ENEM Essay Feedback Using LLM-Augmented Prompts

Flavio Carvalho[1], **Vanessa Soares**[1], **Eduardo Bezerra**[1], **Gustavo Guedes**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – (CEFET/RJ)

flavio.carvalho@eic.cefet-rj.br, vanessa.soares@aluno.cefet-rj.br,
ebezerra@cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *This paper evaluates automated feedback generation for ENEM Competency 5 by augmenting zero-shot prompts with an Evaluator Term Set (CTA), a term set proposed in this work and extracted from human evaluator comments on Competency 5. We compare CTA-augmented feedback against a zero-shot baseline on 46 essays with human reference feedback using BERTScore F1 and the Wilcoxon signed-rank test. CTA augmentation increases mean BERTScore F1 for both models, with statistically significant improvement in semantic similarity to human feedback for both evaluated models at* $\alpha = 0.05$.

1. Introduction

The Brazilian National High School Exam (ENEM) is the primary gateway to higher education in Brazil, taken by millions of students annually [INEP, 2024]. Access to consistent feedback is unevenly distributed, with higher-income students benefiting from supplementary courses that provide iterative feedback cycles [Senkevics and Carvalho, 2023; Medeiros, 2016]. Automated feedback can broaden low-income students' access to formative writing support, aligning with Brazil's National Artificial Intelligence Plan (PBIA), which prioritizes AI for social equity [MCTI, 2024].

Automated essay scoring systems assign grades to ENEM essays [Silva and Araujo, 2024]. However, numerical scores alone do not support writing development [Misgna et al., 2024]. Students need feedback that addresses specific competency criteria, yet automated feedback generation remains underexplored [Anchiêta et al., 2025].

Large Language Models (LLMs) enable competency-specific feedback generation via zero-shot prompting [Anchiêta et al., 2025], but are prone to logical inconsistencies [Hooshyar et al., 2025] and hallucinations for underrepresented content [Bezerra, 2025]. Adding domain-specific context to prompts can mitigate some of these limitations [Schreiter, 2025; Bezerra, 2025]. Evaluator comments offer rubric-aligned terms that generic prompts lack.

In this work, we focus on ENEM Competency 5 (intervention proposal), selected for its complexity and underexploration in the automated feedback literature [Anchiêta et al., 2025], and propose an Evaluator Term Set (CTA, *Conjunto de Termos de Avaliadores*), recurring n-grams extracted from human evaluator comments on Competency 5 in the Automated Essay Score (AES) ENEM dataset [Silveira et al., 2024]. We compare CTA-augmented feedback to a zero-shot baseline [Anchiêta et al., 2025] using BERTScore F1 against human reference feedback and assess statistical significance with the Wilcoxon signed-rank test for two locally-run, open-weight LLMs: `mistral_pt`, a Portuguese-adapted model, and `llama32`, a general-purpose multilingual model.

Our contributions include (i) compiling a CTA from human evaluator comments, (ii) proposing a prompt augmentation strategy based on CTA insertion, and (iii) providing empirical evidence that CTA augmentation increases semantic similarity to human feedback with statistical significance for both evaluated models. The paper reviews related work (Section 2), describes our method (Section 3), presents results (Section 4), and discusses conclusions (Section 5).

2. Related Work

We conducted a literature survey in early 2026 using Google Scholar with the terms "LLM", "feedback", and "ENEM". While the search returned over 100 results on automated essay scoring for ENEM, research on feedback generation is sparse. Focusing specifically on feedback generation, we identified two studies that produce textual feedback for ENEM essays using LLM-based systems [Anchiêta et al., 2025; Silva and Araujo, 2024], with one focusing on feedback generation and the other on scoring with feedback as additional output.

