Data and Metadata. 2024; 3:.359
doi: 10.56294/dm2024.359

ORIGINAL

Optimizing Natural Language Processing: A Comparative Analysis of GPT-3.5, GPT-4, and GPT-4o

Optimización del procesamiento del lenguaje natural: un análisis comparativo de GPT-3.5, GPT-4 y GPT-4o

Manuel Ayala-Chauvin[1], Fátima Avilés-Castillo[1]

[1]Centro de Investigación de Ciencias Humanas y de la Educación, Universidad Indoamérica. Ambato, Ecuador.

Cite as: Ayala-Chauvin M, Avilés-Castillo F. Optimizing Natural Language Processing: A Comparative Analysis of GPT-3.5, GPT-4, and GPT-4o. Data and Metadata. 2024; 3:.359. https://doi.org/10.56294/dm2024.359

Submitted: 21-01-2024 **Revised:** 07-05-2024 **Accepted:** 09-09-2024 **Published:** 10-09-20234

Editor: Adrián Alejandro Vitón-Castillo

Corresponding author: Manuel Ayala-Chauvin

ABSTRACT

In the last decade, the advancement of artificial intelligence has transformed multiple sectors, with natural language processing standing out as one of the most dynamic and promising areas. This study focused on comparing the GPT-3.5, GPT-4 and GPT-4o language models, evaluating their efficiency and performance in Natural Language Processing tasks such as text generation, machine translation and sentiment analysis. Using a controlled experimental design, the response speed and quality of the outputs generated by each model were measured. The results showed that GPT-4o significantly outperforms GPT-4 in terms of speed, completing tasks 25 % faster in text generation and 20 % faster in translation. In sentiment analysis, GPT-4o was 30 % faster than GPT-4. Additionally, analysis of response quality, assessed using human reviews, showed that while GPT-3.5 delivers fast and consistent responses, GPT-4 and GPT-4o produce higher quality and more de-tailed content. The findings suggest that GPT-4o is ideal for applications that require speed and consistency, while GPT-4, although slower, might be preferred in contexts where text accuracy and quality are important. This study highlights the need to balance efficiency and quality in the selection of language models and suggests implementing additional automatic evaluations in future research to complement the current findings.

Keywords: Natural Language Processing; GPT-4o; Response Time; Model Performance; OpenAI API.

RESUMEN

En la última década, el avance de la inteligencia artificial ha transformado múltiples sectores, destacando el procesamiento del lenguaje natural como una de las áreas más dinámicas y prometedoras. Este estudio se centró en comparar los modelos de lenguaje GPT-3.5, GPT-4 y GPT-4o, evaluando su eficiencia y desempeño en tareas de procesamiento del lenguaje natural como generación de texto, traducción automática y análisis de sentimientos. Utilizando un diseño experimental controlado, se midieron la velocidad de respuesta y la calidad de los resultados generados por cada modelo. Los resultados mostraron que GPT-4o supera significativamente a GPT-4 en términos de velocidad, completando tareas un 25 % más rápido en la generación de texto y un 20 % más rápido en la traducción. En el análisis de sentimiento, GPT-4o fue un 30 % más rápido que GPT-4. Además, el análisis de la calidad de la respuesta, evaluado mediante revisiones humanas, mostró que, si bien GPT-3.5 ofrece respuestas rápidas y consistentes, GPT-4 y GPT-4o producen contenido más detallado y de mayor calidad. Los hallazgos sugieren que GPT-4o es ideal para aplicaciones que requieren velocidad y coherencia, mientras que GPT-4, aunque más lento, podría preferirse en contextos donde la precisión y la calidad del texto son importantes. Este estudio destaca la necesidad de equilibrar la eficiencia y la calidad en la selección de modelos lingüísticos y sugiere implementar evaluaciones automáticas adicionales en

futuras investigaciones para complementar los hallazgos actuales.

Palabras clave: Procesamiento de Lenguaje Natural; GPT-4º; Tiempo de Respuesta; Rendimiento del Modelo; API Openai.

