A Zero-Shot Prompting Approach for Automated Feedback Generation on ENEM Essays

Rafael T. Anchiêta
rafael.torres@ifma.edu.br
Federal Institute of Maranhão
Caxias, Maranhão

Anthony I. M. Luz
marquesanthony62@gmail.com
Federal Institute of Piauí
Picos, Piauí

Shara L. C. Lopes
shara.lopes@ifpi.edu.br
Federal Institute of Piauí
Campo Maior, Piauí

Raimundo S. Moura
rsm@ufpi.edu.br
Federal University of Piauí
Teresina, Piauí

ABSTRACT

Automated Essay Scoring (AES) has made significant progress in evaluating written texts, but the generation of constructive feedback for essays, particularly in Portuguese, remains underexplored. This paper proposes a zero-shot prompting approach to automatically generate feedback for ENEM essays, an essential component of formative assessments. We evaluate several Large Language Models (LLMs) – Gemini 2.0 Flash, Sabiá 3, Llama 3-8B, and Qwen 3-8B – to provide feedback on five competencies defined by the ENEM scoring rubric. Using the AES-ENEM dataset, we prompt the models to generate feedback for each competency, comparing their semantic similarity with human feedback using the BERTScore metric. Additionally, a linguist with expertise in ENEM essays evaluates the feedback for its constructiveness and informativeness. Our results demonstrate that while the models perform similarly in BERTScore evaluations, the Qwen model produces the most informative and constructive feedback. This work contributes to the development of automated systems that not only grade essays but also assist in improving student writing skills, potentially reducing teacher workload in large-scale assessments.

KEYWORDS

essays, feedback, language model, zero-shot

1 INTRODUCTION

Automated Essay Scoring (AES) is the computer technology that evaluates and scores the written prose [24]. It aims to provide computational models for automatically grading essays with minimal human involvement [21]. This research area began with Page in 1966 with the Project Essay Grader system [21], which have continuously evolved since then, as noted by Ke and Ng [7].

The AES area has recently gained the attention of the Brazilian community through publicly available corpora [13, 14, 19, 27], methods to grade an essay or its characteristics [3, 15, 18, 25], and robustness analysis [2, 26]. Moreover, the PROPOR'24 Competition recently occurred, aiming to develop computer systems capable of automatically evaluating essays [16].

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

Despite the growing interest in the AES area for Portuguese, researchers have overlooked the generation of feedback for essays. Feedback is a crucial element in evaluating essay writing. Assigning grades to essays alone does not effectively help students improve their essay writing skills [17]. Furthermore, guiding students on areas for improvement and how to enhance their writing skills is essential to the learning process of essay writing [23]. Formative assessments, which aim to improve writing proficiency through ongoing interaction, necessitate providing feedback to essay authors [5, 20]. Thus, developing AES systems that not only grade essays but also provide insightful feedback is crucial for practical formative writing assessments.

To bridge the gap in research concerning feedback generation for Portuguese, this paper contributes a zero-shot prompting strategy designed to generate feedback from ENEM essays. The High School National Exam (ENEM - *Exame Nacional do Ensino Médio*) is used to assess the quality of high school education and serves as an admission test for most public and private universities. ENEM adopts specific competencies to grade essays (cross-prompt trait scoring), analyzing aspects such as adherence to formal language norms, text structure, argument development, and the proposal of solutions, making the development of automatic assessment strategies more challenging.

Our methodology investigated several Large Language Models (LLMs), applied to a selection of essays from the AES-ENEM dataset [27]. We prompted Gemini 2.0 Flash[1], Sabiá 3 [1], Llama 3[2], and Qwen 3 [28] models to generate feedback for each ENEM competency. Then, we compared the automatic and manual feedback using the BERTScore metric [29], which measures the semantic similarity between a generated text and a reference text using contextual embeddings from a BERT model [4]. Furthermore, we enlisted the assistance of one linguist with expertise in evaluating ENEM essays to assess and rank the feedback generated by the LLMs, aiming to identify which responses were the most constructive and informative. It is important to say that the human was unaware that the feedback was automatically generated.