Anchiêta et al. [2025] generate feedback for ENEM essays with zero-shot prompting. They evaluate four LLMs on 190 essays from the AES-ENEM dataset [Silveira et al., 2024]. Each prompt has three sections (context, task, output format) and requests one sentence of feedback per competency. They compare generated feedback to human feedback with BERTScore [Zhang et al., 2019], finding similar scores among models (F1 $\approx$ 0.67-0.69). A linguist with ENEM expertise then ranks feedback from 50 essays, identifying Qwen 3-8B as most informative and constructive.

Silva and Araujo [2024] propose an LLM-based solution structured as a prompt chain following the official ENEM rubric [INEP, 2025]. The study focuses on automated scoring and reports performance results for competency-level and final grades using agreement and correlation measures with human evaluators. The system produces competency-specific feedback as supplementary output, described as detailed and criterion-oriented, but does not evaluate feedback quality similarity to human comments.

This work differs from prior approaches in scope and evaluation method. We augment a zero-shot prompt with a CTA extracted from human evaluator comments in a corpus of ENEM essay practice texts, focusing specifically on Competency 5. We compare CTA-augmented feedback against a zero-shot baseline using BERTScore and apply the Wilcoxon signed-rank test to assess whether domain knowledge augmentation increases semantic similarity to human feedback.

3. Methodology

We compare feedback generated by our CTA-augmented approach against the zero-shot baseline proposed by Anchiêta et al. [2025]. We use the `sourceAOnly` subset of the AES-ENEM dataset [Silveira et al., 2024][1], the only subset containing per-competency evaluator comments, drawn from the Educação UOL platform ("Source A") [Silveira et al., 2024] and following a scoring scale consistent with the official ENEM rubric. This subset is divided into train, validation, and test splits.

Of the 258 essays in `train-sourceAOnly`, 167 contain valid Competency 5 feedback, which are used to construct the CTA. The `validation-sourceAOnly`

[1] Available at `https://huggingface.co/datasets/kamel-usp/aes_enem_dataset`.

subset (66 essays, 46 with valid C5 feedback) provides the essay texts for baseline and CTA-augmented feedback generation, and the corresponding human C5 comments serve as reference for BERTScore evaluation. The `test-sourceAOnly` subset is not used in this work. Since all methodological decisions, including CTA construction and parameter selection, were made using the validation subset, the reported results should be interpreted as validation-set evidence, leaving the test set for future confirmation under a fixed experimental design. Figure 1 illustrates our experimental workflow, showing baseline and CTA-augmented paths, CTA construction from evaluator feedback, and evaluation strategy using BERTScore and Wilcoxon testing.

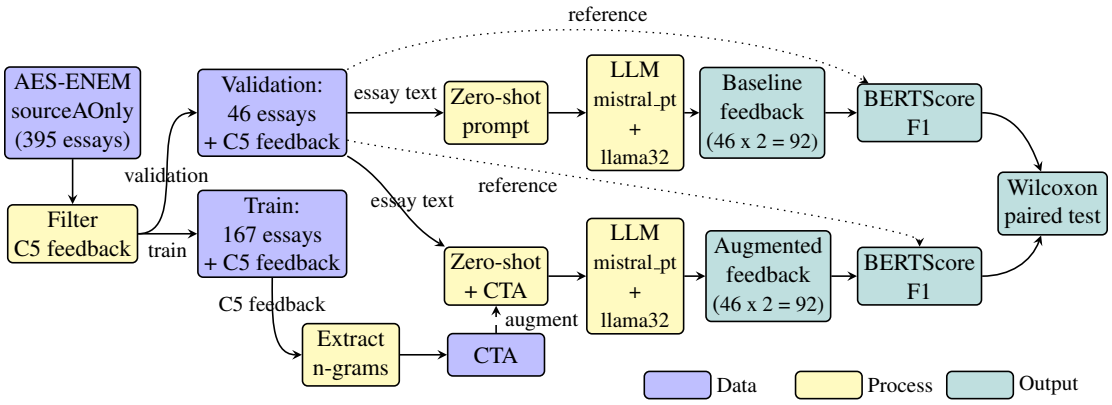

Figure 1. Experimental methodology: CTA construction from training feedback (167 essays), baseline and CTA-augmented generation on the validation subset (46 essays), and BERTScore F1 evaluation with Wilcoxon testing.