INTRODUCTION

In the last decade, the advancement of artificial intelligence (AI) has transformed multiple sectors, from medicine to the entertainment industry.[1] Emerging technologies such as robotics, immersive systems and big data have enabled considerable improvements in today's society.[2] The analysis of large volumes of information allows for more informed and accurate decision-making, optimizing processes and discovering patterns that previously went unnoticed.[1] This feedback has driven significant advances in computing systems.[3] When combined with AI, applications ranging from autonomous driving to personalized recommendation systems can be managed.[4] Among these applications, natural language processing (NLP) has emerged as one of the most dynamic and promising areas.[5] NLP has been a constantly evolving area of research that links computer science and linguistics. Researchers have sought to develop systems capable of understanding and generating human text similarly to how people do.[6]

This goal has led to significant advances over the years, with the emergence of increasingly sophisticated and powerful language models.[7] One of the most notable milestones in this evolution has been the arrival of large language models (LLMs), which have radically transformed our ability to tackle complex tasks in NLP.[8] These models, powered by deep learning techniques and trained on vast amounts of textual data, have demonstrated an impressive ability to understand and generate text with an increasing degree of accuracy and fluency. However, its size and complexity have also posed significant challenges in terms of efficiency and scalability, driving research towards more optimized and sustainable approaches.[7]

Language models have evolved rapidly over the past decade, moving from rule-based approaches to deep learning systems capable of processing large volumes of data. This was driven by significant advances in processing power and the availability of increasingly larger and more diverse data sets.[9] A relevant milestone was the introduction of GPT-3, which captured the world's attention for its unprecedented ability to perform a variety of natural language processing tasks. GPT-3 demonstrated the potential of LLMs in addressing complex NLP challenges, but also pointed out limitations in terms of computational costs and efficiency. Despite these limitations, GPT-3 laid the foundation for future innovations in the field and served as the basis for the development of GPT-4.[10]

With the release of ChatGPT-4, greater capacity and accuracy of language models were achieved.[11] This version, developed by OpenAI , not only surpassed its predecessor in terms of size and complexity, but also introduced significant advances in the quality of the generated text and the accuracy in understanding human language.[12] This improvement was reflected in its ability to generate coherent and relevant content in a variety of contexts, as well as its ability to accurately translate between languages.[13] However, the complexity of GPT-4 posed considerable challenges in terms of efficiency and resource management. The massive size of the model required significant computing power for its training and use, which limited its accessibility and scalability in some environments.[14] In response to these challenges, efforts have emerged to develop optimized versions of GPT-4, such as GPT-4o (released May 13, 2024)[15], which seeks to maintain model quality while reducing its resource footprint and increasing its efficiency in energy use. These initiatives represent an important step towards creating more accessible and sustainable language models for a wide range of real-life applications.

In recent years, there has been research aimed at comparing the GPT-3.5 and GPT-4 language models and their performance relative to human beings. The advancement of artificial intelligence in medicine highlights its potential in medical diagnosis and treatment.[16] When evaluating ChatGPT versions 3.5 and 4.0 for renal cell carcinoma (RCC) queries, improvements were observed through model tuning and optimization. Additionally, OpenAI's GPT-4 has gained attention in the medical research community for its innovative "Data Analyst" feature.[17] ChatGPT-4, used alongside traditional biostatistical software (SAS, SPSS, R), shows promise in the statistical analysis of epidemiological data.

In [18] GPT-4's occupational assessments were compared with a human survey in the United Kingdom. The results showed a high correlation between GPT-4 scores and those of humans, but also revealed that GPT-4 tends to be more generous in its evaluations. However, it significantly underestimates or overestimates the prestige and social value of certain occupations. Similarly, in [19] it was identified that educational materials for patients on bariatric surgery could be very complex to understand without technical knowledge. Therefore, this study evaluated the ability of LLM models to simplify these instructions for any audience.

In [20] the use of these models by Indian students preparing for the university entrance exams in health,

called the National Eligibility cum Entrance Test (NEET), was analyzed. Finally, Stribling et al. found that GPT-4 performed better on biomedical science exams compared to students.[21] However, it shows deficiencies in questions with simulated data and hand-drawn answers, in addition to presenting cases of possible plagiarism and confusion in the responses.

