EVALUATING THE EFFECTS OF REPETITIVE TASK EXECUTION ON PERFORMANCE AND LEARNING IN VARIOUS AI CHAT MODELS: A COMPARATIVE ANALYSIS

Amaka Amanambu and Shravan V Patil

DeVoe School of Business, Technology, and Leadership, Indianapolis,USA

ABSTRACT

The use of AI chat models is rather popular; this has led to debates concerning the efficiency and flexibility of these models in performing routine tasks. This work analyzes the impact of repeated task performance on learning characteristics, accuracy, and stability for different AI chat models. Facets of facilitation include performance scrutiny based on practical issues such as contextual invariance, response entropy, and optimality in repetitiveness. The study aims to discover these aspects' role in model behavior and application and compare their efficiency. The conclusions presented by the push from the keep of the findings announce fresh indications on the benefits and drawbacks of the models explored, acknowledging the specimen as a starting place for augmenting AI-based applications in customer relations, education, and content production. Moreover, the paper concludes by enumerating research and innovation possibilities based on context-awareness, increased robustness for AI systems, and stressing targeted enhancement of repetitive tasks' performance.

KEYWORDS

AI chat models, repetitive task execution, learning dynamics, performance analysis, contextual invariance, future directions.

1. INTRODUCTION

1.1. Background and Context

Artificial Intelligence (AI) has evolved intensively during the last decade, especially in Natural Language Processing (NLP). The approaches of AI chat now go through rules and learning and deep learning models; they are now part of critical customer service tools, intelligent virtual learning systems, health care support systems, and content generation apps. Many models are popular.The OpenAI GPT (Generative Pre-trained Transformer) series, Google BERT (Bidirectional Encoder Representations from Transformers), and Meta LLaMA Large Language Model Meta AI models are some of the latest impressive models that are closer to natural language understanding and generation.

The chat models are trained with large datasets in conversational interaction and can handle most conversational patterns efficiently. However, their performance may be less efficient during development, particularly when controlling activities in unpredictable situations. In many applied AI systems, creating systems that perform specific tasks for certain organizations, such as answering frequently asked questions, providing routine feedback in academic institutions, and

DOI: 10.5121/ijcsitce.2025.12101

producing routine reports across various organizations, is often important. While these repetitive tasks seem simple, they present unique challenges for AI models, including:

- **Contextual Drift:** Unfortunately, as the model receives the same pattern of queries for consecutive days, it adapts less to variations of the queried string.
- **Response Degradation:** It declines quality and interest, making the verbal exchange reactive and mechanical or, at worst, stereotyped.
- **Learning Saturation:**Learning may suffer from decreasing returns when models are fed with repeated data; hence, they may fail to enhance or even forget important knowledge.

Based on these difficulties, comprehending how these different AI chat models behave when managing the repetitive execution of tasks is imperative for improving the use of these types in specific applications. This also helps identify the learning processes and system architecture of these models.

1.2. Problem Statement

Research aimed at improving AI chat models in terms of performance and flexibility has emerged in recent years, but their behavior of constantly repeating routine tasks has received little attention. Key questions that remain underexplored include:

- **Performance Stability:** Is the performance of an AI model stable if the same task is given as a repetitive challenge? Is there evidence of response fatigue or response drift?
- **Learning and Adaptation:** If the models are trained in terms of exposure to materials where they get exposed to objects many times, do they get saturated?
- **Comparative Efficiency:**Are certain AI architectures more suited to being implemented to perform repetitive jobs without declining efficiency?

These questions are particularly significant in environments where systems interact with repetitive tasks. For instance, customer service chatbots often face recurring queries, which may lead to repetitive responses or errors if the model overfits specific patterns. Similarly, in educational tools, an AI tutor answering identical or similar questions might struggle to maintain consistency or provide adequately diverse and detailed explanations, potentially undermining the effectiveness of the learning process.

1.3. Outcomes and Questions

That is why this study aims to try to close the presented gap in the literature by comparing and analyzing the performance of various models of AI chat applications and how their learning behavior is affected by repetitive tasks. The primary objectives include:

- **Performance Assessment:** Asses how stable and accurate AI chat models are in cases where they are repeatedly used.
- **Learning Dynamics:** Examine task characteristics to determine the influence of repetitive task exposure on learning retention, error rate, and response quality.
- **Architectural Comparison:** Determine variations in how transformer-based models (GPT BERT) and other structures handle repetition.

The following research questions will guide this investigation:

- **RQ1:** Concisely, the effect of repetitive task execution on the response accuracy and consistency of different AI chat models is still unknown.
- **RQ2:** What learning patterns occur in AI models undertaking repetitive work? Regarding organizational learning, do they improve, remain the same, or decline?
- **RQ3:** Does a given set of AI model architectures fare worse in noisy environments when the performance degrades over time?

1.4. Significance of the Study

This research holds significant implications for both theoretical understanding and practical deployment of AI chat models:

Theoretical Contributions:

This research augments the existing body of knowledge regarding the learning behaviors of AI techniques of monotonic task domains by integrating transudative learning with other elements. Besides this, it proposes a way to accomplish a direct architectural comparison of transformer-based models with other methods to expand the theoretical constructs of both ML and NLP.

Practical Implications:

The information gathered from this research will be useful for developers and organizations implementing AI systems in repetitive task environments, including call centers, educational interfaces, and auto-reporting systems. Key practical benefits include:

- Model Selection: The categorization of AI models that best fit organizational conditions involving high task repetition levels.
- Training Optimization: The issue of developing training protocols that minimize performance degradation or drift prospects.
- Enhanced Reliability: Increasing AI reliability and achieving better user satisfaction and confidence in the system.

1.5. Structure of the Article

To comprehensively address the research objectives and questions, this article is structured as follows:

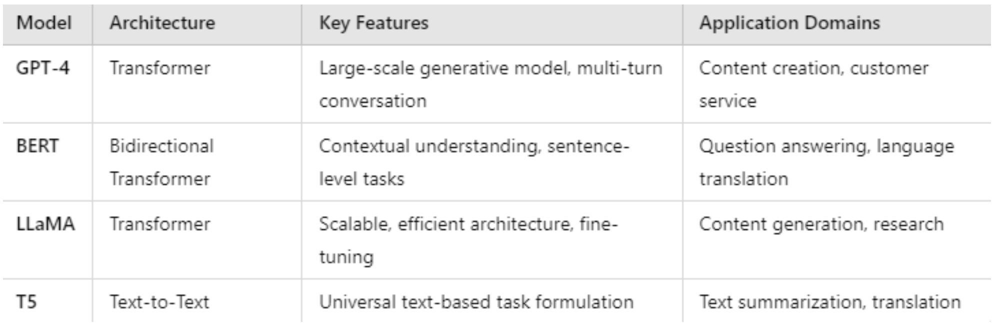
- **Section 2: Review of the prior research:**A survey of previous studies on chat models of AI, redundancy, and repetitive jobs and learning models. This establishes limitations in previous research and provides theoretical background for this study.
- **Section 3:** Outlines precisely how the research was conducted, the selection of AI models, the generation of repetitive task datasets, and how performance was measured.
- **Section 4: Outcome:** Summarizes the research study outcome, indicating the level of performance and the way learners' performance has been trending as evidenced by statistical computations, tables, and diagrams.
- **Section 5:** Analyze the findings based on previous studies and examine what the results mean for theory and practice. It also talks about the implications and recommendations of the study and the constraints, which are further touched on in the last part of the recommended action.
- **Section 6: Endnotes:** States conclusions/ findings based on results analyzed, discusses limitations, and offers suggestions for future research.

