Data-Driven Analysis for Improving Educational Policies: A Case in Brazil's Textbook Program

André Araújo[a], *Rafael* Araújo[a,b], *Luciano* Cabral[a,c], *Luciane* Silva[a,b], *Hilario* Tomaz[a,d], *Emerson* Martins[a,b], *Diego* Dermeval[a,e], *Álvaro* Sobrinho[a,f*], *Alan Pedro da* Silva[a,g], *Leonardo* Marques[a,e], *Filipe* Recch[h], *Sebastian* Munoz-Najar Galvez[e], *Seiji* Isotani[a,e,i], *Ig* Ibert Bittencourt[a,e]

[a]Center for Excellence on Social Technology (NEES), Federal University of Alagoas, Maceió, Alagoas, Brazil.

[b]Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil.

[c]Federal Institute of Pernambuco, Recife, Pernambuco, Brazil.

[d]Federal Institute of Espírito Santo, Serra, Espírito Santo, Brazil

[e]Harvard Graduate School of Education, Cambridge, Massachusetts, United States

[f]Federal University of the Agreste of Pernambuco, Garanhuns, Pernambuco, Brazil. Email: alvaro.alvares@ufape.edu.br.

[g]University of Oxford, Oxford, Oxfordshire, United Kingdom.

[h]University of Pittsburgh, Pittsburgh, PA, United States.

[i]University of São Paulo, São Carlos, São Paulo, Brazil.

Submitted: 31 January 2025, Revised: 26 March 2025, Accepted: 21 April 2025, Published: 23 May 2025

Abstract. Data analytics can support evidence-based decision-making in public policies by enabling the identification of patterns, forecasting needs, and prioritizing actions. Consequently, data-driven analysis can aid policymakers in redesigning and enhancing educational policies, such as textbook distribution for public schools. However, there is no consensus on a structured approach for descriptive analytics in this context. This study presents a descriptive approach to textual data analysis aimed at improving policy implementation and monitoring, with a focus on effort, productivity, and quality of reviews produced by textbook evaluators. Through a case study, we apply natural language processing techniques to analyze thousands of answers to rubrics during the pedagogical evaluation of a public call for literary works under the Brazilian textbook program (Programa Nacional do Livro e do Material Didático - PNLD). The PNLD is one of the most extensive textbook policies, impacting millions of students. Our findings shed light on challenges related to the effort involved and the quality of written reports in the pedagogical evaluation process. Analyzing reports, which reflect some desired and undesired behaviors of evaluators, can offer policymakers insights for making informed decisions and improving textbook programs worldwide. Our descriptive approach to textual data analysis leverages insights to enhance transparency, inform improvements, and guide policy implementation through real-time monitoring.

Keywords. Official Approval, Data Science, Descriptive Analysis, PNLD.
Research paper, DOI: https://doi.org/10.59490/dgo.2025.1023

1. Introduction

A data science process consists of descriptive, predictive, and prescriptive steps, supporting the entire cycle of data-driven analysis and decision-making (Delen, 2019; Sharda et al., 2018). Thus, data-driven analysis has been recognized as a significant approach for supporting governments and public organizations, for instance, in developing and improving educational policies (Badrul Hisham et al., 2024; Kulal et al., 2024). Data analytics enables governments and public organizations to make evidence-based decisions by leveraging large

volumes of data to identify patterns, predict future needs, and prioritize actions (Kulal et al., 2024). Analytical techniques, such as those driven by Artificial Intelligence (AI), can impact public resource allocation and tackle modern governance challenges (Santoso et al., 2024).

In education, data analytics can inform educational policies to reduce inequalities and improve access (Pencheva et al., 2020). Data-driven analysis can also assist policymakers in redesigning and developing educational policies (Badrul Hisham et al., 2024). The Brazilian textbook program (Programa Nacional do Livro e do Material Didático - PNLD) evaluates educational materials, including literary works, to ensure high-quality tools that support learning for millions of public school students. By maintaining a structured pedagogical and technical evaluation process, the PNLD plays an important role in aligning these materials with national educational objectives (Sobrinho et al., 2023). The PNLD is an example of an educational public policy that can benefit from data-driven analysis to enhance monitoring processes and support policy reformulations.

This paper emphasizes descriptive analytics, where we analyze past events to lay the foundation for future predictive and prescriptive steps. While data-driven analysis has proven to be a relevant approach in supporting educational public policies, such as the PNLD, there is no consensus on a structured approach for conducting descriptive analytics in this context. More standardized guidelines are needed for integrating data analysis into policies, resulting in heterogeneous and context-specific approaches (Pencheva et al., 2020). We contribute to the state-of-the-art by proposing a descriptive approach to textual data analysis that supports improving policies for distributing educational materials. Our case study focuses on the PNLD 2024 public call for literary works, analyzed from three key aspects: effort, productivity, and quality. We investigate the PNLD program by analyzing the evaluation process of educational materials.

Providing well-designed educational materials to every student is a significant need that may influence learning performance (Liu, 2023; UNESCO, 2016). Public policies related to educational materials such as textbooks and literary works must include robust evaluation processes to ensure content and editorial quality. The PNLD comprises steps for enrollment/validation, pedagogical evaluation, technical qualification, choice (by public school principals and teachers), negotiation, acquisition, and distribution of educational materials (Sobrinho et al., 2023). This paper depicts the pedagogical evaluation process of PNLD, which considers the quality of contents. We expect that the data analysis from evaluation reports can support policy-makers in making informed decisions to improve this type of program.

Despite a comprehensive and centralized process, which includes steps to evaluate the educational materials rigorously, the PNLD program's pedagogical evaluation process encounters challenges that could compromise the distribution time of educational materials to students and the process quality (Silva et al., 2024).These challenges include high evaluator effort, textual inconsistencies, and report quality issues, which can compromise the program's efficiency and delay the timely distribution of materials. This issue transcends the PNLD program and potentially has implications in the countries that conduct official evaluations with similar characteristics. In the context of the Brazilian textbook policy, the pedagogical evaluation process is started when a public call for educational materials is issued, outlining the evaluation criteria to ensure quality. The call guides a rigorous blind evaluation procedure, where a series of questions and justifications are considered, leading to a partial report. A pedagogical and technical committee then thoroughly reviews this report to ensure compliance with the evaluation criteria, culminating in a final report. The evaluation process teams include five key actors: Evaluator 1, Evaluator 2, Assistant Coordinator, Pedagogical Advisor, and Pedagogical Coordinator. The involvement of multiple materials and actors adds complexity, mainly due to the hierarchical evaluation structure extending from the evaluators to the pedagogical coordinator.

We examine the relationships among the documents generated by each participant to analyze issues related to effort, productivity, and quality. As a case study, we pre-process data and analyze reports from evaluators, assistant coordinators, pedagogical advisors, and pedagogical coordinators involved in the PNLD 2024 public call for literary works. We leverage natural language processing techniques, such as text similarity analysis (Wang & Dong, 2020), to handle thousands of answers to rubrics and gain deeper insights into the challenges encountered during the evaluation process of this particular public call. Few researchers have thoroughly analyzed these complexities before textbooks and literary works receive (or fail to receive) official approval (Clark et al., 2024). Among the potential causes for the lack of data on this process may be the unavailability of comprehensive government data since this process might not be digitalized or be restricted to internal stakeholders. As partners of the Brazilian Ministry of Education, the authors have the authorization to research and innovate in the platform (PNLD-Avaliação) that supports the official approval of the textbooks.