Preprocessing begins with the `train-sourceAOnly` subset (258 essays) with fields for prompt, essay text, supporting text, competency scores, and evaluator comments. We extract Competency 5 feedback and retain essays with a valid non-empty C5 comment, yielding 167 essays. The same preprocessing procedure is applied to `validation-sourceAOnly`, yielding 46 essays used for feedback generation and evaluation. Each essay in both subsets is assigned a stable identifier (`essay_hash`) computed as the SHA-256 hash of the normalized essay text, ensuring consistent pairing across experimental conditions.

We adapt the zero-shot prompting strategy described by Anchiêta et al. [2025] to generate baseline feedback. Following the same three-stage structure of context, task, and output format, we translate the prompt to Portuguese to ensure models generate feedback in the essay's language and focus it on Competency 5 only. Figure 2 presents the complete prompt used for baseline generation.

We apply the prompt individually to each essay using two LLMs running locally via Ollama v0.5.7 [Marcondes et al., 2025]. The models are `cnmoro/mistral_7b_portuguese:q2_K` (7B parameters, Q2_K quantization, digest `321345d18e55`; referred to as `mistral_pt`), which was selected for its Portuguese language adaptation, and `llama3.2:latest` (3.21B parameters, Q4_K_M quantization, digest `a80c4f17acd5`, referred to as `llama32`), selected for its general multilingual capability and availability. Both models are configured following Anchiêta et al. [2025], with `top_p`=0.95, `top_k`=50, and temperatures 0.7 for `mistral_pt` and 0.8 for `llama32`. Experiments were executed on an Intel Core i5-10400 CPU with an

NVIDIA GeForce GTX 1650 GPU (4 GB VRAM) running Ollama v0.5.7.

The baseline generates 92 feedback outputs (46 essays $\times$ 2 models). Using the same essays, models, and parameters, the CTA-augmented condition generates another 92 outputs, yielding paired baseline/CTA feedback for each essay. We compare each condition to the human reference feedback with BERTScore F1 and test paired differences with the Wilcoxon signed-rank test.

(a) Zero-shot baseline prompt

Atue como um revisor especializado em redações dissertativo-argumentativas do ENEM.
A redação a seguir foi escrita por um estudante do ensino médio, e você deverá revisá-la em apenas 1 sentença (somente para a Competência 5): {essay_text}
Use o texto de apoio a seguir para verificar se a redação está alinhada ao tema: {supporting_text}
Você deve fornecer feedback apenas para a Competência 5 do ENEM. Na única sentença, analise a proposta de intervenção apresentada na redação. Mantenha um tom construtivo, claro e detalhado.
IMPORTANTE: Siga o formato de saída EXATAMENTE como especificado abaixo. Não adicione qualquer texto adicional, explicações ou formatação fora do formato especificado. Use APENAS o marcador "## Competency 5##" para apresentar o feedback.
Competency 5##
Sua sentença

(b) CTA-augmented prompt (addition highlighted)

Atue como um revisor especializado em redações dissertativo-argumentativas do ENEM.
A redação a seguir foi escrita por um estudante do ensino médio, e você deverá revisá-la em apenas 1 sentença (somente para a Competência 5): {essay_text}
Use o texto de apoio a seguir para verificar se a redação está alinhada ao tema: {supporting_text}

BASE DE CONHECIMENTO — use os termos abaixo como referência avaliativa para enriquecer sua análise. NÃO os repita mecanicamente.
• term_A • term_B … • term_T

Você deve fornecer feedback apenas para a Competência 5 do ENEM. Na única sentença, analise a proposta de intervenção apresentada na redação. Mantenha um tom construtivo, claro e detalhado.
IMPORTANTE: Siga o formato de saída EXATAMENTE como especificado abaixo.
Não adicione qualquer texto adicional, explicações ou formatação fora do formato especificado.
Use APENAS o marcador "## Competency 5##" para apresentar o feedback.
Competency 5##
Sua sentença

Figure 2. Prompt structures used for feedback generation. (a) Zero-shot baseline. (b) CTA-augmented: a knowledge base block is inserted between the essay content and the output instructions; all other content is identical to (a).

