ReGeo ISSN: 2177-3246

EVALUATING THE PERFORMANCE OF CHATBOTS IN LITERARY CRITICISM ON LIMA BARRETO

AVALIANDO O DESEMPENHO DE CHATBOTS NA CRÍTICA LITERÁRIA SOBRE LIMA BARRETO

EVALUACIÓN DEL DESEMPEÑO DE LOS CHATBOTS EN LA CRÍTICA LITERARIA SOBRE LIMA BARRETO

https://doi.org/10.56238/revgeov17n11-037

Data de submissão: 06/10/2026 **Data de publicação:** 06/11/2026

Raoni Schimitt Huapaya[1], Douglas Almonfrey[2], Gabriel Magalhães Siston[3], Philipe Brito de Moraes[4]

ABSTRACT

In the early 2020s, with the popularization of generative artificial intelligences (GenAIs), discussions intensified regarding the utility of chatbot technology — generative text intelligences — in teaching-learning dynamics. Aiming to verify the capacity of these systems to produce utterances suitable for literary criticism, and to understand their practical limitations. This research compared qualitative and quantitative evaluations of interactions conducted by a human model and by artificial models (ChatGPT, Gemini, and GPT4ALL). The analysis of their performances demonstrated that ChatGPT presented the best results, followed by Gemini and GPT4ALL. Among the discrepant aspects observed in the responses, the following stood out: originality, response length, continuity between responses, vocabulary variability, objectivity of enunciation and the complexity of adapting terms to context. It was also observed that the models' tendency to achieve persuasion through discursive simplification and rhetorical standardization represents a problem for pedagogical dynamics, which should stimulate academic erudition and deep critical thinking.

Keywords: ChatGPT. Generative Artificial Intelligences. Chatbots. Literary Criticism. Teaching and Learning. Language Models.

RESUMO

No início da década de 2020, com a popularização das inteligências artificiais generativas (IAs generativas), intensificaram-se as discussões sobre a utilidade da tecnologia de *chatbots* — inteligências generativas de texto — nas dinâmicas de ensino-aprendizagem. Com o objetivo de verificar a capacidade desses sistemas de produzir enunciados adequados à crítica literária, bem como de compreender suas limitações práticas, esta pesquisa comparou avaliações qualitativas e quantitativas de interações realizadas por um modelo humano e por modelos artificiais (ChatGPT, Gemini e GPT4ALL). A análise de seus desempenhos demonstrou que o ChatGPT apresentou os melhores resultados, seguido pelo Gemini e pelo GPT4ALL. Entre os aspectos divergentes observados nas respostas,

[1] Doutor em literatura pela Universidade Federal do Espírito Santo (UFES). Instituto Federal do Espírito Santo (IFES). E-mail: vitória raoni.huapaya@ifes.edu.br
[2] Doutor em Engenharia Elétrica pela Universidade Federal do Espírito Santo (UFES). Instituto Federal do Espírito Santo (IFES). E-mail: dalmonfrey@ifes.edu.br
[3] Mestre em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Universidade Federal Fluminense (UFF). Niterói. E-mail: gabrielsiston1@gmail.com
[4] Graduando em Engenharia Elétrica pelo Instituto Federal do Espírito Santo (IFES). Vitória. E-mail: philipebm25@gmail.com

destacaram-se: originalidade, extensão da resposta, continuidade entre as respostas, variabilidade vocabular, objetividade do enunciado e a complexidade de adaptar termos ao contexto. Observou-se, ainda, que a tendência dos modelos de alcançar a persuasão por meio da simplificação discursiva e da padronização retórica representa um problema para as dinâmicas pedagógicas, as quais deveriam estimular a erudição acadêmica e o pensamento crítico aprofundado.

Palavras-chave: ChatGPT. Inteligências Artificiais Generativas. Chatbots. Crítica Literária. Ensino e Aprendizagem. Modelos de Linguagem.

RESUMEN

A principios de la década de 2020, con la popularización de las inteligencias artificiales generativas (IA generativas), se intensificaron los debates sobre la utilidad de la tecnología de los chatbots —sistemas de inteligencia artificial para la generación de texto— en las dinámicas de enseñanza-aprendizaje. Con el objetivo de verificar la capacidad de estos sistemas para producir enunciados adecuados para la crítica literaria, así como para comprender sus limitaciones prácticas, esta investigación comparó evaluaciones cualitativas y cuantitativas de interacciones realizadas tanto por un modelo humano como por diversos modelos artificiales (ChatGPT, Gemini y GPT4ALL). El análisis de su desempeño reveló que ChatGPT obtuvo los mejores resultados, seguido por Gemini y GPT4ALL. Entre los aspectos en los que se observaron discrepancias en las respuestas, destacaron los siguientes: originalidad, extensión de la respuesta, continuidad entre intervenciones, variabilidad léxica, objetividad en el enunciado y la complejidad de adaptar los términos al contexto. Asimismo, se observó que la tendencia de estos modelos a lograr la persuasión mediante la simplificación discursiva y la estandarización retórica supone un problema para las dinámicas pedagógicas, las cuales deberían fomentar la erudición académica y un pensamiento crítico profundo.

Palabras clave: ChatGPT. Inteligencias Artificiales Generativas. Chatbots. Crítica Literaria. Enseñanza y Aprendizaje. Modelos de Lenguaje.

1 INITIAL CONSIDERATIONS

The popularization of generative artificial intelligences (GenAIs), known as chatbots or generative text intelligences, grew significantly and began to coexist with inequalities in education. In early 2024, the leading exponent of this technology, ChatGPT, was included in one of the stages of the production of teaching materials in the State of São Paulo. This movement reflects a dissonant expansion of the use of digital technologies in education, following the growing trend of using tools aimed at reducing teaching costs. It is worth noting that the insertion of ChatGPT, which had already been occurring informally in the learning process since its free availability in 2022, acquires relevance upon being made official by a state body, legitimizing its educational use.

Beyond the socio-educational issues that emerge in such a context, this research aimed to empirically discuss whether these “intelligences” would, in fact, generate satisfactory results or, on the contrary, could hinder the pedagogical process surrounding a literary text? Thus, it became necessary to problematize their use, focusing on stressing their potentialities and limits.

Considering these implications, it is relevant to observe that, if literary criticism mediated by artificial intelligence (AI) can offer new interpretative perspectives, this could also transform the student’s relationship with learning to read through literary texts. In this sense, would there be risks of compromising the intellectual autonomy and the critical repertoire of the users?

The results indicate that working with GenAI in teaching dynamics, especially in literary criticism, presents a complex horizon of possibilities. Certainly, problems such as the absence of creativity, the occurrence of hallucinations and the lack of scientific and didactic transparency emerge, which constitute relevant tensions for a possible use of this technology in teaching language and literature.

