REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

THE USE OF CHATGPT IN THE ASSESSMENT OF A PERFECT-SCORE ENEM 2024 ESSAY: A COMPARATIVE ANALYSIS OF HUMAN AND AI-BASED EVALUATION

USO DO CHATGPT NA CORREÇÃO DE UMA REDAÇÃO NOTA MIL DO ENEM 2024: UMA ANÁLISE COMPARATIVA ENTRE A AVALIAÇÃO HUMANA E A ARTIFICIAL

Linguística & Letras e Artes • 25/06/2026

REGISTRO DOI: 10.70773/revistatopicos/782353599

João Gabriel Dias Sousa[1]

Rayla Borges Oliveira[2]

RESUMO

This study analyzes how ChatGPT, in its free and Plus versions, evaluates and comments on a 1,000-point *ENEM* 2024 essay, comparing its responses with the human/official parameters established in the *Participant's Guide*. The theoretical framework draws on the *ENEM* scoring matrix proposed by Inep (2025), studies on formative feedback by Sadler (1989) and Hattie and Timperley (2007), as well as discussions on automated writing assessment developed by Shermis and Burstein (2013), Deane (2013), Chapelle, Cotos and Lee (2015), Ranalli, Link and Chukharev-Hudilainen (2017), Wilson and Roscoe (2020), and Steiss et al. (2024). Methodologically, this is a qualitative, documentary, descriptive-comparative study, configured as a single-case study. The corpus comprises a maximum-score essay, Inep's official comment, and the corrections produced by the two versions of ChatGPT, submitted to the same prompt and analyzed according to categories such as adherence to the scoring matrix, diagnostic accuracy, consistency between score and comment, feedback actionability, and comparison between versions. The results indicate that the free version assigned 960 points and was closer to the official assessment, although with lower analytical density; the Plus version assigned 840 points, providing more detailed feedback, but also adopting a more severe stance and showing less alignment with the official evaluation.

Keywords: ChatGPT; ENEM essay; Automated writing assessment; Formative feedback; Artificial intelligence.

ABSTRACT

Este estudo analisa como o ChatGPT, nas versões gratuita e *Plus*, avalia e comenta uma redação nota 1000 do Enem 2024, comparando suas respostas aos parâmetros humano-oficiais da Cartilha do Participante. A fundamentação teórica articula a matriz

avaliativa do Inep (2025), os estudos sobre *feedback* formativo de Sadler (1989) e Hattie e Timperley (2007), além das discussões sobre avaliação automatizada da escrita em Shermis e Burstein (2013), Deane (2013), Chapelle, Cotos e Lee (2015), Ranalli, Link e Chukharev-Hudilainen (2017), Wilson e Roscoe (2020) e Steiss et al. (2024). Metodologicamente, trata-se de uma pesquisa qualitativa, documental, descritivo-comparativa e configurada como estudo de caso único. O corpus reúne uma redação nota máxima, o comentário oficial do Inep e as correções produzidas pelas duas versões do ChatGPT, submetidas ao mesmo prompt e analisadas segundo categorias como aderência à matriz, acurácia diagnóstica, coerência entre nota e comentário, acionabilidade do *feedback* e comparação entre versões. Os resultados indicam que a versão gratuita atribuiu 960 pontos e se aproximou mais da avaliação oficial, embora com menor densidade analítica; a versão Plus atribuiu 840 pontos, oferecendo *feedback* mais detalhado, porém mais severo e menos alinhado à leitura oficial.
Palavras-chave: ChatGPT; Redação do Enem; Avaliação automatizada da escrita.*Feedback* formativo; Inteligência artificial.

1. INTRODUCTION

The growing presence of generative artificial intelligence systems in educational practices has produced new forms of interaction among students, teachers, texts, and assessment processes. In the field of written production, tools such as ChatGPT have come to be used not only for linguistic revision, but also for correction, score assignment, and the generation of *feedback* on school and academic texts. This scenario becomes especially relevant when one considers the essay of the National High School Exam (Enem), a large-scale assessment that occupies a central place in access to Brazilian higher education

and whose correction is organized according to an analytical matrix composed of five competencies. In this context, investigating how an AI assesses essays considered exemplary by Inep itself makes it possible to discuss both the pedagogical potential of these technologies and their assessment-related limits.

The Enem essay requires the production of a dissertative-argumentative text in the formal written register of Portuguese, in which the participant must defend a point of view, mobilize sociocultural repertoire, organize arguments coherently, and present an intervention proposal that respects human rights. According to Inep (2025), the final essay score can reach 1,000 points, distributed across five assessment competencies, each scored from 0 to 200 points. It is therefore a standardized assessment process, institutionally controlled and conducted by trained human evaluators. By contrast, the assessment performed by AI tools results from an interaction mediated by a *prompt*, which entails considering not only the generated response, but also the model used, the version accessed, and the instructions provided to the system.

In view of this, this article is justified by the need to understand, critically, the extent to which ChatGPT can produce assessments and comments compatible with the official criteria for the Enem essay. This discussion has pedagogical, social, and ethical relevance. Pedagogically, because teachers and students have turned to AI tools as support for teaching, revision, and self-assessment of writing. Socially, because differences between free and paid versions may create asymmetries in access to more detailed or more useful feedback. Ethically, because the partial delegation of assessment practices to automated systems requires caution, especially in high-

stakes contexts such as national examinations. Thus, comparing assessments produced by the free version and the *Plus* version of ChatGPT makes it possible to observe whether there are significant differences in score assignment, in the quality of comments, and in approximation to human/official parameters.

Based on this problem, the study is guided by the following research questions: how does ChatGPT, in its free and *Plus* versions, assess a maximum-score Enem 2024 essay? What is the difference between the assessment made by the paid version and that made by the free version? How does this AI comment on a maximum-score Enem essay? And what differences can be observed between comments produced by AI and those prepared by human/official evaluators? On the basis of these questions, the general objective of this article is to describe how ChatGPT, in its free and *Plus* versions, assesses and comments on a maximum-score Enem 2024 essay. Specifically, it seeks to analyze the scores assigned by the tool in each competency, categorize the comments and feedback generated by the two versions, and compare their convergences with and divergences from the official assessment presented in the Participant's Guide.