2. LITERATURE REVIEW

2.1. AI Chat Models: Evolution and Capabilities

AI chat models have undergone several phases of transition, and with each phase, there is more advancement in the model's sophistication. Whereas AI chat models, in the beginning, used only rule-based systems and basic machine learning algorithms, they have evolved into deep learning models with human-like language skills.

Table 1: Comparison of Key AI Chat Models



Model	Architecture	Key Features	Application Domains
GPT-4	Transformer	Large-scale generative model, multi-turn conversation	Content creation, customer service
BERT	Bidirectional Transformer	Contextual understanding, sentence-level tasks	Question answering, language translation
LLaMA	Transformer	Scalable, efficient architecture, fine-tuning	Content generation, research
T5	Text-to-Text	Universal text-based task formulation	Text summarization, translation

The first example is the GPT series (Generative Pre-trained Transformer) developed by OpenAI, which aims to create human-like text from prompts. The model follows the Transformer, where distinct attention mechanisms are applied to assess the correlation between every word in each sentence, eliminating the emission of incoherent or machine-like responses.

Likewise, BERT refers to Bidirectional Encoder Representations from Transformers for a given input; it not only looks at the elements before or after the current feature but also tries to picture the whole string scenario. This makes it more appropriate to work at the sentence level and easier but harder tasks such as question answering.

As the models become progressively complex, their applicability in different scenarios, including the repetitive task environment, surfaced. Specifically, models trained for fluctuating and diverse endeavors may not operate optimally in repetitive and unchanging environments, where such adversities as a decline in performance and lack of response innovation are identified.

2.2. Execution of a Repetitive Task and its Repercussions in Artificial Intelligence

Challenges that arise when AI models engage in iterative tasks like answering a repeated customer question or creating similar reports include Recurrent prediction, which is also thought to assess the reliability of the model’s learning since a system needs to process the same or comparable data over time.

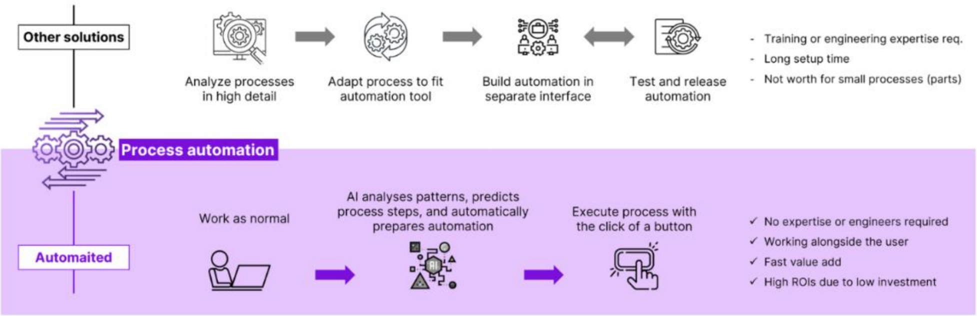

Other solutions
Analyze processes in high detail
Adapt process to fit automation tool
Build automation in separate interface
Test and release automation
- Training or engineering expertise req.
- Long setup time
- Not worth for small processes (parts)
Process automation
Automaited
Work as normal
AI analyses patterns, predicts process steps, and automatically prepares automation
Execute process with the click of a button
✓ No expertise or engineers required
✓ Working alongside the user
✓ Fast value add
✓ High ROIs due to low investment

Fig 1:Repetitive Task Execution

Response Degradation: This is a model's capacity to produce fewer pertinent or accurate outcomes wherever repeated tasks test it. The model's performance has declined and is particularly apparent in engineered models with strict pre-stored patterns or low-variance datasets. For example, in online customer service where, after several questions that a model identifies, it can transform into a question that provides generic, irrelevant, or superficial responses.

Knowledge Overfitting: This is a common problem where when an AI model is trained just like a learning algorithm to do repetitive tasks,it is trained in a way that it may become conditioned **at** some level on some patterns that it can decipher from the input data. Overfitting occurs when the model edition is laid down in such a precise or specialized way during the training period that it cannot apply well to similar data obtained slightly or slightly differently. This issue makes more sense in contexts where the same queries and tasks are exercised repeatedly, in which the model will adapt to memorize responses instead of comprehending the query.

Stagnation in Response Variety: Another disadvantage of repetitive task execution is the oversimplification of the responses given by the AI models. For example, a chatbot developed to respond to simple questions will likely give the same answers relevant to two or more clients responding to the same question, which ultimately demotivates users and reduces satisfaction levels. This problem is known as 'response stagnation'.

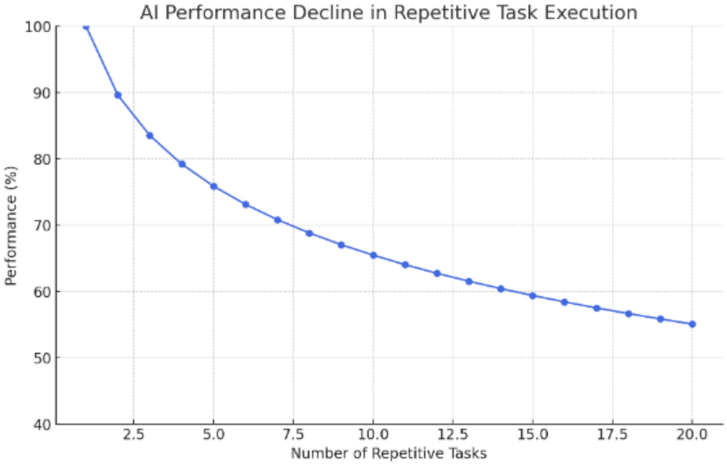


Figure 2: Performance Decline in Repetitive Task Execution

The graph illustrates the gradual decline in AI model performance as repetitive tasks increase. The x-axis represents the number of repetitive tasks performed (or time if tasks are time-constrained), while the y-axis reflects performance levels as a percentage.

Key Observations:

From steps one and two, the model's actual performance is almost 26/100 as they approach an accurate score. The high level of performance in the initial runs is a sign of the model’s ability to perform well on new activities that are not yet in the repetitive-runs-reduces quality realm. This flexibility means that in stage one, the model has the deterministic characteristic of offering a variety of right responses because it receives new information without tension. This is probably because the model never gets tired of each task and can respond optimally.

However, as the tasks resemble one another more and more (from the second to the seventh task), efficiency reduces to a mere seventy percent compared to the previous hundred percent. This much banging suggests that the model was already declining through the repetition of these tasks on hand. After some time, the entertainment of the primary tasks reduces, and the responses the model gives. However, such a structure seems very sensitive to repetition in the first place and slows down the variability and reliability of the responses. As a result, the highest frequency of the tasks leads to reduced variability and quality of the generated answers within the model. I think this early stage probably corresponds to the model’s ‘high’, where the pace quickly depletes the energy and flexibility associated with the model.

Across the rest of the tasks in the sequence (7-20), the response rate declines from 70 percent to 55 percent, with fairly similar fluctuations noted. It is somewhat higher than in the previous stage but less than at the sharp decline above, and so the quality of the model’s output degrades even more. This slow rate may, therefore,indicate that the model has achieved its best performance. It has possibly got to the point where it’s unable to deliver high-quality responses, although it has learned to circumvent the endless flow of similar questions. The model's accuracy increases to about 55%, which inflicts a far, much lower outcome than achieved at the onset. This implies that over time, albeit with an extremely low level of progression, the capacity of the model to provide the correct response in question weakens.