Furthermore, this type of technological mediation alters the form of interaction with texts, even influencing the comprehension of literary works. More specifically, the limitation to chatbots can excessively simplify literary discussions, reducing the formation of repertoire and, consequently, impoverishing the student’s interpretive experience. We understand that the presence of the textual genre is a crucial factor for the development of a notion of identity of the enunciator, who organizes, through this resource, their choices and interpretations.

Initially, this hypothesis is reinforced by Porter and Machery (2024), who invited non-specialists to analyze poems produced by AI and by human authors, both unidentified. The

results indicated that the simplest poems, with specific rhythm and beauty, were attributed to human authors, while the poems with flaws were recognized as produced by AI, for not being “well written”.

These findings highlight how stylistic marks are not seen as acts of erudition or transgression, but rather as inadequacies. This perspective could already characterize chatbots as a rather unstimulating pedagogical technology, since we consider that this repertoire component is fundamental for the emancipation of the student who actively engages in their learning process with literary texts.

Faced with this problematization, we used as an object of analysis the comparison of interactions performed with three artificial intelligence models (ChatGPT, Gemini and GPT4ALL). Our main objective is to contrast the qualitative and quantitative evaluations of these interactions, in order to understand which of them would be most suitable for educational use and, possibly, their implications for the teaching of literature.

2 THEORETICAL FRAMEWORK AND METHODOLOGICAL APPROACH

2.1 DEVELOPMENT OF ANSWER KEYS AND OBTAINING THE MODELS’ RESPONSES

The main strategy of the research is to use multidisciplinary perspectives to analyze the use of artificial intelligence models in literature, their performance and their implications. Firstly, two researchers, x and y, formulated a question for four short stories by Lima Barreto.

The criteria for formulating the questions were not debated between the two researchers. This decision aims to ensure a wide range of possible questions. The primary material, the short stories, were presented to the AIs in order to generate responses to the questions.

The criteria should not be shared. Furthermore, in a second stage, the LLMs were also included in the evaluation of the responses. From this corpus, the first step is to assimilate the information quantitatively in its entirety, to subsequently carry out a qualitative and isolated analysis of the questions and answers. The short stories and questions follow below:

Table 1

Questions

Short story	Questions
"The New California"	Can you suggest a possible interpretation for the presence of intellectuals in the work?
"The New California"	Can you suggest a possible interpretation for the presence of the dead in the work?
"The Man Who Knew Javanese"	How is satire present in the relationship between the title and the story depicted in the short story, considering the conceptions of "knowledge"?
"The Man Who Knew Javanese"	What is the criticism in the short story, and how is it presented?
"The kid"	How is the "*favela*" portrayed in the short story?
"The kid"	In what way are the characters portrayed in the short story?
"Your Excellency"	What does the narrative device of dreams accomplish in the short story?
"Your Excellency"	How is class difference portrayed in the story?

Source: Author's own work

A third researcher, from the field of Systems Development, submitted the questions to the artificial intelligence models GPT, GPT4ALL and Gemini. The researcher's response was considered the answer key. The extracts from these interactions were forwarded to four professors with doctorates in literature to evaluate the results of the four models (human, GPT, GPT4ALL and Gemini), to which they had no identification of authorship. The ratings were given on a scale of 0 to 100, with no shared criteria between them. Concurrently, the LLMs GPT, Gemini, LLAMA3 and Mistral Nemo compared the answer keys with the generated responses, thus producing a similarity index from 0 to 100. A qualitative and statistical analysis of the similarity indices was conducted, carrying out a human vs. AI analysis.

The models were trained using publicly available internet data. It is unknown whether the original texts were included, and therefore, the primary texts were presented to the AIs. The responses from all interactions and the answer key can be accessed in the supplementary material.

2.2 OBTAINING SIMILARITY INDICES BETWEEN IAS' RESPONSES AND THE ANSWER KEYS

In this regard, there is a table similar to Table 1, which contains the responses generated by the GPT's automatic model and the answer key, it is located in the supplementary materials as Table A. It should be noted that the database to which the automatic models have access is not publicly available. Therefore, it is unknown whether they were consulted on the original material or on materials derived from the short stories.

After these procedures, the first two researchers analyze the evaluations and the extracts from the models, from a qualitative perspective through the text analysis methodology (Marconi and Lakatos, 2003), while two other researchers, from the field of NLP (Natural Language Processing), analyze from a quantitative perspective.

2.3 QUANTITATIVE ANALYSIS OF SIMILARITY INDICES

In order to analyze the results obtained in the analysis carried out, NLP techniques were employed. Natural Language Processing, in an attempt to find patterns that justify the similarity indices generated by human consultants and by AIs, is a field of artificial intelligence dedicated to studying ways of translating human communication into a language that computers can understand and reproduce. This includes the interpretation of texts, generation of natural responses, among other applications.

NLP employs various techniques that enable the extraction of information from large volumes of text, making it possible to "quantify" what these texts represent. Therefore, NLP is used to process the existing data in order to extract information capable of explaining the agreement, or lack thereof, between humans and Large Language Models (LLMs), which are machine learning models based on artificial intelligence that use deep learning algorithms to understand and generate text.

Firstly, for this study, the set of responses from the LLMs and the aforementioned answer key were considered as the corpus. Then, to apply the NLP methodology, it was necessary to clean (or preprocess) the database in order to remove all present noise. In this

sense, noise is everything that does not represent relevant information, or that does not characterize the database. Thus, noisy terms can be punctuation marks, prepositions, articles and other words/characters with low semantic value.

To initiate preprocessing, tokenization is necessary, which consists of transforming each word or character of the database into a unique and independent term. Therefore, a text that was once a single entity becomes a set of tokens, which are a smaller division of the original text (Figure 1).

Figure 1

Tokenization example

```
Text before tokenization:  tokenization is text division in smaller pieces

Text after tokenization:  ['tokenization', 'is', 'text', 'division', 'in', 'smaller', 'pieces']
```

Source: Author's own work

Subsequently, to continue cleaning the database, stop words are removed. These are the noise elements abovementioned. This process is shown in Figure 2.

Figure 2

Example of removing stop words

```
Text before removal:  Stop words are useless!! :(

Text after removal:  ['Stop', 'words', 'useless']
```

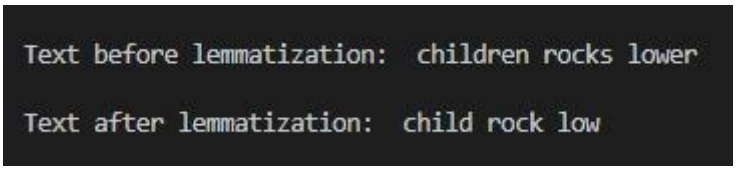
Source: Author's own work

Another technique employed was lemmatization, which consists of reducing a word to its primary and uninflected state. This is an important technique in the comparison between texts, as it can reveal a greater similarity between them (Figure 3).

