InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol. 3 (No.1), 74-90 (2026)
DOI: https://doi.org/10.51173/ijds.v3i1.60
https://ijds.mtu.edu.iq

CHATGPT: PRECISION ANSWER COMPARISON AND EVALUATION MODEL

Aso M. Aladdin [1], Rebwar Khalid Muhammed [2*], Hemin Sardar Abdulla [1], Tarik A. Rashid [3]

[1] *Computer Science Department, College of Science, Charmo University, Sulaimani, Chamchamal 46023, KR, Iraq.*
[2] *Network Department, Computer Science Institute, Sulaimani Polytechnic University, Sulaimani 46001, KR, Iraq.*
[3] *Computer Science and Engineering Department, University of Kurdistan Hewler, Erbil 44001, KR, Iraq.*

** Corresponding author E-mail: rebwar.khalid@spu.edu.iq (Rebwar Khalid Mohammed)*

RESEARCH ARTICLE

ARTICLE INFORMATION

SUBMISSION HISTORY:
Received: 8 September 2025
Revised: 13 December 2025
Accepted: 10 January 2026
Published: 30 January 2026

KEYWORDS:

ChatGPT;
PACEM;
Answer Comparison;
Human-Like Responses;

ABSTRACT

Artificial Intelligence (AI) has made advancements, among other things, OpenAI created the sophisticated model ChatGPT. Conversational, ChatGPT supports natural interactions, providing human-like responses to queries across myriad topics. But it is not infallible, and the degree of accuracy also depends on the complexity of the queries, the context, and how often the prompts are repeated. This work thus proposes a new model, the Precision Answer Comparison and Evaluation Model (PACEM), to systematically address these types of questions and assess ChatGPT's performance. PACEM assesses the correctness and coherence of ChatGPT's answers across numerous fields, including literature, history, law, ethics, and sports. By providing these analyses and comparisons, PACEM goes on record with a detailed understanding of what ChatGPT does well and poorly as a source of reliable information. On top of that, it includes an assessment of response time, considering ChatGPT's speed in producing answers in relation to real or expected ones. The findings show that ChatGPT's answers are usually substantially accurate and often of superior quality compared to those written by the user and other alternatives. Response time generally increases with the complexity or length of the answer. Finally, the study reviews notable takeaways from PACEM's deployment and offers suggestions for future research to address the evolving challenges in AI-driven response assessment.

1. INTRODUCTION

AI is a rapidly growing technology that allows machines to perform tasks that previously required human intelligence. As machine learning and optimisation have improved, modern AI systems can support complex decision-making across multiple domains, including medicine, automation, and data analysis [1], [2]. AI is a rapidly growing field of computer science concerned with developing systems capable of performing tasks that traditionally require human intelligence. AI technologies are providing numerous applications, including medical diagnosis and autonomous vehicle navigation. Plus, AI is emerging alongside new frontier technologies, thereby making its impact cross scientific and industrial domains even more pronounced [3], [4]. Machine learning and hyper-learning algorithms have substantially advanced AI by enhancing its ability to learn from data and make increasingly sophisticated decisions [5], [6]. These algorithms have played a pivotal role in advancing the field by enhancing AI systems' ability to learn from data and perform increasingly sophisticated decision-making tasks. In parallel, optimisation techniques have further advanced AI by improving performance, efficiency, and applicability across real-world problem domains. Collectively, these advancements have strengthened AI systems' capacity to address complex computational challenges with greater accuracy and effectiveness [7], [8], [9]. As a result, the field has progressed toward producing and improving meta-heuristic optimization and hyper-learning algorithms that improve efficiency, scale and performance in complex problem domains. The first

DOI: https://doi.org/10.51173/ijds.v3i1.60

and most widespread use of AI tools was in traditional web search engines and automated site navigation. This meant attention to detail in query formulation, source comparison, and the critical review of content relevance and credibility. It was tedious and highly cognitively intensive; users had to synthesize fragments of information gathered from multiple web pages without the automated contextual support provided by AI-driven systems [10]. Following these advances, OpenAI developed ChatGPT, a large-scale conversational language model designed to generate human-like responses. OpenAI was founded in 2015 and aims to do this safely and economically. In line with this vision, on 30 November 2022, the ChatGPT model was introduced, creating a new level of natural language exchange through extensive training on large, diverse linguistic datasets [1]. ChatGPT is a large-scale language model trained on extensive multilingual conversational datasets to generate coherent and contextually appropriate natural-language responses. It supports a wide range of applications, including conversational agents, machine translation, text summarization, and question answering [12]. Beyond general-purpose generation, ChatGPT can be fine-tuned for specialized downstream tasks such as automated content generation, document abstraction, and creative writing. Owing to its flexible architecture, the model can be integrated across diverse domains, including customer service, education, and information retrieval. Given its deployment in sensitive contexts, the system requires robust encryption and security mechanisms to protect sensitive data [13], [14], [15].

Another important direction for ChatGPT involves personalization, where the model adapts its outputs based on individual user characteristics. Through repeated interactions, ChatGPT can capture linguistic patterns, such as user-specific vocabulary, tone, and stylistic preferences, enabling more contextually appropriate responses. Personalization has practical value in domains such as customer service and education, where tailoring content to user needs can enhance clarity and relevance. Furthermore, the data generated from user interactions can support the development of adaptive language models, improving their alignment with user requirements while maintaining responsible and secure data-handling practices [16], [17]. However, a notable limitation is that ChatGPT does not consistently produce reliable responses. The model may also generate different answers to the same question depending on variations in timing, phrasing, subject matter, or contextual cues, which can introduce ambiguity for users. To address these limitations, the present study employs the PACEM, a framework designed to systematically assess the semantic accuracy and consistency of ChatGPT's responses. Given the rapid evolution of large language models, a rigorous, transparent evaluation approach is essential to understanding their reliability across different contexts. Although ChatGPT continues to undergo frequent updates and improvements, variability in its responses remains an important concern. Therefore, this study aims to evaluate the correctness and stability of ChatGPT's outputs using a diverse set of questions representing both historical and contemporary topics. In doing so, the study contributes to the literature by examining several key dimensions of model performance, as outlined below. This study introduces the PACEM, a set of methods that assesses ChatGPT by using semantic similarity rather than keyword matching or subjective judgment. Unlike other evaluation approaches, PACEM includes repeated-question stability analysis, response-time measurement and paired statistical significance testing. This framework provides a repeatable, quantitatively grounded assessment of large language model performance that enables cross-domain comparison within a single methodology. Additionally, this study contributes to the field by addressing the following key points:

- ChatGPT responds to questions from literature, history, law, and sports. We examine its responses in several fields to determine if the model is accurate and consistent using the Precision Answer Comparison and Evaluation Model (PACEM).
- The PACEM framework guarantees an inclusive evaluation of ChatGPT's strengths and weaknesses in providing trustworthy information on various topics.
- PACEM compares ChatGPT answers to the correct ones and to its own initial responses, examining them in isolation.
- Ask ChatGPT the same question several times and compare the responses to evaluate its accuracy.

- Find out how long each category has been elapsed time in PACEM, and in which ChatGPT does better in the evaluation.

The remainder of the study is organized as follows: Section 2 reviews related works. Section 3 focuses on the methodology and details the evaluation of the problem statements. Section 4 presents the results in tables and charts, and Section 5 discusses these findings. The final section concludes the study, outlines future work, and highlights the necessary appendix tables.