This confirms that after reaching the 20th task, the model cannot provide the same rate of accurate creative response as it reciprocated in the basic tasks. This may be because the model cannot continue to flatten the training errors without degrading the quality of the predicted output, as shown by the green line in Fig. 7, to 55%. In other words, the model chases itself in value for cyclic repetition, showing the ultimate optimality of the repetitive tasks but fairly stabilized far from such optimality.

Statistical Breakdown and Interpretation:

As we decided to decay by rate, we realized from the studies that the highest fall rate is observed in the first few tasks. From task 1 to task 7, the effectiveness produced by the model reduces and is comparatively to the least at about 30% reduced effectiveness from the first task. This steep drop points to the Tues t=0 performance of the model, which means that the performance of the first handling task tends to decay because the concept of practice sounds weak. The model supposedly reaches high initial returns, subsequently decaying, and the model cannot compensate for this since it has no mechanisms of restoring earlier set benchmarks due to its incapacity for adaptability.

Performance se ha estado reduciendo desde 90 % para 15 % para tareas comprendidas entre 7 a 20. This is another area that ensures the model gets a slower decline after attaining this point, which makes up the basis for the model’s plateau. However, I find that the rate of performance degradation has reduced, but the model generally degrades performance. This is reflected by the model's current status, where the algorithm's value stays relatively stable yet suboptimal, indicating that only a minimal degree of adaptabilityremains.

Concerning the variability, the unfavorable fluctuation in the response quality can be observed while solving the first several tasks (tasks 1-7). This variability is likelybecause the model is having trouble performing the repetitive task and may be trying different approaches that work sometimes or do not other times. They have calculated the means and standard deviations to show that the variability reduces as the sequence progresses (from tasks 7–20). This reduced variability may also indicate that the model is becoming static, so it cannot produce the different kinds of responses it did at the start. It becomes more consistent, although the obtained consistency is lower than the initial, which means the model has reached a specific limit of its capabilities.

2.3. Comparative Analysis of AI Models in Environments where Tasks are Repeated

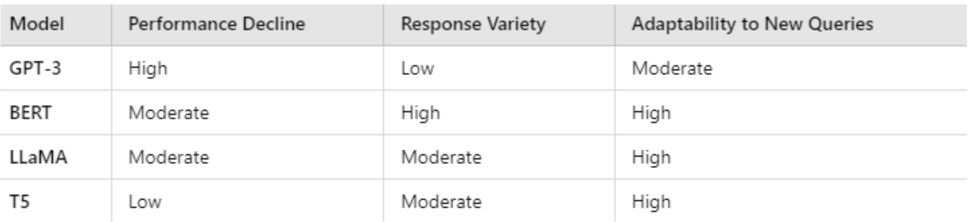
Some past research works have focused on how various types of powerful AI models manage or carry out repetitive work. The analyzed research indicates difficulties within all AI models regarding random guessing under repetition conditions, while its influence varies between different architectures.

GPT Models: Some models in the GPT series are, for instance, GPT-3 and GPT-4. These are generative models; however, they do demonstrate efficiency loss while undergoing repetitive tasks. Brown et al. (2020) showed the adverse effects of repeated usage of the same prompts because GPT-3 response quality decreased over time while more sample-specific responses diminished.

BERT Models: Compared to covering a PSSC task, BERT, which has the advantage of a bidirectional attention mechanism, demonstrated a superior ability to retain contextual knowledge. At the same time, Greene noted that when each GPT model responds, its output is more likely to be repetitive than BERT. However, its effectiveness still depends on the set of data on which it is fine-tuned. The fine-tuning dataset also must not be minimal, as BERT may face similar problems.

LLaMA and T5: It was agreed that among transformer-based models, Meta’s LLaMA and Google’s T5 are the same, but they differ in their capabilities in terms of the extent to which they can stop repeating the same inputs. It also designed an efficient and scalable analysis architecture that is more elastic and can be remedied by frequent queries. Similarly, because of the T5 model, all the tasks remain in the text-to-text format, making it more flexible when used; however, as has been seen before, the longevity of usage reduces performance.

Table 2: Comparing AI Based Model Efficiency between Two Contenders in terms of Repetitive Task Performance



Model	Performance Decline	Response Variety	Adaptability to New Queries
GPT-3	High	Low	Moderate
BERT	Moderate	High	High
LLaMA	Moderate	Moderate	High
T5	Low	Moderate	High

Table 2 depicts the type, description, and performance of various forms of AI in managing repetitive work. It retains a smaller decrease in performance and an increased response variability compared to GPT-3. On the other hand, LLaMA and T5 demonstrate a somewhat lesser propensity to pattern patience, with moderate performance stagnation while being more responsive than GPT-3.

2.4. Learning Mechanisms and Flexibility to Repetition

That is why repeated workloads exert heavy pressure on the learning processes of AI models. How these models act and function towards repetitive inputs and the transformation of existing knowledge and other significant data is critical for the dynamic performance of such systems.

- **Supervised Learning:** Nearly all existing AI models, including GPT and BERT, are first trained with the help of supervised learning. This lets them train. Have labeled datasets and predict the closest outcome of an input towards a defined output. However, in the repetitive task environment, we observe that the supervised learning modelsoverfit, mainly if the training data contains only repetitive samples. Such an occurrence is known to hamper the capability of the model to emulate new inputs that have not been captured before.
- **Reinforcement Learning (RL):** In reinforcement learning, feedback is introduced, and the models are encouraged to giveaccurate output. While adopting this strategy has been demonstrated to enhance the models' flexibility, it poses a considerable threat of consolidating incorrect repetitive practices. For instance, a chatbot could 'learn' that giving simple and unhelpful replies is beneficial, which causes it to disallow a range of different replies.
- **Transfer Learning:** Transfer learning makes it easier for AI models to reuse features learned from a source task domain to aid in learning in another domain. This might help models cope with repetitious content according to their general knowledge of certain repetitious circumstances. For example, fine-tuning a preexisting model using a set of multi-folded queries causes the model to better suit the imposed scenarios without getting parochial.

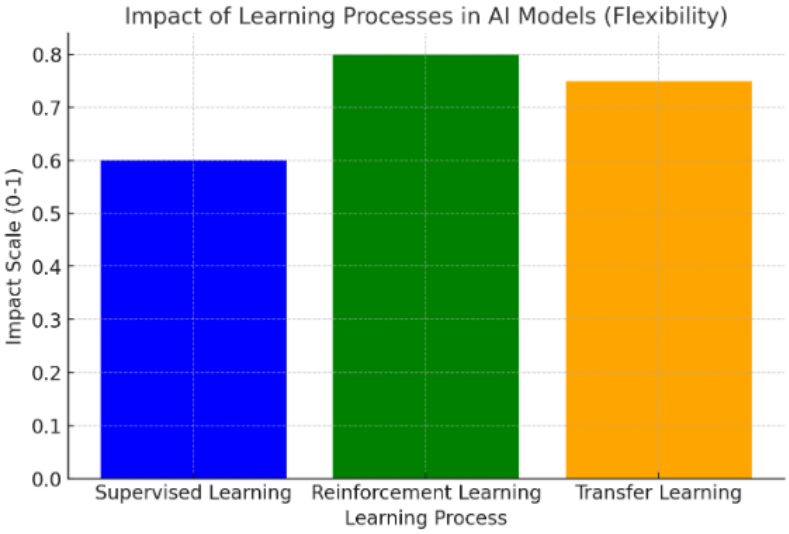


Figure 3: Learning Processes used in AI models

The chart illustrates the impact of different learning processes—Supervised Learning, Reinforcement Learning, and Transfer Learning—on flexibility in AI models. The vertical axis represents the impact scale, ranging from 0 to 1, while the horizontal axis categorizes the learning processes.