The theoretical framework articulates three main axes. The first concerns the scoring matrix for the Enem essay, based on the guidelines of Inep (2025), especially with regard to the five assessment competencies. The second axis involves studies on formative feedback, with emphasis on Sadler (1989), who understands feedback as information about the distance between the performance presented and an expected standard, and Hattie and Timperley (2007), who argue that effective feedback should indicate to the student where they are, where they should go, and

what steps they can take. The third axis encompasses automated writing assessment and the use of AI in education.

Methodologically, the research is characterized as qualitative and descriptive-comparative, since it seeks to interpret and compare assessments and comments generated by two versions of ChatGPT concerning a maximum-score Enem 2024 essay. The corpus consists of a 1,000-point essay selected from the Enem 2025 Participant's Guide, the official comment associated with that essay, and the assessments produced by ChatGPT in two access modalities: the free version and the Plus version. The same prompt was used in both versions, requesting a technical, rigorous correction organized according to the five competencies of the Enem matrix. The analysis was conducted based on categories derived from the theoretical framework: adherence to the matrix, diagnostic accuracy, consistency between score and comment, feedback actionability, prioritization of assessment aspects, and comparison between versions.

In addition to this introduction, the article is organized into four sections. The first presents the theoretical framework, covering the Enem essay scoring matrix, the notion of evaluative feedback, and studies on automated writing assessment. The second describes the methodological procedures adopted in constituting the corpus, preparing the prompt, and analyzing the data. The third presents and discusses the results, comparing the assessments produced by the free and Plus versions of ChatGPT with the official comment in the Participant's Guide. Finally, the conclusions section revisits the research questions, synthesizes the main findings, presents the study's contributions, and indicates its limitations, as well as possibilities for future investigations.

2. THEORETICAL FRAMEWORK

2.1. The Enem Essay And The Scoring Matrix

The Enem essay is a large-scale assessment of school writing, organized by an analytical matrix. According to the Anísio Teixeira National Institute for Educational Studies and Research (2025), the participant must produce a dissertative-argumentative text, in the formal written register of Portuguese, defending a point of view by means of coherent and cohesive arguments and presenting an intervention proposal that respects human rights. The final score can reach 1,000 points and results from the sum of five competencies, each assessed from 0 to 200 points.

Competency I assesses command of the formal written register, covering writing conventions, grammatical aspects, syntactic construction, lexical choice, and register adequacy. Competency II observes understanding of the prompt, development of the theme, command of the dissertative-argumentative text, and productive use of sociocultural repertoire. Competency III focuses on the selection, organization, and interpretation of information, facts, and arguments in defense of a point of view, which involves text planning, thematic progression, and internal coherence. Competency IV addresses the linguistic mechanisms necessary for constructing argumentation, especially referential and sequential cohesion. Competency V assesses the intervention proposal, considering agent, action, means, purpose, detailing, and respect for human rights.

This matrix is the main parameter for examining ChatGPT's responses. An AI assessment can be considered adherent to the

Enem only when each score is justified by criteria corresponding to the competency analyzed. For example, a justification for Competency II must discuss theme, text type, and repertoire; by contrast, a justification for Competency IV must address the articulation among sentences, periods, and paragraphs. If the system assigns a high score but points out flaws incompatible with that score, or if it assigns a low score without explaining what prevents the text from reaching the maximum level, there is a problem of assessment consistency.

Another relevant aspect is that the official Enem score results from an institutional control procedure. According to Inep (2025), the essay is assessed by at least two evaluators with degrees in Language and Literature or Linguistics, independently, with a new assessment conducted when a relevant discrepancy occurs. ChatGPT's assessment, in contrast, results from a specific interaction mediated by a prompt. For this reason, the comparison between human assessment and AI assessment must consider that these are distinct assessment regimes: one institutionally controlled and the other discursively generated by the model in a usage situation.

2.2. Feedback And The Quality Of The Evaluative Comment

The concept of feedback is important for analyzing the comments produced by AI. From a formative perspective, feedback is not equivalent to generic praise or superficial correction; it informs the distance between the performance presented and an expected quality standard. Sadler (1989) understands formative assessment as a process that helps the student recognize quality criteria and bring their production closer to those criteria. Applied to the Enem, this means that the comment is useful only when it explains, based on

the competencies, what the text accomplishes adequately, what needs to be improved, and how such improvement can occur.

Hattie and Timperley (2007) argue that effective feedback answers three questions: where the student should go, how they are doing, and what the next step is. This formulation makes it possible to differentiate merely descriptive comments, justificatory comments, and guiding comments. Feedback that only states that the essay is "well structured" has low diagnostic usefulness; by contrast, a comment that relates thesis, repertoire, argumentative progression, and assigned score offers greater analytical value. When AI indicates specific rewriting strategies, it approaches actionable feedback.

Steiss et al. (2024) are especially relevant to this study because they compared human feedback and feedback produced by ChatGPT on students' texts. The authors analyzed dimensions such as adherence to criteria, clarity of guidance, accuracy, prioritization of essential aspects, and supportive tone. These dimensions can be adapted to the Enem essay: a good AI comment should be based on the competencies, be understandable, point to real evidence from the text, prioritize aspects that affect the score, and avoid both empty praise and unfounded criticism.

2.3. Automated Writing Assessment And Chatgpt

Automated writing assessment, known as Automated Writing Evaluation (AWE) or Automated Essay Evaluation (AEE), predates current generative systems. Shermis and Burstein (2013) show that this field brings together technologies aimed at score assignment, identification of linguistic features, and generation of feedback. However, the speed and initial standardization offered by such

systems do not eliminate the need for validation, especially when the writing being assessed involves argumentation, authorship, coherence, sociocultural repertoire, and genre adequacy.

