposés, with the mean word count of the final exposés being 1509.4 words as compared to 1361.9 words for the draft exposés. This suggests significant text expansion during the revision process. The small increase in MATTR-100 from 0.717 to 0.724 (shown in Figure 3b) demonstrates an improvement in length-normalized lexical richness. Lexical density rose from 0.639 to 0.645 in the final submissions (see Figure 3c), indicating that a slightly higher proportion of the final submissions included academic language that was content-bearing. The pairwise standardized effect sizes are summarized in Figure 3d. The changes were most favorable for word count, MATTR-100, and lexical density, and were least favorable for average sentence length and raw type–token ratio.

Table 4. Draft-to-final changes in AI-derived writing metrics.

Metric	<i>n</i>	Draft Mean	Final Mean	Mean Gain	95% CI	<i>p</i> Value	Cohen <i>dz</i>	Interpretation
Word count	55	1361.909	1509.436	147.527	85.358 to 209.696	<0.001	0.642	Significant text expansion
Average sentence length	55	23.396	22.734	−0.662	−1.211 to −0.113	0.019	−0.326	Slight reduction in sentence length
Type–token ratio	55	0.396	0.387	−0.009	−0.016 to −0.001	0.030	−0.300	Expected decline due to longer texts
MATTR 100	55	0.717	0.724	0.006	0.002 to 0.011	0.011	0.355	Improved normalized lexical richness
Lexical density	55	0.639	0.645	0.006	<0.001 to 0.012	0.032	0.297	Improved academic information density
Adjacent-sentence cohesion	55	0.060	0.062	0.001	−0.002 to 0.005	0.393	0.116	Small non-significant change
Flesch reading ease	55	20.406	19.903	−0.503	−1.985 to 0.980	0.500	−0.092	No meaningful readability shift
AI writing quality index	55	84.547	85.564	1.017	−4.946 to 6.980	0.734	0.046	Small non-significant increase

**Figure 3.** Draft-to-final writing development in the matched exposé text-pair sample ($n = 55$): (a) word count, (b) MATTR-100, (c) lexical density, and (d) paired effect sizes across selected AI-derived writing metrics.

4.3. Human Assessment Score Improvement

The strongest evidence of research writing development was obtained from the score-matched subset. As shown in Table 5, the mean human assessment score increased from 5.26 in the draft stage to 7.55 in the final stage. A 95% confidence interval for the mean gain was (1.77, 2.82). This improvement was confirmed to be statistically significant by the paired-sample *t*-test: $t(27) = 9.04$, $p < 0.001$. The effect size was large (Cohen *dz* = 1.709). The Wilcoxon signed-rank test also upheld the robustness of the improvement, $p < 0.001$. The results show that the latter were longer and had more lexical development, and human evaluators rated them significantly higher than the earlier ones.

Table 5. Human assessment, machine learning performance, and model interpretation results.

Experiment	Model or Test	<i>n</i>	Target	Main Statistics	Performance Metrics	Interpretation
Human assessment improvement	Paired-sample <i>t</i> -test	28	Human assessment score	Draft mean = 5.26, final mean = 7.55, gain = 2.295, $t(27) = 9.04$, $p < 0.001$, Cohen $d_z = 1.709$	Wilcoxon $p < 0.001$	Strong evidence of draft-to-final writing improvement
Automated research writing scoring	Ridge Regression	2437	human_score_mean	Cross-validated regression	MAE = 0.648, RMSE = 1.422, $R^2 = 0.248$	Moderate linear baseline
Automated research writing scoring	Random Forest	2437	human_score_mean	Cross-validated regression	MAE = 0.278, RMSE = 1.242, $R^2 = 0.426$	Best scoring model in this study
Peer feedback impact prediction	Random Forest	28	impact_target	Exploratory prediction	MAE = 0.976, RMSE = 1.364, $R^2 = -0.069$	Not reliable due to the feedback on the revision alignment limitation
Feature importance	Random Forest	2437	human_score_mean	Lexical density = 0.529, word count = 0.173, Flesch reading ease = 0.119	Feature importance ranking	Human scores were most associated with lexical density and text development