ISSN: 2177-3246

Figure 3

Example of lemmatization

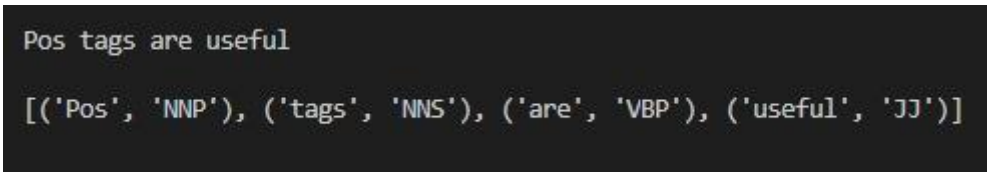


Source: Author's own work

In parallel, another essential technique is to identify part-of-speech (POS) tags, which consists of identifying all tokens of a given grammatical class. In this sense, when studying a database, it is necessary to discover the factors that characterize it. Therefore, one of the factors that may fit this objective is to determine the predominant grammatical class of the database, as shown in Figure 4.

Figure 4

Example of POS tagging



Source: Author's own work

In Figure 3, it is possible to analyze the database according to the grammatical class of the tokens. Where “NN” refers to a common noun in the singular, “NNS” to a common noun in the plural, “VBP” to a verb in the present indicative tense, and “JJ” to an adjective or ordinal numeral.

Another test conducted in this investigation involved a new round of collecting responses from the LLMs to the questions in Table 1. However, in this case, the extension of the responses was limited to the extension of the responses found in the answer key, in order to investigate the effect of the number of words in the comparison between the responses and the answer key, carried out by the LLMs. Finally, in addition to the NLP techniques, statistical tools were used in order to analyze the results presented in this research. The extracts and tables with similarity indices and NLP evaluations are included in the appendix of the text.

2.4 QUALITATIVE ANALYSIS OF TEXTS

As mentioned earlier, qualitative analysis is an analysis of texts. This means that we are not affiliated with a specific theory. However, the research that underpins our theoretical framework leads us to work with specific premises regarding AI. We discussed intrasemiotic adaptation through the work of and Santaella (2023), to discuss the functioning of AI and originality. In addition, text analysis by Marconi and Lakatos (2001) was employed.

3 APPLICATION OF THE METHOD

Initially, after conducting the experiment, the results were organized in order to extract relevant information about the quantitative performance. Therefore, in order to understand the nuances between the evaluations of the human consultants, two groups were created: similarity indices for the human consultants and similarity indices for the automatic models (LLMs). In this sense, it is possible to compare the behavior of one group against the other for each of the questions asked.

The analysis is focused on the quantitative performance of artificial intelligences and how they behave in relation to human thought and reasoning, seeking to compare these behaviors using NLP and statistical tools in order to understand the reason for the existence (or not) of agreement between human responses and the results of LLMs.

Firstly, after separation into groups (humans and LLMs), the mean and standard deviation of both were calculated. Thus, using this data, it was possible to draft the normal distribution of each group, through an algorithm in the Python programming language. The aim was to graphically compare the behavior of both, thus evidencing the level of similarity between them.

In this sense, it was necessary to compare the responses of the LLMs with the answer key, in order to determine a possible correlation between them. In addition, NLP techniques were adopted. Therefore, as it was done for the normal distribution, these techniques were applied through a Python programming environment, and the results obtained were explained graphically, highlighting the points where there was greater or lesser compatibility between the automatic models and the answer key. Including, the figures were made in python.

Furthermore, the attempt to reduce the GPT's responses to match the length of the answer key was conducted using LLMs. To achieve this, the questions were resubmitted to the model, with the condition that the response had approximately the same number of characters as the answer key.

4 DISCUSSION AND RESULTS

The discussion and results are divided into two subsections titled “Quantitative Performance Analysis” and “Qualitative Performance Analysis”, which provide a more in-depth examination of the issues within the corpus.

4.1 QUANTITATIVE PERFORMANCE ANALYSIS

First, the quantitative analysis focuses on the similarity indices of the responses generated by the LLM “GPT”. Therefore, the tables, figures and graphs in this subsection refer to GPT.

Table 2

GPT’s results

LLM MODELS				CONSULTANTS				STATISTICAL INDICATORS			
GEMINI	GPT	LLAMA 3	MISTRAL NEMO	1	2	3	4	LLM AVERAGE	LLM STD	CONSULTANT AVERAGE	CONSULTANT STD
70	70	40	60	20	60	0	70	60,00	14,14	37,50	33,04
20	40	20	50	20	50	0	40	32,50	15,00	27,50	22,17
92,5	85	80	95	100	100	100	90	88,13	6,88	97,50	5,00
90	85	80	85	90	100	100	50	85,00	4,08	85,00	23,80
25	60	20	60	100	50	20	15	41,25	21,75	46,25	39,02
20	60	0	20	90	20	10	20	25,00	25,17	35,00	36,97
65	60	20	80	100	40	10	30	56,25	25,62	45,00	38,73
90	85	85	60	100	60	50	60	80,00	13,54	67,50	22,17
TOTAL AVERAGE								58,52	15,77	55,16	27,61

Source: Author's own work

As previously stated, there are two groups: the first four columns of Table 2 refer to the group of automatic models (GPT, Gemini, LLAMA3 and Mistral Nemo). While the following four refer to the consultants. It is possible to note that, for each of these columns, there are eight rows, and each entry corresponds to one of the questions asked about the short stories.

The last four columns of the table are statistical data calculated from the results obtained. Furthermore, after reducing the length of the GPT-generated responses, the result presented in Table 3 was obtained.

Table 3

GPT's results after length reduction

LLM MODELS				STATISTICAL INDICATORS	
GEMINI	GPT	LLAMA3	MISTRAL NEMO	LLM AVERAGE	LLM STD
80	70	70	60	70,00	8,16
25	30	25	40	30,00	7,07
92,5	85	90	95	90,63	4,27
92,5	85	85	92,5	88,75	4,33
40	60	40	95	58,75	25,94
40	60	20	30	37,50	17,08
40	30	20	50	35,00	12,91
75	70	80	70	73,75	4,79
TOTAL AVERAGE				60,55	10,57

Source: Author's own work

Firstly, the consultants' evaluations, and consequently their statistical data, were excluded from Table 3, as the reduction test was conducted aiming only at altering the results of the automatic models. Subsequently, it is notable, when comparing Table 3 with Table 2, that the influence of the corpus reduction is minimal, slightly increasing the overall average, and decreasing the average of the standard deviations. We continue the analysis based on Table 2.

Most responses showed a difference of up to points between the consultants' average and the LLM models' average. The only exception was the first response, where the LLMs' average was 60 and the consultants' average was 37.50. Even so, this is considered a normal result, as the variation was present only in the LLAMA3 response.