Our findings revealed that all LLMs achieved similar results when compared with human feedback using the BERT-Score metric. When assessed by a human, the Qwen model ranked highest across all competencies. Human evaluation has also indicated that this

[1] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash
[2] https://ai.meta.com/blog/meta-llama-3/

automatically generated feedback may be helpful in an educational context, helping teachers reduce their workload. To the best of our knowledge, this is the first initiative on Automated Feedback Generation on ENEM Essays.

The rest of this paper is organized as follows. Section 2 briefly presents the related work. In Section 3, we introduce the corpus used to generate automatic feedback. In Section 4, we detail our zero-shot prompt strategy to condition LLMs to provide feedback. Section 5 discusses the results achieved. Finally, Section 6 concludes the paper and presents future directions.

2 RELATED WORK

As mentioned, researchers have neglected the generation of feedback for essays in Portuguese. Here, we concisely present the works related to the English language, which is the most studied.

Hellman et al. [6] proposed an approach called kNN-MIL, which employs k-nearest neighbor (kNN) and multi-instance learning (MIL) models for determining the impact of each sentence in the overall score of the essay. The aggregated score from the sentences is used to create a document-level holistic score. By framing the AES task as MIL, this work presented a sentence annotation mechanism to enhance the feedback process in building an explainable AES model.

Kumar and Boulanger [8] proposed a deep learning approach to automated essay scoring, which provides feedback using explanation. The proposed approach aims to predict the quality of the writing style and expose the decision process that led to these predictions. To this end, the authors utilized the SHAP tool [12] to generate local and global explanations.

Liu et al. [11] developed an encoder-decoder neural network model called GEEF (Generate Essay Feedback), which addresses different aspects of essay writing, such as fluency, coherence, richness, and literary talent. The authors compared their approach with baseline methods using BLEU [22] and ROUGE [9] metrics, achieving promising results.

Our strategy differs from others because we leverage the text generation capability of language models to provide feedback on essays written in Portuguese. In the next section, we present the corpus used in the experiments.

3 CORPUS

We utilized a selection of essays from the AES-ENEM dataset [27]. This corpus is composed of essays from the *Vestibular UOL*[3] and *UOL Educação*[4] websites, and it consists of 3,586 essays. Besides the graded essays, the corpus includes a prompt, supporting text, and feedback for each essay. However, only 190 essays contain feedback for each ENEM competency, which we used in our experiments. From these 190 essays, we computed some statistics about feedback for each competency, as shown in Table 1.

As we can see, the competencies with the most tokens and sentences are 3 and 1, respectively. The latter is related to adherence to the formal written norm of Portuguese, while the former refers to argumentation in defense of a point of view. We suggest consulting the participant's handbook for more details about the ENEM essay assessment criterion[5].

Table 1: Statistics for each competency.

Competency	Tokens		Sentences	
	Mean	Std	Mean	Std
C1	25.31	12.65	2.83	1.81
C2	19.00	9.29	2.03	1.11
C3	28.15	16.91	3.16	2.22
C4	13.58	6.03	1.52	0.75
C5	15.16	8.62	1.75	1.03

Additionally, we calculated the distribution of essay grades, as illustrated in Figure 1. The distribution is somewhat similar to the normal distribution, with a concentration of essay grades between 480 and 600.

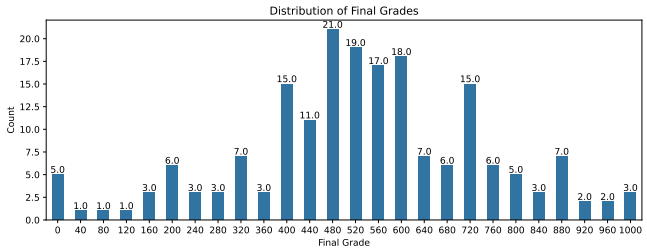

Figure 1: Distribution of final grades.

In the following section, we outline our strategy for generating feedback.