4.4. Descriptive Analysis of Peer Feedback

The feedback analytics feature consolidated 16,068 feedback/comment records. The mean feedback item was around 30.47 words, but the distribution was very skewed, with a median of 11 words and a maximum of 3696 words. This means that there are inline comments and longer feedback records within this data set. The average lexical density of feedback was 0.792, indicating that most feedback was content-oriented. In the meantime, the average number of words used as an action was 0.188, and the average for questions was 0.062. This pattern indicates that many of the comments were short evaluative comments rather than detailed revision instructions. Hence, feedback quantity must not be considered to be synonymous with feedback quality.

The feedback results are useful to understand the exploratory feedback impact model later. Descriptive evidence indicates that feedback records varied greatly in length, revision density, and revision orientation. However, the automatic feedback-to-writing merge was not entirely parallel with the adjustment of the text span in the targets, so the features of the feedback are described and not used as causal predictors of improvement. This interpretation follows the methodological caveat that the impact analysis of feedback is not sufficiently aligned to draw sound causal inferences.

4.5. Automated Research Writing Scoring and Feature Importance

The automated scoring test aimed to determine whether AI-calculated writing features could be used to estimate human assessment scores. The results of the human score improvement test and the machine learning experiments are summarized in Table 5. Random Forest was found to outperform Ridge Regression in automated research writing scores. Ridge Regression achieved MAE = 0.648, RMSE = 1.422, and $R^2 = 0.248$, while Random Forest achieved MAE = 0.278, RMSE = 1.242, and $R^2 = 0.426$. The results suggest that the correlation between the linguistic features derived from AI and human scores was not strictly linear. Random Forest was able to model nonlinear feature interactions more effectively, thereby providing a more robust baseline model for automated research writing evaluation.

An exploratory feedback impact prediction model failed to generalize properly. The model was trained on 28 matched cases but obtained a negative R^2 value of -0.069 . This indicates that the model was not as successful as a baseline prediction model based on a simple mean. This is an important result to note, as it does not mean that peer

feedback is ineffective. Instead, it means that there is a data integration restriction in the current pipeline. To model the impact of feedback robustly, there needs to be a more precise mapping between individual comments, individual spans of revisions, and the final changes in the text.

Lexical density was the highest predictor of human writing score, based on the importance value obtained from the Random Forest score model, with an importance value of 0.529. This was followed by word count, Flesch reading ease, average sentence length, and number of unique words. The findings indicated that the scores of human assessment were best correlated with information-rich writing, adequate text development, and appropriate readability. The remaining factors, cohesion and MATTR 100, contributed less to the baseline model, and citation count did not make any meaningful contribution.

Figure 4 shows the performance results of the machine learning models and feature importance scores for automated research writing scoring. The figure is broken into four subplots. The performances of the models in terms of mean absolute error are compared against each other in Figure 4a, with lower values indicating more precise prediction. The root mean squared error, which penalizes larger prediction errors, is compared for the two in Figure 4b. Figure 4c displays the R^2 values for the scoring model and the exploratory feedback impact model, along with the negative R^2 value for the feedback impact prediction model. The Random Forest feature importance values are provided in Figure 4d to identify the most important AI-based writing features that have the greatest impact on the prediction of human scores.