2. LITERATURE REVIEW

ChatGPT's future aims to continuously refine its language models through better training algorithms and larger datasets. These models are still prone to inaccuracies but are better able to understand and answer complex questions as they are exposed to more data [18]. That could open the door to developing entirely new and innovative uses in operating system platforms, healthcare, sport, literature and finance [19], [20], [21], where processing and understanding vast amounts of data is fundamental. Although addressing different technological domains, they collectively illustrate parallel trajectories in the evolution of technological systems. Each demonstrates how technologies progressively become more capable, user-friendly, and widely adopted. Moreover, both the OS-platform literature and the AI/ChatGPT studies converge on key evaluative dimensions, including human technology interaction, usability, efficiency, reliability, and the limitations influencing user adoption. Plus, the evolution of programming technology in creative writing and gaming comprises an exciting dimension that would profoundly change human-chaological interaction [22].

Table 1. Summary of selected references providing context and clarity for the results

No.	Title of Study	Varieties of Data Employed	Published year	Ref.
1	Evaluating the Accuracy, Comprehensiveness, and Validity of ChatGPT Compared to Evidence-Based Sources Regarding Common Surgical Conditions: Surgeons	Common surgical conditions	2023	**[23]**
2	Evaluating the efficacy of ChatGPT in addressing patient queries about acne and atopic dermatitis	Acne, atopic dermatitis	2023	**[24]**
3	ChatGPT for Tinnitus Information and Support: Response Accuracy and Retest after Three and Six Months	Tinnitus	2023	**[25]**
4	Evaluating ChatGPT -4's historical accuracy: a case study on the origins of SWOT analysis	Historical details	2023	**[26]**
5	AI in practice: measuring its medical accuracy in oculoplastic consultations	Oculoplastic consultations	2024	**[27]**
6	Evaluating ChatGPT's Accuracy in Providing Screening Mammography Recommendations among Older Women: AI and Cancer Communication	Screening mammography recommendations	2024	**[28]**
7	Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument	Healthcare information	2023	**[29]**
8	Evaluating AI Language Models in News Retrieval: A Comparative Study of ChatGPT-Plus and DeepSeek	News articles	2025	**[30]**

Some examine its performance across several fields and assess the accuracy of the information it provides in both medical and non-medical areas. Studies have shown that ChatGPT typically provides accurate and comprehensive responses, particularly for common surgical conditions [23], and it is also effective for acne and atopic dermatitis [24]. It is not easy to read, lacks information, and is also structured in a way that might omit newer treatments. In addition to conceptual studies on the potential of ChatGPT for specialised fields, such as tinnitus, reviews in these areas conclude

that this tool also exhibits inaccuracies and requires further refinement in the hands of experts in those fields [25].In historical contexts, such as tracing the origins and evolution of SWOT analysis, ChatGPT-4 shows proficiency in general concepts but exhibits discrepancies and hallucinations in specific historical details [26]. An additional research paper evaluates the medical accuracy of AI models, specifically ChatGPT-4 and DALL·E, in oculoplastic consultations [27]. Another research paper assesses ChatGPT's accuracy in providing screening mammography recommendations for older women, specifically those aged 75 years and older [28]. Another focus of the research papers assessing the reliability of medical information provided by ChatGPT-4, an AI chatbot, in the context of healthcare [29]. The study evaluates model performance in real information-seeking scenarios using real-world articles and retrieval queries. Additionally, it uses comparison-answer evaluation and relevance-judgment datasets to assess retrieval quality, accuracy, and contextual precision [30].

In evaluating the results of the new study, continue to take a closer look at the references that are discussed in Table 1. Together, these results indicate that although ChatGPT can offer useful information, continuous improvement and professional validation are necessary to improve its overall accuracy and reliability across a variety of fields.

3. METHODS

The PACEM study evaluates the semantic accuracy of text-generation models based on the degree to which they are meaningfully related to the correct answers. While GPT-2 is cited as a starting point to illustrate the conceptual foundations of the PACEM structure, this experimental study was conducted with ChatGPT-4.0 to ensure it is appropriate for contemporary language models. GPT-4o was used to generate responses, identify flaws, and identify gaps that could be improved and used in the future to refine the model further. The assessment process includes setup, model initialization, response generation, semantic similarity calculation, and accurate measurement. These steps are repeated over multiple iterations to ensure the strength and reliability of the results.

3.1. Dataset Preparation

The dataset used in this study contains prompts and their corresponding correct answers. The dataset is loaded into a panda DataFrame for further processing. The columns are structured as follows:

- Prompt: The input text or question provided to the model.
- Correct Answer: The expected response to the prompt.

Let the D dataset be represented in eq.1, where P_i is the prompt, C_i is the correct answer, and n is the total number of samples.

$$D = \{(P_i, C_i)\}_{i=1}^{n} \quad \dots (1)$$

3.2. Model Initialization

Two models are initialized:

- Text Generation Model: A pre-trained model (GPT-4o) s used to generate responses.
- Sentence Transformer Model: A pre-trained sentence transformer (paraphrase-MiniLM-L6-v2) is initialized using the sentence-transformers library to encode both the generated response and the correct answer into vector representations.

Let M_{gen}represent the text generation model and M_{sim}represent the sentence transformer model.

3.3 Response Generation and Similarity Calculation

For each prompt P_i in the dataset D , the text generation model M_{gen}produces a response R_i. The similarity between $R_i$and the correct answer C_i is then computed using cosine similarity, as defined in eq.2.

$$Cosine\_similarity(A,B) = \frac{A.B}{||A||||B||} \quad \text{... (2)}$$

Where A and B are vector representations of the sentences (generated response and correct answer, respectively). The steps involved are:

- Encode $C_i$and R_i using the sentence transformer model $M_{sim}$to obtain their vector representations VC_i and VR_i
- Let the cosine similarity between VC_i and VR_i be denoted as eq.3.

$$S_i = \frac{VC_i \cdot VR_i}{||VC_i||||VR_i||} \quad \text{... (3)}$$

A response R_i is considered correct if the similarity score S_i exceeds a predefined threshold T. In this study, the threshold is set to 0.6, as shown in eq. 4.

$$Correct(R_i) = \begin{cases} 1 & if\ S_i > T \\ 0 & if\ S_i \leq T \end{cases} \quad \text{... (4)}$$

In this study, the similarity threshold was set to T = 0.6 to determine whether a generated response is considered semantically correct. This value was selected based on empirical observations and prior semantic-similarity evaluation practices, where cosine similarity scores above 0.6 generally indicate strong semantic alignment rather than superficial lexical overlap. A lower threshold may incorrectly classify partially related or vague responses as correct, whereas a higher threshold (e.g., ≥ 0.75) may penalize valid paraphrases and conceptually accurate answers expressed using different linguistic structures. Therefore, T = 0.6 provides a balanced trade-off between precision and recall, ensuring robust and fair evaluation across diverse question domains.

3.4. Accuracy Computation

The accuracy of one iteration is defined as the ratio of correct responses to the total number of prompts in eq. 5, where the denominator is the number of prompts in the dataset.

$$Accuracy = \frac{\sum_{i=1}^{n} Correct(R_i)}{n} \times 100 \quad \text{... (5)}$$

3.5. Iterative Evaluation

The evaluation is then repeated 30 times to ensure the reliability. These metrics are recorded for each iteration:

- Accuracy: The percentage of correct responses.
- Elapsed Time: The time taken to complete the iteration.