As may be seen in the responses from Tables 2 and 3, regardless of the questions, the variation was minimal. It can be ascertained that the response with the greatest variation among the scores was the first, about the short story "The Kid": "How is the "*favela*" portrayed in the short story?" (referring to line 5 of Table 2). In a way, it is the most concise question asked in this experiment. It explicitly identifies the spatial referent (short story), the temporal referent (-ed, indicating the past perfect tense), and only one noun.

In contrast, the response that obtained the least variation was from the short story “The Man Who Knew Javanese”. The question “How is satire present in the relationship between the title and the story depicted in the short story, considering the conceptions of ‘knowledge’?” (referring to line 3 of Table 2) demonstrates greater complexity in its interpretative trail. Its writing makes use of a present local referent, which can be misinterpreted due to its similarity with the present temporal referent. It is also considered that it offered a wider range of criteria, naturally contributing to its difficulty. The interaction was also the one with the highest average among consultants and LLMs.

It should be noted that the scores were assigned based on the complexity of the response, even though the average of short responses versus longer responses did not interfere with the evaluation, regardless of these results.

As for the averages, the interaction with the highest average is explained by the same reason as the standard deviation. The complexity of the response may be a reason why the third interaction in Tables 2 and 3 achieved the highest average among the scores. We also consider that the question provided sufficient criteria to guide it with text construction parameters. For example, it includes the comparison between the title and the text, and the consideration of the idea of knowledge. It is noted that the fourth interaction, also conducted for the same question, achieved a similar score. Curiously, the interaction with the lowest score is notable for having a smaller number of words and, naturally, less demonstration of criteria for construction.

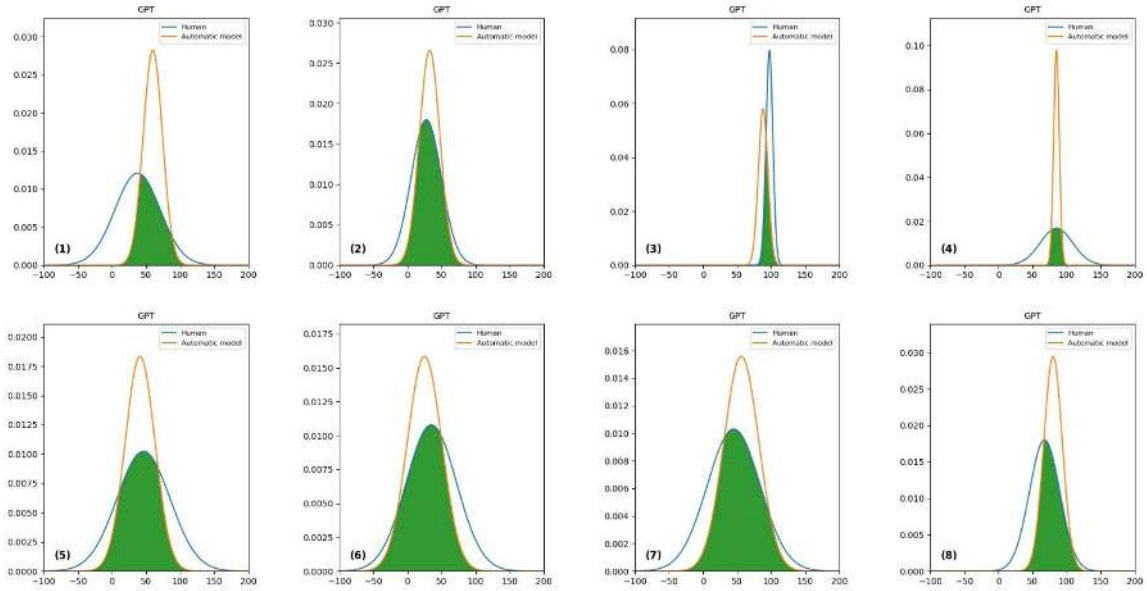
Next, we address an analysis of the Gaussian distribution, or normal distribution. It is a symmetric curve whose peak lies at the midpoint, presenting a characteristic bell shape. This distribution is defined by the average and the standard deviation of the samples obtained in an experiment. The area under the curve determines the probability of a given event occurring, with the area of a specific interval representing the probability that the variable will take on values within that interval.

In the context of this experiment, the Gaussian distribution represents the behavior of the similarity indices generated both by the LLMs group and by the consultant group, in relation to each question and each short story.

Furthermore, the area in which the normal distribution of the consultants coincides with the area of the LLMs indicates the degree of agreement between the behaviors of the two groups. This intersection highlights instances in which artificial intelligence aligns with human assessments, demonstrating the compatibility between their behaviors.