OpenAI discusses the security improvements and challenges associated with the GPT-4 model, highlighting how the mitigations and processes implemented reduce certain types of misuse, although they remain limited and fragile in some cases.[22] Through a series of qualitative and quantitative evaluations, the analysis of GPT-4 shows an increase in performance in areas such as reasoning and knowledge retention compared to GPT-3.5, but also identifies emerging risks and the need for advanced planning and governance to address these challenges.

Although GPT-3.5 and GPT-4 have been widely studied and used, there is a lack of specific research on the advantages and limitations of optimized versions such as GPT-4o. Most studies have focused on the general capabilities of large language models, without addressing in depth how optimizations can affect efficiency and performance in practical applications. This study seeks to fill that gap, providing empirical data on how GPT-4o can overcome the limitations of GPT-4 in specific contexts.

METHOD

The methodology of this study focuses on the comparative evaluation of the efficiency and performance of the GPT-3.5, GPT-4, and GPT-4o models. Several stages and evaluation criteria are followed, which are detailed below.

Study Design

The study design involved the use of three ChatGPT models—GPT-3.5, GPT-4, and GPT-4o—accessed via the OpenAI API, leveraging Python to facilitate structured interactions with these models. This setup allowed for the consistent execution of representative NLP tasks such as text generation, machine translation, sentiment analysis, and others. Data collected included model responses, response times, word count, sentence count, and paragraph count, which were systematically recorded in a CSV file for analysis.

Prompts

Table 1 presents a detailed description of the eight prompts that were integral to this study. These prompts were designed to test a variety of language processing capabilities of the GPT-3.5, GPT-4, and GPT-4o models. Each prompt targets specific skills ranging from basic text generation to complex thought synthesis, offering insights into each model's proficiency in different aspects of natural language processing (NLP).

Table 1. Summary of the generated prompts

No.	Description	Prompt
1	Essay Writing	Write a 500-word essay on the importance of artificial intelligence in modern medicine. Be sure to include specific examples of applications and the benefits they bring to the diagnosis and treatment of diseases.
2	Scientific Article Abstract	Read the following excerpt from a scientific article and provide a 150-word abstract. Excerpt: "Artificial intelligence has revolutionized numerous fields, including medicine, engineering, and social sciences. In particular, deep learning models have proven to be powerful tools for analyzing large volumes of data and predicting clinical outcomes."
3	Technical Question Response	Explain what reinforcement learning is and how it is used in training autonomous agents. Provide examples of practical applications in the industry.
4	Short Story Generation	Write a 300-word short story about an explorer who discovers a lost civilization on a remote island. Describe his adventures and the challenges he faces.
5	Automatic Translation	Translate the following paragraph from English to Spanish: "Artificial intelligence and machine learning are transforming various industries by enabling automation, improving efficiency, and providing deep insights through data analysis."
6	Sentiment Analysis	Analyze the sentiment of the following text and provide a brief explanation: "Although the movie had an interesting plot, the characters were a bit flat, and the ending was predictable."
7	Code Generation	Write a Python program that takes a list of integers and returns the list sorted in ascending order. Include comments to explain each part of the code.
8	Poem Creation	Write a four-stanza poem about the beauty of nature in spring. Use descriptive language and metaphors.

Evaluation Criteria

Response Speed

The response speed of each model was evaluated by measuring the processing time required to complete the selected tasks. The times from receipt of input to generation of output were recorded. These tests were repeated twenty times to obtain a representative average and minimize the impact of possible specific

variations. Speed is a relevant factor in real-time applications, such as virtual assistants and chatbots, where users expect fast and accurate responses.

Response Quality

The quality of the responses generated by the three models analyzed was evaluated using a set of qualitative and quantitative metrics. Assessments of coherence, relevance, and accuracy of responses across various tasks were included. These evaluations were carried out by reviewing the complexity of the text generated conducted by a team of expert evaluators. Each evaluator has a background in linguistics and natural language processing tools, ensuring a high level of scrutiny and expertise in the assessment process. A total of four evaluators participated in this study, each independently reviewing the output from the ChatGPT responses to ensure objectivity.