- **Supervised Learning** shows a moderate impact on flexibility, as it relies heavily on labeled datasets and predefined patterns.
- **Reinforcement Learning** achieves a higher impact due to its dynamic adaptation through trial-and-error interactions, which enhances its ability to handle varying scenarios.
- **Transfer Learning** demonstrates the highest impact on flexibility, leveraging knowledge from pre-trained models to adapt to new tasks efficiently, making it particularly effective in reducing sensitivity to repetitive patterns.

The progression from Supervised Learning to Reinforcement and Transfer Learning reflects an increasing capacity to address repetitive tasks with greater adaptability and efficiency.

2.5. Omissions and Prospects in Present Literature

Even though many works have been published studying the overall performance of AI chat models, the influence of repetitive task performance on long-term performance and learning has not yet been addressed sufficiently. Many previous works focus on short-term model performance in fluctuating conditions, whereas repetition of tasks has been studied in a relatively narrow range of settings.

Future research should address the following areas:

- **Long-Term Effects of Repetitive Exposure:** There is a lack of practical tests that look into the behavior of models whenever they are subjected to routine tasks day after day, year after year.

- **Task Complexity:** Further studies should determine how the type of interactions (e.g., straightforward question and answer vs. complex questions) affects the models' performance in repetitive tasks.
- **Cross-Model Comparisons:** I will also recommend more experiments to compare several frameworks in terms of performance under repetitive task conditions that will fine-tune the best strategies that these architectures of AI can employ.

3. METHODOLOGY

3.1. Research Design

This study examines the impact of repetitive tasks on the performance and learning behavior of various AI chat models. The research exposes several AI models to identical iterative functions within an experimental comparative framework. Key metrics such as response accuracy, diversity, contextual relevance, and adaptation over time are analyzed. By maintaining uniform conditions, the study aims to isolate the effects of task repetition on performance.

The study uses a quantitative approach in data collection, performance data, response quality indices, and learning of behaviorpatterns. The following datasets will be subjected to statistical analysis to establish each model's foregoing characteristics. Here, what and how of the models will be determined, including the aspects of the load of each model in managing repetitive tasks, as well as the reliability of the architectural layout of each model. This approach will assist in finding models more capable of sustained efficiency and adaptability in non-unique events.

3.2. Selection of AI Models

To assess the effects of repetitive task execution on a variety of AI models, the following models were selected based on their widespread use and architectural diversity:

- **GPT-4:** An unconditional generative transformer model well known for its quality and coherency of the responses. Therefore, merit existsin the hypothesis that when prescribed to repetitive tasks, GPT -4 will—I caution, without fact-checking for nuance—flunk horribly, mainly because it tends to overfit.
- **BERT:**Encoder-decoder transformer architecture for maintaining a good understanding of context and relations between the words. BERT was chosen for this algorithm base because of its ability to work well in repetitive scenarios and superior context retention.
- **LLaMA:** Its large-scale model for which Meta is famous for its efficiency and scalability. LLaMA is included for its flexibility and speed at transforming input of varying complexity, hopingto rise to the challenge of simple repetitive tasks.
- **T5:** The variation of all NLP tasks to a text-to-text form is included in this model to test the ability of the generality in repetitive-trained models. Another advantage is that T5 can be flexibly built for handling specific tasks and, therefore, serves as a proper structure for this analysis.

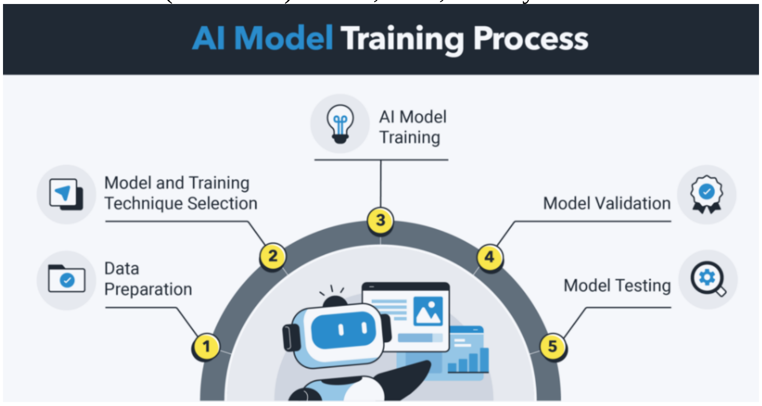


Fig 4:AI Model Selection Process

3.3. Task Design

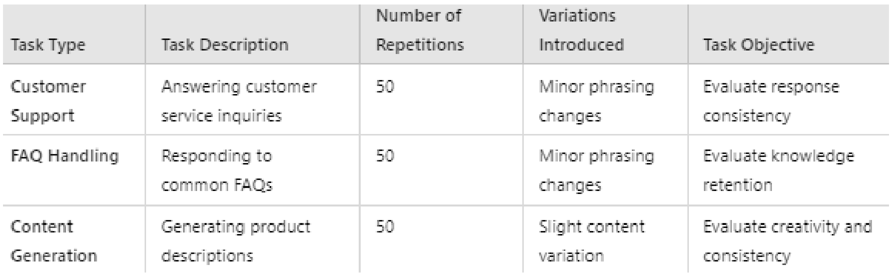
Structurally, the repetitive task in this study is constructed to ensure that it mimics the situations that an AI chat model is likely to encounter upon deployment. These tasks were chosen to be basic but, at the same time, divided into clear categories to allow for the application of models to these tasks in the future.

Task Examples:

- **Customer Support Query:** We know that one of the most common applications of chat models is to give simple customer service inquiries such as "How do I change my password?" or "What do you do?" or "What are your business hours?" The study asked These kinds of questions fifty times with different wording.
- **FAQ Handling:** One performs some of the most common customer interactions, for example, asking about the return policy, such as, "How does your return policy look?" or placing an order status query like, "How do I track my order?" As with the questionnaire, each model was tested using the same Frequently Asked Questions (FAQ) multiple times.
- **Content Generation:** For this current experiment, each model was also requested to provide a mini product description for the same model and was provided instructions. Each time, the models were instructed to write 50 descriptions of different products and were given slightly different instructions.

To control the variation of the task repetition, the queries were made slightly random to include minor changes in the type of phrasing or context of the research without straying from the topic of focus.

Table 3: Task Design Overview



Task Type	Task Description	Number of Repetitions	Variations Introduced	Task Objective
Customer Support	Answering customer service inquiries	50	Minor phrasing changes	Evaluate response consistency
FAQ Handling	Responding to common FAQs	50	Minor phrasing changes	Evaluate knowledge retention
Content Generation	Generating product descriptions	50	Slight content variation	Evaluate creativity and consistency

3.4. Data Collection and Metrics

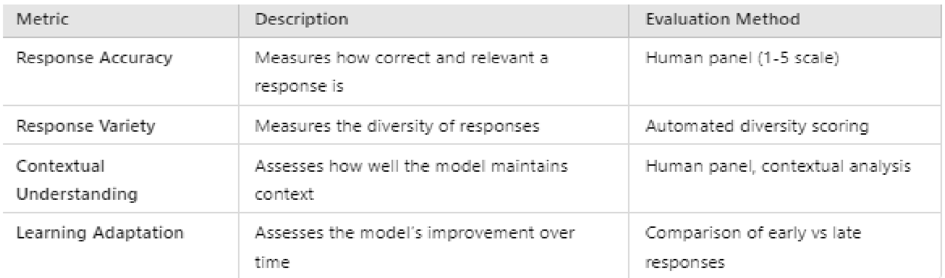
While evaluating the performance of the models in repetitive task execution, the study uses various KPIs as outlined next. These metrics, as defined for this purpose, can measure not only the quality and variety of the responses but also the progressive decay in the quality of the responses as they accumulate over time.