The CTA is a knowledge injection strategy, distinct from few-shot prompting. Providing full essay and feedback pairs as in-context examples would require selecting representative instances and risks introducing topic-specific bias, and it also increases prompt length substantially for models with limited context windows. The CTA instead provides a compact, topic-agnostic vocabulary extracted from the evaluator community, compatible with the zero-shot design of the baseline [Anchiêta et al., 2025].

We compile a *Conjunto de Termos de Avaliadores* (CTA, Evaluator Term Set) to augment prompts with domain knowledge specific to Competency 5 (intervention proposal), which is evaluated on five aspects: concrete action, executing agent, means of implementation, detailing, and respect for human rights [INEP, 2025]. The CTA com-

prises recurring n-grams extracted from human evaluator comments on Competency 5 in the AES-ENEM dataset.

We extract n-grams of length 2 to 4 tokens from the 167 human evaluator comments on Competency 5 using the Natural Language Toolkit (NLTK) [Bird et al., 2009], after removing standard Portuguese stopwords. Additionally, we removed 13 high-frequency terms identified as non-discriminative for Competency 5 assessment through empirical frequency analysis of the 167 training comments. These include terms of structural reference (e.g., *autor*, *parágrafo*), topic-specific terms tied to particular essay themes rather than C5 criteria (e.g., *hino*, *armas*), and evaluator format artifacts (e.g., *competência*, *vermelho*). We retain only n-grams present in at least three distinct comments and select the top 20 by document frequency, yielding a single term set of evaluator expressions applied uniformly to all essays. The term extraction procedure results in a ranked list serving as the CTA, and Table 1 shows the 10 highest-ranked terms.

Table 1. 10 highest-ranked terms of the CTA, ranked by document frequency (DF).

Rank	Term	DF	Rank	Term	DF
1	sugestão intervenção	18	6	além disso	5
2	conclusão coerente	16	7	proposta solução	4
3	resolver problema	10	8	sugestões intervenção	4
4	proposta intervenção	6	9	solução problema	4
5	quer dizer	6	10	conclusão confusa	4

The terms reflect expressions tied to the intervention construct (e.g., *sugestão intervenção*, *proposta intervenção*, *resolver problema*) and evaluative quality markers used to signal proposal weaknesses or strengths (e.g., *intervenção genérica*, *conclusão confusa*, *conclusão coerente*). Those two types of evaluator vocabulary provide the LLM with register characteristic of human comments, without prescribing essay content.

We augment the zero-shot prompt with CTA to provide evaluator-grounded context for Competency 5 feedback generation. For each essay, we insert a list of the 20 most frequent terms into the prompt, with one term per line, as shown in Figure 2. The CTA block is inserted between the essay content and the output format instructions, without modifying any other part of the prompt. All 46 essays receive the same CTA block.

We maintain experimental consistency by using the same models (`mistral_pt` and `llama32`) and parameters employed in baseline generation. The CTA block is the only difference between conditions, and all other prompt content and generation parameters are identical. This procedure generates 92 CTA-augmented feedback instances (46 essays $\times$ 2 models) for paired comparison with the baseline.