2. Methodology

2.1. Descriptive Approach to Textual Data Analysis

We structured our descriptive approach to textual data analysis by defining our goals, Research Questions (RQs), and metrics to support the answers to the RQs. Our initial goal (Table 1) focuses on comprehending the report writing efforts of the pedagogical and technical committee. The final report serves as a written explanation of the committee's decision regarding the literary works' approval (or rejection). We identified six metrics based on text similarity (Wang & Dong, 2020) and text length, aimed at answering three RQs.

Tab. 1 – Goal 1. Understand report writing efforts from analyzing previous works of the assistant coordinator, pedagogical advisor, and pedagogical coordinator.

Question Description	Metric
(RQ1.1) What is the effort level of the assistant coordinator when reusing and editing the peer revision?	**(M1.1)** Similarity between reports from double-blind revisions and assistant coordinator.
	(M2.1) Total length of reports from double-blind revisions and assistant coordinator.
(RQ2.1) What is the effort level of the pedagogical advisor when reusing and editing the report from the assistant coordinator?	**(M3.1)** Similarity between reports from the assistant coordinator and pedagogical advisor.
	(M4.1) Total length of reports from the assistant coordinator and pedagogical advisor.
(RQ3.1) What is the effort level of the pedagogical coordinator when reusing and editing the report from the pedagogical advisor?	**(M5.1)** Similarity between reports from the pedagogical advisor and coordinator.
	(M6.1) Total length of reports from the pedagogical advisor and coordinator.

Our second goal (Table 2) centers on assessing the productivity of the pedagogical and technical committee during the report writing process. We established three metrics based on the time taken to complete tasks to achieve this goal, aiming to address three specific RQs.

Tab. 2 – Goal 2. Understand the productivity in report writing from analyzing previous works of the assistant coordinator, pedagogical advisor, and pedagogical coordinator.

Question Description	Metric
(RQ1.2) What is the productivity level of the pedagogical coordinator?	**(M1.2)** The time the pedagogical coordinator completes tasks.
(RQ2.2) What is the productivity level of the pedagogical advisor?	**(M2.2)** The time the pedagogical advisor completes tasks.
(RQ3.2) What is the productivity level of the assistant coordinator?	**(M3.2)** The time the assistant coordinator completes tasks.

Our third goal (Table 3) concentrates on evaluating the quality of partial reports provided by evaluators regarding compliance, partial compliance, and non-compliance with evaluation criteria. These partial reports serve as textual explanations of the evaluator's decisions. Quality refers to how easy the text is to understand, providing grammatically correct evaluations. To address the four RQs related to this goal, we established seven metrics based on text readability[1], complexity[2], spelling quality, length, and similarity.

Tab. 3 – Goal 3. Understand the quality of partial reports from analyzing previous works of evaluators.

Question Description	Metric
(RQ1.3) What is the writing quality of partial reports?	**(M1.3)** Readability of partial reports.
	(M2.3) The spelling quality of partial reports.
(RQ2.3) What is the level of detail in partial reports?	**(M3.3)** The total length of partial reports.
	(M4.3) The level of complexity of partial reports.
(RQ3.3) What is the level of uniqueness of partial report contents?	**(M5.3)** Similarity of reports of the same evaluator across different documents.
(RQ4.3) What is the agreement level between evaluators?	**(M7.3)** Similarity between partial reports of two evaluators of the same document.

Our fourth and fifth goals (Table 4), respectively, concentrate on the efforts and productivity of evaluators when writing partial reports regarding compliance, partial compliance, and non-compliance with evaluation criteria. We consider four levels of effort, which refer to how laborious it is to produce a text based on previous reports. No effort occurs when no changes are required from the previous reports. For instance, the assistant coordinator reused all text produced by a specific evaluator (i.e., 100% similarity, the same total length, and zero difference between the strings). Minor effort occurs when some changes are required from the previous reports. For instance, the assistant coordinator changed the writing style from the text produced by a specific evaluator. Major effort occurs when significant changes are required from the previous reports. For instance, the assistant coordinator changed the report by mentioning specific external documents, such as laws. Full

[1] Text readability refers to how easily a reader can comprehend a reporting text.
[2] The complexity of the report text is influenced by external references (e.g., laws and national curriculum).

effort occurs when changes in decisions are required from the previous reports. For instance, the assistant coordinator changed the decision from conditionally approved to reproved. Conversely, productivity measures the value generated per hour worked. To address the two RQs associated with these goals, we established three metrics based on task completion time, text length, and text complexity.

Tab. 4 – Goals 4 and 5. Understand the efforts and productivity of writing partial reports from analyzing previous works of evaluators.

Question Description	Metric
(RQ1.4) What is the effort level when writing partial reports?	**(M1.4)** Total length of partial reports.
	(M2.4) The level of complexity of partial reports.
(RQ1.5) What is the productivity level of the evaluators?	**(M1.5)** The time the evaluator completes tasks.

Our sixth goal (Table 5) centers on assessing evaluators' quality of question responses regarding compliance, partial compliance, and non-compliance with evaluation criteria. This analysis is essential because evaluators objectively state one of these three fixed possibilities for each evaluation criterion. To address the two RQs related to this goal, we defined two metrics based on the Kappa index (Cohen, 1960; McHugh, 2012).

Tab. 5 – Goal 6. Understand the quality of questions responses from analyzing previous works of evaluators.

Question Description	Metric
(RQ1.6) What is the agreement level between evaluators?	**(M1.6)** Kappa concordance between two evaluators' answers to the same literary work questions.
(RQ2.6) What is the agreement level of each evaluator with the pedagogical coordinator?	**(M2.6)** Kappa concordance between the evaluator's answer to questions and the final report.

Our seventh and eighth goals (Table 6) concentrate on evaluating the quality of justifications (associated with questions from the sixth goal) and the efforts expended by evaluators in writing them, respectively. Each evaluator produces a textual justification for the responses to the questions. To address the three RQs associated with these goals, we established six metrics based on text readability, spelling quality, length, and complexity.

Tab. 6 – Goals 7 and 8. Understand the quality of justifications and efforts in writing them from analyzing previous works of evaluators.

Question Description	Metric
(RQ1.7) What is the writing quality of justifications?	**(M1.7)** Readability of justifications.
	(M2.7) The spelling quality of justifications.
(RQ2.7) What is the level of detail in justifications?	**(M3.7)** The total length of justifications.
	(M4.7) The level of complexity of justifications.
(RQ1.8) What is the effort level when writing justifications?	**(M1.8)** Total length of justifications.
	(M2.8) The level of complexity of justifications.