Graph 1

GPT's Gaussian distributions



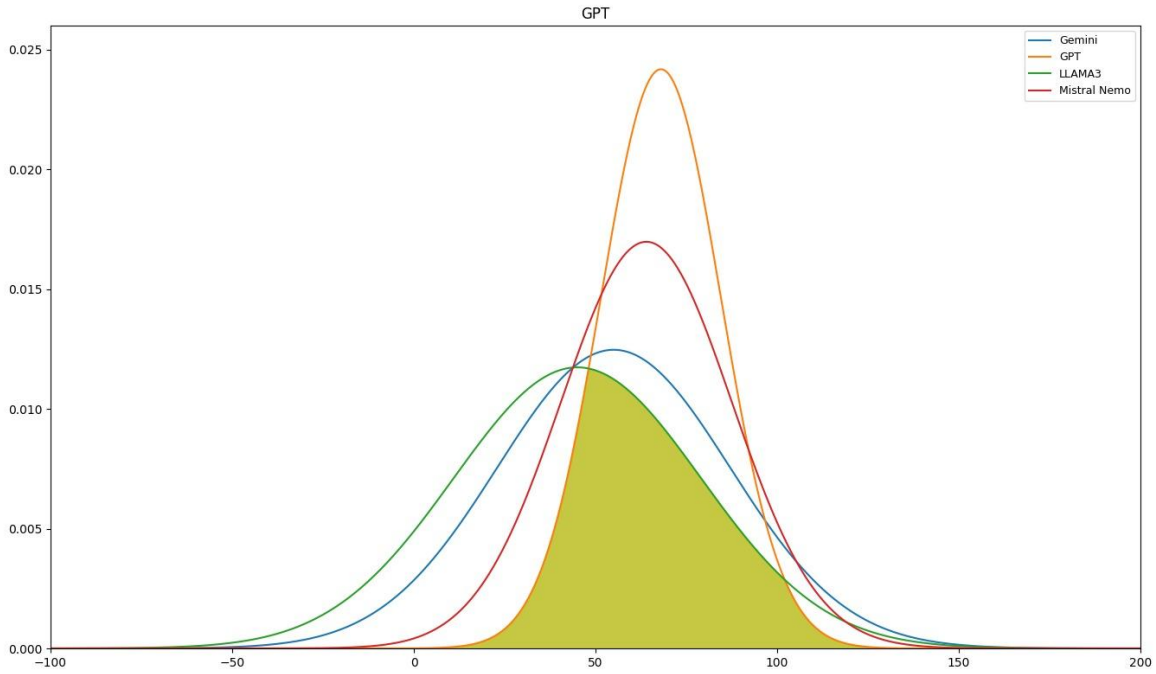
Source: Author's own work

As can be seen in Graph 1, there are two curves, one blue and one orange, distributed across two numerical scales, which represent quantified variables for the consultants' responses. The blue curve represents the consultants' responses, while the orange curve represents the LLMs' responses. Additionally, the green area symbolizes the interval in which the groups agree with each other. Thus, it is possible to see that there are questions where there is a high degree of overlap between the groups, which means that the consultants largely agree with the LLMs. On other questions, however, there is little overlap, indicating low agreement between them.

Similarly, it is possible to express the extent to which each group agreed with each other, for the entire set of responses. This is done again using means and standard deviations, but now for each column of Table 2, rather than each row. The overlap can be observed in Figures 1 and 2. This is done again, through the averages and standard deviations, but now for each column of Table 2, instead of each row of it. The intersection can be observed in Graphs 2 and 3.

Graph 2

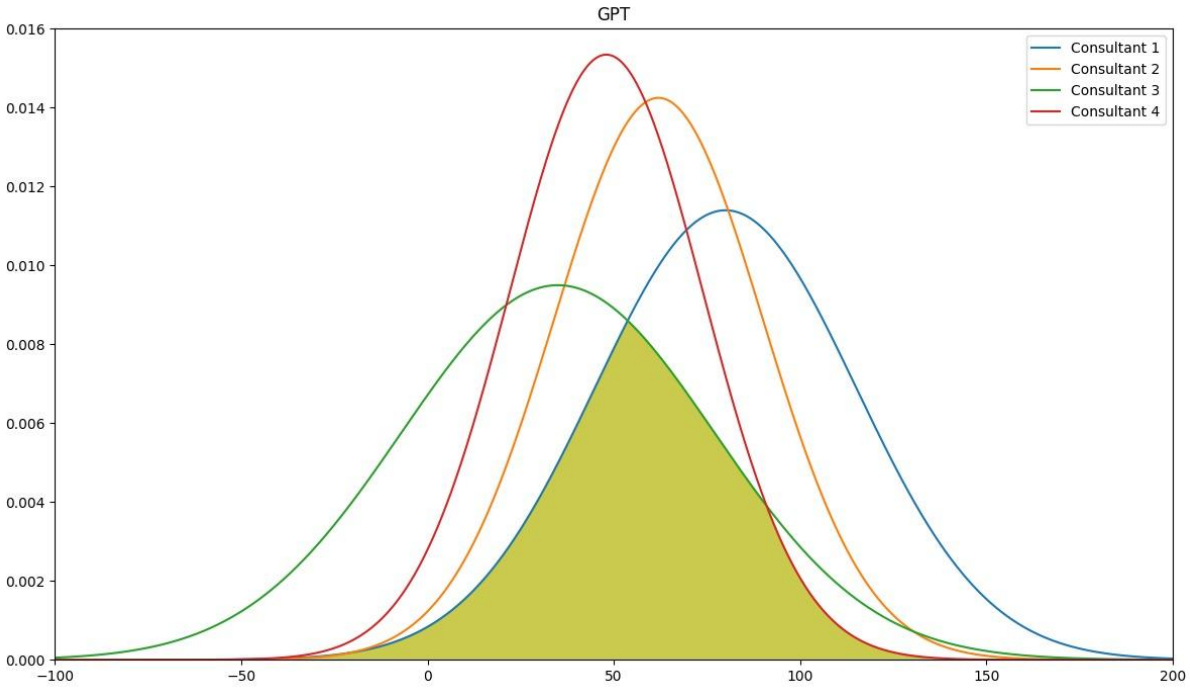
LLMs' Gaussian distributions.



Source: Author's own work.

Graph 3

Consultants' Gaussian distributions.

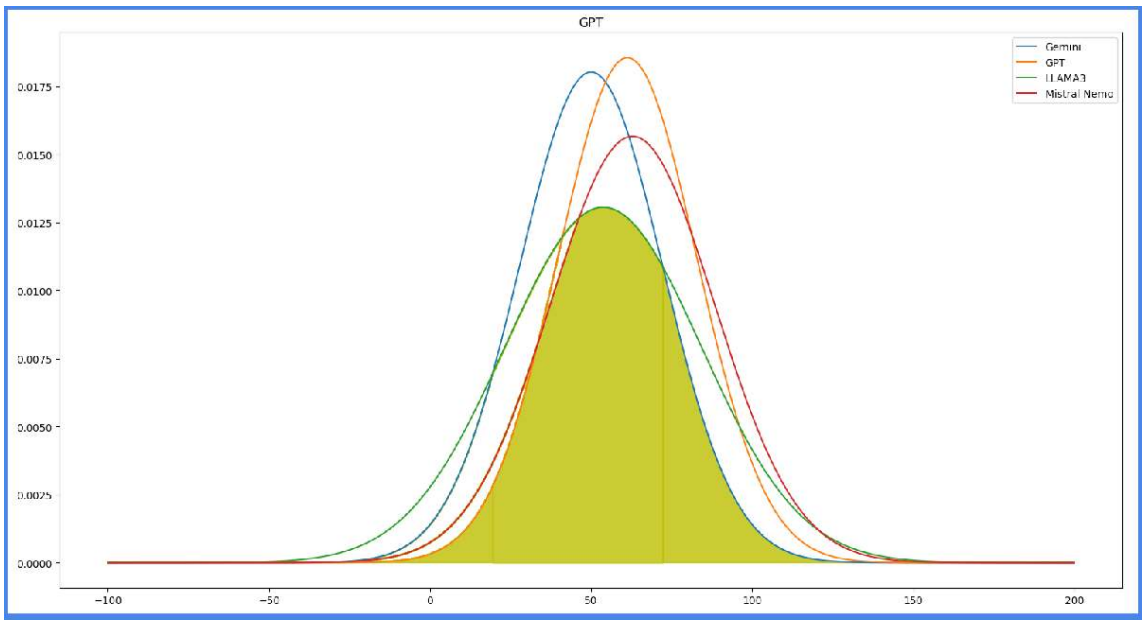


Source: Author's own work.