Deane (2013) warns that writing cannot be reduced to easily measurable formal markers, because it involves social practices, rhetorical choices, and meaning construction. This observation is decisive for the Enem, since the matrix assesses dimensions that go beyond grammar and normative correctness. ChatGPT can produce fluent and technically plausible comments, but the verbal fluency of the response does not guarantee that the judgment is valid. Thus, it is necessary to investigate whether the system actually assesses the construct provided for by the matrix or whether it shifts the analysis to generic notions of “good writing”.

Chapelle, Cotos, and Lee (2015) contribute by discussing validity in diagnostic assessments based on AWE. For the authors, validity depends on the interpretation and use of the information produced by the system. In the case of this study, this implies asking whether ChatGPT's scores and comments help to understand the essay's performance according to the Enem competencies. Williamson, Xi, and Breyer (2012) also emphasize that automated scoring systems should be evaluated by considering context of use, implementation, and consequences. Therefore, an AI correction should not be judged only by its speed or apparent objectivity.

Ranalli, Link, and Chukharev-Hudilainen (2017) argue that the usefulness of automated feedback must be examined in terms of accuracy and formative uptake. Wilson and Roscoe (2020) reinforce that the efficacy of these systems should be observed through multiple metrics, such as performance, users' perceptions, and

pedagogical integration. From this perspective, ChatGPT should be analyzed as an assessment support tool, not as a full substitute for the human evaluator. The decisive point is to verify in which dimensions AI can assist teachers and students and in which aspects it requires human mediation.

The specificity of ChatGPT lies in the fact that it generates natural-language responses based on commands. The prompt, therefore, becomes part of the methodological design. Instructions such as “follow the Guide,” “assign scores by competency,” and “do not invent problems” function as attempts to control the judgment produced by AI. Even so, different versions of the system may vary in the length, level of detail, stability, and accuracy of responses. The comparison between the free version and the Plus version should observe whether there is a real difference in adherence to the criteria, diagnostic quality, and coherence among score, comment, and suggestion.

2.4. Analytical Categories For The Research

Based on the authors discussed, the analysis of ChatGPT's corrections can be organized by categories. These categories make it possible to examine the assigned scores, classify the generated comments, and compare official human assessments with AI assessments produced in different versions of the tool.

Table 1. Analytical categories.

Category	Analytical question	Theoretical basis	Use in the analysis

Adherence to the matrix	Does the comment correspond to the competency assessed?	Inep (2025); Chapelle, Cotos, and Lee (2015)	Verify whether the justification mobilizes the correct descriptors.
Diagnostic accuracy	Does the AI point out real problems in the text?	Steiss et al. (2024); Ranalli, Link, and Chukharev-Hudilainen (2017)	Compare comment, essay, and official assessment.
Score-comment consistency	Does the score match the justification?	Sadler (1989); Hattie and Timperley (2007)	Identify contradictions between score, criticism, and praise.
Actionability	Does the feedback guide concrete improvement?	Hattie and Timperley (2007); Wilson and Roscoe (2020)	Distinguish generic suggestions from applicable guidance.
Prioritization	Does the AI highlight what most affects the score?	Steiss et al. (2024); Deane (2013)	Verify whether the comment emphasizes central or peripheral aspects.
Comparison between versions	Does the paid version differ from the free version?	Shermis and Burstein (2013); Zawacki-Richter et al. (2019)	Compare length, density, evidence, and stability.

Source: authors' own elaboration, based on the theoretical framework.

These categories make it possible to answer the research questions without reducing the analysis to score differences. Adherence to the

matrix indicates whether the AI effectively follows the official criteria; diagnostic accuracy assesses whether the comment is faithful to the text; score-comment consistency reveals internal coherence; actionability shows the pedagogical value of the feedback; prioritization verifies whether the AI distinguishes central problems from peripheral observations; and comparison between versions makes it possible to discuss whether paid access produces a relevant assessment advantage.

3. METHODOLOGY

This research is characterized as a qualitative, documentary, descriptive-comparative study configured as a single-case study. The qualitative approach is justified because, according to Paiva (2019), research in linguistic studies makes it possible to interpret language phenomena by considering their contexts of production, circulation, and signification. The documentary nature derives from the fact that the corpus consists of previously produced documents—the maximum-score essay selected from the Participant's Guide, the official comment by Inep, and the assessments generated by ChatGPT—which approximates Cellard's (2008) conception, according to which documentary analysis requires selection, contextualization, and critical examination of documents. The descriptive-comparative character, in turn, is related to the description and contrast between the scores and comments produced by the tool's free version and Plus version, taking as a parameter the human/official assessment presented by Inep. Finally, the investigation is configured as a single-case study, since it focuses on a delimited analytical situation: the assessment, by two versions of ChatGPT, of a maximum-score Enem 2024 essay selected from the Participant's Guide. This delimitation is compatible with Yin

(2015), for whom the case study allows a contemporary phenomenon to be investigated in depth within its specific context.

The research corpus was constituted from the Enem 2025 Participant's Guide, published by the Anísio Teixeira National Institute for Educational Studies and Research, a document that presents guidance on the Enem essay, explains the five assessment competencies, and brings together twelve 1,000-point essays from Enem 2024. From this material, a maximum-score essay was selected, as well as the official comments associated with it and information related to the assessment criteria for each competency. The choice of this material is justified by the fact that the Guide represents an institutional and normative reference source for scoring the Enem essay, allowing the assessments produced by AI to be confronted with human/official parameters.

The unit of analysis of this study is therefore composed of three main sets of data: the selected 1,000-point essay; the official comment present in the Participant's Guide; and the assessments produced by ChatGPT in two access modalities, the free version and the Plus version. The comparison among these three sets makes it possible to observe the extent to which AI approaches or distances itself from the official criteria, both in score assignment and in the formulation of evaluative comments.