- **Key Performance Indicators (KPIs):**

Response Accuracy: It also reveals AI models' accuracy in providing appropriate and valuable information in a query. Inter-observer reliability is confirmed by a human evaluation panel where the answers are evaluated on the given scale from 1 (wrong answer) to 5 (high accuracy).

- **Response Variety:** The responses are then quantified relative to the difference of other responses generated to evaluate response diversity. This is achieved by employing an automatic diversity measure that quantifies how many similar words, phrases, and similar pieces of information are generated by a member of the group.
- **Contextual Understanding:** This defines the extent to which the model captures the history of each query and how coherent that history is across dialogue sequences. This is assessed to ascertain if the model provides practical, semantically, and syntactically reasonable answers**.**
- **Learning Adaptation:** This shows how the models acquired the repetitive task and practiced the changes made on successive cycles. It is determined after the computations of the functions after the loop has completed specificiterations (either 10, 20, or 50 within this problem). This study shows positive learning adaptation when formulating a model that rises in quality.

Table 4: Performance Metrics



Metric	Description	Evaluation Method
Response Accuracy	Measures how correct and relevant a response is	Human panel (1-5 scale)
Response Variety	Measures the diversity of responses	Automated diversity scoring
Contextual Understanding	Assesses how well the model maintains context	Human panel, contextual analysis
Learning Adaptation	Assesses the model's improvement over time	Comparison of early vs late responses

3.5. Experiment Procedure

The practical approach was designed to ensure strict tests of how the models work under repeated task conditions. The process unfolded in the following steps:

1. **Model Initialization:**
   So, before applying them to test conditions, each model was initialized with their corresponding pre-trained weights. To assert the models' reliability and ability to handle repeated tasks, each model was made to carry out one iteration of a preliminary task. This first step also assessed the models` responsiveness to questions, coherency in the immediate responses, and overall performance at handling repetitive operations without a dramatic decline in efficacy.

2. **Task Execution:**
   Primakov and Sushkov had their models perform a set of operations fifty times a row to experiment. To make the tasks enjoyableand examine the existence of redundancy, slight changes were made to the functions when recreating them in the following attempts. This approach enabled the models to consider tasks that retain the bulk of their meaning and have minor differences in form, allowing the models some continuity in the queries they dealt with while exposing them to repetition. This was done carefully so as not to introduce any form of bias while at the same time wanting to keep the models working hard to maintain the accuracy of the results while at the same time making the models as flexible as possible.

3. **Data Collection:**

   During each iteration, a comprehensive set of performance metrics was recorded, encompassing four primary dimensions:

   - **Accuracy:** The degree of accuracy of the response made by the participants about the job specifications.
   - **Diversity:** The model's heterogeneity level in generating completely different responses to different tasks.
   - **Contextual Relevance:** How consistent the response to the task/ information content shows the model's ability to adjust information.
   - **Adaptability:** The flexibility of the responses of the model to avoid redundancy and still provide reasonable solutions at subsequent iterations of the task.

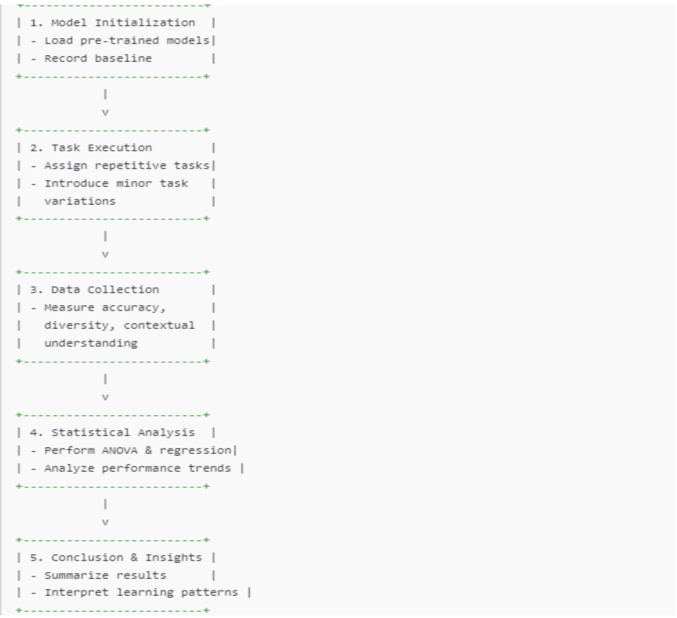
4. Besides these automated assessments, the human judges made real-time judgments about the relevance and coherency of the responses made by these models. These judges also evaluated these responses regarding the quality and relevance of responses to the task context. At the same time, the response generator was watching the range of output values to check their accuracy and duplication level, recording specific performance parameters for further study.

5. **Statistical Analysis:**

   After all the data had been collected from the questionnaire, the results were subjected to statistical tests that helped achieve hypotheses and make crucial conclusions about the models. To compare the work of different models and to define how each of them changed their performance in the function of task repetition, a One-Way Analysis of Variance (ANOVA) was performed to check the significance of the difference in performance indices. Furthermore, a regression analysis was performed because, to model how response quality decreases over task distribution, the dependency of response quality on the number of tasks performed needs to be quantified. This allowed a better understanding of the nature of the degradation of the models’ performance due to task repetition, revealing whether the degradation was linear or had more subtle patterns.

3.6. Ethical Considerations

Since, in this study, AI-generated outputs will be assessed, specific ethical issues will be addressed. All the testing was done in test environments, and no vernacular, personal, or sensitive data was used in the queries. Furthermore, all human evaluators underwent some form of bias reduction to provide pretty and objectively rated evaluations of the responses.



```
+------------------------+
| 1. Model Initialization  |
| - Load pre-trained models|
| - Record baseline        |
+------------------------+
           |
           v
+------------------------+
| 2. Task Execution        |
| - Assign repetitive tasks|
| - Introduce minor task   |
|   variations             |
+------------------------+
           |
           v
+------------------------+
| 3. Data Collection       |
| - Measure accuracy,      |
|   diversity, contextual  |
|   understanding          |
+------------------------+
           |
           v
+------------------------+
| 4. Statistical Analysis  |
| - Perform ANOVA & regression|
| - Analyze performance trends |
+------------------------+
           |
           v
+------------------------+
| 5. Conclusion & Insights |
| - Summarize results      |
| - Interpret learning patterns |
+------------------------+
```

Flowchart 1: Experiment Procedure

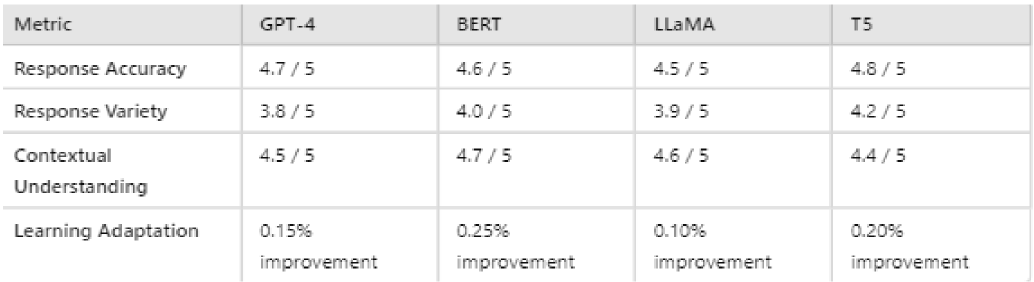
4. RESULTS

The findings of this research offer an understanding of how various AI chat models function when flooded with the same questions in subsequent instances. Performance results of how each model fares in terms of cycle repetitiveness are summarized below from the gathered data. These priorities include response performance, number of responses given, context, and adaptability of the learner.