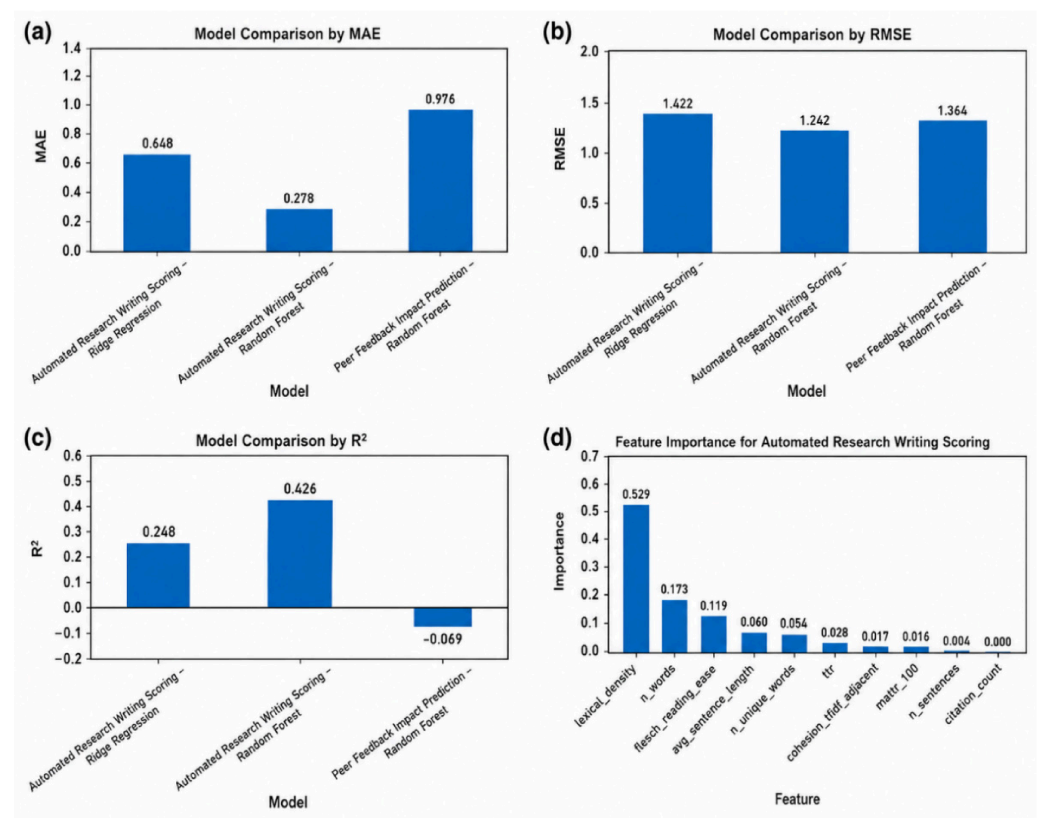


Figure 4. Machine learning model performance and feature importance. (a) Mean absolute error comparison across models. (b) Root mean squared error comparison across models. (c) R^2 comparison for automated writing scoring and exploratory feedback impact prediction. (d) Random Forest feature importance for predicting human writing scores from AI-derived writing features.

The overall findings indicate that AI tools for writing analytics can identify meaningful research writing development associated with revision. The subsequent exposés were

longer and slightly higher in normalized lexical variety as well as slightly more lexically dense than the draft exposés. Human assessment results showed the largest improvement (large and statistically significant), with a substantial increase from draft-to-final scores. In addition, the automated scoring experiment demonstrated that the machine learning models (Random Forest and Ridge Regression) can moderately approximate the human writing score. But the direct prediction of improved writing from the feedback features remained tentative and unreliable. The conclusions cannot thus be understood as causal evidence that peer feedback itself led to the observed improvements in revision-associated writing development.

A total of 16,068 peer feedback and comment records were summarized in the feedback analytics module. The mean size of the feedback items was around 30.47 words. However, the distribution was skewed (median = 11 words, maximum = 3696 words), indicating that the data include both short inline comments and longer feedback records. Lexical density averaged 0.79, indicating that many comments were content-related. On average, however, the numbers of action words and questions are low (0.19 and 0.06, respectively), suggesting that many of the feedback comments may have been brief evaluative comments rather than specific instructions for revision. This is important for the discussion because it clearly shows why the quality of the feedback matters and should be modeled carefully, not just as a number of comments received [7,22].