The final accuracy is computed as the average accuracy across all iterations in eq.6, where m is the number of iterations.

$$Final\ Accuracy = \frac{1}{m}\sum_{j=1}^{m} Accuracy_j \quad \text{... (6)}$$

3.6. Statistical Analysis and Significance Testing

To enhance the statistical validity of the descriptive analysis and the inferred statistical validity, additional analyses were conducted on the accuracy results across question categories. The mean accuracy and sample standard deviation were calculated for each category to represent central tendency and variability, respectively. Let $x_i$denote the accuracy obtained for the i-th question template and $n$denote the total number of templates. The mean accuracy was calculated as shown in Eq. (7):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{... (7)}$$

The corresponding sample standard deviation (STD) was computed as shown in Eq. (8):

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad \ldots (8)$$

To assess whether performance differences between ChatGPT and actual answers were statistically significant, a paired t-test was employed, as both methods were evaluated on the same matched set of question templates. Let $x_{i,c}$and $x_{i,a}$denote the accuracies of ChatGPT and actual answers, respectively, for the i-th template. The paired differences were defined as $d_i = x_{i,c} - x_{i,a}$. The t-statistic was computed as shown in Eq. (9):

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \qquad \ldots (9)$$

Where $\bar{d}$ denotes the mean of the paired differences and $s_d$represents their standard deviation. The associated two-tailed p-value was obtained from the student's t-distribution with $n - 1$degrees of freedom. A significance level of $\alpha = 0.05$was adopted.

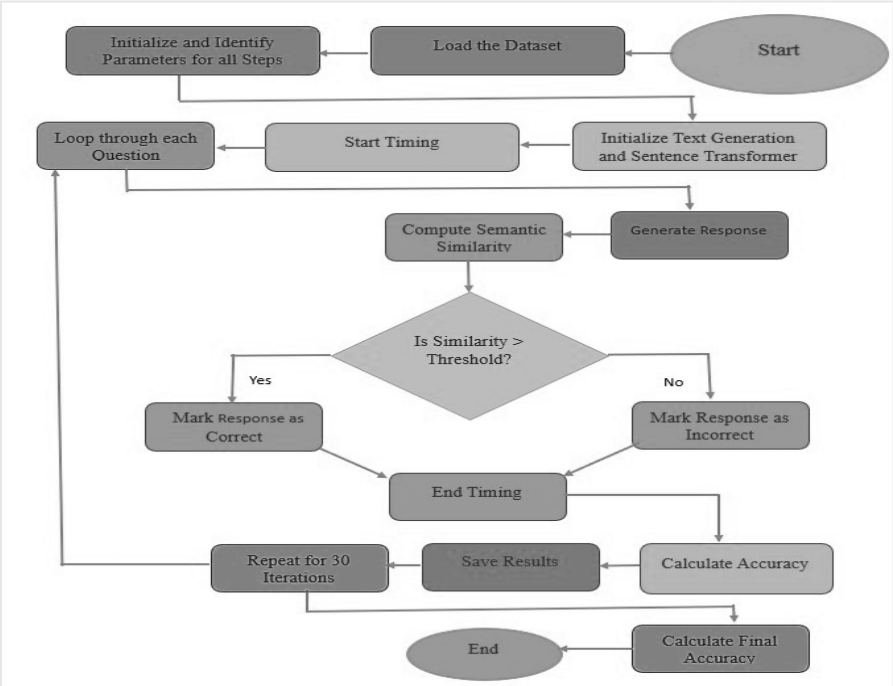

Figure 1. Evaluating Text Generation Model Accuracy Using Semantic Similarity

3.7. Saving Results

After the evaluation, the results are saved to two CSV files:

- accuracy_time_results.csv: Contains the accuracy and elapsed time for each iteration.
- generated_responses.csv: Contains the generated responses, correct answers, and their similarity scores for further analysis.

Also, as explained, this makes it a better assessment of the text generation model's strengths and weaknesses. As shown in Fig. 1, the evaluation begins by establishing the parameters at each step, which are maintained throughout the process. The set of questions and answers is then imported into the system. In the loop that iterates over each question, the "$start" time is logged to measure response generation speed. ChatGPT is used as a text-generation pipeline to produce responses to questions. These are then used to compare responses against the correct ones using a sentence-transformer model to measure semantic similarity.

The calculated similarity score is then compared to a predefined cut-off score. A response is flagged as correct if the score is above the cut-off, while a response is flagged as incorrect if the score is below the cut-off. The timing is also stopped, and the procedure is repeated 30 times to ensure

consistent results. The results from each iteration are then saved for analysis. Finally, it computes accuracy from the correct responses and, after all iterations, computes the final accuracy. Such a holistic approach enables the assessment of how well the text-generation model performs. As shown in Fig. 2, the algorithm estimates the accuracy of a text generation model by comparing its generated responses to the correct answer using semantic similarity. The approach is iterative, ensuring accurate solutions are retained and saved for the overall analysis.

1. **Initialize Resources:**
   - Load the dataset containing question and correct answers.
   - Initialize the text generation pipeline (e.g., GPT-2).
   - Initialize the sentence transformer model for computing semantic similarity.
2. **Set Parameters:**
   - Set the similarity threshold (e.g., 0.6) for marking responses as correct.
3. **Start Evaluation:**
   - Initialize variables to track total correct answers, total questions, and time.
4. **Iterate Over Dataset:**
   - For each question in the dataset:
     - Generate a response using the text generation model.
     - Compute the semantic similarity between the generated response and the correct answer using the sentence transformer model.
     - If the similarity score exceeds the threshold, mark the response as correct.
     - Record the generated response, similarity score, and correctness.
5. **Repeat for Multiple Iterations:**
   - Repeat the evaluation process for a predefined number of iterations (e.g., 30 times).
   - Accumulate the results, including accuracy and timing for each iteration.
6. **Calculate Final Accuracy:**
   - After all iterations, compute the final accuracy as the percentage of correct responses across all iterations.
7. **Save Results:**
   - Save the accuracy, elapsed time, and generated responses for each iteration to CSV files for further analysis.
8. **End Process:**
   - Print the final accuracy and confirm that results have been saved successfully.

Figure 2. Algorithm for Iterative Evaluation of Text Generation Accuracy

4. RESULTS

Initially, ChatGPT's conversational response accuracy was illegible because the test focused on the detection module in the newly updated model. Several subjects were tested on performance, including comparing ChatGPT's answers to the correct answers and measuring response time. Possible biases in the data construction process may impact the evaluation of ChatGPT's performance. These biases can affect outcomes, such as imbalanced topic coverage, difficult questions, or inconsistent human annotations. For instance, a dataset dominated by structured, rule-based questions, such as Legal and Ethical, can be accidentally favoured by ChatGPT models. In contrast, open-ended or rapidly changing questions, such as Sports, are more problematic. These limitations are important to understand because they can influence how model accuracy and generalizability are interpreted. The assessment included categories ranging from literature questions to history, law and ethics, and sports. Results included accuracy and the time elapsed in ChatGPT's responses, measured using a Python library. The results, gathered by asking and answering one question at a time, were obtained by repeating each question 30 times so that each answer could be compared with the correct one. These questions and answers are in the appendix.