4 ZERO-SHOT PROMPTING

Zero-shot prompting is a technique for conditioning a language model with instructions to perform a task without any examples. This method of conditioning is called "prompting", and it must not be part of the training data of the model [10].

With a zero-shot strategy, we prompted several language models to generate feedback for each ENEM competency. We designed a prompt and organized it into three steps. First, we explain the task context, as depicted in Figure 2. In this context, we requested the model act as a specialized proofreader for ENEM essays, providing it with an essay and supporting text.

```
Act as a specialized proofreader in dissertative-argumentative essays for ENEM.

The following essay was written by a high school student,
and you will need to revise it in just 5 sentences: {essay}

Use the following supporting text to check whether
the writing is in line with the topic: {supporting_text}
```

Figure 2: Context for the models.

[3]https://vestibular.brasilescola.uol.com.br/banco-de-redacoes
[4]https://educacao.uol.com.br/bancoderedacoes/
[5]https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2024_cartilha_do_participante.pdf

Second, we detail the model task, as shown in Figure 3. In the task, the model should provide feedback for each ENEM competency based on the ENEM criteria. Each feedback should be one sentence long and constructive, clear, and detailed. Although the feedback in the corpus is, on average, longer than one sentence, we requested language models to generate one sentence to avoid hallucinations.

```
You must provide feedback for each ENEM competency.

In the first sentence, provide feedback related to the formal use of the Portuguese language.
Take into account the grammatical and spelling aspects, suitability for formal writing,
and choice of vocabulary.

In the second sentence, analyze whether the essay is related to the proposed theme
and the presence of sociocultural repertoire.

In the third sentence, analyze whether the essay selects, relates, organizes,
and interprets information, facts, opinions, and arguments in defense of the chosen point of view.

In the fourth sentence, analyze the cohesion and coherence of the essay.

In the fifth sentence, analyze the proposed intervention in the essay.

Keep a constructive, clear, and detailed tone.
```

Figure 3: Task to be performed by the models.

In the third step, we asked the model for an output, as presented in Figure 4. This output indicates the sentences for each ENEM competency.

```
The output should be as follows:

## Competency 1 ##
First sentence
## Competency 2 ##
Second sentence
## Competency 3 ##
Third sentence
## Competency 4 ##
Fourth sentence
## Competency 5 ##
Fifth sentence
```

Figure 4: Example of output of the model.

From this prompt, we conditioned the following models: Gemini 2.0 Flash, Sabiá 3, Llama 3-8B, and Qwen 3-8B. We chose these models because they have a free Application Programming Interface (API) or are open-source. For example, the first one is a Google model, which has a free-of-charge API that is sufficient for our experiments. The second is a Maritaca AI model. It was trained on documents written in Portuguese, with a special focus on Brazil-related resources. Moreover, it offers a free API for teaching and research purposes. The third and fourth are open-source models from Meta AI and Alibaba Cloud, respectively.

We used the default parameters to prompt the models. Table 2 shows values for `temperature`, `top_p`, and `top_k` parameters. The first controls the randomness and creativity of the generated text. The second controls the cumulative probability threshold for token selection, and the last limits the number of tokens the model considers at each step. The `top_k` value for Sabiá is not informed in its documentation. However, since this model is based on Llama, we believe the value of this parameter is the same as in the Llama model.

Table 2: Parameters for the LLMs.

Model	Parameter		
	Temperature	Top_p	Top_k
Gemini	1.0	0.95	64
Sabiá	0.7	0.95	-
Llama	0.8	0.95	50
Qwen	0.6	0.95	20

In the following section, we detail our experiments and results.

5 EXPERIMENTS AND RESULTS

First, we computed the number of tokens and sentences produced by the models for each competency. Table 3 presents the results for Gemini, Sabiá, while Table 4 shows Llama and Qwen values. The symbol "↑" denotes that the model generated more information than humans (Table 1), whereas "↓" indicates the opposite. Regarding tokens, only Gemini produced fewer tokens (C1, C3) than humans. Concerning the sentences, Gemini (all competencies), Sabiá (C1, C2, and C3), and Llama (C1 and C3) generated fewer sentences than humans. The Qwen model generated significantly more tokens and sentences than the other models, indicating a higher level of verbosity.