5. Discussion

This study's findings support converging evidence of measurable growth in writing between the draft and final versions of the Exposía research exposé in the Exposía corpus. The most significant individual indicator is the human assessment score improvement (mean gain = 2.295 and $dz = 1.709$), which indicates that trained evaluators believe the scores showed significant improvements in quality. The results of the text metrics derived from the AI are complementary: final exposés were significantly longer, slightly more lexically complex (normalized lexical variety, MATTR-100), and slightly more lexically dense. Considering the results together, it can be suspected that students made substantive revisions rather than cosmetic changes.

This is not in contradiction with the observed reduction in raw TTR. The drawback of raw TTR is that it is sensitive to the length of the text [23,24]—as the number of words increases in a document naturally, the type–token ratio decreases. Additionally, the rise in MATTR-100 (which controls for text length) indicates that vocabulary richness was not reduced when measured appropriately. This separation emphasizes that multiple, complementary lexical diversity measures in writing analytics research are important [27].

The automated scoring results show that, when using NLP-derived features alone, the machine learning models can moderately replicate human writing scores. The higher R^2 value of the Random Forest model ($R^2 = 0.426$) compared to the Ridge Regression model ($R^2 = 0.248$) suggests that the relationship between linguistic indicators and human scores is nonlinear, which aligns with previous studies in the field of Educational Data Mining [29,31]. Lexical density emerged as the most important feature, with human evaluators more strongly favoring passages with high information density, suggesting that academic prose is particularly valued for its information density. This result is consistent with the work conducted by several researchers, which revealed that among the most significant factors that predict writing quality in academic settings are content density and vocabulary sophistication [24,26].

The exploratory feedback impact prediction experiment yielded a negative R^2 (-0.069), indicating that the writing improvement attributable to feedback features could not be predicted by the model. It is important to note, however, that these negative findings

do not indicate that peer feedback has no impact on writing achievement. Instead, it is a data alignment restriction: during the initial automatic merge of feedback records and specific revision changes, the granularity was insufficient. In addition, there are a lot of shorter, non-correspondential textual annotations and/or evaluation comments made in many of the records of the Exposía corpus. This is a limitation that should be overcome in future research, which can be achieved using the approach of comment-span alignment, connecting individual feedback comments to the corresponding text passages being commented on, and analyzing the quality of the feedback comments in terms of rhetorical modes (directive, evaluative, epistemic), and then modeling the connection between the comment quality and the revision outcomes [7,8].

Lexical density and word count were the two most influential variables in the Random Forest model, and automated scoring performance should be interpreted with caution. The features may indicate information-rich academic writing and sufficient text development, but may also lead the model to prefer longer, denser texts. This may undermine the performance of concise writers, multilingual learners, and learners with different rhetorical styles from those found in the training corpus. Furthermore, the models were trained on human assessment scores, which can be biased based on the assessment scores or the composition of the assessment dataset. This proposed framework should then only be applied as a decision support tool and not as a stand-alone grading mechanism. Specific practical remedies are to carry out error and calibration audits for each of the subgroups; evaluate within text-length and linguistic-background groups; include length-normalized and rubric-aligned features; remove or down-weight features that create systematic differences; add uncertainty estimates and abstention for low-confidence predictions; revalidate periodically across disciplines and learning groups; and require human review for high-stakes decisions. Demographic or linguistic-background information that is ethically and legally available should be reported on relevant learner subgroups to provide fairness metrics, but with consideration of data-protection requirements and privacy concerns.

The clear reporting of successful and unsuccessful experiments is an important methodological contribution. It is crucial to publish negative or null results in machine learning research to avoid publication bias and to inform future research design [36]. The incorporation of the failed feedback impact model, with its detailed interpretation, is a useful resource for researchers who wish to undertake feedback–revision alignment research.