Graph 2 shows the Gaussian distributions describing the level of agreement among the LLMs for all eight responses generated by GPT, and their area of intersection is 0.5282. Similarly, Graph 3 presents the same information, but for the human consultants, with an intersection area of 0.5357. This analysis demonstrates that humans, overall, agree slightly more among themselves than the automatic models. Conversely, the attempt to reduce the length of the GPT's responses to match that of the answer key increased the intersection area of the LLMs group to 0.7131. The new value demonstrates greater agreement among the LLMs when analyzing a reduced response, as shown in Graph 4.

Graph 4

LLMs' Gaussian distributions for reduced responses.



Source: Author's own work.

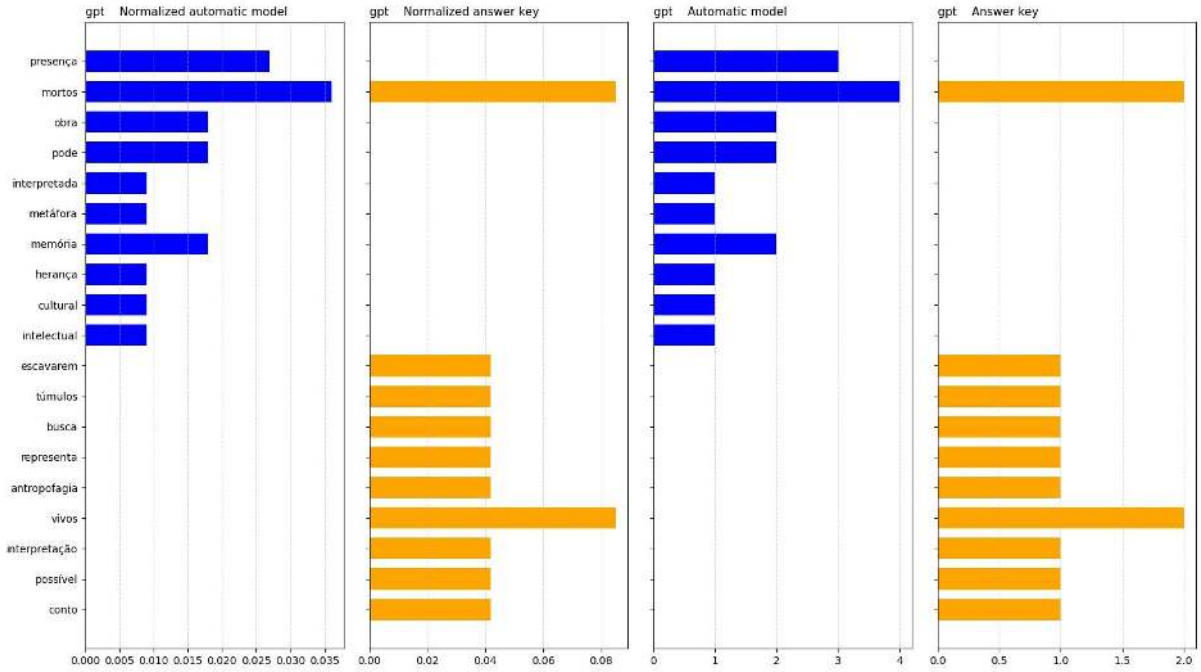
In order to compare the corpora of responses generated by LLMs with those in the answer key, NLP techniques were applied to these corpora. The goal was to identify information that explains the similarities and differences found between these two distinct corpora.

As presented in Graph 5, the normalized histogram results from the number of times a word appears, divided by the total number of words in the text. In this regard, the reason for presenting a normalized histogram stems from the difference in size between the responses from the automatic models and the answer key. Therefore, to highlight the weight a word holds in each corpus, relative to its length, it became necessary to normalize them. The non-

normalized version would simply present the number of times the word appeared, in order to demonstrate the repetition of words in each corpus.

Graph 5

Histogram after preprocessing.

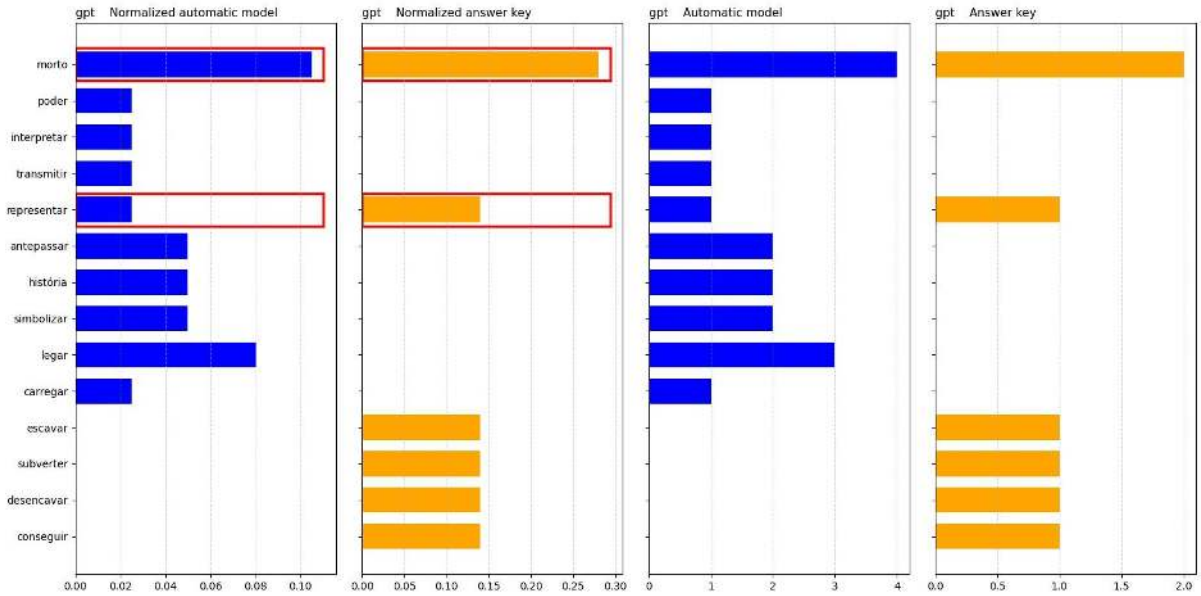


Source: Author's own work.

It is evident that the automatic model presents greater repetition of words, as demonstrated in Graph 5. For example, the word “mortos” appears in all four samples, both in the integral and in the normalized version. It is also clear that the word “vivos” appears only in the answer key, this is understood to be a strategy employed by the researchers to underscore the story’s central conflict, with life and death serving as its central themes.

Graph 6

Histogram after lemmatization + POS tagging.



Source: Author's own work.

It may be seen, through Graph 6, that the intersection between the texts worsened. Before the application of lemmatization and POS tagging techniques, there was coincidence in the usage of certain words, however, upon restoring the original meaning of the terms, they became restricted to exclusive use within the model or within the answer key.