2.2. Datasets

To answer our RQs, we gathered data from reports submitted during the 2024 PNLD public call for literary works (focused on the final educational years). Thus, based on the data gathered from the Brazilian Ministry of Education database, we defined four datasets to conduct our analysis: **2024_3_functions_report dataset** (1450 instances, comprising 396 reports from assistant coordinators, 491 from pedagogical advisors, and 563 from pedagogical coordinators), **2024_3_evaluators_report dataset** (927 instances, comprising reports from 181 evaluators), **2024_3_functions_question dataset** (125465 instances, comprising 42226 responses from assistant coordinators, 34821 from pedagogical advisors, and 48418 from pedagogical coordinators), and **2024_3_evaluators_question dataset** (79838 instances, comprising responses from 181 evaluators).

The original datasets contained more than 30 data fields. However, we reduced this to 15 fields to better understand the efforts and capacities of evaluators, assistant coordinators, pedagogical advisors, and pedagogical coordinators. For each dataset, we used the following data fields to conduct our analyses: **CO_call** (identification of the public call), **DS_call** (description of the public call), **CO_object** (identification of the object related to the public call), **DS_object** (description of the object related to the public call), **CO_evaluation_team** (evaluation team alphanumeric code), **DS_function** (description of function - i.e., assistant coordinators, pedagogical advisors, or pedagogical coordinators), **CO_work** (identification of the literary work), **CO_form_block** (identification of a block of an evaluation form), **CO_form_eval** (identification of an evaluation form), **CO_question** (identification of an objective question of the evaluation form), **DS_question** (description of an objective question of the evaluation form), **CO_question_response** (identification of a response to an objective question of

the evaluation form), **DS_question_response** (description of a response to an objective question of the evaluation form), **DS_question_justification** (justification for responding to an objective question of the evaluation form), and **CO_evaluator** (identification of the education professional - i.e., evaluator).

2.3. Data Pre-processing and Analysis

Our data pre-processing procedure comprises seven steps (Figure 1): data verification, data adequacy, dimensionality reduction, encoding, content verification, justification adjustment, and creation of auxiliary data fields. These steps were crucial to enable our data analysis.

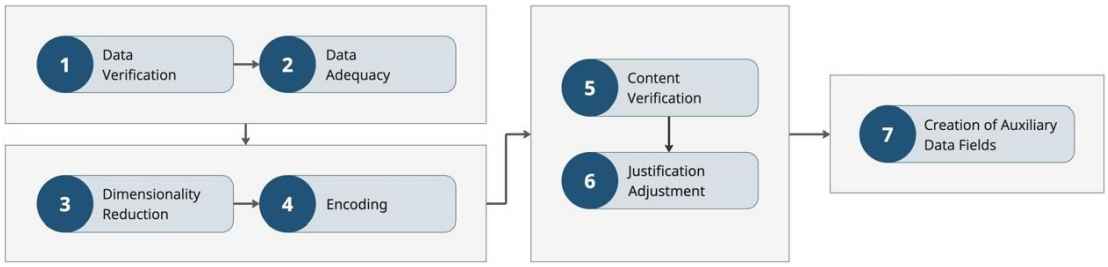


Fig. 1 – Data pre-processing steps.

We checked if the data conformed to the database structure (Step 1). A missing column label was almost always, which caused data inconsistency. We identified the issue and added two labels after the CREATED_AT and UPDATED_AT fields, ensuring the structure is correct (Step 2). Another important task was removing columns with zero or null data (Step 3). We conducted encoding tests to maintain the standard, which is almost always UTF-8. However, there were cases where the encoding only worked with Latin1, making it the extraction encoding used (Step 4). Even with the correct encoding, some columns contained terms with strange characters, which were corrected (Step 5). When a question was not justified (recorded a single space or a period), we included a default text: "Justification not recorded by evaluator" (Step 6). Additionally, when a question was unanswered, we marked it as "Not Answered". Finally, we included new columns to assist in subsequent data analyses (Step 7): **DS_JUST_CLEAN** (content without the tags in cases where the justification is in a mixed HTML + STRING format), **CHAR_COUNT** (number of characters in the justification), **WORD_COUNT** (number of words in the justification), **SENTENCE_COUNT** (number of sentences in the justification), **DS_DAYS_SPENT** (a delta between the "UPDATED_AT" and "CREATED_AT" fields, indicating the number of days worked), and **Q_EVALUATIONS_WORKED** (number of evaluations the stakeholder worked on in the public call).

Our textual transformation to numeric data relied on the Tfidfvectorizer function from the sci-kit-learn library. Additionally, we applied the Cosine, Jaccard, Spacy, and SBERT methods to analyze similarities. The Cosine similarity represents the degree of similarity between two documents (Kwangil Park & Kim, 2020). The Jaccard similarity represents the ratio of the size of the intersection over union (Verma & Aggarwal, 2020). We relied on the Cosine and Jaccard as syntactic similarity measures. Conversely, Spacy relies on word vectors to capture semantic relationships (Honnibal & Montani, 2017). Embedding is usually pre-trained on large corpora, and the Cosine calculates the similarity (Equation 2). Finally, SBERT similarity relies on sentence embeddings from a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model (Reimers & Gurevych, 2019). Similarly to Spacy, calculating similarity relies on the Cosine method. We relied on the Spacy and SBERT similarities as semantic similarity measures.

3. Results and Discussion

3.1. Effort Levels by Function

3.1.1. Assistant Coordinators

The assistant coordinators review the reports from the blind evaluations conducted by the evaluators. To understand the similarity between an assistant coordinator's report and each evaluator assigned to the same literary work, grouping by conditionally approved and reproved, we applied the Cosine, Jaccard, Spacy, and

SBERT methods. We also calculated the mean, median, and Standard Deviation (SD) by grouping all assistant coordinators, considering the public call and evaluation results.

Based on Table 7, the text similarities between assistant coordinators and evaluators' reports for reproved and conditionally approved literary works were compared using four different methods. For the Cosine method, the conditionally approved literary works show a higher mean similarity (0.98) than the reproved literary works (0.66). The median is also higher for conditionally approved literary works (1.00 vs. 0.67). However, the SD is lower for the reproved literary works (0.26 compared to 0.13), indicating that, while the mean similarity is lower, there is greater variability in the conditionally approved group. For the Jaccard method, the conditionally approved literary works again show a higher mean similarity (0.99) compared to reproved literary works (0.83), and the median is significantly higher (1.00 vs. 0.86). The SD is also slightly lower for conditionally approved literary works (0.08 vs. 0.14), indicating more consistent results in the conditionally approved group. The Spacy method shows that conditionally approved literary works have higher similarity metrics, with a mean of 0.99 and a median of 1.00, compared to a mean of 0.94 and a median of 0.97 for the reproved literary works. Additionally, the SD is smaller for conditionally approved literary works (0.04 vs. 0.07), reflecting a higher consistency in the results. Finally, for the SBERT method, the conditionally approved literary works exhibit a higher mean (0.99) and median (1.00) compared to the reproved literary works (mean of 0.91 and median of 0.93). The SD is lower for the conditionally approved literary works (0.05 vs. 0.10), indicating that the similarities for conditionally approved literary works are not only higher but also more consistent. Across all methods, the conditionally approved literary works generally show higher and more consistent similarity scores compared to the reproved literary works.