6. Limitations

This study is limited in several aspects of the research design, data construction, technical model, and external validity. First, this study is not causal because it does not prove that peer feedback causes students to learn more. The Exposía corpus contains a peer-feedback-supported revision context without a random assignment, a no-feedback control group, or a counterfactual design. Thus, observed writing gains could be attributed to peer feedback, teacher feedback, self-directed revision, additional research, or the effects of time on task.

Secondly, the analytical subsets are small in magnitude. A total of 55 pairs of draft and final exposé and 28 pairs of scores were used for the matched text-pair analysis and human score analysis, respectively. These subsets are adequate for exploratory paired analysis but would not be suitable for robust generalization or for significant predictions of feedback impact. The negative R^2 for the exploratory feedback-improvement model should be interpreted as a limitation of the diagnostic data for this data linkage and sample size, rather than as an indication that feedback is ineffective.

Third, the NLP features that detect representable surface-level and local textual features (e.g., length, lexical richness, readability, cohesion) do not directly evaluate rhetorical

structure, argument quality, methodological appropriateness, citation integration, or disciplinary relevance. The higher-level dimensions are important to research writing and should be covered in future models.

Fourth, the automatic feedback–revision merge was not granular enough. Much of the feedback is brief comments or annotations that do not correspond directly to specific changes in the text. Further improvement in the alignment between the comment span and the revision span is needed to enable reliable feedback-to-revision modeling.

Fifth, automated scoring can perpetuate human-rating bias or favor certain writing styles. Conciseness in the model, use of lexical density, and word count might be problematic for multilingual writers, writers whose academic style differs from the patterns found in the corpus, and concise writers.

Lastly, the Exposía corpus can be thought of as a certain academic genre, institutional context, and disciplinary context. While the model could be applicable in other academic writing environments, preprocessing rules, feature extraction, scoring rubrics, and feedback classifications would need to be adapted for use across disciplines, languages, genres, and peer review.

7. Conclusions and Future Work

This study aimed to present an AI-supported model for evaluating research writing development during revision in a corpus of peer feedback. The framework, based on this Ec, combined NLP-based feature extraction, statistical pair analysis, and machine learning to analyze changes in draft-to-final writing, approximate human assessment scores, and explore the boundaries of predicting feedback-improved writing. It was found that human assessment scores increased significantly from the draft to the final submission, with a large effect size. AI-driven metrics also found that final exposés were longer and had moderate improvements in normalized lexical richness and lexical density. Random Forest also performed better than Ridge Regression in automatic scoring, indicating that nonlinear models might be better suited to modeling the relationship between linguistic features and human scores. The exploratory feedback-improvement prediction model, however, failed to generalize, largely due to the small score-matched sample and lack of feedback–revision alignment. Thus, the study’s findings support using AI to aid the evaluation of writing development associated with revision, rather than providing causal evidence that peers alone led to the observed improvement.

Future investigations should be undertaken in five ways. For one, comment-span alignment models should relate each peer feedback comment to the draft passage under consideration and to the revised passage in the final text. This would allow direct measurement of feedback uptake and revision success. Second, the classification of feedback needs to be expanded from surface indicators of feedback to include evaluative, directive, interrogative, epistemic, and revision-oriented feedback. Third, there is a need to represent higher-level dimensions of research writing through discourse parsing, argument mining, citation-function analysis, methodological-consistency assessment, and section-structure modeling. Fourth, a controlled comparison experiment should be performed between the current interpretable NLP feature set and transformer embeddings/LLM-aligned features, as well as a hybrid feature set (both interpretable and transformer or LLM-aligned). Embeddings from the documents should be extracted by a fixed pre-trained document encoder; embeddings from the LLM should be generated by a fixed version of the LLM, a predefined, rubric-aligned prompt, and a set of fully reported or deterministic generation settings. For all configurations, the same author- or exposé group-based cross-validation folds should be used, and results compared using MAE, RMSE, R^2 , calibration, and error measures for the subgroups. If there is a difference in performance, paired bootstrap confidence intervals or

an appropriate paired resampling test should be used to test the difference. It should also include computational cost, model and prompt versions, feature dimensionality, uncertainty and reproducibility settings. Finally, a larger, multi-institutional, multi-disciplinary, multilingual dataset should be used to assess the framework comprehensively across genres, disciplines, learner groups, and peer feedback designs.