4.1 Literature Template Problem Evaluation and Result Accuracy

Here, we analyse ChatGPT's answers against established models and the evaluations reported in the literature. This is based on the questions/answers provided in Appendix Table 6. Template 1 includes 10 questions, chosen at random as mentioned in the methodology, that address both older information and current themes pertinent to the literature subjects. Template 2 consists of another set of questions and answers on the same topic to further challenge ChatGPT, which are listed in Appendix Table 7. Templates 3 and 4 originated from Templates 1 and 2, respectively, and were

created to test ChatGPT's consistency and rigorously measure response confusion. More specifically, Template 3 asks the first five questions from Template 1 and the last five questions from Template 2. Template 4, in contrast with Template 3, is intended to be its reverse; questions are asked in the opposite order. These supplementary templates aim to test ChatGPT's ability to handle fast-paced, diverse queries and to identify potential inconsistencies in its replies. The obtained accuracy was format-dependent and also dependent on the quality of the responses. The performance of ChatGPT's answers was then analysed in detail against the researchers' answers, using factual data. ChatGPT was also more accurate than the researchers in all question types, as seen in Table 2.

On top of that, producing quality answers was more time-consuming, as it required additional processing, especially when analysing ChatGPT's responses in relation to the researchers' answers. As shown in Table 2, ChatGPT outperformed human responses in terms of accuracy across all four templates. For example, Template 1 achieves an accuracy of 72.33%, which is a significant increase compared to the 20.67% correct responses from real answers. The same goes for the rest of the templates: Template 2 shows 38.67% accuracy for ChatGPT and 12.67% for humans; Template 3 shows 49.00% for ChatGPT versus 20.00% for human answers; and Template 4 shows 68.33% accuracy for ChatGPT and 11.67% for humans. These results indicate ChatGPT's aptitude for producing accurate, precise responses, confirming its ability to comprehend and articulate subjects. A second study assessed processing time and showed a speed-accuracy trade-off. The downside of ChatGPT is that it is more accurate but takes longer to process. Template 1, which achieved the highest accuracy of 72.33, took 1155.231 seconds, or 38.51 seconds per response.

In comparison, human responses had a total response time of 507.37 seconds, averaging 16.91 seconds per response. Template 3, with a moderate accuracy of 49.00%, took an average of 1234.22 seconds, 41.14 seconds per response. This means that ChatGPT has the potential to generate more correct answers than humans, but at the cost of longer processing time.

Table 2. ChatGPT's performance in answering questions on literature-related subjects.

Literature Question Templates	Results Compared to Actual Answers			Results Compared to ChatGPT Answers		
	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)
Template_1	20.67%	507.37	16.91	72.33%	1155.231	38.51
Template_2	12.67%	547.14	18.24	38.67%	1224.18	40.81
Template_3	20.00%	560.71	18.69	49.00%	1234.22	41.14
Template_4	11.67%	577.07	19.24	68.33%	1332.274	44.41

4.2 Problem Analysis and Result Accuracy of History Templates

In this section, we evaluate the accuracy of ChatGPT's responses against the pre-established models and assessments located in history. This assessment is taken from the questions and answers in Appendix Table 8. Template 1 consists of 10 randomly selected questions, as previously discussed. These questions use both historical material and current issues relevant to the topics in history. Template 2 contains a unique set of questions and answers on the same topic. It is intended to serve as a supplemental analysis of ChatGPT's, detailed in Appendix Table 9. To evaluate ChatGPT's consistency and potential confusion, Templates 3 and 4 were generated from Template 1 and Template 2, respectively, as mentioned in the previous subsection on literature template evaluation. These additional templates aim to assess ChatGPT's ability to handle fast, diverse questions and to identify potential contradictions in its answers.

Nevertheless, the study found that accuracy varied according to the question structure and the overall quality of the answers. An extensive analysis was conducted to compare the precision of results obtained from ChatGPT with that of results derived from researchers' responses based on confirmed information. Table 3 demonstrates that ChatGPT consistently achieved higher accuracy than the experimenters across all question structures. Furthermore, producing responses of

superior quality required additional time for processing, especially when assessing ChatGPT's achievements relative to the researchers' responses. As shown in Table 3, ChatGPT responses are consistently more accurate than human responses across the four templates. The maximum obtained with Template 2, exactly 64.33%, represents a considerable gain compared to the minimum accuracy of true answers, which is 9.67%. This pattern repeats, as seen in other templates. For example, Template 1 indicates that the accuracy of deriving from ChatGPT's responses is 31.00%, while the accuracy of actual responses is 21.00%. Using Template 3, the accuracy of human responses is 15.00%, while ChatGPT's is 33.67%. Template 4 states that "ChatGPT's accuracy of 58.67% was significantly above the 16.33% accuracy of human answers. The provided results highlight ChatGPT's ability to generate correct answers, as it is quite proficient at understanding and articulating information appropriately.

In addition, the study assessed processing time and explored the trade-off between accuracy and efficacy. Increasing ChatGPT's processing time improves accuracy. Template 2, which achieved the highest accuracy of 64.33%, took 1310.251 seconds, yielding an average response time of 43.68 seconds. On the other hand, the total time taken for actual responses was 611.87 seconds, with an average of 20.40 seconds per answer. Similarly, Template 4 achieved an adequate accuracy of 58.67% and required 1284.44 seconds, resulting in an average response time of 42.81 seconds. These findings suggest that although ChatGPT can generate more accurate responses than humans, it consequently has a trade-off: longer processing times. Template 1 has a lower accuracy rate of 21.00% compared to the actual responses to the history questions. The total time to answer all the questions is 630.29 seconds. Template 3 exhibits a lower accuracy of 15.00%, but the overall time for all inquiries is 612.36 seconds. The average time spent responding to each question was 20.41 seconds. Therefore, compared with ChatGPT's responses, it achieves an accuracy of 33.67%, indicating greater correspondence with ChatGPT's answers than with the actual answers. The cumulative duration to respond to all responses is 1346.816 seconds, with an average of 44.89 seconds per answer.

Table 3. The performance of ChatGPT in answering questions relating to History themes.

History Question Templates	Results Compared to Actual Answers			Results Compared to ChatGPT Answers		
	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)
Template_1	21.00%	630.29	21.00	31.00%	1306.006	43.53
Template_2	9.67%	611.87	20.40	64.33%	1310.251	43.68
Template_3	15.00%	612.36	20.41	33.67%	1346.816	44.89
Template_4	16.33%	615.02	20.50	58.67%	1284.44	42.81

4.3 Legal And Ethical Templates Problem Analysis and Result Accuracy

As we saw in the previous section, the factual accuracy of ChatGPT's responses to legal and ethical inquiries about law is questionable. In Appendix Table 10, a sample from this is analyzed in terms of questions and responses. The first template consists of 10 questions that pertain specifically to elements inherent to the field of law. The questions aim to assess understanding of key concepts, ranging from historical material to present-day issues relevant to the field of Law. The second template is composed of 10 legal and ethical questions and their answers. The questions posed aim to provide additional examination of ChatGPT, as summarized in Appendix Table 11. Also, to evaluate the reliability and possibilities of ChatGPT, Templates 3 and 4 were derived from Templates 1 and 2. As noted above in the literature and History themes template review. Outcomes were mixed, contingent upon question format and quality of the answers. As shown in Table 4, ChatGPT's accuracy was consistently higher than the experimenters' across all question types. This increased accuracy, but was at the cost of longer computing times to produce high-quality answers.