Tab. 7 – Comparison of text similarities between assistant coordinators and evaluators reports for conditionally approved and rejected literary works.

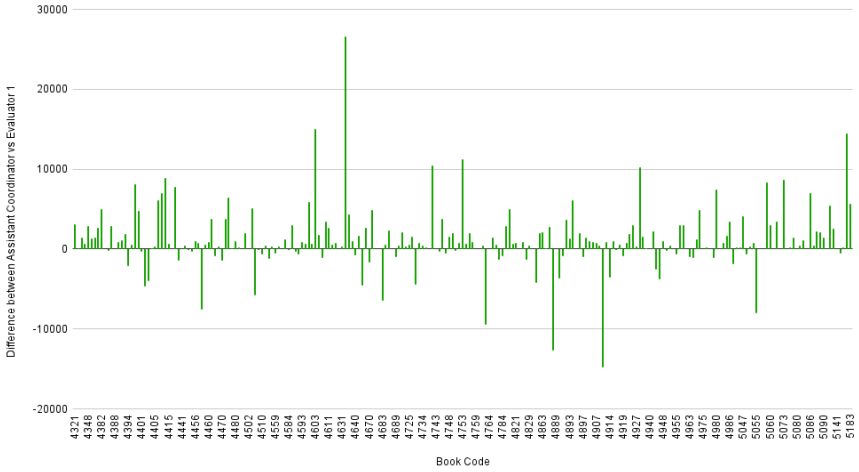
	Reproved			Conditionally Approved		
Metric	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
Cosine	0.66	0.67	0.26	0.98	1.00	0.13
Jaccard	0.83	0.86	0.14	0.99	1.00	0.08
Spacy	0.94	0.97	0.07	0.99	1.00	0.04
SBERT	0.91	0.93	0.10	0.99	1.00	0.05

We applied the Student's t-test to determine whether there are statistically significant differences between the Cosine and Jaccard similarities of the reproved and conditionally approved literary works. For the reproved literary works, we obtained a p-value of 0 for both the Cosine vs. Jaccard and Spacy vs. BERT similarities at a significance level of 0.05. This indicates a significant difference between the means of the two samples, suggesting that the observed variations in similarity metrics are statistically significant. The difference in similarity between the Spacy and SBERT methods can be attributed to two factors. The first is the different approaches each method uses to calculate text similarity. While Spacy relies on basic syntactic and semantic features of words and sentences, SBERT uses pre-trained sentence representations in high-dimensional vector spaces, capturing more complex semantic nuances. These differences result in distinct assessments of how semantically close the texts are, with Spacy perceiving greater proximity and SBERT indicating a more considerable distance. The second factor is the writing style of the assistant coordinator and the evaluator.

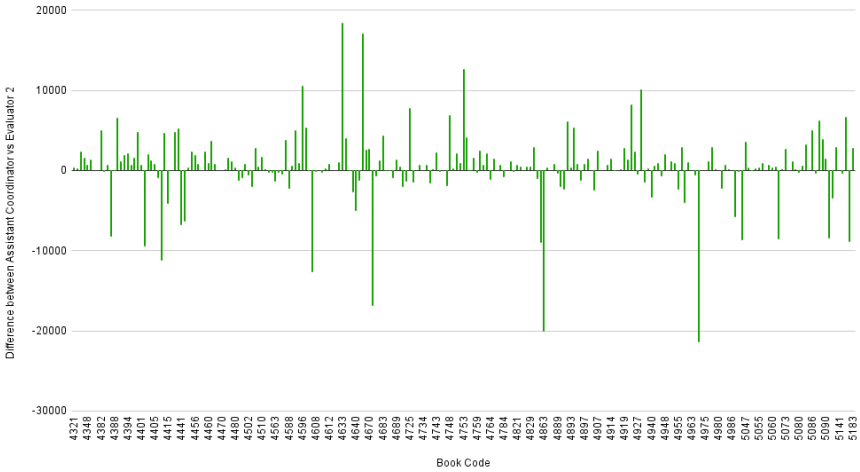
For the conditionally approved literary works, we obtained a moderate p-value (0.4987) for the Cosine and Jaccard similarities based on the Student's t-test. This indicates no significant difference between the means of the two samples for these similarity metrics. Furthermore, the calculated t-statistic value of -0.0156 and the corresponding t-distribution value of 0.4938 support the conclusion that the observed differences are not statistically significant. Conversely, the Student's t-test results, when comparing the Spacy and BERT similarities, presented a high p-value (0.8031), indicating no significant difference between the means of the two samples. The calculated t-statistic value of -0.0151 and the corresponding t-distribution value of 0.4940 confirm that the observed differences are not statistically significant. In general, the high mean values of the semantic similarity metrics indicate relatively high agreement between the assistant coordinators' and evaluators' reasoning. We further explored the Cosine and Jaccard results.

We also analyzed the textual differences between the assistant coordinators' and evaluators' reports by comparing the contents of the reproved literary works. We filtered out literary works with only one evaluation and removed duplicated reports. Figure 2 illustrates the differences between the assistant coordinators' answers to rubrics (for specific literary work) and the evaluators' during the peer review process. For 65 literary

works (Figure 2a), the assistant coordinator either removed or modified some text produced by the evaluator.



(a) Assistant coordinator vs evaluator 1.



(b) Assistant coordinator vs evaluator 2.

Fig. 2 – Difference between the texts of assistant coordinators and the evaluators for reproved literary works.

The assistant coordinator added new text for 156 literary works by incorporating text from another evaluator or creating new content. There was no difference in the answers to rubrics for 12 literary works. In contrast, when considering the second evaluator (Figure 2b), the assistant coordinator removed or modified text from the evaluator for 76 literary works and added text for 147 literary works. For instance, for literary work 4415, the assistant coordinator reused part of Evaluator 1's text (resulting in a text similarity of almost 1.0 across all metrics) while removing or altering some text from Evaluator 2. The texts were also similar.

The previous discussion on text similarities and total length (word count) suggests varying levels of effort by assistant coordinators in adjusting the evaluators' answers to rubrics: no effort, minor effort, major effort, and full effort. However, to increase confidence in answering RQ1.1, it is necessary to investigate specific instances of literary works evaluations to understand better what typically leads to a major or full effort. Thus, we conducted an n-gram analysis to guide our investigation, with n up to 7.

Figure 3 combines a Kernel Density Estimate (KDE) plot (center) with a bivariate contour plot (top and side) for reproved literary works. It displays the joint distribution of two variables: n_gram (X-axis) and jaccard_sims_ngrams (Y-axis). In the central KDE plot, shades of blue indicate varying density levels, with darker areas representing higher data point density. The KDE plot reveals a higher density of points within the 0.0 to 0.2 range for Jaccard similarity, regardless of the n-gram.