Comparative Analysis

Finally, a comparative analysis of the collected data was conducted to identify the strengths and weaknesses of each model. Statistical methods were used to determine the significance of the differences observed between each of them.

RESULTS

Script Details

The experiment utilized a Python script tailored only by the model identifier—GPT-3.5, GPT-4, and GPT-4o—to maintain uniform evaluation conditions. This approach allowed for accurate comparisons across key metrics such as response time and text complexity, ensuring consistency in performance evaluation. Data was automatically recorded into separate CSV files for each model, enhancing efficiency and minimizing manual errors.

The script settings included a *max_tokens* limit to standardize response lengths and a *temperature* to optimally balance predictability with creativity in the outputs. Additionally, sophisticated error handling was integrated to manage API constraints effectively and ensure uninterrupted data collection, providing a reliable basis for detailed analysis.

Response Speed

GPT-4o demonstrated enhanced processing speeds across NLP tasks like text generation, machine translation, and sentiment analysis compared to GPT-4, completing tasks up to 30 % faster. For instance, in text generation, GPT-4o was 25 % quicker than GPT-4, averaging 1,2 seconds per response. In translation, GPT-4o's responses were 20 % faster, and for sentiment analysis, it improved response times by 30 %. These findings, illustrated in Figure 1, highlight GPT-4o's speed and lo wer variability in response times, making it suitable for time-sensitive applications. Despite its speed, GPT-3.5 consistently delivered faster responses, albeit with occasional slower outliers, whereas GPT-4 showed significantly varied response times, suggesting potential advantages in quality not fully explored by this study's metrics.

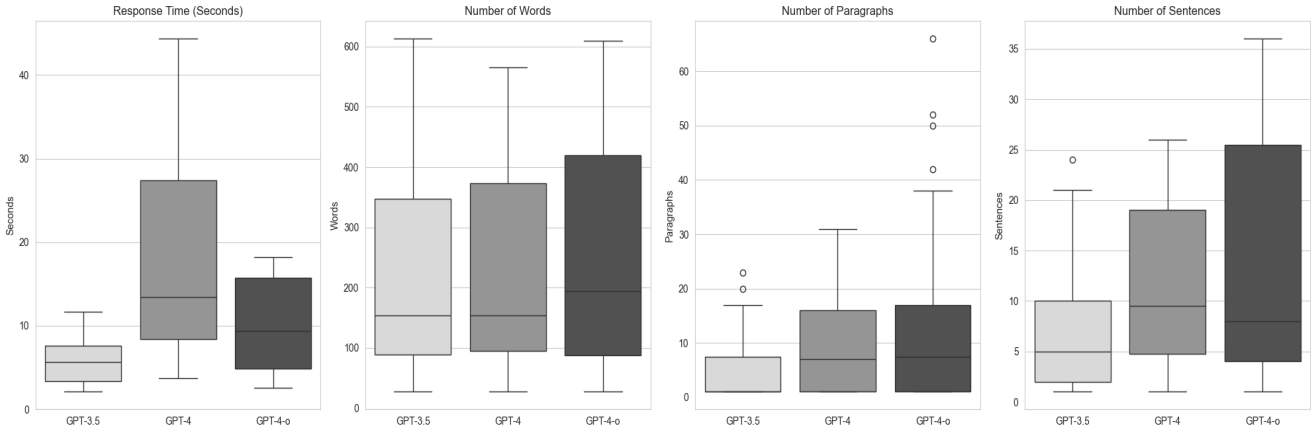

Figure 1. Box plot of response time for GPT-3.5, GPT-4 and GPT-4o models

Analysis of the number of words, sentences, and paragraphs

The box plot in figure 2 presents the analysis of the number of words, sentences, and paragraphs generated by each of the models. It was found that GPT-3.5 produces texts with a consistent and moderate number of words, with some outliers at both extremes. GPT-4 generates texts with a similar word count to GPT-3.5,

although slightly less variable and with some outliers towards the lower side. GPT-4o tends to produce longer texts on average than the other two models, with less variability and fewer outliers.