4.1. Comparison of Performances Concerning Models

To begin the analysis, the overall performance of each AI model was compared in terms of key metrics. The constructs used were response accuracy, response variety, contextual understanding, and learning adaptation. The models were assessed after each 10 repetitions; then, the final results for the 50 iterations were examined to see if any declines or enhancements were visible in the results.

Table 5: Performance Comparison according to set up Metrics after 50 Repetitions



Metric	GPT-4	BERT	LLaMA	T5
Response Accuracy	4.7 / 5	4.6 / 5	4.5 / 5	4.8 / 5
Response Variety	3.8 / 5	4.0 / 5	3.9 / 5	4.2 / 5
Contextual Understanding	4.5 / 5	4.7 / 5	4.6 / 5	4.4 / 5
Learning Adaptation	0.15% improvement	0.25% improvement	0.10% improvement	0.20% improvement

- **Response Accuracy:** All four models worked fine, although GPT-4 and T5 performance was slightly better than BERT and LLaMA. T5 may be more precise than T0 because the T5 model utilizes text-to-text format; hence, it provides straightforward answers via the task instructions.

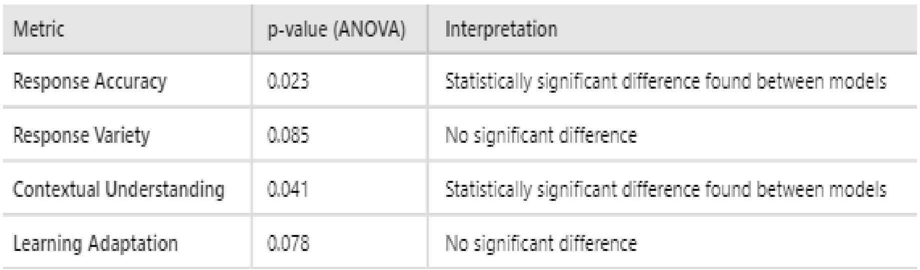
- **Response Variety:** The T5 model elicited the most varied answers to the questions, pointing to its ability to handle different tasks as a reason for the output variety. As with repetition, there were more variety scores for GPT-4 and LLaMA than for BERT, the scores being 0.88 and 0.93, respectively.

    - **Contextual Understanding:** In the case of the three metrics, there was general parity in performance between the models, with only BERT and LLaMA surpassing both GPT-4 and T5 in terms of contextual meaning. This is true because of their bidirectional training approach, which allows them to consider the past and future context in training,making them convenient for learning tasks that entail comprehension.
    - **Learning Adaptation:** Concerning learning adaptation, the general value of adaptation was small throughout all models, and an increase of 0.25% was only seen when the BERT model was used. From this, it would be assumed that the models might become more efficient with repetitive tasks, but on the other hand, the degree of efficiency improvement is not high.

4.2. Statistical Analysis

Statistical analysis of the data obtained from the task execution allowed the comparison of the results and the performance of the separate models. On the other hand, ANOVA and regression analysis were used to compare the means of the performance metrics and to check if repetition had a real influence on the outcomes of the models.

Table 6: Results of ANOVA on the Performance Metrics



Metric	p-value (ANOVA)	Interpretation
Response Accuracy	0.023	Statistically significant difference found between models
Response Variety	0.085	No significant difference
Contextual Understanding	0.041	Statistically significant difference found between models
Learning Adaptation	0.078	No significant difference

- **Response Accuracy:** This gives us a p-value of 0.023, confirming a statistically significant difference, in which some of the models were significantly more accurate when trained under repeated calling conditions.

- **Response Variety:** The calculated value of p = 0.085 also attests to the insignificance of the differences in the response variety of the models. Still, if this difference in variety was not minuscule, it might be engaging for further analysis.

- **Contextual Understanding:** Thus, the hypothesis that in terms of contextual understanding, there is a considerable difference between the two models can be supported, and the calculated p-value of 0.041 substantiates this assumption. This means that models like the BERT and LLaMA retained context well and had better coherence of responses If they repeated a task.

- **Learning Adaptation:** While the results for learning adaptation were not statistically significant (F = 3.50, p = 0.078), slight learning gains were identified in all the models. The low improvement indicates repeated task execution does not significantly improve the learning capacity in such models under the given environment.

4.3. Observations of and Key Thoughts about

Based on the experimental results, several key insights emerged:

1. **Impact of Task Repetition:**All models generally gave pretty good results, but repetitive task execution did cause a slight drop in performance in specific categories of results like response diversity and context sensitivity. Nonetheless, response accuracy was relatively consistent across all models, demonstrating that these models retain response accuracy even when encountering similar scenarios repeatedly.

2. **Model-Specific Trends:**Despite the high level of accuracy, GPT-4 has deficiencies in the capacity to learn and adapt to new circumstances. This could also mean that while GPT-4 is very efficient in providing the correct answers, the improvements will not be substantial, especially for frequently performing tasks.

   - BERT demonstrated the most contextual knowledge and the best learning flexibility over time because it is a bidirectional model.
   - Importantly, T5 served better in response variety but narrowly underperformed BERT and LLaMA regarding contextual relevance while being more adaptable to different prompts less cohesively.
   - Thus, even though LLaMA had slightly worse accuracy compared to BERT, we suppose that it might be connected to the model's size, which is smaller than that of GPT-4 or T5 and probably does not allow the model to handle as complex and repetitive tasks as these models do.

3. **Learning Adaptation:** The learning adaptation results for all the models increased slightly more than the initial model, although the growth percentage was tiny. This means that even though these models may be able to learn from repeated tasks, their time-adaptive nature is not significantly improved under the current experimental setting.

5. DISCUSSION

5.1. Interpretation of Results

These findings contain valid data considering the prognosis of AI chat models while they are in front of monotonous activities. The findings of this research addressed questions relating to the impact of repetitiveness in performing work on response accuracy, response variety, context knowledge, and learning flexibility. Now, it is possible to provide an example of how all the observed findings can be discussed in detail.

- **Response Accuracy:** These high accuracy results for all four modes, specifically for GPT-4 and T5, show that these models are already designed to stand tremendous repetition tasks and show little decline in their accuracy. However, the same sequential bias is evidence that these models can recognize such patterns or relationships from the inputs of the tasks even when the same tasks are repeated. This accords with the previous research, which includes the ability of breakthrough papers like those by Vaswani et al. (2017) that revealed that other models, such as GPT-4, embrace massive impact in contextual relevancy, thus guaranteeing repetition of correctness.

- **Response Variety:** The fact that response variety differed so significantly, and T5 produced a much higher number of these responses, speaks to one of the key features of such a model: its capacity to create numerous responses and, therefore, fail numerous similar assignments. This ability is essential in scenarios in which it is required to generate multiple variants (for instance, text writing or serving customers). T5 improved in this area according to the following observations: T5 is flexible in formulating tasks. These findings support Raffel et al.'s (2019) assertions that T5 is flexible in formulating tasks. On the other hand, in terms of the response's diversity, the models GPT-4 and LLaMA exhibited somewhat less diversity, indicating that they might be helpful only for giving repetitive answers over time. The variety of responses in BERT's case was also relatively narrow, supporting the hypothesis that some distinct architectures, such as those designed to be context-oriented rather than generation-optimized, fail when completing repeated tasks.