Table 3: Statistics for each competency - Gemini and Sabiá.

Competency	Gemini				Sabiá			
	Tokens		Sentences		Tokens		Sentences	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
C1	20.25 ↓	3.67	1.03 ↓	0.18	29.71 ↑	6.16	2.12 ↓	0.46
C2	23.09 ↑	3.73	1.01 ↓	0.07	27.34 ↑	4.29	1.85 ↓	0.38
C3	22.97 ↓	3.27	1.02 ↓	0.14	29.84 ↑	3.61	2.03 ↓	0.25
C4	18.97 ↑	3.02	1.02 ↓	0.14	28.08 ↑	4.66	2.72 ↑	0.58
C5	23.36 ↑	5.64	1.02 ↓	0.12	30.63 ↑	4.58	2.24 ↑	0.48

Table 4: Statistics for each competency - Llama and Qwen.

Competency	Llama				Qwen			
	Tokens		Sentences		Tokens		Sentences	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
C1	28.37 ↑	8.63	2.56 ↓	0.71	49.42 ↑	9.36	3.58 ↑	0.91
C2	26.79 ↑	5.68	2.25 ↑	0.47	45.02 ↑	6.64	3.21 ↑	0.66
C3	28.81 ↑	5.91	2.55 ↓	0.54	47.24 ↑	7.31	3.74 ↑	0.72
C4	20.72 ↑	4.35	2.29 ↑	0.48	44.18 ↑	7.93	3.88 ↑	0.74
C5	30.87 ↑	8.20	2.59 ↑	0.78	44.33 ↑	7.40	3.51 ↑	0.77

In addition to the number of tokens and sentences, we compared automatic and manual feedback using the BERTScore metric [29]. As a BERT model, we utilized the multilingual BERT-based model. One can see in Table 5 that the results are very similar. Only the

Qwen model achieved a slightly lower result than the other models. We believe it is due to the number of tokens and sentences generated by the model.

Table 5: Comparison between automatic and manual feedback.

Competency	BERTScore - F1			
	Gemini	Sabiá	Llama	Qwen
C1	0.67	0.67	0.67	0.66
C2	0.69	0.69	0.69	0.67
C3	0.68	0.68	0.69	0.68
C4	0.69	0.68	0.69	0.67
C5	0.68	0.67	0.68	0.67

To better understand the results, we organized the corpus into intervals of 200 per score, as presented in Table 6. We are interested in knowing the score interval where the automatic feedback agrees with human feedback. However, as in Table 5, the results for each interval and competency are very similar.

Table 6: Detailed comparison between automatic and manual feedback.

Competency	Interval	BERTScore - F1			
		Gemini	Sabiá	Llama	Qwen
C1	[0, 200]	0.67	0.67	0.68	0.66
	(200, 400]	0.68	0.67	0.67	0.67
	(400, 600]	0.68	0.68	0.68	0.67
	(600, 800]	0.68	0.67	0.68	0.66
	(800, 1000]	0.69	0.68	0.69	0.67
C2	[0, 200]	0.70	0.69	0.70	0.68
	(200, 400]	0.70	0.69	0.69	0.68
	(400, 600]	0.69	0.69	0.69	0.68
	(600, 800]	0.69	0.68	0.69	0.68
	(800, 1000]	0.68	0.69	0.69	0.67
C3	[0, 200]	0.68	0.68	0.67	0.68
	(200, 400]	0.69	0.68	0.68	0.68
	(400, 600]	0.68	0.68	0.67	0.68
	(600, 800]	0.68	0.68	0.67	0.68
	(800, 1000]	0.69	0.69	0.69	0.68
C4	[0, 200]	0.69	0.67	0.68	0.67
	(200, 400]	0.69	0.68	0.70	0.68
	(400, 600]	0.70	0.68	0.70	0.68
	(600, 800]	0.70	0.68	0.70	0.67
	(800, 1000]	0.70	0.68	0.69	0.67
C5	[0, 200]	0.70	0.68	0.69	0.67
	(200, 400]	0.69	0.68	0.68	0.68
	(400, 600]	0.68	0.68	0.68	0.68
	(600, 800]	0.68	0.68	0.68	0.68
	(800, 1000]	0.69	0.68	0.69	0.68