This suggests a possible discrepancy in reports produced between the assistant coordinators and evaluators for disapproved works, as the Jaccard similarity is dense. Additionally, a minor density around 1.0 indicates

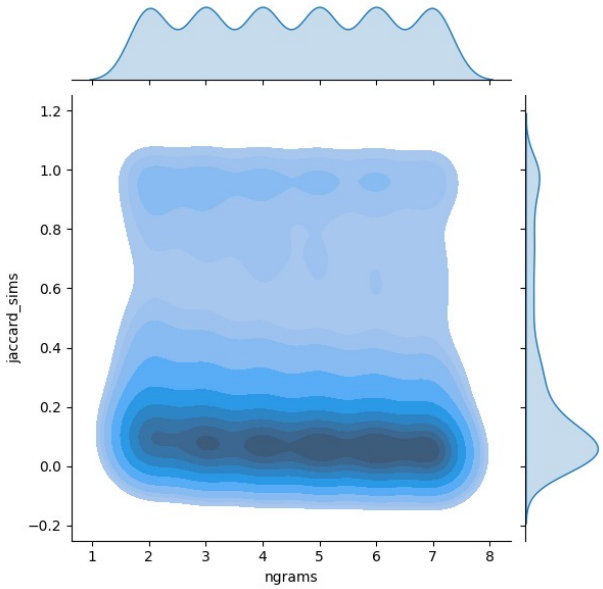


Fig. 3 – KDE plot (center) with a bivariate contour plot (top and side) for reproved literary works.

that some reports are similar, with some being literal copies. We used the visualizations presented in Figures 2 and 3 to focus our analysis on specific problematic cases, aiming to understand the evaluation issues better.

We used the same procedure to analyze the conditionally approved literary works. In the same sense as Figure 2, Figure 4 illustrates the differences between the assistant coordinators' answers to rubrics and the evaluators' during the peer review process, considering conditionally approved literary works. Some of the assistant coordinators' reports (49 in total) removed some text produced by Evaluator 1. Similarly, for Evaluator 2, 44 reports had such removals. Considering both evaluators, for 129 reports, there was no difference between the assistant coordinators' and evaluators' answers to rubrics, showing that, in many cases, assistant coordinators reuse all of the evaluators' produced text. Figure 5 also combines a KDE plot (center) with a bivariate contour plot (top and side) for conditionally approved literary works.

We used the visualizations in Figures 4 and 5 to focus our analysis on specific problematic cases from the conditionally approved. For instance, in literary works 4398 and 4449, the assistant coordinator removed the text from both evaluators and produced new text. The evaluators reproved or approved the literary work. However, the assistant coordinator reverted the decision, producing a different report (i.e., low text similarity).

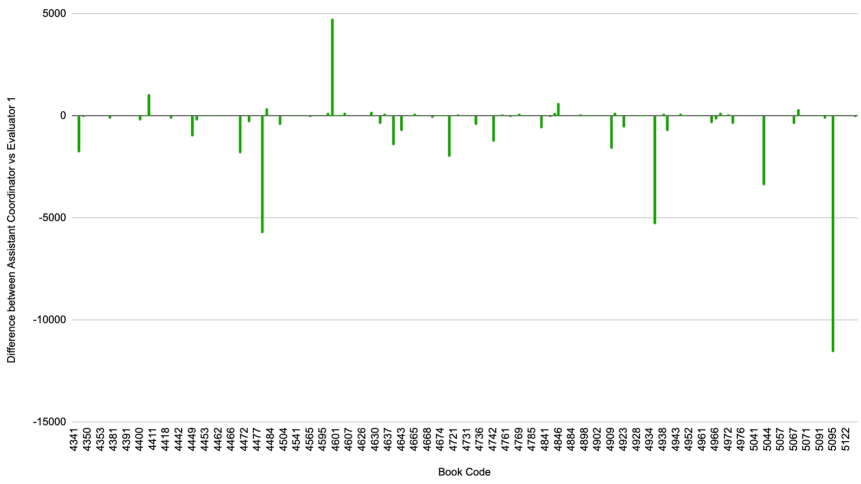
Assistant coordinators' efforts tended to be significant (sometimes full effort) for the reproved literary works, involving removing or adding text from the evaluators' reports. Conversely, no effort or minor effort tends to be required when reusing at least one of the evaluators' reports for conditionally approved literary works.

3.1.2. Pedagogical Advisors and Coordinators

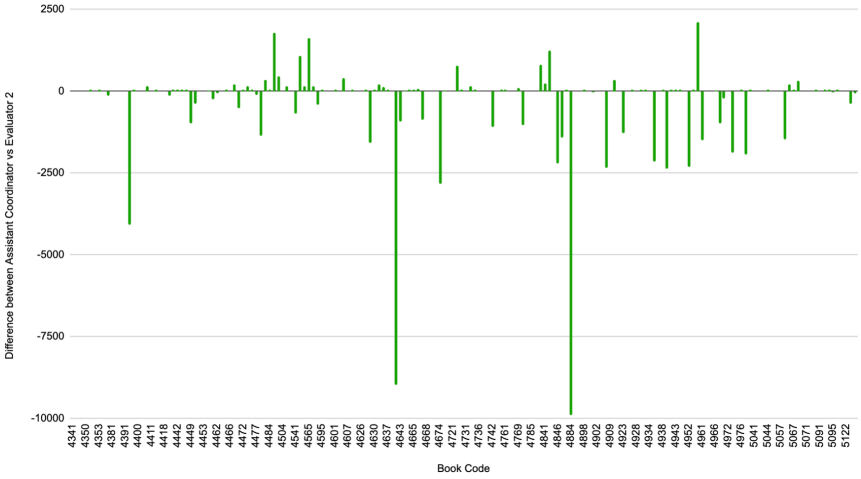
After the assistant coordinator finished analyzing the blind evaluations, 50 literary works were approved, 233 were reproved, and 180 were conditionally approved. Table 8 compares the similarities of the assistant coordinators' and pedagogical advisors' mean, median, and SD. Such results evidence that assistant coordinators and pedagogical advisors agree in their evaluations, with very few differences in the answers to rubrics.

Moreover, Table 9 compares the mean, median, and SD of similarities between the pedagogical advisors' and coordinators' reports. Similar to the previous case, the pedagogical coordinators usually reuse most of the text the pedagogical advisors produce, with only a few adjustments. However, the mean similarities of the reproved literary works slightly decreased for the pedagogical advisors' and coordinators' reports.

Thus, these two phases of the evaluation process typically require a minor effort to consolidate the final report



(a) Assistant coordinator vs evaluator 1.



(b) Assistant coordinator vs evaluator 2.

Fig. 4 – Difference of the answers of assistant coordinators and the evaluators for conditionally approved.

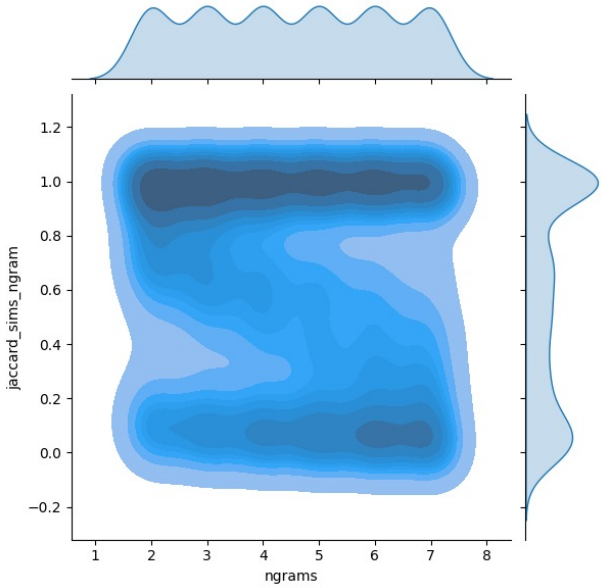


Fig. 5 – KDE plot (center) with a bivariate contour plot (top and side) for conditionally approved literary works.