The first methodological procedure consisted of collecting the material available in the Participant's Guide. Initially, the 1,000-point essays present in the document were identified and, subsequently, one of them was selected to compose the initial sample of the analysis. In addition to the essay itself, the official comments prepared by Inep and the descriptions of the five competencies of

the scoring matrix were extracted. This procedure was necessary so that the evaluation of the responses generated by ChatGPT would not be based solely on the researchers' subjective impressions, but on criteria previously defined by the institution responsible for the examination itself.

Next, the material was prepared for insertion into the AI tools. The original file was divided so that only the information necessary for correction would be used: the selected essay, the criteria of the assessment matrix, and the relevant guidelines extracted from the Participant's Guide. This stage aimed to avoid the inclusion of excessive or irrelevant information that could interfere with the models' response. At the same time, the aim was to ensure that ChatGPT had access to sufficient official criteria to carry out an assessment compatible with the Enem matrix.

Subsequently, a specific prompt was prepared to guide the assessment of the essay. This prompt was constructed based on the criteria of the Participant's Guide and requested that ChatGPT assume the role of a specialized evaluator of the Enem 2024 essay. The command determined that the assessment should be rigorous, technical, and transparent, with a separate analysis of the five assessment competencies. It was also requested that scores be assigned at the levels provided for in the matrix, that is, 0, 40, 80, 120, 160, or 200 points in each competency, with a final sum on a scale from 0 to 1,000 points.

Before the correction itself, the prompt instructed the verification of possible zero-score conditions, such as complete departure from the theme, failure to meet the dissertative-argumentative text type, insufficient text, or disrespect for human rights. In addition, it

specified the aspects that should be observed in each competency: command of the formal written register of Portuguese, understanding of the prompt, use of sociocultural repertoire, argumentative organization, textual progression, cohesion mechanisms, and intervention proposal. Finally, it established a fixed structure for the response, composed of preliminary analysis, comments by competency, score assignment, final justification, and suggestions for improvement. For reasons of length, it was decided not to transcribe the full prompt in the body of the article, since it is a long command; however, its description is presented to ensure methodological transparency.

After the prompt was prepared, the essay was submitted to ChatGPT in two distinct situations: first in the free version and then in the Plus version. The two submissions took place on April 14, 2026. In the free version, the model then available to the user was used, identified in the research material as GPT-5.3 Instant. In the Plus version, the GPT-5.4 Thinking model was used. In both situations, the same prompt and the same essay were used in order to reduce interference in the procedure and allow a more controlled comparison between the responses.

After the assessments generated by the two versions of ChatGPT were collected, the data were organized into comparative tables. Initially, the scores assigned by competency and the total score indicated by each version were systematized. Next, the comments were analyzed qualitatively, considering their relation to the official Enem matrix and to the human comment presented in the Participant's Guide. This stage made it possible to observe not only the numerical difference between the assessments, but also the ways in which each version of AI justified its evaluative decisions.

The analysis was conducted based on categories derived from the study's theoretical framework. The first category was adherence to the matrix, Inep (2025) and Chapelle, Cotos, and Lee (2015), through which it was observed whether the AI comments effectively corresponded to the competency assessed. The second was diagnostic accuracy, Steiss et al. (2024) and Ranalli, Link, and Chukharev-Hudilainen (2017), aimed at verifying whether the problems or qualities pointed out by ChatGPT were present in the text analyzed. The third category was consistency between score and comment, Sadler (1989) and Hattie and Timperley (2007), used to examine whether the score assigned was coherent with the justification presented. The fourth was feedback actionability, Hattie and Timperley (2007) and Wilson and Roscoe (2020), that is, the capacity of the comment to offer concrete guidance for textual improvement. The fifth was prioritization, Steiss et al. (2024) and Deane (2013), related to the identification of the aspects most relevant to the score. Finally, the sixth category corresponded to comparison between versions, Shermis and Burstein (2013) and Zawacki-Richter et al. (2019), with the objective of verifying differences between the assessment produced by the free version and that produced by the Plus version.

These categories allowed the data to be analyzed at two complementary levels. At the first level, the scores assigned by the two versions of ChatGPT were compared, observing convergences and divergences in each competency. At the second level, the evaluative comments were examined, considering their analytical density, their adequacy to the official criteria, and their proximity to or distance from the human/official assessment. Thus, this research was not limited to verifying which version assigned a higher or lower score, but sought to understand the evaluative profile of each one.

Finally, it should be noted that this study has an important methodological delimitation: the analysis focuses on a 1,000-point essay selected from the Participant's Guide. Thus, the results do not intend to generalize the functioning of all assessments produced by ChatGPT in contexts of text correction. The objective was to carry out an initial, situated, and qualitative analysis capable of evidencing possibilities, limits, and differences between versions of the tool in the specific context of assessing Enem essays. This delimitation, however, does not reduce the relevance of the study, because it allows observation, in a controlled manner, of how AI interprets an essay officially recognized as exemplary and how its comments approach or diverge from human assessment parameters.

4. RESULTS AND DISCUSSION

In this section, the assessments produced by ChatGPT in its free version and in its Plus version are analyzed for the first essay in the sample, taken from the Enem 2025 Participant's Guide, referring to the theme “Challenges for valuing African heritage in Brazil.” The criterion for choosing this essay is that it is the first one that appears in the Guide.

According to the official comment in the Guide, the essay analyzed presents a well-conducted text plan, with organized argumentation, legitimate, pertinent, and productive sociocultural repertoire, as well as an intervention proposal articulated with the problem discussed. The Guide highlights that the text mobilizes references such as Carolina Maria de Jesus, Elijah Anderson, Atlas of Violence, and Florestan Fernandes, articulating them with the axes of structural racism and social neglect.

