Foundation University
Journal of Engineering and
Applied Sciences

FUJEAS
Vol. 4, Issue 2, 2023.
DOI:10.33897/fujeas.v4i2.873

Research Article

Potential of Large Language Models (LLMs) as Supplementary Tools for Historical Learning: Users' Interaction and Knowledge Acquisition

Muhammad Hasnain [a], Sardar Usman [b]

[a] Department of Computer Science, Lahore Leads University, Lahore, Pakistan
[b] Department of Computer Science, Grand Asian University, Sialkot, Pakistan

* **Corresponding author**: drhasnain.it@leads.edu.pk

Abstract:

This study explores the strengths and limits of large language models (LLMs) in exploring the information on history, an area unexplored in the existing literature. ChatGPT and Gemini, as LLMs, have demonstrated superior performance in education, healthcare, and business. This study proposes utilizing the ChatGPT (ver. 3.5) and Gemini applications to acquire information on historical figures like Sher Shah Suri and Mughal Emperors and Sikhs in the subcontinent. To evaluate the proposed study, this study used two data sets: the first data set comprised a set of questions (n = 26) and the second data set contained questions (n = 35). The results indicate that ChatGPT provides concise answers to the questions of both datasets compared to the Gemini application. However, Gemini exhibited a higher accuracy (92.30%) than ChatGPT with accuracy (76.92%) for dataset 1. For the dataset 2, ChatGPT showed better accuracy (68.57%) than Gemini with accuracy (65.71%). Further research could expand on this study by employing additional artificial intelligence (AI) tools on large-scale datasets from diverse domains.

Keywords: Large Language Models; Historical Figure; Prompt Engineering; Performance.

1. Introduction

Advances in neural networks, deep learning, and artificial intelligence have transformed the way we approach a wide range of tasks, from finance to academia and industry. The ability of AI models to analyze textual, audio, and video data has allowed users to adopt applications such as near-human-level text translation, speech recognition, and image caption generation [1]. Neural networks have been used to address various natural language processing (NLP) problems. However, recurrent neural networks exhibit limitations in capturing distant dependencies in data sequences, such as information occurring at the end or beginning of the text [2] [3]. Recurrent neural networks excel in predicting future values based solely on the basis of past values.

Generative Pre-Training (GPT) as one of the first autoregressive generative models uses the transformer architecture [4]. GPT works by decoding the stack from the original transformer and converts bidirectional self-attention into unidirectional. Therefore, the GPT model is better suited for tasks, including summarization, generation, and translation of text. Variants of the GPT model, such as GPT2 and GPT3, are variants of the GPT model that demonstrate the capability of the models to tackle

Article Citation:
Hasnain et al. (2023). "Potential of Large Language Models (LLMs) as Supplementary Tools for Historical Learning: Users' Interaction and Knowledge Acquisition". *Foundation University Journal of Engineering and Applied Sciences*
DOI:10.33897/fujeas.v4i2.873

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright
Copyright © 2023 Hasnain et al.

Published by
Foundation University
Islamabad.
Web: https://fui.edu.pk/

complicated tasks without performing supervised training on large datasets. 45TB textual data have been used to train the GPT3 model information. As a result, the parameters increased to 175B, producing outstanding results, particularly in the scenario of few shots and zero shots [5].

Consider a scenario where students from various disciplines want to utilize emerging artificial intelligence (AI) applications; facilitating this is a complex task. However, ChatGPT and other large language models (LLMs) have made this scenario increasingly feasible [6]. ChatGPT showed a high success rate to confront the psychosomatic medicine multiple-choice questions [7]. This article explores the potential of ChatGPT and Gemini applications in the area of historical learning.

The contribution of this article is as follows:

- This article contributes in providing insight into the potential of Chatbots such as ChatGPT and Gemini to support historical learning.
- This article contributes to informing educators and learners on the appropriate use of technology as a complement to conventional learning methods.

The layout of the article is structured as follows:

Section 2 presents the materials and methods used in this study. Section 3 provides the results obtained, while Section 4 presents a discussion of these results. The conclusion, Section 5, summarizes the main points of the study.

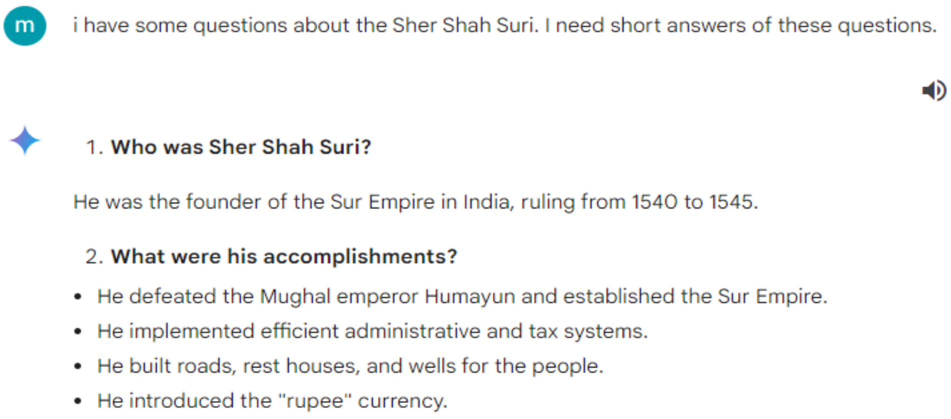
2. Materials and Methods

2.1 History Education Data Sets

To assess the Chatbots' understanding of historical data, two datasets on history education were created in this study. We selected a subset of 26 questions from various online sources, such as testbook.com, on the topic "Sher Shah Suri" and examined the study to assess the performance of Chatbots [8]. Each question had a short answer in one or two lines. Furthermore, we also selected another subset of 35 questions from the central superior services (CSS) exam [9].