Funding: This research received no external funding.

Institutional Review Board Statement: This study used a publicly available dataset and did not involve direct human subject data collection by the authors. All student names in the Exposía corpus are pseudonymized in the publicly released version. No institutional review board approval was required because the analysis is based on secondary use of an existing, de-identified public dataset.

Informed Consent Statement: Participant consent was waived because this study used a publicly available dataset and did not involve direct human subject data collection by the authors.

Data Availability Statement: This study used the Exposía academic writing and peer-feedback corpus, which is available for academic research through the official dataset access procedure [35]. The processed analytical outputs, figures, and computational pipeline used in this study can be made available by the authors upon reasonable request, subject to the dataset license and access conditions.

Acknowledgments: This paper is part of the project BFP/RGP/ICT/24/368 funded by the Ministry of Higher Education, Research and Innovation.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Crusan, D.; Plakans, L.; Gebril, A. Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assess. Writ.* **2016**, *28*, 43–56. [\[CrossRef\]](#)
2. Saricaoglu, A.; Bilki, Z. The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annu. Rev. Appl. Linguist.* **2025**, *45*, 253–273. [\[CrossRef\]](#)
3. Misgna, H.; On, B.-W.; Lee, I.; Choi, G.S. A survey on deep learning-based automated essay scoring and feedback generation. *Artif. Intell. Rev.* **2024**, *58*, 36. [\[CrossRef\]](#)
4. Kochmar, E.; Vu, D.D.; Belfer, R.; Gupta, V.; Serban, I.V.; Pineau, J. Automated personalized feedback improves learning gains in an intelligent tutoring system. In *Proceedings of the International Conference on Artificial Intelligence in Education*; Springer: Cham, Switzerland, 2020; pp. 140–146.
5. Topping, K.J.; Gehringer, E.; Khosravi, H.; Gudipati, S.; Jadhav, K.; Susarla, S. Enhancing peer assessment with artificial intelligence. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 3. [\[CrossRef\]](#)
6. Topping, K. Peer assessment: Channels of operation. *Educ. Sci.* **2021**, *11*, 91. [\[CrossRef\]](#)
7. Kerman, N.T.; Banihashem, S.K.; Karami, M.; Er, E.; Van Ginkel, S.; Noroozi, O. Online peer feedback in higher education: A synthesis of the literature. *Educ. Inf. Technol.* **2024**, *29*, 763–813. [\[CrossRef\]](#) [\[PubMed\]](#)
8. An, Z.; Lai, C.; Gan, Z. Motivation in self-directed use of technology for English learning among high, average, and low achievers. *System* **2023**, *115*, 103051. [\[CrossRef\]](#)
9. Huisman, B.; Saab, N.; Van Driel, J.; Van Den Broek, P. Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assess. Eval. High. Educ.* **2018**, *43*, 955–968. [\[CrossRef\]](#)
10. Li, D.; Qian, J. Engaging in dialogic peer feedback in L2 writing development: A microgenetic approach. *System* **2024**, *126*, 103513. [\[CrossRef\]](#)
11. Banihashem, S.K.; Kerman, N.T.; Noroozi, O.; Moon, J.; Drachsler, H. Feedback sources in essay writing: Peer-generated or AI-generated feedback? *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 23. [\[CrossRef\]](#)
12. Jongasma, M.; Meeter, M.; van Muijlwijk-Koezen, J.; Scholten, D. The role of platform design in online peer feedback: A comparison of two digital environments. *Educ. Inf. Technol.* **2025**, *30*, 20797–20820. [\[CrossRef\]](#)
13. Malhotra, M.; Chhabra, I. A comprehensive review of the changing landscape of academic dishonesty in automated proctoring in the era of artificial intelligence. *Discov. Educ.* **2026**, *5*, 236. [\[CrossRef\]](#)
14. Zhu, Y.; Chen, X.; Shu, L.; Tan, B.; Song, X.; Liu, L.; Wang, M.; Chen, J.; Ruan, N. An Efficient Conversational Smart Compose System. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 456–462.
15. Ellis, R. A typology of written corrective feedback types. *ELT J.* **2009**, *63*, 97–107.