Table 2. Comparison of the scores assigned by the two versions of ChatGPT

Competency	Free ChatGPT	ChatGPT Plus	Difference
Competency I	160	160	0
Competency II	200	160	-40
Competency III	200	160	-40
Competency IV	200	160	-40
Competency V	200	200	0
Total	**960**	**840**	**-120**

Source: prepared by the authors

Initially, a significant difference is observed between the two assessments. The free version assigns, based on the sum of the explicit competencies, 960 points to the essay, whereas the Plus version assigns 840 points. This difference of 120 points is relevant because, in the Enem assessment logic, the essay score is composed of five competencies, each assessed from 0 to 200 points, according to the reference matrix presented by Inep (2025). The Guide itself clarifies that the essay is judged by five competencies and that each evaluator assigns up to 200 points per competency.

The free version presents an assessment closer to the general valuation made by the Guide. In the preliminary analysis, it states that the text “fully meets the proposed theme” and that it is “well structured,” with “pertinent sociocultural repertoire” and a “detailed social intervention aligned with human rights.” This initial reading already anticipates the evaluative profile of the free version: it is a

predominantly confirmatory correction, which recognizes the essay as a high-performance production and penalizes only Competency I.

The Plus version, in turn, also recognizes the overall quality of the essay, but adopts a more severe stance. In the preliminary analysis, it states that the text presents an introduction with a thesis, two development paragraphs, and a conclusion with an intervention proposal; in addition, it recognizes that the essay remains within the thematic scope by discussing structural racism and social neglect as obstacles to valuing African heritage. However, despite this recognition, it assigns 160 points to Competencies II, III, and IV, justifying that the essay would present room for refinement regarding the analytical yield of the repertoire, the density of argumentative links, and the use of cohesive mechanisms.

In light of Chapelle, Cotos, and Lee (2015), this finding is important because the validity of an automated assessment does not depend only on the production of a score, but on the relation between that score, the construct assessed, and the use intended for the result. In this case, the divergence between 960 and 840 shows that the two versions of ChatGPT do not operate in a fully stable manner when faced with the same text and the same matrix. Thus, although both mobilize terminology compatible with the Guide, their interpretation of the criteria varies considerably.

4.1. Adherence To The Matrix And Diagnostic Accuracy

With regard to adherence to the matrix, the two versions organize the assessment by competencies, mention the dissertative-argumentative type, discuss adherence to the theme, and assign separate scores for each dimension assessed.

In Competency II, the free version states that the essay “demonstrates excellent understanding of the thematic prompt” and that it “remains entirely within the thematic scope.” In addition, it mentions as productive repertoires the work *Quarto de Despejo*, by Carolina Maria de Jesus, the sociological concept of ghettoization, by Elijah Anderson, and Florestan Fernandes's critique of the idea of racial democracy. In the end, it concludes that “the repertoire is well articulated with the arguments and is not merely cited,” thus justifying the maximum score.

This excerpt reveals that the free version correctly recognizes the central elements provided for by Competency II: adherence to the theme, dissertative-argumentative structure, and sociocultural repertoire. However, the analysis remains at a more declarative than demonstrative level. The AI states that the repertoire is productive, but does not explain, with greater density, how each reference operates in sustaining the thesis. At this point, the assessment approaches what Steiss et al. (2024) classify as feedback that is clear and adherent to the criteria, but not necessarily sufficiently precise in terms of prioritization and textual evidence.

The Plus version, on the other hand, presents a more detailed diagnosis. It recognizes that “the sociocultural repertoire is pertinent and, in general, productive,” mentioning Carolina Maria de Jesus, Elijah Anderson, Atlas of Violence, and Florestan Fernandes. However, unlike the free version, it argues that the score does not reach 200 because the repertoire “is not always explored with maximum analytical yield in relation to the exact core of the theme.” According to this assessment, at some points, the text would focus more on racism against the Black population in broad terms than on the specific mechanisms of devaluation of African heritage.

This difference is decisive. The Plus version demonstrates greater diagnostic concern and greater capacity to make explicit the distance between the observed performance and maximum performance. However, its reading partially distances itself from the Guide, because the official comment considers that the essay presents a well-conducted argumentative plan and legitimate, pertinent, and productive repertoire, especially through references to literature and sociology. Thus, the Plus version is more analytical, but not necessarily more aligned with the official assessment.

4.2. Consistency Between Score And Comment

As for consistency between score and comment, the free version presents more evident internal coherence in Competency I. By assigning 160 points, the AI points out problems of parallelism, use of preposition, syntactic construction with the gerund, and punctuation. One of the highlighted excerpts is “— and preventing the recognition of the cultural aspects of these people,” on which the free version comments that “the gerund does not articulate adequately with the main clause,” suggesting replacement with “— which prevents recognition...”.

In this case, there is correspondence between the score assigned and the justification presented. The free version recognizes good command of the standard norm, but identifies punctual slips that would prevent the maximum score. This assessment approximates the Guide's own reading, which also recognizes excellent command of the standard norm, but points out lexical imprecision in the use of “exhibited” and truncation in the formulation “and the lower salary of these individuals”.

However, in Competencies II, III, and IV, the free version assigns 200 points and presents mostly laudatory comments. In Competency III, for example, it states that the text evidences “a well-defined text plan,” with a clear thesis, coherent thematic progression, adequate selection of information, internal coherence, and indications of authorship. It then concludes that “the argumentation is in-depth and well supported, justifying the maximum score”.

Internal consistency exists, since the maximum score stems from an equally positive justification. However, the comment does little to differentiate what would be a “very good” text from a genuinely “excellent” text. In Sadler's (1989) terms, there is less explicitness regarding the distance between observed performance and the standard of excellence, because the AI does not test the limits of the text or make fine assessment criteria explicit.