Given the similarity in results across models, using BERTScore, we invited one linguist with expertise in evaluating ENEM essays to rank the automatically generated feedback for each competency. Our objective was to identify which model produces the most informative and constructive feedback in each case. To accomplish this, we randomly selected 50 essays, 10 essays for each 200-point interval. Next, the linguist ranked the automatic feedback for each competency. This ranking was based on the ENEM criteria for each competency, analyzing whether the feedback helps a student improve their writing skills. Finally, the linguist chose between the best-ranked feedback of each competency and the human feedback to identify which is more informative and constructive. To prevent bias, the linguist was unaware that the feedback originated from LLM models and also did not know which feedback was produced by a human.

Figure 5 presents the LLMs ranked by the linguist. Notably, the Qwen model consistently achieved the highest ranking across all competencies, producing feedback that is more informative and constructive. We believe that one contributing factor to this result is the number of tokens and sentences generated by the model, even though we prompted it to generate only one sentence.

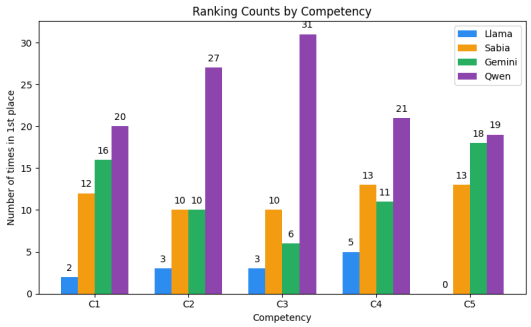

Figure 5: Ranking of the LLM models.

Gemini and Sabiá models were both ranked second. Gemini won Sabiá in the first and fifth competencies, while Sabiá beat Gemini in the third and fourth competencies. This result suggests that Gemini was better able to handle the standard norm of Portuguese (C1) and the intervention proposal for the problem addressed in the essay (C5). In contrast, Sabiá was better at identifying if an essay presents a coherent and cohesive argumentation (C3 and C4). Finally, the Llama model was ranked last in all competencies.

After ranking the models, the linguist voted between the best-ranked feedback of each competency and the human feedback. Table 7 shows the comparison between human feedback and the best-ranked model.

Table 7: Comparison between human feedback and the best-ranked model.

Competency	Human victories	Qwen victories
C1	9	17
C2	2	25
C3	3	28
C4	1	21
C5	3	17

We can see from this table that, out of 50 essays, few human feedbacks were better than those generated by LLMs, indicating that, most of the time, automatic feedback was more constructive and informative than human feedback.

6 FINAL REMARKS

This paper explored a zero-shot prompting strategy for generating automated feedback on ENEM essays, addressing a significant gap in the Automated Essay Scoring (AES) research, particularly for Portuguese. Our findings demonstrated that while quantitative evaluation using BERTScore revealed broadly similar performance across the models, a qualitative assessment highlighted Qwen 3 as the top-performing model, consistently generating feedback deemed more informative and constructive across all ENEM competencies.

For future work, we intend to create a large corpus with human feedback for each ENEM competency. We will also explore other strategies, such as few-shot learning, chain-of-thought, and fine-tuning, using models developed exclusively for Portuguese, such as SoberanIA[6]. The source code and dataset used are available at https://github.com/rafaelanchieta/Feedback-essay.

ACKNOWLEDGMENTS

The authors are grateful to Maritaca AI for providing credits for experiments with Sabiá 3.

REFERENCES

[1] Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. *Sabiá-3*. Technical Report. Maritaca AI.