Regarding the number of paragraphs, we noticed that GPT-3.5 tends to generate texts with a consistent and low number of paragraphs, with some outliers towards the higher side. In contrast, GPT-4 produces texts with more paragraphs on average than GPT-3.5, although with similar variability and some outliers. GPT-4o is like GPT-4 but with an even more concentrated distribution around a medium number of paragraphs, indicating greater consistency.

Regarding the number of sentences, GPT-3.5 generates a consistent number of sentences. GPT-4 is like GPT-3.5 but with greater variability and more outliers. GPT-4o generates a higher number of sentences on average than the other two models, with less variability and fewer outliers.

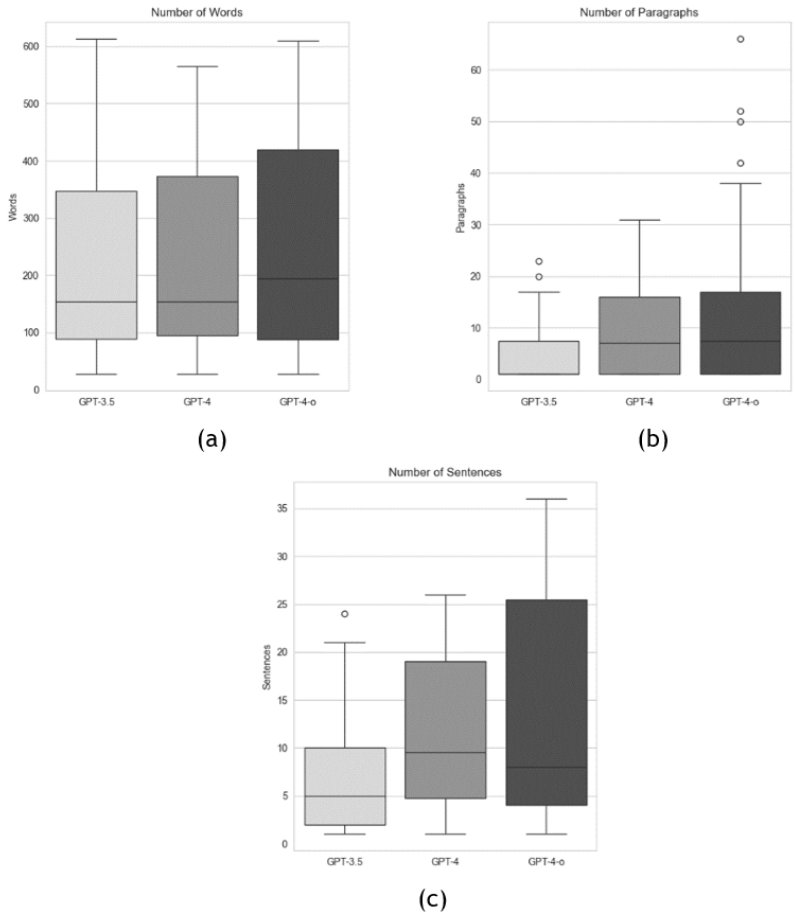

Figure 2. Box plot of number of (a) words, (b) paragraphs and (c) sentences generated by GPT-3.5, GPT-4 and GPT-4o models

Quality of Responses

The quality of the responses generated by the three models was evaluated using both qualitative and quantitative metrics, focusing on aspects such as relevance and accuracy. The evaluation encompassed seven distinct features (table 2), each rated on a scale from 0 to 10 using a rubric designed to ensure objective and consistent scoring. These parameters included clarity, coherence, grammar, and other critical characteristics that contribute to the overall effectiveness of the responses.

Table 2. Parameters for human evaluation

No.	Evaluated features	Assessment
1	Clarity coherence	0-10
2	Development of ideas	
3	Organization structure	
4	Grammar spelling	
5	Originality creativity	
6	Relevance content	
7	Use of sources evidence	

https://doi.org/10.56294/dm2024.359

Comparative Analysis

In this study, the responses of the GPT-3, GPT-4, and GPT-4o language models were evaluated in eight different tasks, with the objective of determining their performance in generating coherent and accurate content. The scores were calculated from 20 repetitions of each task for each model, obtaining specific means and standard deviations for each prompt, as seen in table 3. The results show that GPT-3 had consistent performance, with means ranging between 8,02 and 8,11 and relatively low standard deviations, indicating stable performance.