The Plus version presents greater internal consistency precisely because it associates the 160-point scores with specific justifications. In Competency III, it recognizes that there is a text plan and that the essay follows a clear argumentative line: introduction with thesis, development on structural racism, development on social neglect, and conclusion with intervention. However, it states that the score does not rise to 200 because “argumentative consistency, although good, still presents some leaps in explanation.” As an example, it mentions that the transition between social segregation, exclusion from “erudite practices,” and inattention to African legacies could be “more densely made explicit”.

In this respect, the Plus assessment is closer to the formative feedback model proposed by Hattie and Timperley (2007), because it answers, albeit implicitly, three questions: where the text is, why it

would not reach the maximum level, and what could be improved. The problem, however, is that this internal consistency does not guarantee external consistency with the official assessment, since the Guide considers the argumentative plan well conducted and coherent.

4.3. Quality And Actionability Of Feedback

The main qualitative difference between the two versions appears in the actionability of feedback. The free version provides understandable and organized comments, but its most concrete guidance is concentrated in Competency I. It suggests, for example, replacing “the lower salary of these individuals” with a construction with greater parallelism, such as “lower salaries”; it also proposes changing the gerund construction to “which prevents recognition...”.

These suggestions are useful, but restricted. In the other competencies, the comment assumes a predominantly confirmatory tone. In Competency IV, for example, the free version states that the essay presents “excellent command of cohesion and coherence mechanisms,” with the use of connectives such as *“Assim”, “Nesse sentido”, “Além disso”, “Dessa forma”, “Portanto”, “Ademais*”, and “*Em última instância*”.” It further adds that the connectives are used “naturally” and contribute to the fluency of the text.

The Plus version, by contrast, offers more operational guidance. At the end of the assessment, it recommends: “Improve thematic precision,” “Densify the analysis of the repertoire,” “Refine syntactic construction,” and “Use connectives more discreetly.” It also suggests that the essay relate each argument more directly to the

devaluation of African heritage, mentioning historical memory, culture, religiosity, aesthetics, intellectual production, and social practices of African origin.

This type of comment presents greater actionability, because it not only identifies problems but also guides pathways for rewriting. In this direction, it approaches Hattie and Timperley's (2007) conception, according to which effective feedback should indicate next steps, and also the perspective of Ranalli, Link, and Chukharev-Hudilainen (2017), for whom the usefulness of automated feedback should be examined in terms of accuracy and formative uptake.

Even so, it is necessary to relativize this advantage. More detailed feedback may be pedagogically useful, but, if it is misaligned with the official matrix or if it is more severe than human assessment in the case of a maximum-score essay, it may lead the student to a distorted perception of their writing. At this point, Wilson and Roscoe (2020) are relevant in arguing that the efficacy of automated systems should be observed through multiple metrics, and not only through the length, fluency, or apparent sophistication of the comment.

4.4. Comparison Between The Guide's Arguments And The Arguments Of The Two AI Versions

Table 3 synthesizes the comparison between the arguments mobilized by the Guide, by the free version, and by the Plus version. The objective is not only to compare scores, but to observe how each source interprets the same textual performance.

Table 3. Comparison between the arguments of the Guide, ChatGPT free, and ChatGPT Plus

Analyzed dimension	Participant's Guide/Inep	Free ChatGPT	ChatGPT Plus	Analytical interpretation
Adherence to the theme and thematic scope	Considers that the theme was addressed adequately, with articulation between structural racism, social neglect, and the valuing of African heritage.	States that the text "fully meets the proposed theme," addressing the challenges for valuing African heritage with a focus on structural racism and social neglect.	Recognizes that the text addresses the theme, but states that the focus could be more precise in relation to the core "African heritage".	There is convergence in recognizing adherence to the theme, but divergence in the degree of exigency. The Plus version restricts the concept of thematic focus more strongly.
Sociocultural repertoire	Assesses the repertoire as legitimate, pertinent, and productive, highlighting Carolina Maria de Jesus, Elijah Anderson, and Florestan Fernandes.	Also considers the repertoire productive and legitimized, stating that it is not merely cited.	Recognizes that the repertoire is pertinent, but states that it is not always explored with maximum analytical yield.	The free version is closer to the Guide; the Plus version presents a more exacting reading, but one more distant from the official comment.
Text plan and argumentative	Considers that there is a well-conducted argumentati	States that there is a well-defined text plan, clear thesis,	Recognizes the textual plan, but points to "leaps in	The Plus version offers a more refined diagnosis,

progression	ve plan, with thematic progression and coherent articulation among information, facts, and opinions.	coherent thematic progression, and indications of authorship.	explanation" and argumentative links that could be developed further.	but downgrades a dimension that the Guide evaluates positively.
Cohesion and linguistic mechanisms	States that the text presents thematic continuity and a varied repertoire of cohesive resources employed without inadequacies.	Assigns 200 points, highlighting varied and adequate use of connectives and efficient referential cohesion.	Assigns 160 points, alleging that the connectives, although correct, sound formulaic and little sophisticated.	The divergence is expressive: the Guide and the free version value cohesion; the Plus version penalizes what it interprets as discursive predictability.
Command of the standard norm	Recognizes excellent command of the standard norm, but points out lexical imprecision and truncated formulation.	Assigns 160 points, pointing out parallelism, preposition, gerund, and punctuation.	Also assigns 160 points, pointing out syntactic construction, dash, and lexical redundancy.	The two versions converge on the score of 160, but the Guide seems less severe, as it frames the problems as punctual in a high-performance text.
Intervention proposal	Considers the proposals	Assigns 200 points, stating that	Also assigns 200 points, identifying	There is strong convergence

	detailed, viable, articulated with the discussion, and respectful of human rights.	the proposal is complete, detailed, and articulated with the problem.	agent, action, means, purpose, and detailing.	among the three assessments. Competency V is the dimension with the greatest assessment stability.