[2] Rafael T Anchiêta, Rogério F de Sousa, and Raimundo S Moura. 2024. A Robustness Analysis of Automated Essay Scoring Methods. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, Belém, Brazil, 75–80.

[3] Rogério F. de Sousa, Jeziel C. Marinho, Francisco A. R. Neto, Rafael T. Anchiêta, and Raimundo S. Moura. 2024. PiLN at PROPOR: A BERT-Based Strategy for Grading Narrative Essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*. Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, 10–13.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[5] Anton Havnes, Kari Smith, Olga Dysthe, and Kristine Ludvigsen. 2012. Formative assessment and feedback: Making learning visible. *Studies in educational evaluation* 38, 1 (2012), 21–27.

[6] Scott Hellman, William Murray, Adam Wiemerslage, Mark Rosenstein, Peter Foltz, Lee Becker, and Marcia Derr. 2020. Multiple Instance Learning for Content Feedback Localization without Annotation. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Seattle, WA, USA → Online, 30–40.

[7] Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, Macao, China, 6300–6308.

[8] Vivekanandan Kumar and David Boulanger. 2020. Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education* 5 (2020), 22.

[9] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[10] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys* 55, 9 (2023), 1–35.

[6]https://soberania.ai/

[11] Yuanchao Liu, Jiawei Han, Alexander Sboev, and Ilya Makarov. 2024. Geef: a neural network model for automatic essay feedback generation by integrating writing skills assessment. *Expert Systems with Applications* 245 (2024), 123043.

[12] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Long Beach, CA, USA, 4765–4774.

[13] Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2021. Essay-BR: a Brazilian Corpus of Essays. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2021*. SBC, Online, 53–64.

[14] Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2022. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. *Journal of Information and Data Management* 13, 1 (2022), 65–76.

[15] Jeziel C. Marinho, Fábio C., Rafael T. Anchiêta, and Raimundo S. Moura. 2022. Automated Essay Scoring: An approach based on ENEM competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*. SBC, Campinas, Brazil, 49–60.

[16] Rafael Ferreira Mello, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotanif. 2024. PROPOR'24 Competition on Automatic Essay Scoring of Portuguese Narrative Essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*. Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, 1–5.

[17] Haile Misgna, Byung-Won On, Ingyu Lee, and Gyu Sang Choi. 2025. A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review* 58, 2 (2025), 1–40.

[18] Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023. Towards explainable prediction of essay cohesion in portuguese and english. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, Arlington TX USA, 509–519.

[19] Hilário Oliveira, Rafael Ferreira Mello, Péricles Miranda, Hyan Batista, Moésio Wenceslau da Silva Filho, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotani. 2025. A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in Brazilian Portuguese. *Data in Brief* 60 (2025), 111526.

[20] Leanne Owen. 2016. The Impact of Feedback as Formative Assessment on Student Performance. *International Journal of Teaching and Learning in Higher Education* 28, 2 (2016), 168–175.

[21] Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan* 47, 5 (1966), 238–243.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318.

[23] Melissa M Patchan, Christian D Schunn, and Richard J Correnti. 2016. The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology* 108, 8 (2016), 1098.

[24] Mark D Shermis and Felicia D Barrera. 2002. Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring. In *Annual Meeting of the American Educational Research Association*. ERIC, New Orleans, LA, 1–30.

[25] Joyce M Silva, Rafael T Anchiêta, Rogério F de Sousa, and Raimundo S Moura. 2024. Investigating Methods to Detect Off-Topic Essays. In *Proceedings of the 34th Brazilian Conference on Intelligent Systems*. Springer, Belém, Brazil, 346–357.

[26] Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá. 2024. Investigating Universal Adversarial Attacks Against Transformers-Based Automatic Essay Scoring Systems. In *Proceedings of the 34th Brazilian Conference on Intelligent Systems*. Springer, Belém, Brazil, 169–183.

[27] Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024. A New Benchmark for Automatic Essay Scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, 228–237.

[28] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[29] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations*. OpenReview.net, Online.