Table 3. Scores for each prompt with human evaluation

Prompt	GPT-3 Mean±SD	GPT-4 Mean±SD	GPT-4o Mean±SD
1	8,07±0,125	9,52±0,0573	9,10±0,0600
2	8,02±0,0733	9,53±0,0571	9,12±0,0566
3	8,09±0,0809	9,54±0,0535	9,14±0,0567
4	8,11±0,0700	9,55±0,0550	9,14±0,0550
5	8,07±0,0574	9,55±0,0550	9,14±0,0550
6	8,02±0,0548	9,54±0,0548	9,15±0,0548
7	8,10±0,0678	9,55±0,0550	9,15±0,0548
8	8,04±0,0394	9,55±0,0550	9,15±0,0548

On the other hand, GPT-4 proved to be the most robust model, with means ranging from 9,52 to 9,55 and even lower standard deviations, reflecting superior consistency and precision compared to GPT-3. GPT-4o, while also showing strong performance, ranked slightly below GPT-4, with means ranging from 9,10 to 9,15 and similar standard deviations. This suggests that, although GPT-4o is very effective, GPT-4 is still the leading model in terms of overall performance on the tasks tested.

The analysis of variance (ANOVA) performed to compare the scores of the three models revealed significant differences in performance ($F(2, 477) = 26303$, $p < 2e\text{-}16$). This highly significant result indicates that the differences observed between the models are not a product of chance and that there are real discrepancies in their abilities to handle the assigned tasks. Post-hoc tests confirmed that there are statistically significant differences between the GPT-4 and GPT-3.5 models, but none existed between the GPT-4 and GPT-4o models.

ANOVA Analysis

The Tukey test is used to make multiple comparisons between groups and determine which of them differ significantly from each other. In Table 4, the Tukey test results for comparisons between the models:

Table 4. ANOVA Analysis between models

Comparison	Mean Difference	Standard error	z value	p-value	Difference Significant
GPT-4 vs GPT-3	1,47	0,0138	106,52	< 0,001	Yes
GPT-4o vs GPT-3	1,07	0,0138	77,54	< 0,001	Yes
GPT-4 vs GPT-4o	0,40	0,0138	28,99	< 0,001	Yes

Analysis of variance (ANOVA) showed a significant difference in scores between models ($F(2, 477) = 26303$, $p < 2e^{-16}$). This result indicates that there are significant differences in the performance of the GPT-3, GPT-4, and GPT-4o models. The mean difference between GPT-4 and GPT-3 is 1,47, which is statistically significant ($p < 0,001$). This indicates that GPT-4 performs significantly better than GPT-3 on the tested tasks. The difference in means between GPT-4o and GPT-3 is 1,07, also statistically significant ($p < 0,001$). This shows that GPT-4o also performs better than GPT-3, although the improvement is smaller compared to GPT-4. The mean difference between GPT-4 and GPT-4o is 0,40, which is significant ($p < 0,001$). This suggests that GPT-4 performs better than GPT-4o, but the difference is minor compared to the difference between GPT-4 and GPT-3.

DISCUSSION

Previous literature has shown that GPT-4 significantly outperforms GPT-3.5 and Bard. This was demonstrated in [20], where there was greater accuracy in answering NEET-2023 questions. This underlines the importance of carefully selecting the models used in education. In our study, the optimized version of GPT-4 showed the best results in terms of accuracy, establishing a benchmark for evaluating and improving the performance of LLMs in educational tasks, thus promoting their responsible and informed use in various learning environments.

Similarly, previous studies have identified that GPT-4 had higher accuracy in responses regarding the detection of cancer cells compared to version 3.5.[16] A better presentation of understandable content for the public without technical knowledge was also observed.[19] However, both versions showed instability in the responses. Our study, which included GPT-4o, found that iterative optimization of the model allowed responses to be stabilized and accuracy improved. This approach has the potential to optimize responses generated by artificial intelligence for the benefit of science.