Source: prepared by the authors

Table 3 shows that the three assessments converge regarding the overall quality of the essay, especially in recognizing adherence to the theme, the presence of sociocultural repertoire, and the completeness of the intervention proposal. However, they diverge regarding the degree of exigency applied to the intermediate competencies. The Guide interprets the text as exemplary, even though it recognizes small lexical and syntactic imprecisions; the free version follows this overall positive assessment, assigning the maximum score in four of the five competencies; the Plus version, in turn, carries out a more restrictive reading, distinguishing "good execution" from "excellent execution".

This difference is clear in Competency IV. The Guide states that the text presents "thematic continuity" and a "varied repertoire of cohesive resources, used without inadequacies," citing expressions such as *"Nesse sentido", "Nesse contexto", "Dessa forma", "Desse modo", "Portanto", "Logo", "Ademais*", and "*Em última instância*". The free version follows a similar interpretation, considering these same mechanisms as evidence of "excellent command" of cohesion. The Plus version, on the other hand, reinterprets the recurrence of these connectives as a sign of formulaicity, stating that they "sound visibly

formulaic" and approach a "marker of school organization" more than sophisticated discursive articulation.

This divergence is analytically relevant. It shows that the paid AI not only identifies elements of the text, but assigns them a different evaluative value from that attributed by the Guide. The same textual phenomenon, the recurrent use of argumentative operators, is read by the Guide and the free version as an adequate cohesive resource; by the Plus version, as a trait of predictability and lower refinement. Thus, the difference between the versions lies not only in the length of the comment, but in the interpretation of what counts as excellence.

4.5. Comparison With The Human/official Assessment

The comparison with the official comment in the Guide reveals that the free version is closer to the human assessment in terms of the overall valuation of the essay, although it presents lower analytical density. The free version recognizes the essay as a high-performance text and, similarly to the Guide, values the repertoire, argumentative progression, cohesion, and intervention proposal. Its main limitation is low problematization: the comments tend to confirm the quality of the text more than to explain, in detail, how this quality is constructed.

The Plus version, in turn, offers a more critical and pedagogically actionable reading, but distances itself from the official interpretation at relevant points. This occurs especially when it penalizes Competency II for supposed insufficient thematic focus, Competency III for argumentative leaps, and Competency IV for formulaic use of connectives. The Guide, by contrast, interprets these

aspects as sufficiently productive for the composition of an exemplary essay.

This result confirms the need to treat ChatGPT as an assessment support tool, and not as a full substitute for the human evaluator. According to Deane (2013), writing cannot be reduced to easily measurable formal markers, because it involves social practices, rhetorical choices, and meaning construction. In the case analyzed, both the free version and the Plus version recognize formal and argumentative features of the essay, but assign different weights to these features.

In addition, Shermis and Burstein (2013) recall that automated writing assessment involves score assignment, identification of linguistic features, and generation of feedback. However, the existence of these resources does not eliminate the need for validation. The present analysis shows precisely this: the free version seems numerically closer to the official assessment, but is less analytical; the Plus version is more analytical and actionable, but assigns a considerably lower score to an essay recognized by the Guide as exemplary.

4.6. Partial Discussion Of The Findings

The results of this first analysis suggest that access to the paid version of ChatGPT does not necessarily represent greater approximation to the official Enem assessment. The advantage of the Plus version is more associated with the level of detail, the argumentative organization of the comment, and the provision of rewriting suggestions. However, this qualitative advantage is not

converted, in this case, into greater adherence to the human/official reading of the essay.

On the other hand, the free version presents a more favorable assessment and is closer to the general valuation made by the Guide, but its comment tends to be less rigorous in distinguishing between high performance and maximum performance. Thus, from the standpoint of the research, the difference between the versions should not be described only as a difference between a “better” assessment and a “worse” one. What is observed is a difference in evaluative profile: the free version operates in a more confirmatory and laudatory way; the Plus version operates in a more critical, diagnostic, and prescriptive way.

This finding directly dialogues with the ethical and pedagogical concern of this study, especially with regard to the possibility of a paid version offering a significant advantage to students. In this sample, the Plus version did not assign a higher score nor did it prove closer to the official assessment; its advantage lay in the density of the feedback. Therefore, the possible benefit of the paid version is not in the score itself, but in the greater capacity to transform the assessment into rewriting guidance.

Even so, this feedback requires teacher mediation, because it may present severity criteria that do not coincide with the official parameters of the Enem. In summary, both versions of ChatGPT are capable of producing assessments organized according to the Enem essay competencies, but they present distinct limitations: the free version tends to be more generous and less analytical; the Plus version tends to be more rigorous and actionable, but more severe than the official assessment. This result reinforces the need for

validation, systematic comparison, and human mediation, as argued by Chapelle, Cotos, and Lee (2015), Williamson, Xi, and Breyer (2012), Ranalli, Link, and Chukharev-Hudilainen (2017), and Wilson and Roscoe (2020).

5. CONCLUSIONS

This article aimed to describe how ChatGPT, in its free and Plus versions, assesses and comments on a maximum-score Enem 2024 essay, observing the convergences and divergences between the assessments produced by artificial intelligence and the human/official parameters presented in the Participant's Guide. Based on the analysis carried out, it was possible to verify that both versions of the tool can organize the correction according to the Enem competency matrix, mobilizing terms compatible with the official evaluative discourse, such as sociocultural repertoire, text plan, cohesion, standard norm, and intervention proposal. However, the results also show that the formal organization of the response does not, by itself, guarantee assessment equivalence in relation to human/official correction.

Regarding the first research question—how does ChatGPT, in the free and Plus versions, assess a maximum-score Enem 2024 essay? —the data indicate that AI recognizes the overall quality of the essay analyzed, but assigns different weights to the assessed aspects. The free version produced an assessment closer to the official valuation of the Guide, assigning 960 points to the text and recognizing the maximum score in four of the five competencies. This assessment assumed a predominantly confirmatory profile, since it highlighted the merits of the essay and identified only punctual problems in Competency I. The Plus version, in turn, assigned 840 points,

adopting a more severe stance toward the same text and the same prompt. Although it also recognized the quality of the essay, it penalized Competencies II, III, and IV, pointing to a supposed need for greater thematic precision, greater argumentative density, and less formulaic use of cohesive resources.