Tab. 8 – Comparison of text similarities between assistant coordinators and pedagogical advisors' reports for conditionally approved and rejected literary works.

	Reproved			Conditionally Approved		
Metric	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
Cosine	0.90	1.00	0.18	0.93	1.00	0.20
Jaccard	0.95	1.00	0.08	0.97	1.00	0.10
Spacy	0.99	1.00	0.03	0.98	1.00	0.06
SBERT	0.97	1.00	0.05	0.98	1.00	0.06

Tab. 9 – Comparison of text similarities between pedagogical advisors and pedagogical coordinators' reports for conditionally approved and rejected literary works.

	Reproved			Conditionally Approved		
Metric	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
Cosine	0.83	0.90	0.18	0.93	1.00	0.18
Jaccard	0.92	0.95	0.09	0.96	1.00	0.11
Spacy	0.97	0.99	0.04	0.98	1.00	0.05
SBERT	0.95	0.98	0.07	0.98	1.00	0.07

(RQ2.1 and RQ3.1). Only a few specific cases demand a significant effort.

3.2. Productivity Levels by Function

We address RQ1.2, RQ2.2, and RQ3.2 (productivity in report writing) by analyzing the previous works of the assistant coordinator, pedagogical advisor, and pedagogical coordinators. Thus, we computed the value generated per hour worked as the ratio of the number of words of reports and the number of worked days. We analyzed the productivity of the assistant coordinators, pedagogical advisors, and pedagogical coordinators who worked in the 2024 PNLD public call for literary works. Our datasets contain software logs that record the coordinators' access from the first update to the last. The assistant coordinators' value generated per hour worked had a mean of 8.32, a median of 2.69, and an SD of 15.01. The pedagogical advisors' value generated per hour worked had a mean of 4.88, a median of 1.74, and an SD of 8.33. Finally, the pedagogical coordinators' value generated per hour worked had a mean of 3.84, a median of 1.43, and an SD of 6.11. We cannot affirm that such values evidence low productivity. This data does not accurately represent the daily work distribution. A low output value per hour worked may indicate that some evaluations are concentrated on answers to rubrics on specific days by a particular role, which is not the desired behavior (RQ1.2, RQ2.2, and RQ3.2). The Brazilian Ministry of Education staff oversees the pedagogical evaluation process team and requires frequent updates from the assistant coordinator, pedagogical advisor, and pedagogical coordinators. Higher-level functions can request the removal of lower-level functions. For instance, a pedagogical coordinator can ask for the assistant coordinator to be removed. Any function can request to be removed. During monitoring, the Ministry of Education staff or a specific function decides whether to maintain or remove personnel.

3.3. Writing Quality by Evaluators

3.3.1. Readability

We applied the Flesch and Gunning Fog indexes to understand better the readability levels of the evaluators' answers to rubrics. The Flesch index indicates reading levels as follows: 5th to 6th grade (scores 5-6), 7th to 8th grade (scores 7-8), 9th to 10th grade (scores 9-10), 11th to 12th grade (scores 11-12), college-level (scores 13-16), and graduate level (scores 17+). In contrast, the Gunning Fog index classifies texts as very easy (scores 90-100), easy (scores 80-89), fairly easy (scores 70-79), standard (scores 60-69), moderately difficult (scores 50-59), difficult (scores 30-49), and very difficult (scores 0-29).

Table 10 presents an overview of our readability analysis results using the above two measures. Values in parentheses represent the SD, and N is the number of partial reports. The readability of the partial reports varies significantly, with a difference in complexity levels. For instance, the reports for approved literary works are relatively less complex, as indicated by their lower Gunning Fog Index (mean = 19.53, SD = 3.43),

compared to the conditionally approved literary works (mean = 21.61, SD = 3.89). Reports for rejected literary works show a moderate level of complexity (mean = 16.07, SD = 5.49). This variation suggests that the less complex reports for approved and conditionally approved literary works tend to include more direct texts stating the approval, whereas the more complex ones, particularly for conditionally approved literary works, tend to be longer and include excessive detail that may not add substantial value to the decision-making process. Overall, the variation in readability indexes highlights that some texts are more challenging to read than others. This disparity arises from the inconsistent levels of detail in the reports, coupled with the evaluators' diverse writing styles. Consequently, this increases the risk of writing answers to rubrics that are less accessible and harder to interpret, which could hinder the evaluation process and decision-making.

Tab. 10 – Summary of the mean (and SD) readability measures using the Flesch and Gunning Fog indexes.

Result	N	Flesch Index	Gunning Fog Index
Rejected	66	39.6 (19.07)	16.07 (5.49)
Approved	176	34.63 (9.12)	19.53 (3.43)
Conditionally Approved	22	31.9 (19.5)	21.61 (3.89)

When analyzing spelling, the 17 answers to rubrics for approved literary works contained the fewest errors, with only three punctuation mistakes. The 115 answers to rubrics for conditionally approved literary works contained 16 errors, averaging approximately 0.14 errors per text: seven orthographic, six grammatical, and three punctuation errors. The 675 answers to rubrics for rejected literary works exhibited the highest number of errors, totaling 107 errors, averaging about 0.16 errors per text: 53 orthographic, 32 grammatical, and 22 punctuation errors. While the errors increased with the number of reports, the average errors per text remained low, indicating good writing quality in terms of orthographic, grammatical, and punctuation accuracy.

3.3.2. Level of Detail

To enhance our discussion regarding the quality, we used metrics for total length and complexity. Complexity refers to the number of citations of external documents, such as laws and the national curriculum. The total length of evaluation reports for approved literary works had a mean of 269.68, a median of 176.00, and an SD of 350.10. For conditionally approved literary works, the evaluation reports had a mean of 565.52, a median of 220.00, and an SD of 1253.28. The evaluation reports for rejected literary works had a mean of 2656.64, a median of 1219.00, and an SD of 4098.70. Thus, the evaluation reports for approved literary works are relatively short. The conditionally approved literary works show more detailed partial reports than the approved ones, while the rejected literary works partial reports are longer. This higher level of detail in rejection reports is expected, as it enhances confidence in evaluators' decisions and provides robust justification, especially considering that rejections may be subject to appeals. However, a high total length value may indicate redundant or less relevant text. We complement this analysis by using the complexity of reports, as the criteria for approvals and rejections rely on different external documents. We used regular expressions to identify mentions of external documents in the reports, relying on specific keywords (e.g., ABNT NBR) and regular expressions.