The analysis of the number of words, sentences, and paragraphs generated by each model revealed notable differences. Compared to GPT-3.5 and GPT-4, GPT-4o typically yields lengthier texts with lower variability and fewer outliers. This consistency in text length and organization suggests that GPT-4o can provide more thorough and organized responses, which could be advantageous for applications needing extensive text output. These results are in line with earlier research that demonstrate GPT-4's sophisticated ability to produce well-written, detailed texts,[23] as well as Nakajima et al.[24] which show GPT-4 performs better when producing responses of a high quality in certain areas of study.

It is important to recognize the limitations when technology imitates human reasoning. The use of AI can lead to inconsistent or fabricated academic responses, suggesting the need for improvements in its implementation for future assessments.[21] There is also a risk of latent bias, highlighting the need for appropriate policies for the integration of LLM tools.[18] Other limitations are related to the consistency of results and the application of advanced statistical methods.[17] Therefore, the use of these tools is recommended as support for researchers with intermediate experience in data analysis, which can reduce operational barriers and improve the interpretation of the results of epidemiological analysis.

CONCLUSIONS

Compared to GPT-4 and GPT-3.5, GPT-4o improves response speed, enabling more quickly and reliably completed tasks such text generation, translation, and sentiment analysis. Because of this, GPT-4o is particularly suitable for time-sensitive applications. The use of the OpenAI API was instrumental in ensuring accurate, uniform task execution and reducing manual errors.

GPT-4o also offers advanced functionalities like file handling and audio processing. However, this study maintained a focus on basic text processing features to align with the capabilities of GPT-3.5 and GPT-4, facilitating a direct comparison. While GPT-3.5 is adept at delivering quick, reliable responses, GPT-4 might provide enhanced quality, a potential not fully explored within this study's scope. Future research should integrate automated evaluation metrics like BLEU, ROUGE, and BERT scores to complement human reviews and provide a more nuanced understanding of model performance.

REFERENCES

1. Rama Krishna S, Rathor K, Ranga J, Soni A, Srinivas D, Anil Kumar N. Artificial Intelligence Integrated with Big Data Analytics for Enhanced Marketing. 6th Int. Conf. Inven. Comput. Technol. ICICT 2023 - Proc., Institute of Electrical and Electronics Engineers Inc.; 2023, p. 1073–7. https://doi.org/10.1109/ICICT57646.2023.10134043.

2. Ayala-Chauvin M, Avilés-Castillo F, Buele J. Exploring the Landscape of Data Analysis: A Review of Its Application and Impact in Ecuador. Computers 2023;12. https://doi.org/10.3390/computers12070146.

3. Jan Z, Ahamed F, Mayer W, Patel N, Grossmann G, Stumptner M, et al. Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. Expert Syst Appl 2023;216:119456. https://doi.org/10.1016/j.eswa.2022.119456.

4. Hwang MH, Lee GS, Kim E, Kim HW, Yoon S, Talluri T, et al. Regenerative Braking Control Strategy Based on AI Algorithm to Improve Driving Comfort of Autonomous Vehicles. Appl Sci Switz 2023;13:946. https://doi.org/10.3390/app13020946.

5. Castillo-González W, Lepez CO, Bonardi MC. Chat GPT: a promising tool for academic editing. Data Metadata 2022;1. https://doi.org/10.56294/dm202223.

6. Johri P, Khatri SK, Al-Taani AT, Sabharwal M, Suvanov S, Kumar A. Natural Language Processing: History, Evolution, Application, and Future Work. Lect. Notes Netw. Syst., vol. 167, Springer, Singapore; 2021, p. 365–75. https://doi.org/10.1007/978-981-15-9712-1_31.

7. Tang R, Chuang YN, Hu X. The Science of Detecting LLM-Generated Text. Commun ACM 2024;67:50–9. https://doi.org/10.1145/3624725.

8. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can Large Language Models Transform Computational Social Science? Comput Linguist 2024;50:1–55. https://doi.org/10.1162/coli_a_00502.

9. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, et al. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access 2024;12:26839–74. https://doi.org/10.1109/ACCESS.2024.3365742.