As for the second question—what is the difference in the assessment made by the paid AI and by the free one?—the analysis allows concluding that the main difference is not only in the final score, but in the evaluative profile of each version. The free version proved more numerically aligned with the official assessment, but presented comments that were less dense and less problematizing. The Plus version, by contrast, elaborated a more analytical, detailed, and prescriptive assessment, offering more concrete rewriting guidance. However, this greater density did not mean greater adherence to the Guide, because the Plus version negatively reinterpreted aspects that the official comment evaluated as productive or adequate, especially with regard to the use of sociocultural repertoire, argumentative progression, and cohesion mechanisms.

With regard to the third question—how does this AI comment on a maximum-score Enem essay?—it was found that ChatGPT tends to produce comments that are organized, technically plausible, and structured according to the five competencies of the matrix. However, the comments vary in their degree of diagnostic accuracy and actionability. The free version comments on the essay in a more laudatory way, emphasizing adherence to the theme, the presence of pertinent repertoire, the argumentative structure, and the intervention proposal. Its suggestions for improvement appear more concretely above all in Competency I. The Plus version, on the other

hand, comments on the essay in a more critical and formative way, making possible paths for improvement explicit. Even so, this more rigorous stance may produce a reading more distant from the official assessment, especially when it transforms typical traits of high-performance school writing into indications of textual limitation.

Regarding the fourth question—what is the difference between comments made by AI and comments made by humans in relation to a maximum-score essay?—the results show that the human/official comment in the Guide operates as an institutionally situated assessment, aimed at explaining the reasons why the text achieves exemplary performance. AI, in turn, although it can reproduce part of the assessment terminology, does not necessarily assign the same value to the textual phenomena identified. The free version is closer to the official reading, but tends to offer lower analytical density; the Plus version offers more developed comments, but adopts severity criteria that do not fully coincide with human/official parameters. Thus, the difference lies not only between human assessment and AI assessment, but also among the very modes of operation of the AI versions.

In this way, it is concluded that ChatGPT can function as an auxiliary tool in the process of assessing and revising essays, especially when used to organize comments by competency, identify formal aspects, and suggest rewriting possibilities. However, its results must be interpreted with caution. The paid version did not prove, in this case, necessarily closer to the official assessment, although it produced more detailed and actionable feedback. The free version, in turn, presented greater proximity to the expected score for a 1,000-point

essay, but less capacity to explicate finely the criteria that justify this performance.

Finally, it is emphasized that the findings of this study must be understood within the methodological limits of the research. The analysis focused on a maximum-score essay, submitted to two versions of ChatGPT in a controlled usage situation. Therefore, the results cannot be generalized to all essays, all themes, or all possible interactions with AI systems. Even so, the study evidences the need for teacher mediation and human validation in the pedagogical use of these tools, especially in high-stakes assessment contexts such as the Enem. Future research may expand the corpus, compare different AI models, test prompt variations, and investigate whether the observed patterns remain in essays with intermediate or low scores.

REFERENCES

CELLARD, André. A análise documental. In: POUPART, Jean et al. **A pesquisa qualitativa: enfoques epistemológicos e metodológicos.** Petrópolis: Vozes, 2008. p. 295-316.

CHAPELLE, Carol A.; COTOS, Elena; LEE, Jooyoung. Validity arguments for diagnostic assessment using automated writing evaluation. **Language Testing,** [s. l.], v. 32, n. 3, p. 385-405, 2015. DOI: 10.1177/0265532214565386.

DEANE, Paul. On the relation between automated essay scoring and modern views of the writing construct. **Assessing Writing,** [s. l.], v. 18, n. 1, p. 7-24, 2013. DOI: 10.1016/j.asw.2012.10.002.

HATTIE, John; TIMPERLEY, Helen. The power of feedback. **Review of Educational Research,** [s. l.], v. 77, n. 1, p. 81-112, 2007. DOI: 10.3102/003465430298487.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **A redação do Enem 2025:** cartilha do(a) participante. Brasília, DF: Inep/MEC, 2025.

PAIVA, Vera Lúcia Menezes de Oliveira e. **Manual de pesquisa em estudos linguísticos.** 1 ed. São Paulo: Parábola Editorial, 2019.

RANALLI, Jim; LINK, Stephanie; CHUKHAREV-HUDILAINEN, Evgeny. Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. **Educational Psychology,** [s. l.], v. 37, n. 1, p. 8-25, 2017. DOI: 10.1080/01443410.2015.1136407.

SADLER, D. Royce. Formative assessment and the design of instructional systems. **Instructional Science,** [s. l.], v. 18, n. 2, p. 119-144, 1989. DOI: 10.1007/BF00117714.

SHERMIS, Mark D.; BURSTEIN, Jill (org.). **Handbook of automated essay evaluation:** current applications and new directions. New York: Routledge, 2013.

STEISS, Jacob; TATE, Tamara; GRAHAM, Steve; CRUZ, Jazmin; HEBERT, Michael; WANG, Jiali; MOON, Youngsun; TSENG, Waverly; WARSCHAUER, Mark; OLSON, Carol Booth. Comparing the quality of human and ChatGPT feedback of students' writing. **Learning and Instruction,** [s. l.], v. 91, art. 101894, 2024. DOI: 10.1016/j.learninstruc.2024.101894.

[1] Mestre em Letras pela Universidade Estadual do Piauí e doutorando em Letras pela Universidade Federal do Piauí. E-mail: acesse o artigo original para visualizar o e-mail

[2] Discente do Curso Superior de Letras – Língua Portuguesa e Literatura de Língua Portuguesa do Universidade Federal do Piauí. E-mail: acesse o artigo original para visualizar o e-mail