As shown in Table 4, the operators' results achieved remarkable accuracy, even though human

responses were significantly lower than ChatGPT's across all templates. While the accuracy of actual replies is only 35.33%, Template 3's accuracy is 90.33%, thus showing a considerable improvement. This impact is regularly observed in the other templates. For instance, template 2 of ChatGPT's model achieved an accuracy of 84.67% compared with 68.00% for a human answer. For template 1, ChatGPT's accuracy was 83.67%, while the human answer was 34.67%. For template 4, the human answer scored above 37.33%, while ChatGPT's accuracy was 78.33%. Also, ChatGPT's use regarding its ability to produce accurate responses raises legal and ethical issues. The analysis of processing time is another dimension to consider in this regard, along with the trade-off between accuracy and efficiency. We hypothesized that the longer the stimuli processed by ChatGPT, the higher their accuracy would be. Template 3 is the most accurate, requiring 1281.32 seconds to achieve 90.33% accuracy, corresponding to an average time of 42.71 seconds. Conversely, the processing time for the real answers was longer than 580.44 seconds, averaging 19.35 seconds. On the other hand, Template 2 achieved 84.67% accuracy and a total time of 1402.83 seconds. Mean response time was 46.76 seconds. It was shown that ChatGPT can achieve higher precision in question-answering than the actual answer. Thus, it comes at the cost of longer processing times.

Table 4. The evaluation performance of ChatGPT in answering questions on Legal and Ethical.

Legal and Ethical Question Templates	Results Compared to Actual Answers			Results Compared to ChatGPT Answers		
	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)
Template_1	34.67%	576.01	19.20	83.67%	1364.16	45.47
Template_2	68.00%	550.97	18.37	84.67%	1402.83	46.76
Template_3	35.33%	580.44	19.35	90.33%	1281.32	42.71
Template_4	37.33%	586.04	19.53	78.33%	1444.36	48.15

4.4 Evaluation And Result Accuracy of the Sports Template Problem

Our study was interested in the ultimate performance of ChatGPT's responses to questions of sport. It demonstrates ChatGPT's effectiveness by reporting the accuracy percentages for each template. This is conducted using the questions and answers in Appendix Table 12, below. The assessment used template 1, which comprised 10 random questions, as outlined in the methodology, and centered on both historical content and current issues in sports. Template 2 was presented in Table 13 of the Appendix and featured several of the same standard questions, with responses intended to continue testing ChatGPT. At the same time, we extract information from templates 1 and 2 to create templates 3 and 4, respectively. The details limiting templates identical to Literature Related subjects are detailed in the previous three sub-sections. Table 5 presents the results from four templates, used as examples. They found varying levels of correctness in the questions, depending on how they were constructed. This indicates that a relationship exists between ChatGPT's recommendation accuracy and participants' answers, as supported by validated evidence. The average accuracy of ChatGPT was consistently better than that of the researchers across all questions, indicating that ChatGPT was superior regardless of question type. The ability of ChatGPT to provide valid answers was demonstrated across all cases, underscoring the superiority of the results over those from human answers. Plus, higher-quality answers also required more time spent in processing, particularly when assessing ChatGPT's performance relative to human answers.

The collected data are shown in Table 4. Even though ChatGPT's performance has reached an extremely high level of accuracy in the processes, human responses were much lower in accuracy, regardless of the template. For instance, taking template 1 as a case study, the accuracy of ChatGPT's answers, which is also the percentage of human responses, is 69.00 per cent, and that of humans' responses is 38.33 per cent. Clearly, in relation to template 3, underscoring that the resulting accuracies are 67.00 percentages for ChatGPT's responses and 33.33 percentage regarding the actual answer. As seen, template 4 is unique in that its evaluation is based on the percentages of ChatGPT's answer and the human's actual answer, at 62.00 and 32.27, respectively. In general

statistics and in terms of sample size, the accuracy of the answers given by ChatGPT using template 2 resulted in 59.00 percent, in contrast, the accuracy of human responses is 33.67 percent. More generally, given that humans have contributed to high-accuracy responses and improvements in accuracy, it's clear that ChatGPT will be able to provide better, more relevant answers to questions and create more accurate responses than humans. The analysis was conducted using a "comparison analysis" and processing time, which was found to be a compromise between accuracy and speed. Essentially, ChatGPT's purpose is to generate human-like conversational responses to user input by processing massive amounts of data. It can deliver this information in a search-engine style or with more literary, would-be novelist flair.

This study is also concerned with finding a good balance between precision and time processing speed. Longer processing times are a result of ChatGPT's superior accuracy. Template 1 also had the highest correct score (69.00%), a total time of 1244.921 seconds, and an average reaction time of 41.50 seconds. Conversely, the time humans take to respond is 572.09 seconds, averaging 19.07 per response. The average response time for template3, which was less than accurate (67.00%), was 1280.24 seconds, of which 42.67 seconds were needed to complete the task. The rationale is that ChatGPT can provide more precise answers than humans, but it takes longer to process the data it's fed.

Table 5. A comparative analysis of the performance of sport-related question templates.

Sport Question Templates	Results Compared to Actual Answers			Results Compared to ChatGPT Answers		
	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)	Accuracy	Sum Elapsed Time (s)	Average Elapsed Time (s)
Template_1	38.33%	572.09	19.07	69.00%	1244.921	41.50
Template_2	33.67%	543.29	18.10	59.00%	1288.77	42.96
Template_3	33.33%	549.42	18.31	67.00%	1280.24	42.67
Template_4	32.67%	604.99	20.17	62.00%	1303.47	43.45

The comparison between ChatGPT and actual answers in responding to literature-related questions indicates that ChatGPT achieved a substantially higher mean accuracy (57.08%) than actual answers (16.25%), demonstrating a marked improvement in its performance on literature questions. Although the standard deviation of ChatGPT responses (15.95%) was higher than that of actual answers (4.74%), this variability reflects differences in the complexity of the literature question templates rather than inconsistent performance. A paired t-test for the four matched literature question templates revealed a statistically significant difference between the two approaches ($t = 5.24$, $p = 0.0135$). These results show that ChatGPT far exceeds actual answers in terms of answer quality, demonstrating its usefulness as an aid to learning about the literature of study, as represented in Table 6.

Table 6. Statistical Analysis of ChatGPT's Performance on Literature-Related Questions

Metric	Actual (%)	ChatGPT (%)
Mean	16.25	57.08
Standard Deviation (STD)	4.74	15.95
Sample Size (n)	4	4
t-value (Paired t-test)	5.25	
p-value	0.0135	

The comparison between ChatGPT and actual answers in responding to history-related questions shows that ChatGPT achieved a considerably higher mean accuracy (46.92%) than actual answers (15.50%), indicating improved performance across all evaluated history question templates. Although the standard deviation of ChatGPT responses (17.03%) exceeded that of actual answers (4.66%), this variation reflects differences in historical question complexity rather than inconsistency in response quality. A paired t-test conducted on the four matched templates revealed that, despite the observed performance improvement, the difference did not reach statistical significance at the 0.05 level ($t = 3.04$, $p = 0.0558$). These findings suggest that while ChatGPT shows

strong potential to answer history-related questions, further evaluation with a larger sample is required to confirm statistical significance, as shown in Table 7.

Table 7. Statistical Analysis of ChatGPT's Performance on History-Related Questions

Metric	Actual (%)	ChatGPT (%)
Mean	15.50	46.92
STD	4.66	17.03
Sample Size (n)	4	4
t-value (Paired t-test)	3.04	
p-value	0.0558	

The comparison of ChatGPT's responses with actual answers for legal and ethical questions shows that ChatGPT has a much higher mean accuracy (84.25%) than the actual answers (43.83%), demonstrating significant improvement across all evaluated legal and ethical question templates. In addition, ChatGPT responses showed greater variability (STD = 4.92%) than the actual responses (STD = 16.15%), suggesting greater consistency and reliability in this area. Two methods differed statistically in a paired t-test with the four matched templates (t = 4.80, p = 0.172). These findings indicate that ChatGPT is far superior to actual responses in legal and ethical question-answering and also illustrate its utility and robustness as a support tool in academic contexts of formal, rule-based study, as shown in Table 8.