We evaluate semantic similarity between generated and human reference feedback for Competency 5 using BERTScore F1 [Zhang et al., 2019], measured with `bert-base-multilingual-cased` via the `bert-score` library (`v0.3.13`), setting `lang="pt"` without IDF weighting or baseline rescaling. We conduct pairwise comparisons matching each generated feedback instance with the corresponding human C5 feedback by `essay_hash`, retaining only pairs where both texts exceed 10 characters. BERTScore F1 values are organized by model and condition (baseline $\times$ CTA-augmented) and used as input to the Wilcoxon signed-rank test at $\alpha = 0.05$. All

code, generated feedbacks, CTA term list, and evaluation results are publicly available at `https://github.com/LaCAfe/enem-feedback-cta`, while the AES-ENEM dataset is available separately from its original source [Silveira et al., 2024].

4. Results

Table 2 presents mean BERTScore F1 values for Competency 5, comparing the zero-shot baseline and the CTA-augmented approach for each model. Mean BERTScore F1 values range from 0.673 to 0.686, with standard deviations between 0.020 and 0.028 across models and conditions. The analysis considers only paired comparisons between human and automatic comments for which both conditions produced valid outputs.

Table 2. BERTScore F1 for Competency 5 feedback, comparing the zero-shot baseline and the CTA-augmented approach. Values are mean ± standard deviation over valid paired instances (N = 46) per model.

Model	N	Zero-shot (mean±std)	CTA (mean±std)
`mistral_pt`	46	0.675±0.020	**0.686**±0.028
`llama32`	46	0.673±0.025	**0.683**±0.027

We apply a one-sided paired Wilcoxon signed-rank test (CTA $>$ baseline), justified by the directional hypothesis that CTA augmentation increases semantic similarity to human feedback. Both models show significant improvement at $\alpha = 0.05$: $p = 0.0050$ for `mistral_pt` and $p = 0.0263$ for `llama32` over $N = 46$ paired instances. Bonferroni correction is not applied, as the two models constitute independent replications of the same hypothesis [Perneger, 1998].

To assess whether the observed gains reflect lexical overlap rather than generation quality, we computed BERTScore F1 for synthetic strings formed by concatenating the top-N CTA terms ($N \in \{5, 10, 15, 20\}$) against the same 46 human references. All synthetic conditions scored below the zero-shot baseline (maximum 0.607 ± 0.029 at $N = 5$, versus baseline 0.673 and 0.675), indicating that lexical overlap alone does not account for the CTA-augmented gains.

5. Conclusion

We investigate the use of a *Conjunto de Termos de Avaliadores* (CTA, Evaluator Term Set) as additional knowledge for automated feedback generation on ENEM essays, focusing on Competency 5. We compare a zero-shot baseline with a CTA-augmented approach by measuring semantic similarity between automatic and human feedback using BERTScore and statistical testing. Results indicate that CTA augmentation increases semantic similarity relative to the zero-shot baseline, with both evaluated language models showing statistically significant improvement at $\alpha = 0.05$.

This work suggests that incorporating evaluator-derived terminology from human feedback comments can increase BERTScore semantic similarity between LLM-generated and human feedback, providing a replicable framework for augmenting zero-shot prompts with evaluator content knowledge. Future work should extend evaluation to the remaining ENEM competencies and complement automatic metrics with qualitative analysis by human evaluators to assess feedback qualities not captured by semantic similarity.

Limitations

This study has methodological limitations that should be considered when interpreting results. The quantitative approach using BERTScore as the primary comparison criterion does not capture qualitative aspects of generated feedback such as clarity, pedagogical appropriateness, tone, or actionability. While BERTScore measures semantic similarity to human feedback, it does not assess whether feedback effectively supports student learning or addresses specific writing deficiencies.

Both experimental conditions use a single generation per essay per model. Because LLM outputs are stochastic at temperatures above zero, the observed BERTScore differences reflect one sample from the generation distribution. Multiple independent runs per condition, or fixing the Ollama `seed` parameter alongside the model digest, would be required to confirm the stability of the reported gains; this remains an avenue for future validation.