16. Burstein, J. The e-rater scoring engine: Automated essay scoring with natural language processing. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2003; Volume 113121.
17. Landauer, T.K.; Laham, D.; Foltz, P.W. Automated scoring and annotation of essays with the Intelligent Essay AssessorTM. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2003; pp. 87–112.
18. Uto, M. A review of deep-neural automated essay scoring models. *Behaviormetrika* **2021**, *48*, 459–484. [\[CrossRef\]](#)
19. Mizumoto, A.; Eguchi, M. Exploring the potential of using an AI language model for automated essay scoring. *Res. Methods Appl. Linguist.* **2023**, *2*, 100050. [\[CrossRef\]](#)
20. Lillo-Fuentes, F.; Venegas, R.; Lobos, I. Evaluación automatizada y semiautomatizada de la calidad de textos escritos: Una revisión sistemática. *Perspect. Educ.* **2023**, *62*, 5–36. [\[CrossRef\]](#)
21. Zhang, H.; Costley, J.; Courtney, M.; Shulgina, G.; Fanguy, M. The impact of different peer feedback types on student academic writing performance from dyadic and individual analyses. *Educ. Inf. Technol.* **2025**, *30*, 6339–6366. [\[CrossRef\]](#)
22. Cho, K.; MacArthur, C. Student revision with peer and expert reviewing. *Learn. Instr.* **2010**, *20*, 328–338. [\[CrossRef\]](#)
23. Graesser, A.C.; McNamara, D.S.; Kulikowich, J.M. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educ. Res.* **2011**, *40*, 223–234.
24. Kyle, K.; Crossley, S.A. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Q.* **2015**, *49*, 757–786. [\[CrossRef\]](#)
25. Kyle, K. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA, 2016.
26. Crossley, S.A.; Kyle, K.; McNamara, D.S. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behav. Res. Methods* **2016**, *48*, 1227–1237. [\[CrossRef\]](#) [\[PubMed\]](#)
27. McCarthy, P.M.; Jarvis, S. MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* **2010**, *42*, 381–392. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Lu, X. Automatic analysis of syntactic complexity in second language writing. *Int. J. Corpus Linguist.* **2010**, *15*, 474–496. [\[CrossRef\]](#)
29. Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [\[CrossRef\]](#)
30. K G, S.; Kurni, M. Educational data mining & learning analytics. In *A Beginner's Guide to Learning Analytics*; Springer: Cham, Switzerland, 2021; pp. 29–60.
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
32. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [\[CrossRef\]](#)
33. Al Husaini, Y.N.S.; Al Yahmadi, H.H.N.; Al Husaini, M.A.S. A systematic review of the utilization of artificial intelligence applications within the educational Oman systems. *Int. J. Eval. Res. Educ.* **2025**, *14*, 4758. [\[CrossRef\]](#)
34. Al Yahmadi, H.H.N.; Al Wahaibi, I.; Al Hunaini, S.; Al Husaini, Y.N. The scientific research competencies among postgraduate students in Omani universities. *Int. J. Innov. Res. Sci. Stud.* **2025**, *8*, 2735–2746. [\[CrossRef\]](#)
35. Zyska, D.; Rozovskaya, A.; Kuznetsov, I.; Gurevych, I. Exposía: Teaching and Assessment of Academic Writing Skills for Research Project Proposals and Peer Feedback. *arXiv* **2026**, arXiv:2601.06536.
36. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **2012**, *90*, 891–904. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.