Finally, the analysis indicated that the average scores for each group varied considerably, although the difference between the consultants' evaluation and that of the LLMs was not relevant. Some groups could be evaluated as more or less complex based on their scores, and the standard deviations indicate that the responses also did not vary substantially. Regarding the NLP techniques, greater lexical variety is indicated for the answer key responses. We will continue the analysis below.

4.2 ANALYSIS OF QUALITATIVE RESULTS

At this initial stage, the main comparative results are contextualized so they can be better explained subsequently. We begin with GPT. As shown in Table 2, the GPT's responses are more extensive than the answer key version produced by the researchers. This is not a pattern for all LLMs, as others occasionally responded briefly.

The structuring of the text into topics, however, can serve as a parameter for the consultants' analysis, as it can be interpreted both as a didactic resource and as a distribution. It is observed that this topic-based structuring is widely used beyond GPT, as can be seen in

the supplementary material when analyzing all the responses generated for the questions in the table.

However, it is understood that the enumeration does not follow a logic of consecutiveness, but rather one of didacticism, since these are six important points about the story, even if they do not possess a unity as a whole. It is also noted that the responses do not include delusions.

The questions evade objective character and require an indication of the manner through which the short stories proceed to achieve meaning, through words such as “how” and “in what way”. For the question “How is class difference portrayed in the story?” Chat GPT presents a response with narrative movements of synthesis of the short story and demonstration in topics of the stages of concretization of this meaning.

Furthermore, the last two short stories, numbers 5 and 6, demonstrate two perceptions of the short story’s stylistics. In them, irony and social criticism are understood as elements related to the manner of delivering the content. The term “social criticism” was cited twice by GPT, while it was cited four times by Gemini and only once by GPT4ALL. We consider that social criticism may be the definition of a phenomenon, and when repeated several times it symbolizes a linguistic crutch that loses its meaning. It also becomes possible to note the absence of creativity in the use of synonyms.

As for the response developed by the researchers, it serves to identify keywords from an analysis considered appropriate for literary criticism. The term “how” in the question enables two approaches: one of an objective nature, concerning the object, and another subjective, depending on the subject who interprets them.

The question “How is the ‘favela’ portrayed in the short story?” was answered by the human model and by the LLM in different ways. While the human model uses the lexicons “non-homogeneous”, “generalizations” and “idealizations” to thematize Lima Barreto’s approach as complex, the artificial version achieved a response of a descriptive nature about the object. It is divided into topics of interest such as architecture, vegetation, social relations, and cultural aspects.

It is noted that GPT’s response projected the term “how” from the question onto an image of the setting and the psychology of the characters, producing a simulacrum of truth that is proposed by Lima Barreto in the fictionalization of the text. However, it disregards the role of co-enunciator of the readers and their connections with the world. Ultimately, this

shows that “how” takes into account only the interaction between the text’s interlocutors, disregarding the writer’s, Lima Barreto, perspective and the text’s relationship to society.

In terms of similarity, GPT achieved the best performance in scoring, with 1765 points, approximately 300 points more than Gemini and 324 more than GPT4ALL, which ranked second and third, respectively. GPT was better evaluated in six responses, while Gemini was the best in 2. GPT4ALL did not achieve numerical superiority in any response.

Based on the results, we suppose that the criterion of response length may serve as an analytical criterion for the quality attributed to the responses. This was already expected, as an analytical question requires greater complexity of argumentation to be addressed. Therefore, it is expected that consultants assign scores according to the length of the responses.

Table 4

GPT4ALL scores

LLM MODELS				CONSULTANTS				STATISTICAL INDICATORS			
GEMINI	GPT	LLAMA 3	MIXTRAL	1	2	3	4	LLM AVERAGE	LLM STD	CONSULTANT AVERAGE	CONSULTANT STD
75	60	60	60	50	40	0	70	63,75	7,50	40,00	29,44
25	40	20	20	30	30	0	40	26,25	9,46	25,00	17,32
90	85	95	95	89	70	100	40	91,25	4,79	74,75	26,27
30	30	20	30	90	20	0	90	27,50	5,00	50,00	46,90
70	75	60	60	30	40	10	30	66,25	7,50	27,50	12,58
25	70	20	30	10	40	0	30	36,25	22,87	20,00	18,26
70	60	20	60	90	20	10	30	52,50	22,17	37,50	35,94
87.5	70	70	80	100	60	50	40	73,33	5,77	62,50	26,30
TOTAL AVERAGE								54,64	10,63	42,16	26,63

Source: Author's own work

A question for which all three intelligences presented adequate responses, according to the consultants, was: “How is satire present in the relationship between the title and the story depicted in the short story, considering the conceptions of ‘knowledge’?” (referring to line 3 of Tables 2, 4, and 5). Of twelve scores, two were considered low (20 and 40), while the remaining were above 80. This question also demonstrated complexity in the quantitative analysis.

Curiously, it is evident that the question addressed above draws on more theoretically concrete references, as opposed to those based on thematic references. Those that, to conduct a textual analysis, draw upon broad and potentially ambiguous concepts, as can be seen in the meanings of satire, a concept extensively developed by theories of humor (Lins and Gonçalves, 2012, pp. 19–34), and the etymological or theoretical concept of the word "knowledge". In other words, a theoretical framework is one that is supported by a third-party theory. Focusing solely on the treatment of humor, it is known that it has been treated as a scientific object for several centuries, and that there is little consensus regarding the definition of satire, to the point that it can be understood in its literal sense or in a scientific sense.

Table 5

Gemini scores

LLM MODELS				CONSULTANTS				STATISTICAL INDICATORS			
GEMINI	GPT	LLAMA 3	MIXTRAL	1	2	3	4	LLM AVERAGE	LLM STD	CONSULTANT AVERAGE	CONSULTANT STD
65	60	20	60	90	40	0	50	51,25	20,97	45,00	36,97
25	40	20	40	45	30	0	30	31,25	10,31	26,25	18,87
90	85	92	95	90	80	100	95	90,50	4,20	91,25	8,54
85	80	85	70	100	80	90	25	80,00	7,07	73,75	33,51
35	30	20	60	0	30	0	20	36,25	17,02	12,50	15,00
35	70	20	60	0	20	0	20	46,25	22,87	10,00	11,55
45	40	60	45	80	30	10	20	47,50	8,66	35,00	31,09
85	70	80	60	100	80	90	20	73,75	11,09	72,50	35,94
TOTAL AVERAGE								57,09	12,77	45,78	23,93

Source: Author's own work

Gemini's response to the question "Can you suggest a possible interpretation for the presence of intellectuals in the work?" (referring to line 1 of Table 5) suggests that the meaning of "intellectuals" was not retrieved by the AI. In this case, it reproduced common stereotype of the word. Thus, it is inferred that the LLM's performance is unsatisfactory when it is supported by thematic referentials.