10. Zhang M, Li J. A commentary of GPT-3 in MIT Technology Review 2021. Fundam Res 2021;1:831–3. https://doi.org/10.1016/j.fmre.2021.11.011.

11. Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Nat Lang Process J 2024;6:100048. https://doi.org/10.1016/j.nlp.2023.100048.

12. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report 2023.

13. Egli A. ChatGPT, GPT-4, and Other Large Language Models: The Next Revolution for Clinical Microbiology? Clin Infect Dis 2023;77:1322–8. https://doi.org/10.1093/cid/ciad407.

14. Hatakeyama-Sato K, Yamane N, Igarashi Y, Nabae Y, Hayakawa T. Prompt engineering of GPT-4 for chemical research: what can/cannot be done? Sci Technol Adv Mater Methods 2023;3. https://doi.org/10.1080/27660400.2023.2260300.

15. OpenAI. Hello GPT-4o n.d. https://openai.com/index/hello-gpt-4o/ (accessed May 21, 2024).

16. Liang R, Zhao A, Peng L, Xu X, Zhong J, Wu F, et al. Enhanced Artificial Intelligence Strategies in Renal Oncology: Iterative Optimization and Comparative Analysis of GPT 3.5 Versus 4.0. Ann Surg Oncol 2024;31:3887–93. https://doi.org/10.1245/s10434-024-15107-0.

17. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. J Glob Health 2024;14:04070. https://doi.org/10.7189/jogh.14.04070.

18. Gmyrek P, Lutz C, Newlands G. A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK. SSRN Electron J 2024. https://doi.org/10.2139/ssrn.4700366.

19. Srinivasan N, Samaan JS, Rajeev ND, Kanu MU, Yeo YH, Samakar K. Large language models and bariatric surgery patient education: a comparative readability analysis of GPT-3.5, GPT-4, Bard, and online institutional resources. Surg Endosc 2024;38:2522–32. https://doi.org/10.1007/s00464-024-10720-2.

20. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DØ. Evaluating Large Language Models for the National Premedical Exam in India: Comparative Analysis of GPT-3.5, GPT-4, and Bard. JMIR Med Educ 2024;10:e51523. https://doi.org/10.2196/51523.

21. Stribling D, Xia Y, Amer MK, Graim KS, Mulligan CJ, Renne R. The model student: GPT-4 performance on graduate biomedical science exams. Sci Rep 2024;14:1–11. https://doi.org/10.1038/s41598-024-55568-7.

22. GPT-4 System Card OpenAI 2023.

23. Rahaman MS, Ahsan MMT, Anjum N, Terano HJR, Rahman MM. From ChatGPT-3 to GPT-4: A Significant Advancement in AI-Driven NLP Tools | Journal of Engineering and Emerging Technologies 2023.

24. Nakajima N, Fujimori T, Furuya M, Kanie Y, Imai H, Kita K, et al. A Comparison Between GPT-3.5, GPT-4, and GPT-4V: Can the Large Language Model (ChatGPT) Pass the Japanese Board of Orthopaedic Surgery Examination? Cureus 2024;16. https://doi.org/10.7759/cureus.56402.

FINANCING

This research was supported by the Universidad Indoamérica under project No. 295.244.2022, entitled "Big data análisis y su impacto en la sociedad, educación e industria".

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Ignacio Ayala-Chauvin.
Data curation: Fátima Avilés-Castillo.
Formal analysis: Ignacio Ayala-Chauvin.
Acquisition of funds: Ignacio Ayala-Chauvin.
Research: Fátima Avilés-Castillo, Ignacio Ayala-Chauvin.
Methodology: Ignacio Ayala-Chauvin.
Project management: Ignacio Ayala-Chauvin.
Resources: Ignacio Ayala-Chauvin.
Software: Fátima Avilés-Castillo.
Supervision: Ignacio Ayala-Chauvin.
Validation: Ignacio Ayala-Chauvin.
Display: Fátima Avilés-Castillo.
Drafting - original draft: Fátima Avilés-Castillo.
Writing - proofreading and editing: Ignacio Ayala-Chauvin.

https://doi.org/10.56294/dm2024.359