Table 8. Statistical Analysis of ChatGPT's Performance on Legal and Ethical Questions

Metric	Actual (%)	ChatGPT (%)
Mean	43.83	84.25
STD	16.15	4.92
Sample Size (n)	4	4
t-value (Paired t-test)	4.80	
p-value	0.0172	

ChatGPT was significantly faster than actual answers in responses to the sport question, with a mean accuracy of 64.25% compared to 34.50% for actual answers, indicating that all the sport question templates tested had much higher mean accuracy. Though ChatGPT responses were slightly more variable (STD = 4.57%; actual response = 2.59%), both measures were reasonably consistent across procedures, with ChatGPT prevailing as the stronger method. Two approaches showed a statistically significant difference on the four matched templates in a paired t-test. t = 17.19, p = 0.0004. This indicates that ChatGPT is far superior to actual responses in answering questions about sport and is well-equipped to answer structured factual questions in this domain, as shown in Table 9.

Table 9. Statistical Analysis of ChatGPT's Performance on Sport-Related Questions

Metric	Actual (%)	ChatGPT (%)
Mean	34.50	64.25
STD	2.59	4.57
Sample Size (n)	4	4
t-value (Paired t-test)	17.19	
p-value	0.0004	

5. DISCUSSION

GPT is hyper-optimized for human interaction and the use of super-advanced AI. But ChatGPT's answers are not necessarily correct or perfect, particularly when asked questions quickly or repeatedly, such as 20 times in a short period. The present study sought to examine these questions, with an emphasis on the accuracy and speed of ChatGPT's answers. Plus, the research explores how answers may vary in quality and how confusion may arise depending on the type of category or question, as well as the question's structure, when asking ChatGPT. Moreover, the study demonstrates that while ChatGPT is generally accurate and consistent in generating responses, its accuracy is lower than that of the actual user-provided answers. ChatGPT and similar AI-based tools

are already being deployed across higher education for teaching, learning, research, and institutional purposes. They also play a major role in public health, primarily through medical assistance. Sections 1 and 2 of the present study detail this issue, with additional emphasis from Atlas in 2023 [31]. Therefore, ChatGPT needs to enhance its ability to answer the prompt questions more accurately. The present study conducts a specific evaluation of ChatGPT's accuracy to establish its reliability, particularly in educational and medical contexts.

For the practical application of AI tools such as ChatGPT in education and research, outputs should be carefully checked, as these models may produce errors or reflect the training data. It reveals transparency on when and how AI tools are used, along with proper attribution and documentation of any fine-tuning. Education needs to develop a clear policy on how to use AI responsibly and ethically, be consistent with learning goals, and provide instruction to help students use AI responsibly and ethically. AI should support, not replace, human judgment in education and research, especially in tutoring, assessment, and summarizing research. In the paragraphs, Bahak and his colleagues discussed an alternative method for assessing the accuracy of ChatGPT's responses. For the measure of accuracy, they used a model called Question Answering System (QAS, also known as news and prompts. This model yielded accuracy rates between 50-75% on all tested questions [32]. But this model was also more accurate than ChatGPT's responses in this study, particularly on ethical prompt templates, achieving over 90% accuracy. The study showed that ChatGPT's responses are not stable but rather depend on several variables, including the types of questions being prompted, the degree of specificity required of the responses, variations in the placement of the prompting questions, and the blending of question types. But ChatGPT's answers are, for the most part, normative despite this variance. There are also discrepancies in these results: Template 2 history questions show a 9.67% difference in accuracy, and Template 3 shows a 15.00% difference between the actual responses and ChatGPT's modified responses. These differences are further exemplified in several of the other examples within the study. Also, the length of prompt questions and answers significantly influences ChatGPT's accuracy and its ability to address the issue at hand.

Elapsed time in testing prompt responses and a comparison of two cases was also of interest in this study model. Findings suggest that the question category has a significant effect on ChatGPT, while question or response length influences the time taken. ChatGPT does generate responses autonomously, but it is often time-consuming because its answers are lengthy, as shown in the Appendix tables. Conversely, authors' answers to the prompts took significantly less time and were brief, succinct replies, such as two- to three-word responses, as presented in Fig 3. This highlights that a future version of ChatGPT should be able to process and provide answers faster. While the PACEM framework offers a rigorous and replicable approach for evaluating ChatGPT's semantic accuracy, several methodological limitations must be acknowledged. First, PACEM's scoring is sensitive to the predefined similarity threshold (T = 0.6); altering this cut-off can meaningfully shift accuracy outcomes, especially for borderline or domain-specific responses. Second, reliance on cosine similarity may introduce bias, as this metric captures surface-level semantic overlap but may not fully reflect factual correctness, depth of reasoning, or contextual nuances, particularly when responses are lexically rich but partially incorrect. Third, PACEM depends heavily on the performance of the chosen sentence-transformer (paraphrase-MiniLM-L6-v2), which may struggle with specialized terminology in legal, historical, or medical domains. Fourth, since ChatGPT is periodically updated, the embedding of behaviors and the generated answers may vary over time, limiting perfect reproducibility. Finally, semantic similarity alone cannot assess logical validity or evidence-based accuracy, indicating that future versions of PACEM should integrate factuality metrics or reasoning-quality evaluation modules.

Fig. 3 also provides a plot comparing ChatGPT's accuracy in providing correct human or true answers across four categories: Literature, History, Legal and Ethical, and Sport. The stated accuracies in ChatGPT's responses are higher than the true ones across all categories. The largest absolute differences were in Literature and History, where ChatGPT's responses were 72.33% and 64.33% correct, compared with 20.67% and 21.00% for the actual answers in these subjects. The Law (Legal and Ethical) category is a draw, with ChatGPT's accuracy at 90.33% and the human-

provided answers' accuracy at 68.00%. While there is also a clear difference in Sports, ChatGPT's accuracy of 69.00% in this case is much higher than the accuracy of the actual answers, which is 38.33%, suggesting a major overestimation of its capacity to address sports questions. In our study, ChatGPT outperformed human responses in accuracy across the examined domains, though the magnitude of this advantage varied by category. ChatGPT outperformed humans in the Law (Legal and Ethical) domain, achieving 90.33% accuracy compared to 68.00%, suggesting robust performance on structured, rule-based questions, while human performance remains substantial. In contrast, ChatGPT was 69.00% accurate in the Sports domain, up from 38.33% accuracy for human participants. This indicates that while ChatGPT may appear confident in its answers to sports-related questions, it does not always match correct responses.