The partial reports for approving literary works had a mean of 0.42, with an SD of 0.50. Conversely, the partial reports for conditionally approving literary works had a mean of 0.50, with an SD of 0.50. The reports for rejecting works had a mean of 2.40, with an SD of 3.35. These results suggest that reports for rejecting works provide higher levels of detail due to their greater total length and number of mentions of external documents. Although the conditionally approved generally had longer partial reports than the approved literary works, they contained fewer mentions of external documents. The results suggest that evaluations could provide more details on the criteria used to approve, conditionally approve, or reject literary works (RQ2.3).

3.3.3. Level of Uniqueness

We examined the similarities in partial reports provided by the same evaluator across different literary works (RQ3.3). We organized the reports by grouping those from the same evaluator according to the classifications

of approval, conditional approval, and rejection. Grouping all reports related to approvals and the evaluators who wrote them, we observed that reports had a mean Jaccard similarity of 0.83, a median of 1, and a standard deviation (SD) of 0.29. For conditionally approved, the reports showed a mean similarity of 0.85, a median of 1, and an SD of 0.28. On the other hand, reports for rejected works presented a mean similarity of 0.48, a median of 0.45, and an SD of 0.19. Based on the results from the previous RQs, the high similarity in reports for approvals and conditional approvals was expected, as they typically consist of straightforward statements of approval or partial approval, with minimal details and nearly identical content. However, the moderate mean similarity observed in reports for rejections (approximately 50%) was unexpected, as these reports are generally the most detailed and should be unique, displaying low similarity (RQ3.3).

3.3.4. Agreement Level Between Two Evaluators

Table 11 presents the evaluators' answers to rubrics similarities for approved, conditionally approved, and reproved literary works during the peer review. The metrics indicate that the partial report texts for approved literary works have varying levels of similarity. The Cosine similarity metric shows moderate similarity, while Jaccard, Spacy, and SBERT metrics indicate higher similarities, with Spacy reaching the highest values. The SD is relatively low for the Cosine, Spacy, and SBERT metrics, suggesting a consistent pattern in the text similarities. This indicates that while the texts may differ in structure or specific terms (Cosine and Jaccard), they share semantic content (Spacy and SBERT). For conditionally approved literary works, the similarity levels are generally high, with Cosine, Jaccard, Spacy, and SBERT metrics showing comparable results. The SD values for Cosine suggest more variation in text structure and specific term usage, but the other metrics indicate consistent semantic overlap across the texts. Regarding reproved literary works, the similarity metrics also show high levels of agreement for Jaccard, Spacy, and SBERT. The SD values for the metrics suggest consistency among the reports. Overall, the level of agreement is high during the peer review process (RQ4.3).

Tab. 11 – Evaluators' answers to rubrics similarities for approved, conditionally approved, and reproved literary works during the peer review.

	Approved			Conditionally Approved			Reproved		
Metric	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**	**Mean**	**Median**	**SD**
Cosine	0.65	0.85	0.32	0.67	0.85	0.30	0.48	0.48	0.15
Jaccard	0.82	0.95	0.21	0.82	0.95	0.20	0.75	0.77	0.13
Spacy	0.92	0.97	0.13	0.91	0.96	0.13	0.90	0.93	0.10
SBERT	0.84	0.91	0.15	0.88	0.91	0.14	0.86	0.88	0.11

3.4. Efforts and Productivity of Evaluators

By analyzing evaluators' previous works, we address RQ1.4 and RQ1.5 (efforts and productivity in writing partial reports). Thus, we reused the total length and complexity of the partial reports previously discussed to assess the effort involved in answers to rubrics. The partial reports tend to be brief and less complex for approved literary works. The partial reports for rejected literary works are generally longer. Their complexity is higher than that of approved and conditionally approved literary works. As the evaluation criteria become more stringent and the need for explanations increases, the partial reports' length and complexity rise significantly, indicating a greater effort required from the evaluators.

Moreover, we computed the value generated per hour worked as the ratio of the number of words of partial reports and the number of worked days. The evaluator 1 presented a mean of 8.30, a median of 2.55, and an SD of 19.69. The evaluator 2 presented a mean of 9.97, a median of 2.97, and an SD of 23.19. Again, this data may not accurately reflect the daily work distribution, as evaluators might concentrate tasks on specific days.

3.5. Quality of Questions Response by Evaluators

We address RQs 1.6 and 2.6 (quality in answering objective questions) by analyzing the previous work of evaluators. We calculated the Cohen's Kappa index to assess the level of agreement during the peer review

process. The average Kappa index (with SD) between the first and second evaluators was 0.61 (0.49). We also calculated the index to evaluate the agreement between the evaluators' opinions and the final decisions. The average Kappa index (with SD) between the first evaluators and the final results was 0.70 (0.46). Similarly, the average Kappa index (with SD) between the second evaluators and the final results for the same literary work was 0.68 (0.47). These results indicate a reduced level of agreement across the analyses of evaluators vs evaluators of the same literary work. There is an increase in concordance when considering the evaluators and the final decision. This finding is consistent with our earlier results, highlighting existing disagreements in evaluations, such as cases where a rejection was later changed to a conditional approval (or the contrary).

3.6. Quality of Justifications and Effort from Evaluators

We address RQ1.7, RQ2.7, and RQ1.8 by analyzing previous works of evaluators. We analyzed 79,838 questions across 467 literary works to assess the quality and effort in writing justifications. Among these, 12,363 questions lacked justifications. For approved literary works, the readability of the justifications showed a mean Flesch index of 42.48 (22.11) and a mean Gunning Fog index of 8.39 (4.59). For conditionally approved literary works, the readability scores had a mean Flesch index of 25.60 (21.18) and a mean Gunning Fog index of 9.71 (5.12). For rejected literary works, the justifications had a mean Flesch index of 33.61 (18.59) and a mean Gunning Fog index of 9.10 (5.15). The readability scores indicate that justifications for approved, conditionally approved, and rejected literary works are generally complex to read.

When analyzing spelling quality, the reproved literary works related to 2480 questions, with a mean total length of 283,28 words. The justifications presented 2165 mistakes, of which 1672 were orthographic, 411 were grammatical, and 82 were punctuation. The conditionally approved literary works were related to 307 questions (mean total length of 51,07 words), and the justifications presented 28 mistakes, of which 17 were orthographic, nine were grammatical, and two were punctuation. The approved literary works related to more than 63,845 questions, of which we analyzed 10% of them as a significant sample (mean total length of 109,6 words). The justifications presented 5712 mistakes, of which 4341 were orthographic, 1211 were grammatical, and 161 were punctuation. The number of errors increased with the number of questions. However, similar to our previous results, the number of errors was relatively low, indicating good writing quality regarding orthographic, grammatical, and punctuation accuracy.

Considering the level of detail and effort, we used total length and complexity. For approved literary works (48 and 2 outliers), the mean total length of the justifications was 152,77 words (419.02). These justifications generally exhibited low complexity, with an average of 0.40 (0.49) mentions. For conditionally approved works (181), the mean total length was 140.07 (264,81) words, and the justifications were slightly more complex than those for approved works, with an average of 0.44 (0,63) mentions. For rejected works (235), the mean total length of the justifications was 738,68 (1012,39) words. These justifications were more complex than those for approved and conditionally approved works, with an average of 1.47 (2.69) mentions. The level of detail and effort in justifications seems to increase as the decision is towards a rejection. However, the justifications present a few mentions of external documents, which may decrease clarity.