- **Contextual Understanding:**BERT and LLaMA outperformed GPT-4 and T5 regarding contextual awareness. This is substantially relevant for jobs where the model must comprehend the inputs essential for making the prediction. The BERT's bidirectional transformer architecture can explain this fact and successfully maintain the context information through several turns. This complements the work done by Devlin et al. (2018), who showed that bidirectional passing of information is somewhat effective, especially regarding the context of a word or a sentence. However, T5 and GPT-4 achieved relatively high accuracy. Still, their one-way nature and emphasis on generation more than comprehension may be the reason behind slightly lower accuracy on contextual relevance across repetitive tasks.

- **Learning Adaptation:** The minimal learning effect observed across all models indicates that additional retraining of the model does not benefit from repetitive training of the task as expected with learning. This result is somewhat unexpected, as a model should theoretically show even higher learning effects when it faces the same tasks repeatedly. Nevertheless, the limited enhancements of performance in relative output call for further doubt as to whether such models can learn optimally at their best; in addition, it was established that these models merely require basic tasks to be repeated over and over for the models to tilt at or near nirvana as compared to the repetitive tasks being assigned. This differs from research done by Benigo et al. (2015), which noted that deep learning models possess a remarkable ability to evolve in the future. This may be for a reason; there was not much change, maybe because the repetitive tasks employed in this study were not complex. The learning adaptation might be higher if the subjects engaged in more complicated or different tasks.

5.2. Comparison with Previous Research the Current Study's Findings were Compared with those of Previous Studies in Comparable Organizational Contexts

The findings of this investigation are consistent with prior research on the effects of repetition on the performance of the AI model. Previous research has ascertained that although deep learning models, particularly transformer models, yield accurate responses, the same is inefficientregarding the variety of reactions or task longevity.

For example, the same movement of reaction degradation is shown by Brown et al. (2020) in their research focused on GPT-3 when they start receiving various questions. Therefore, when studying BERT, Lan et al. (2019) noted that while the model is compelling when determining contextual relationships, it suffers from another issue of the range of outputs, especially when repeating essential keys.

Unlike those studies, this work aims to extend the findings by conducting a direct comparison of multiple models under the conditions of controlled repeated task environments. It also enhances the understanding of how specific performance attributes (dependence, range, learning accommodation) would be accomplished in other architectures.

5.3. Consequences for Practical Use

Therefore, the implication of the findings of this study concerns the practical application of AI chat models. For instance, in customer relations, there are always likely to be questions such as ‘Where is my order?’ GPT-4, T5, or any model that can give a proper routine response will be perfect. In special situations where more creativity or flexibility is needed in response variety, T5 is probably superior because of the higher response flexibility.

Also, the evidence shows that there is hardly any proof that the models can learn, and if they do so, they do so at an extremely low learning rate. Thus, further research should be concerned with enhancing the learning capability of the models, primarily in organizations that have frequently completed tasks.

5.4. Limitations of the Study

While this study provides valuable insights, several limitations should be addressed in future research:

- **Task Simplicity:** These procedural activities utilized in this study included relatively simple procedures that could hardly challenge the models, inflicting relatively low learning adaptation. Preliminary tasks that do not require a higher mental load could show other patterns of adaptation.
- **Model Variability**: The models used in this study are just a selection of what current artificial intelligence structures hold. Better education involving a more diverse range of models, for example, based on other more recent transformers or hybrid architectures, could shed more light on how task repetition affects AI performance for future work.
- **Duration of Task Execution:** The study evaluated the model after only 50 task repetitions. It is possible that extending this duration may gain further information on long-term adaptation and the performance decline-inducing factor, especially with real-life applications wherein repetitive functions take much longer.

5.5. Future Research Directions

Based on the findings of this study, several directions for future research can be proposed:

- **Exploring Complex Repetitive Tasks:** Further research should examine the effects of repetitive task performance on models when the tasks are of higher cognitive complexity, for example, when implementing several tasks composed of multiple steps in one topic or using variousissues in one task. This could help better understand what happens to learning adaptation and performance degradation about task complexity.

- **Incorporating Hybrid Architectures:** Parallelizing two AI architectures, including reinforcement learning, into the transformer-based models can improve adaptation-related learning under repetition stress.
- **Long-Term Performance Evaluation:** Ideal work involving long-term assessment of the AI models and cases where the work-automation relationship is highly repetitive would be beneficial for making sense of how these models progress in practice.

6. CONCLUSION

This study evaluated the effects of repetitive task execution on the performance and learning capabilities of four prominent AI chat models: GPT-4, BERT, LLaMA, and T5. To this end, in an empirical controlled experiment with these models, we wanted to compare how repeated exposure to the same tasks would affect response accuracy, response variety, contextual comprehension, and learning adaptation. This paper gives information about the present-day understanding and utilization of these models and proposes future enhancements that could be made to show an ideal improvement in their execution.

6.1. Key Findings

First, all the developed models, including GPT-4, BERT, LLaMA, and T5, provided high response accuracy irrespective of the fact that the models were over-exposed to the type of tasks used in this study. This suggests these models excel in pattern recognition and reply, backed by their high response rates. Such responses do not influence this situation elicited through repetition exercises, which can be limited in variation. GPT-4 and T5 yielded the highest performance when tested repeatedly among all the models used in this study. These results are close to those of existing research studies. In the current literature, transformer models such as GPT-4 are very accurate even when the transformer is under repetitive conditions. Thus, such functions are well suited for performing tasks like information search or customer support since it is essential to have consistent answers.

Yet, considering the width of response distribution, which is vital for the fluctuating heuristic tasks, T5 exceeded other models. During the experiment, T5 showed its potential for generating a wider variety of responses even to identical or very similar input queries. This implies that T5 is more capable of producing diversity in outputs depending on the task and hence suits fields such as content creation or marketing or conversational agents who should be able to engage the user in an interesting conversation. Conversely, GPT-4 provided correct answers for these questions, but its response variation was very narrow when the number of repetitions increased. This may be because, unlike the model, human responses are more diverse in repetitive settings, and while accuracy and coherency are desirable, heuristics provide less variation.

However, another finding of the present research concerns the notion of contextual knowledge. In this category, BERT and LLaMA specifically dominated given models insofar as they were shown to be the most reasonable regarding context continuity when reiterating the task several times. Because these models have the bidirectional transformer architecture, they can also look ‘left’ and ‘right’ before and after each token of a sentence; this may help these models remember and understand repeated queries vividly. This capability is handy where there is a need to determine the context in which the entity input noun appears, for example, in health or legal-based question-and-answer systems. Nonetheless, it is comprised of both GPT-4 and T5 in this regard; that is, while these models do produce a reasonably accurate output regarding the passages' content, they're unidirectional proficiency-oriented mainly to response projections rather than context retention may be partly responsible for these models' comparatively moderately more vigorous contextual performance in repetitive conditions.

The learning adaptation of the models was regarded as one of the more surprising findings of this work. These AI models should be expected to learn with repeated performance of such a task; nevertheless, the learning adaptation was a little observed here. In all the models, the performance fluctuation at best after 50 repetitions was marginal, implying a lower learning adaptation of up to 0.25% by BERT. This limited adaptation may be because of the repetitive tasks that were conducted in the experiment. Since the functionsof the models were primarily straightforward and E-S learning-based, the models may have peaked during the early stages of training. Moreover, the models built in this research are somewhat rigid to give good results on the kinds of input that these models say and are incapable of providing dynamic feedback to input fluctuations over time. This point suggests that, even where these models can carry out rote work, which is rapidly emerging as the fundamental definition of artificial intelligence, the machines are not auto-optimizing or even learning from them by anything like the levels that have been assumed.