These thematic referentials would be those that refer to the object of analysis itself in a broad manner, the text. In the absence of the work itself, due to copyright issues, even

relying on analyses carried out on the internet, the LLM's response does not retrieve the meaning of the word used in the short story.

Reflecting on the application of these theoretical vs. thematic concepts and their relationship with the database, it is inferred that working with literary text is not a problem, after all, responses to "What is the criticism in the short story, and how is it presented?" (referring to line 4 of Tables 2, 4, and 5) were evaluated and obtained good scores. However, it is possible to perceive the difficulty of the AIs in creating a valuation based solely on secondary materials. It is advisable, in order to improve their performance, to conceptualize what is meant by criticism, through examples or referencing a scientific method or author.

This endorses the expectation that AIs would have difficulty in developing value-based inferences regarding the work. Regarding this topic, it is important to explain that the theoretical framework developed until now has reflected the difficulty of AIs in presenting originality (Santaella, 2023). As demonstrated thus far, further evidence of this difficulty is observed, nevertheless, this "original" character is problematized.

The etymological meaning of "original", according to the Michaelis dictionary (2024), would be related to the origin of a primitive meaning for the element discussed. This kind of essence that is subjective to something would be, in the case of the LLM, an inference that maintains aspects of the trajectory through which the response is formulated.

As mentioned in Santaella (2023), the conceptualization of originality is not transparent in Chatbot interactions. It is assumed that it is not evident because it does not restore the origin of the utterance. Perhaps due to the absence of utterances that could support this theoretical construction.

This is confirmed by the result found in the consultants' evaluation for the question: "In what way are the characters portrayed in the short story?" (referring to lines 5 and 6 of Table 2). The low score attributed to it endorses the inference presented by Santaella. There is a certain difficulty on the part of LLM in adapting to reading different genres, in this case the short story, and in developing a theory-based analysis based on external sources. After all, the question does not point to a specific perspective, and without a definitive view, it offers a general analysis. In this sense, we can be concerned with the description of what creation is, and how it occurs through the LLM.

When thinking about which textual construction to use to achieve a better result, we come across the indication of a referential that supports a theoretical application, and that had the possibility of presenting inference, in order to be original in this situation. When

reflecting further on other resources to use in this construction, we decide to compare the difference in scores assigned to the three intelligences.

The question: “What is the criticism in the short story, and how is it presented?” (referring to line 4 of Tables 2, 4, and 5) is noteworthy. In this case, the disparity in scores, likely attributable to the question’s complexity, given that it contained an additional embedded question, resulted in low evaluations across all three models. Even so, other questions also presented this complexity. Perhaps the indication of a singular reading. A more attentive reading may perceive that the word “criticism” appears only in this question, which makes it special, as it is evidence that the LLM may not understand the meaning of the word criticism for a human and linguistic theory. As for the question that was contemplated by more objective lexicons, they resulted in more adequate responses.

Upon reflection of the quantitative and qualitative results, we perceive that, in fact, the consultants took into account that short responses did not contain the necessary information for a satisfactory answer. While hallucinations were not observed in the responses, the quantitative analysis demonstrates that the answer key (human) presented greater lexical variety, which naturally makes the LLM’s responses limited.

It was ascertained that the LLM does not retrieve the meaning of certain words used in these questions and potentially cannot articulate them with the responses from the literary texts. This caused its analyses to be descriptive and adequate, but at times confusing and redundant.

5 FINAL REMARKS

In summary, aspects that were identified in the discrepant responses were originality, response length, response consecutiveness, vocabulary variability, objectivity of enunciation and the complexity of adapting a term to the context of the utterance.

Considering all the inferences from the analysis, when contrasting these results with our perceptions, it is understood that the artificial models develop objective strategies to persuade the reader. The absence of synonyms demonstrates an objective interpretation, contextual understanding at the interlocutor’s instance and the inability to verify information from the database.

Their attempt to achieve persuasion through the reproduction of terms may represent a didactic strategy of simplifying the result, even when not requested. Meanwhile, the

expectation that they would not demonstrate repertoire and erudition is confirmed through this analysis, although this was not a crucial point in the evaluators' assessment.

With respect to this objectivism, we do not understand them as utilitarianism, initially. However, when contrasted with primary lexical choices, that is, a textual interpretation that did not comprehend polysemic meanings, we understand that AI performs simplified interactions through these strategies. The alterity that is formed between the human and artificial model, however, does not manifest itself through stylistic questions.

At the end of the writing, we realized that the experiment could have explored different situations in prompting content from the machine, perhaps questions related to gender and extreme situations could reveal some variation in the LLM's performance. However, we believe we made the correct decision in demonstrating usage situations.

ACKNOWLEDGMENTS

We would like to thank the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) for its support.

AUTHOR CONTRIBUTIONS

Douglas Almonfrey and Philipe Brito de Moraes were responsible for the experiments with the chatbots, while Raoni Schimitt Huapaya and Gabriel Magalhães Siston were responsible for analyzing the results.

CLINICAL TRIAL REGISTRATION

Not applicable (this study is not a clinical trial).

CONSENT TO PARTICIPATE

As the study exclusively used secondary and anonymized data from consultants, with no possibility of individual identification or direct contact, obtaining Informed Consent was not applicable.

CONSENT FOR PUBLICATION

As the manuscript does not contain identifiable personal data, images, or individual details of any participant, formal consent for publication is not applicable.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests or financial conflicts that could have influenced the work reported in this paper.

ETHICAL STATEMENT

This study utilized exclusively data from anonymous consultants, guaranteeing complete confidentiality and non-identifiability of the participants. As it did not involve the direct collection of identifiable personal data or interventions on human subjects, submission to an Institutional Review Board (IRB) / Ethics Committee was not required, in accordance with current ethical guidelines.

DATA AVAILABILITY STATEMENTS

The datasets consist of primary data and are available upon reasonable request from the corresponding author, Raoni Schimitt Huapaya. Contact: raoni.huapaya@ifes.edu.br.

REFERENCES

Editora Melhoramentos. (2024). Original. In *Michaelis dicionário brasileiro da língua portuguesa*. UOL. https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/original

Lakatos, E. M., & Marconi, M. de A. (2003). *Fundamentos de metodologia científica* (5ª ed.). Atlas.

Lins, M. da P. P., & Gonçalves, L. S. (2012). *O humor como discurso de prevenção: O cartum sob a ótica da pragmática* (Dissertação de Mestrado). Universidade Federal do Espírito Santo, Vitória, ES, Brasil.

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14, Artigo 27419. https://doi.org/10.1038/s41598-024-76900-1

Santaella, L. (2023). Balanço crítico preliminar do ChatGPT. *Revista FAMECOS*, 30. https://doi.org/10.15448/1980-3729.2023.1.44877