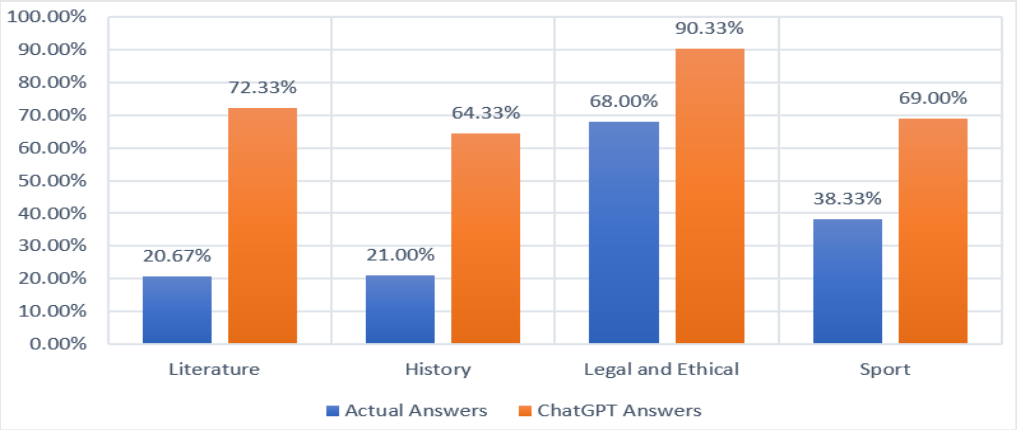

Figure 3. Accuracy discrepancies between ChatGPT-generated answers and actual outcomes across different knowledge categories.

Fig. 4 compares the total time required for ChatGPT's answers and human answers across four categories: Literature, History, Legal and Ethical, and Sport. For the Literature category, ChatGPT's responses took 1155.231 seconds to complete, which is more than double the time of the actual responses at 507.37 seconds. When looking at overall time duration, ChatGPT's experiences led to significantly longer answer generation times across all categories, with the largest time disparities in History and Legal and Ethical at 1284.44 and 1281.32, respectively. Finally, in the sports discipline, the required time for the ChatGPT answer was 1244.921 seconds, while the actual response time was 543.29 seconds. From these differences, it can be inferred that ChatGPT's answers consistently take longer across all categories, especially in History and Legal and Ethical, to a much greater degree. The data indicate that the difference between time spent and precision gained is negligible. Finally, greater accuracy is positively associated with greater time spent compared with lower accuracy.

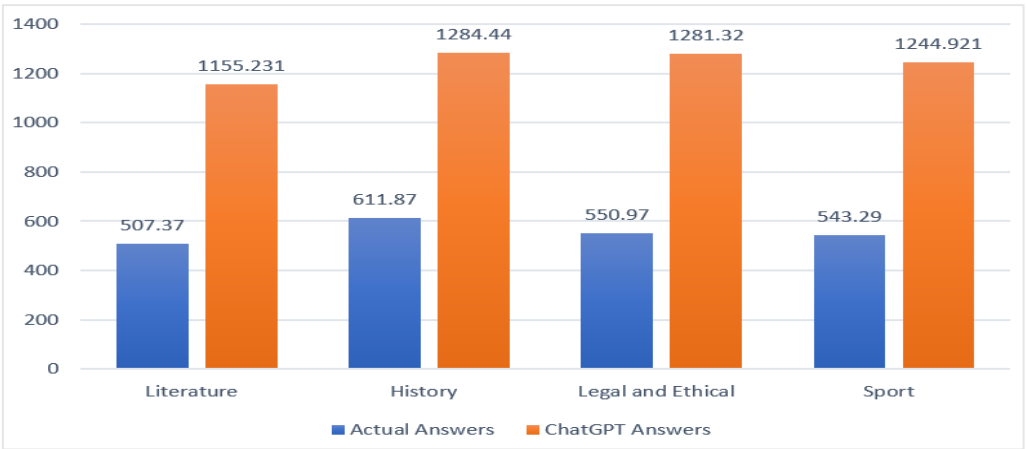

Figure 4. Total elapsed time (seconds) required to generate ChatGPT's answers versus the actual answers across four distinct categories.

6. CONCLUSION

These results suggest that PACEM provides a reliable baseline for evaluating ChatGPT's accuracy, consistency, and response behavior across various domains. In general, ChatGPT achieved the highest accuracy scores of any human-generated response across the literature, history, legal/ethical, and sports categories. This demonstrates ChatGPT's high degree of accuracy in providing data-rich, semantically similar responses, as measured by semantic-similarity comparisons. They also show a clear trade-off: higher accuracy is accompanied by considerably longer processing cycles. This suggests that ChatGPT's strengths lie in depth and completeness rather than speed, especially with complex or multi-step prompts. The results also show that ChatGPT's performance may vary according to question type, prompt formulation, and domain specificity. The advantages of this approach are evident, but fluctuations in consistency and occasional response drift underscore the need for cautious use in high-stakes situations that require precision. Thus, these findings suggest that PACEM can be used to identify such strengths and weaknesses systematically and is a viable means of monitoring generative-AI behavior across disciplines. In practice, these results indicate that ChatGPT is well-suited to applications where accuracy, explanation quality, and interpretability are the main concerns, such as education, legal reasoning support, and medical information review. However, human responses may still be effective in time-sensitive situations. The time-accuracy trade-off outlined in this study is therefore indispensable for understanding the application and use of ChatGPT.

Future studies should explore combining PACEM with new assessment domains such as factual reliability benchmarks, temporal consistency testing, and prompt-sensitivity analysis. A path to improving ChatGPT's speed without sacrificing semantic accuracy is worth exploring. A broader range of testing across professional interest domains will also help validate PACEM's utility and clarify ChatGPT's performance limits. Importantly, a major limitation of this study is the relatively small dataset. The dataset does not constitute big data, and generating a large, high-quality dataset that meets these specific criteria is inherently challenging. Consequently, only a limited number of questions were included in each category, limiting the generalizability of the findings across complex and diverse domains such as literature, history, and law. Additionally, the responses were collected from ChatGPT using different question samples, and the analysis relied on descriptive and comparative statistical models to select and evaluate the data. While this approach allows identification of patterns and relative weaknesses, the limited question diversity and small sample size reduce the robustness of the conclusions.

CONFLICT OF INTEREST

The authors declare that there is *no* conflict *of interest* regarding the publication of this paper.

ACKNOWLEDGEMENTS

The authors thank their institutions for providing the resources and ongoing assistance necessary to undertake this research.

FUNDING

This research received no external funding.

REFERENCES

[1] O. M. Alyasiri, D. Akhtom, and M. N. Alrasheedy, "An Overview of GPT -4's Characteristics through the Lens of 10V's of Big Data," in 2023 3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), IEEE, Dec. 2023, pp. 201–206. doi: 10.1109/ICICyTA60173.2023.10429032.

[2] A. A. Hassan, H. S. Abdulla, T. Y. Mawlood, R. K. Muhammed, A. M. Aladdin, and T. A. Rashid, "A Multi-Account Statistical Evaluation of ChatGPT Proficiency in the Kurdish Sorani Language," UHD Journal of Science and Technology, vol. 9, no. 2, pp. 319–334, Nov. 2025, doi: 10.21928/uhdjst.v9n2y2025.pp319-334.

[3] J. Shabbir and T. Anwer, "Artificial intelligence: A powerful paradigm for scientific research," The Innovation, 2018, doi: 10.1016/j.xinn.2021.100179.

[4] B. K. Arif and A. M. Aladdin, "A Comparative Analysis of ChatGPT and Traditional Machine Learning Algorithms on Real-World Data," Kurdistan Journal of Applied Research, vol. 10, no. 2, pp. 93–118, Sep. 2025, doi: 10.24017/science.2025.2.8.

[5] A. Shollo, K. Hopf, T. Thiess, and O. Müller, "Shifting ML value creation mechanisms: A process model of ML value creation," The Journal of Strategic Information Systems, vol. 31, no. 3, p. 101734, 2022, doi: 10.1016/j.jsis.2022.101734.