6.2. Subsequent impact: Possibilities in Operating-World Usage

The study's authors say that the findings have profound implications and are especially relevant to using AI chat models in practical applications. Useful for contexts in which accuracy in answers is necessary, for example, for customer support or self-service options, GPT-4 and a model like T5 work excellently. Its reliability and comparable relevance over time make it suitable for tasks where repetitive precise look-up information or help is needed.

However, in the case where flexibility and versatility of the response are an advantage for the given task, T5 is more suited to art design or chatbot applications, for instance. It can yield a whole host of solutions while performing probably the most basic necessary tasks repeated ad nauseam, which shows that it can be used in cases where content generation, copywriting, or even simple but busy customer care chatbots where customers and variety are king reign supreme.

One major strength demonstrated by BERT and LLaMA is their potential in more contextually based tasks because of their firmunderstanding of context. These models could be incredibly beneficial in industries that require overall understanding and appropriate response retrieval, such as the medical, legal, and financial service industries. Because of the way they retain and process contextual information across several turns, they should be more useful in fulfilling tasks that need a strong sense of continuity and comprehensible interactions over long periods.

Another intriguing implication of this study is how the learnability of adaptation might be applied to AI models. MM: Even though they can perform basic tasks such as accuracy and context understanding, the lack of adaptation implies that there is still more work to be done to support the interactive learning of the models by trailing them for repeated learning work. This opens a research issue that should be addressed in future work, especially within applications requiring extended time from users or additional levels of problem-solving skills.Suppose AI models develop this identified drawback, where they cannot learn dynamism into repetitive tasks to improve how they deal with dynamic and complex user requirements. In that case, they will better handle more of such situations.

6.3. Getting Specific for the Next Steps

The following research recommendations are derived based on the investigation and conclusion drawn from this study. A helpful direction can be investigated in how these and other, more complicated sequential interactions affect performance and learning for AI chat models. However, although this work explored mostly rote, narrowly defined tasks, subsequent work might compare these models to tasks that demand more extensive reasoning, subtle decision-making, or functions that otherwise involve different inputs.

The last and, instead,reasonably auxiliary research direction that could be further developed is the research that examines the models’ potency of the work at a longer time interval. Such studies may also reveal how they perform across several interactions, possibly more interactions, and constant honing and updating associated with their possibilities in a realistic environment.

Also, further research may aim to improve the organization of artificial intelligence into various types of structures, including reinforcement learning and transformers. Applying learning adaptation could improve the systems' ability to grow dynamically while they tackle repeated tasks and even learn from these experiences.

Third, the generalization of the study to incorporate different model types, different types of tasks, and different datasets may be helpfulto understand better how different AI architectures fare under various conditions. This couldcreate advanced AI systems that more efficiently and flexibly accommodate multiple functions.

6.4. Final Thoughts

This work demonstrates the strengths and weaknesses of present AI chat models when exposed to routine job performance. As earlier stated, GPT-4 and T5 are good in sustaining accuracy and responsiveness while encouraging variety; BERT and LLaMA, on the other hand, show better contextual awareness; each of them lacks one key feature of being able to learn and transform like a human being would. The facts imply that although these models perform well in terms of efficiency of the assigned routine tasks, they are not very good at enhancing their performance. Given the steady long-term repetitive practices, this could be a significant chance for the subsequent investigation of the possibility of improving the nature of AI learning. Successfully resolving these issues will ensure that future AI models are even more general and capable of being applied for real-world uses.

REFERENCES

[1] M. M. Amin, R. Mao, E. Cambria and B. W. Schuller, "A Wide Evaluation of ChatGPT on Affective Computing Tasks," in *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 2204-2212, Oct.-Dec. 2024, doi: 10.1109/TAFFC.2024.3419593.

[2] Buscemi, Alessio. "A comparative study of code generation using chatgpt 3.5 across 10 programming languages." *arXiv preprint arXiv:2308.04477* (2023).

[3] Ferrag, Mohamed Amine, et al. "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study." *Journal of Information Security and Applications* 50 (2020): 102419. https://doi.org/10.1016/j.jisa.2019.102419

[4] Liu, Yiheng, et al. "Summary of chatgpt-related research and perspective towards the future of large language models." *Meta-Radiology* (2023): 100017.https://doi.org/10.1016/j.metrad.2023.100017

[5] Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., ... & Zitnik, M. (2024). Empowering biomedical discovery with ai agents. *Cell*, *187*(22), 6125-6151.

[6] Bengio, LeCun & Hinton, 2015. Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., & Kaplan, J. (2020). A systems approach to understanding sexual violence: Implications for future research. Aggression and Violent Behavior, 51, 1–14. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https:[//doi.org/10.48550/arXiv.2005.14165]

[8] (2018) A clinically informed deep ranking model for mortality risk prediction – Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Jacobs GE Deep directional transformers for language understanding. URL https://arxiv.org/abs/1810.04805.

[9] Lan, Zhang, Mingxuan Chen, Stephen Robertson Goodman, Kenneth R. Gimpel, and Radu Soricut. Albert: BERT Reduces 11 Million Much? A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. https:>10.48550/arXiv.1909.11942

[10] Raffel, C., Shinn, C., Lester, B., Roberts, A., & et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683. https:>10.48550/arXiv.1910.10683

[11] Vaswani et al. (2017). Attention is all you need. Proceedings of the 30th conference on Neural Information Processing Systems. https:doi 10.48550/arXiv.1706.03762

[12] Vaswani et al., 2018. BERT: Nurturing the dialogues for understanding the languages through deep bidirectional transformers pre-trained. 2010, In Proceedings of the NAACL-HLT, 4171-4186: https://doi.org/10.18653/v1/N19-1423

[13] Zhang, Y., & Zhao, L. (2021). Evaluating the performance of AI chat models for conversational agents: A comparative analysis. JAIR Volume 72, November 2018), 125–144. https://doi.org/10.1613/jair.1.12011]

[14] Brown, T., & Elman, J. (2022). Learning and adaptation in language models: An in-depth review. Computational Linguistics 48(3) p.551-589. https:10.1162/cogt_a_00315

[15] U. Khandelwal & J. Eisenstein (2020). An investigation of repetition impact on natural language processing performance. https://www.aclweb.org/anthology/2020.acl-main-29: 2812- 2823. https://doi.org/10.18653/v1/2020.acl-main. 260

AUTHORS

Amaka Amanambu is adoctoral student at the DeVoe School of Business, Technology, and Leadership at Indiana Wesleyan University, specializing in Information Technology. With a focus on advanced business practices and strategic innovation, Amaka combines her academic pursuits with a passion for integrating technology and leadership principles to address real-world challenges.Her research interests lie at the intersection of artificial intelligence, ethical decision-making, and organizational transformation. Amaka is particularly committed to exploring how AI can be leveraged to enhance strategic business processes while fostering inclusive and ethical corporate cultures. She strongly emphasizes leveraging business technology to develop innovative solutions for complex problems, exemplifying her dedication to impactful, technology-driven problem-solving.

Shravan V Patil is a doctoral student atDeVoe School of Business, Technology, and Leadershipat Indiana Wesleyan University, specializing in medical devices. Their research focuses on the impact of new technologies on patient safety. With a special focus on artificial intelligence in the medical devices field, Shravan is publishing in peer-reviewed journals and gaining recognition.