3.7. Threats to Validity

Evaluating the quality, productivity, and effort involved in writing responses to rubrics entails a degree of subjectivity. To address this problem, we provided clear definitions of these concepts. Productivity is the most challenging aspect to assess, as measuring value generated per hour—based on the number of words written—may not fully capture the complexity of evaluators' behavior. Future research may further refine these definitions to enhance objectivity. Moreover, generalizing these findings to other contexts requires a broader range of analyses, including different public calls for educational materials and textbook programs. As such, some results may not be directly applicable to educational material evaluation programs in other countries or cultural contexts without a more in-depth analysis that takes local specificities into account.

4. Conclusions

Our descriptive approach to textual data analysis enables valuable insights that can drive innovations in public policies and provide strategies to support policymakers in their decision-making processes for educational

material distribution programs. The insights fall within three key dimensions of analysis: effort, productivity, and quality of textual productions/evaluations. Answers to rubrics can reveal desirable or undesirable behaviors that directly influence the quality and efficiency of educational policies, either positively or negatively. Thus, the insights acquired from these answers to rubrics help identify critical points in evaluation processes, enhance transparency, and guide adjustments to improve policy implementation. Additionally, the approach supports monitoring, providing actionable data to inform improvements. The proposed metrics—focused on effort, productivity, and quality—support the principles of digital government by enabling real-time monitoring of the evaluation process through business intelligence tools. This contributes to increased transparency, data-driven decision-making, and greater efficiency in the management of public educational programs. When applied to actual policy implementation, our approach can help shape new processes, procedures, and rules that encourage desired behaviors while discouraging undesired ones.

The case study revealed that the peer review process for evaluating literary works during the 2024 public call was effective despite reduced agreement levels between evaluators. Assistant coordinators showed slightly higher alignment with evaluators' reasoning but played a significant role in refining reports for rejected works, often requiring substantial modifications. While productivity levels were challenging to interpret due to uneven daily work distribution, text readability emerged as a concern, particularly for rejected works, where diverse writing styles and detailed justifications increased complexity. Reports for rejections were more prolonged and more thorough, reflecting the need to enhance confidence in decisions subject to appeals. In contrast, less detailed reports for approvals and conditional approvals may risk reducing transparency, a critical aspect of public policy. Overall, collaboration between roles like assistant coordinators, pedagogical advisors, and coordinators exhibited high agreement, indicating fewer issues in this level of evaluation.

These findings present new research opportunities for process improvement, human resource training, and AI applications. For example, policymakers could leverage AI interventions to enhance the capabilities of evaluators and other roles while also reducing their workload. Generative AI-based chatbots could assist with various tasks, from fundamental analyses like spelling checks to providing text suggestions for report writing. Future research can explore the use of generative AI to improve the quality of evaluation reports.

Funding

The authors acknowledge the support of the Brazilian Ministry of Education through the Decentralized Execution Term (Termo de Execução Descentralizada – TED) No. 10965.

Data Access Statement

Data generated during the study are available from the corresponding author upon reasonable request.

Contributor Statement

André Araújo: Conceptualisation, Data Curation, Funding Acquisition, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Rafael Araújo: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Luciano Cabral: Conceptualisation, Data Curation, Formal Analysis, Investigation, Validation Methodology, Writing-Original Draft, Writing - Review & Editing. Luciane Silva: Conceptualisation, Data Curation, Formal Analysis, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Hilario Tomaz: Conceptualisation, Data Curation, Formal Analysis, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Emerson Martins: Data Curation, Validation, Writing - Review & Editing. Diego Dermeval: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Álvaro Sobrinho: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Alan Pedro da Silva: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Leonardo Marques: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Filipe Recch: Data Curation, Validation, Writing - Review & Editing. Sebastian Munoz-Najar Galvez: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing. Ig Ibert Bittencourt: Conceptualisation, Data Curation, Investigation, Validation, Methodology,

Writing-Original Draft, Writing - Review & Editing. Seiji Isotani: Conceptualisation, Data Curation, Investigation, Validation, Methodology, Writing-Original Draft, Writing - Review & Editing.

Use of AI

During the preparation of this work, the author(s) used ChatGPT in order to conduct orthographic correction. After using this tool/service, the authors reviewed, edited, made the content their own and validated the outcome as needed, and take full responsibility for the content of the publication.

Conflict Of Interest (COI)

There is no conflict of interest.

References

Badrul Hisham, A. A., Mohamed Yusof, N. A., Salleh, S. H., & Abas, H. (2024). Transforming governance: A systematic review of ai applications in policymaking. *Journal of Science, Technology and Innovation Policy*, *10*(1), 7–15.

Clark, P., Llewellyn, K., Capó García, R., & Clifford, S. (2024). Context matters in history textbook studies: A call to address the socio-political landscape of textbook production. *Historical Encounters*, *11*(1), 136–150.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Delen, D. (2019). *Prescriptive analytics: The final frontier for evidence-based management and optimal decision making*. FT Press.

Honnibal, M., & Montani, I. (2017). Spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Kulal, A., Rahiman, H. U., Suvarna, H., Abhishek, N., & Dinesh, S. (2024). Enhancing public service delivery efficiency: Exploring the impact of ai. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*, 100329.

Kwangil Park, J. S. H., & Kim, W. (2020). A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, *34*(5), 396–411.

Liu, X. (2023). Effects of free textbooks on academic performance: Evidence from china's compulsory education. *Review of Development Economics*, *27*(4), 2518–2537.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochem Med (Zagreb)*, *22*(3), 276–282.

Pencheva, I., Esteve, M., & Mikhaylov, S. J. (2020). Big data and ai in education policy: Addressing inequality through technological adoption. *Policy and Internet*, *12*(3), 256–276.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Santoso, A., Kartika, D., Lestari, P., & Yusof, Z. B. (2024). Reimagining public resource allocation: Insights into ai and analytical techniques to address governance challenges in the 21st century. *Journal of Applied Smart Healthcare Informatics*, 66–72.

Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective*. pearson.

Silva, A. S., Araújo, A., Palomino, P., Araújo, R., & Dermeval, D. (2024). Mastering requirements volatility: Strategies for dynamic environments and multiple stakeholders in government software projects. *IEEE Access*, *12*, 183060–183077.

Sobrinho, A., Ibert Bittencourt, I., Carvalho Melo da Silveira, A., Pedro da Silva, A., Dermeval, D., Brandão Marques, L., Cezar Ianzer Rodrigues, N., Carolina Silva e Souza, A., Ferreira, R., & Isotani, S. (2023). Towards digital transformation of the validation and triage process of textbooks in the brazilian educational policy. *Sustainability*, *15*(7).

UNESCO. (2016). Every child should have a textbook.

Verma, V., & Aggarwal, R. K. (2020). A comparative analysis of similarity measures akin to the jaccard index in collaborative recommendations: Empirical and theoretical perspective. *Social Network Analysis and Mining*, *10*(1), 43.

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information*, *11*(9).