[6] R. Schmidt, A. Zimmermann, M. Möhring, and B. Keller, "Value creation in connectionist artificial intelligence–a research agenda," AMCIS 2020 proceedings-Advancings in information systems research: August 10-14, 2020, Online, pp. 1–10, 2020.

[7] J. Tang, G. Liu, and Q. Pan, "A Review on Representative Swarm Intelligence Algorithms for Solving Optimization Problems: Applications and Trends," IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 10, pp. 1627–1643, 2021, doi: 10.1109/JAS.2021.1004129.

[8] A. M. Aladdin and T. A. Rashid, "A New Lagrangian Problem Crossover—A Systematic Review and Meta-Analysis of Crossover Standards," Systems, vol. 11, no. 3, p. 144, Mar. 2023, doi: 10.3390/systems11030144.

[9] A. A. H. Amin, A. M. Aladdin, D. O. Hasan, S. R. Mohammed-Taha, and T. A. Rashid, "Enhancing Algorithm Selection through Comprehensive Performance Evaluation: Statistical Analysis of Stochastic Algorithms," Computation, vol. 11, no. 11, p. 231, 2023.

[10] A. M. Aladdin, C. M. Rahman, and M. S. Abdulkarim, "The Scientific Comparison between Web-Based Site and Web-Builder (Open Source) Project: Functionalities, Usability, Design and Security," International Journal of Scientific Research and Management (IJSRM), vol. 6, no. 06, Jun. 2018, doi: 10.18535/ijsrm/v6i6.ec05.

[11] V. Taecharungroj, "'What can ChatGPT do?' Analyzing early reactions to the innovative AI chatbot on Twitter," Big Data and Cognitive Computing, vol. 7, no. 1, p. 35, 2023, doi: 10.3390/bdcc7010035.

[12] A. Nazir and Z. Wang, "A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges," Meta-Radiology, vol. 1, no. 2, p. 100022, 2023, doi: 10.1016/j.metrad.2023.100022.

[13] M. Mijwil, Mohammad Aljanabi, and Ahmed Hussein Ali, "ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information," Mesopotamian Journal of CyberSecurity, vol. 2023, pp. 18–21, Feb. 2023, doi: 10.58496/MJCS/2023/004.

[14] R. K. Muhammed et al., "Comparative Analysis of AES, Blowfish, Twofish, Salsa20, and ChaCha20 for Image Encryption," Kurdistan Journal of Applied Research, vol. 9, no. 1, pp. 52–65, 2024, doi: 10.24017/science.2024.1.5.

[15] R. K. Muhammed, K. H. Ali Faraj, J. F. G. Mohammed, Ahmad Al Attar Tara Nawzad, S. J. Saydah5, and D. A. Rashid, "Automated Performance Analysis E-services by AES-Based Hybrid Cryptosystems with RSA, ElGamal, and ECC," Advances in Science, Technology and Engineering Systems Journal, vol. 9, no. 3, pp. 84–91, 2024, doi: 10.25046/aj090308.

[16] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, 2023, doi: 10.1016/j.iotcps.2023.04.003.

[17] M. Mijwil and M. Aljanabi, "Towards artificial intelligence-based cybersecurity: The practices and ChatGPT generated ways to combat cybercrime," Iraqi Journal For Computer Science and Mathematics, vol. 4, no. 1, pp. 65–70, 2023, doi: 10.52866/ijcsm.2023.01.01.0019.

[18] N. Khan, Z. Khan, A. Koubaa, M. K. Khan, and R. bin Salleh, "Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis," Conn Sci, vol. 36, no. 1, p. 2353630, Dec. 2024, doi: 10.1080/09540091.2024.2353630.

[19] M. C. Keiper, G. Fried, J. Lupinek, and H. Nordstrom, "Artificial intelligence in sport management education: Playing the AI game with ChatGPT," J Hosp Leis Sport Tour Educ, vol. 33, p. 100456, 2023, doi: 10.1016/j.jhlste.2023.100456.

[20] R. Xu and Z. Wang, "ChatGPT in Healthcare from the Perspective of Digital Media: Applications, Opportunities and Challenges," Heliyon, 2024, doi: 10.1016/j.heliyon.2024.e32364.

[21] A. M. Aladdin, Y. N. Bakir, and S. I. Saeed, "The effects to trend the suitable os platform," Journal: Journal of Advances in Natural Sciences, vol. 5, no. 01, 2018, doi: 10.24297/jns.v5i1.7528.

[22] M. Aljanabi, "ChatGPT: Future directions and open possibilities," Mesopotamian journal of Cybersecurity, vol. 2023, pp. 16–17, 2023, doi: 10.58496/MJCS/2023/003.

[23] H. Nasef et al., "Evaluating the Accuracy, Comprehensiveness, and Validity of ChatGPT Compared to Evidence-Based Sources Regarding Common Surgical Conditions: Surgeons' Perspectives," Am Surg, p. 00031348241256075, 2024, doi: 10.1177/000313482412560.

[24] C. B. Lau, E. Lilly, J. Yu, and G. P. Smith, "Evaluating the efficacy of ChatGPT in addressing patient queries about acne and atopic dermatitis," Clin Exp Dermatol, p. llae187, 2024, doi: doi.org/10.1093/ced/llae187.

[25] W. W. Jedrzejczak, P. H. Skarzynski, D. Raj-Koziak, M. D. Sanfins, S. Hatzopoulos, and K. Kochanek, "ChatGPT for Tinnitus Information and Support: Response Accuracy and Retest after Three and Six Months," Brain Sci, vol. 14, no. 5, p. 465, 2024, doi: 10.3390/brainsci14050465.

[26] R. W. Puyt and D. Ø. Madsen, "Evaluating ChatGPT -4's historical accuracy: a case study on the origins of SWOT analysis," Front Artif Intell, vol. 7, p. 1402047, 2024, doi: 10.3389/frai.2024.1402047.

[27] A. J. Neuhouser, A. Kamboj, A. Mokhtarzadeh, and A. R. Harrison, "Artificial intelligence in practice: measuring its medical accuracy in oculoplastics consultations," Modeling and Artificial Intelligence in Ophthalmology, vol. 6, no. 1, pp. 1–11, May 2024, doi: 10.35119/maio.v6i1.137.

[28] D. Braithwaite et al., "Evaluating ChatGPT's Accuracy in Providing Screening Mammography Recommendations among Older Women: Artificial Intelligence and Cancer Communication," Res Sq, 2024, doi: 10.21203/rs.3.rs-3911155/v1.

[29] H. L. Walker et al., "Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument," J Med Internet Res, vol. 25, p. e47479, 2023, doi: 10.2196/47479.

[30] O. Al-Janabi, O. M. Alyasiri, E. A. Jebur, and S. M. Nafl, "Evaluating AI Language Models in News Retrieval: A Comparative Study Of ChatGPT-Plus and DeepSeek (R1)," InfoTech Spectrum: Iraqi Journal of Data Science, vol. 2, no. 2, pp. 14–20, Jun. 2024, doi: 10.51173/ijds.v2i2.33.

[31] S. Atlas, "ChatGPT for higher education and professional development: A guide to conversational AI," University of Rhode Island, 2023, Accessed: Aug. 24, 2024. [Online]. Available: : https://digitalcommons.uri.edu/cba_facpubs/548

[32] H. Bahak, F. Taheri, Z. Zojaji, and A. Kazemi, "Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models," arXiv preprint arXiv:2312.07592, 2023.