Regarding scope limitations, our analysis focuses exclusively on Competency 5 (intervention proposal) of ENEM essays. Results may not generalize to other competencies, which assess different writing aspects such as grammatical correctness, argumentation structure, or textual cohesion. The CTA augmentation strategy targets terminology specific to intervention proposals and would require adaptation for other competencies.

The selection of additional stopwords for CTA refinement involved researcher judgment based on empirical frequency analysis. The complete list and the extraction code are publicly available, ensuring full reproducibility of the reported results. However, applying this approach to a different dataset would require a new curation step, as the additional stopwords are specific to the distribution of terms in the training comments used here.

Ethics Statement

The data used in this study come from publicly available sources. No primary data collection involving human participants was conducted. This study analyzes student essays and evaluator comments from two existing public datasets compiled by third parties for research purposes and made available in open repositories and educational platforms.

The research focuses on computational analysis of automated feedback generation for educational assessment. No operational system affecting students' academic outcomes was deployed. Generated feedback was analyzed solely for research comparison against existing human evaluator comments. The study acknowledges that automated methods may reflect biases present in source materials, and results are reported with methodological and scope limitations.

Disclosure of Generative AI Usage

The authors declare the use of generative AI tools during the preparation of this manuscript. Specifically, we used Claude Sonnet 4.5 (Anthropic) to correct the grammar, without changing the tone, style or adding any new information, as well as to assist in writing code followed by revisions and necessary modifications. All final editorial decisions, including verification of consistency, originality, and technical correctness of the content, remained under the sole responsibility of the authors.

References

Anchiêta, R. T., Luz, A. I., Lopes, S. L., and Moura, R. S. (2025). A zero-shot prompting approach for automated feedback generation on ENEM essays. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, pages 511–515. SBC.

Bezerra, E. (2025). Introduction to LLM-Based agents. In Duarte, D., Timbó, F., Schreiner, G., and Chaves, I., editors, *Tópicos em Gerenciamento de Dados e Informações: Minicursos do SBBD 2025*, chapter 2. Sociedade Brasileira de Computação, Porto Alegre.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Hooshyar, D., Yang, Y., Šír, G., Kärkkäinen, T., Hämäläinen, R., Cukurova, M., and Azevedo, R. (2025). Problems with large language models for learner modelling: Why LLMs alone fall short for responsible tutoring in K–12 education. *arXiv preprint arXiv:2512.23036*.

INEP, I. (2024). ENEM 2024 tem 4,3 milhões de inscritos confirmados. Acesso em: 2 nov. 2024.

INEP, I. (2025). A redação no ENEM 2025: cartilha do(a) participante. Publicado em: 03 out. 2025 08:46. Acesso em: 30 mar. 2026.

Marcondes, F. S., Gala, A., Magalhães, R., Perez de Britto, F., Durães, D., and Novais, P. (2025). Using ollama. In *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*, pages 23–35. Springer.

MCTI, M. (2024). Plano brasileiro de inteligência artificial (PBIA) 2024-2028. Gov.br. Acesso em: 26 set. 2024.

Medeiros, M. (2016). Income inequality in Brazil: new evidence from combined tax and survey data. *World Social Science Report*, page 107.

Misgna, H., On, B.-W., Lee, I., and Choi, G. S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, 58(2):36.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *Bmj*, 316(7139):1236–1238.

Schreiter, D. (2025). Prompt engineering: How prompt vocabulary affects domain knowledge. *arXiv preprint arXiv:2505.17037*.

Senkevics, A. S. and Carvalho, M. P. D. (2023). Youth and access to higher education: On the non-place of the preparatory course student. *Educação Em Revista*, 39.

Silva, W. A. d. and Araujo, C. C. d. (2024). Automated ENEM essay scoring and feedbacks: A prompt-driven LLM approach. Trabalho de Graduação, Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Recife, Brasil.

Silveira, I. C., Barbosa, A., and Mauá, D. D. (2024). A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 228–237.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.