2.2 Prompt Engineering

Prompt engineering has shown a significant impact on the output of generative large language models (LLMs). Therefore, we standardized the format of the input data. First, we excluded questions that include images, since ChatGPT version 3.5 only accepts textual data as input. Additionally, we removed tables-formatted questions to allow Chatbots to better understand historical information. Each question was posed as a separate instance to provide a comprehensive understanding. We used both direct questions and multiple-choice questions. An example of direct questions posed and their responses is shown in Figure 1.



i have some questions about the Sher Shah Suri. I need short answers of these questions.

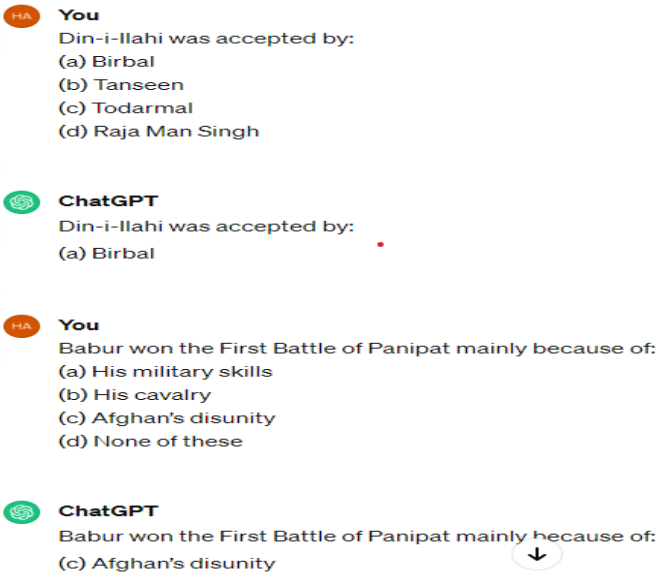
1. **Who was Sher Shah Suri?**

He was the founder of the Sur Empire in India, ruling from 1540 to 1545.

2. **What were his accomplishments?**

- He defeated the Mughal emperor Humayun and established the Sur Empire.
- He implemented efficient administrative and tax systems.
- He built roads, rest houses, and wells for the people.
- He introduced the "rupee" currency.

Figure 1: A template of direct questions posed to the ChatGPT and Gemini LLMs



You
Din-i-Ilahi was accepted by:
(a) Birbal
(b) Tanseen
(c) Todarmal
(d) Raja Man Singh

ChatGPT
Din-i-Ilahi was accepted by:
(a) Birbal

You
Babur won the First Battle of Panipat mainly because of:
(a) His military skills
(b) His cavalry
(c) Afghan's disunity
(d) None of these

ChatGPT
Babur won the First Battle of Panipat mainly because of:
(c) Afghan's disunity

Figure 2: A template of multiple-choice questions posed to the ChatGPT application

Figures 1 and 2 are illustrations of templates used to pose direct and MCQs to Chatbots. However, both Chatbots learned the questions and responded accordingly. However, in some cases, both models provided lengthy and confusing responses. To know the exact answer to the questions, we provide the questions with four options. We directed Chatbots to give us one more closely related option as an answer to a question.

2.3 Overview of Data

The responses generated by both Chatbots were assessed to verify whether (1) the answer to each question was correct; (2) the responses to the questions were consistent with the number of prompts; and (3) the Chatbots provided scientific results about the questions posed. The responses were ensured to be short and accurate for direct questions. It was also checked that the answers to the MCQs were correct and among the list of multiple choices for each question.

2.4 Statistical Analysis

The strength of the models was assessed using the relative proportion of correct answers given by the Chatbots and mapping them with the sources. A value >70% indicated a satisfactory score of correct responses given by Chatbots. However, scores below <70% show below-average performance of Chatbots. We used Excel and Word applications to organize and analyze the data.

3. Results

Table 1 illustrates the performance of two LLMs, ChatGPT and Gemini, on historical data tests examined in this study. The scores for both models are displayed below. Gemini achieved a relatively

Table 1: Correct and incorrect response of two models in data set 1

Model	Correct answered (%)	Incorrect answered (%)	Total
ChatGPT	20(76.92)	6(23.08)	26(100%)
Gemini	24(92.30)	2(7.69)	26(100%)

higher accuracy rate compared to ChatGPT, with Gemini scoring a precision of 92.30% and ChatGPT a precision of 76.92%. Gemini demonstrated superior performance in the selected data set (Sher Shah Suri). Additionally, Gemini provided fewer incorrect answers compared to ChatGPT, with Gemini answering only two incorrectly while ChatGPT answered six incorrectly. On average, Gemini outperformed the ChatGPT application by 15.38%.

Table 2: Responses of two models on data set 2

Model	Correctly answered (%)	Incorrectly answered (%)	Total
ChatGPT	24(68.57)	11(31.43)	35(100%)
Gemini	23(65.71)	12(34.29)	35(100%)

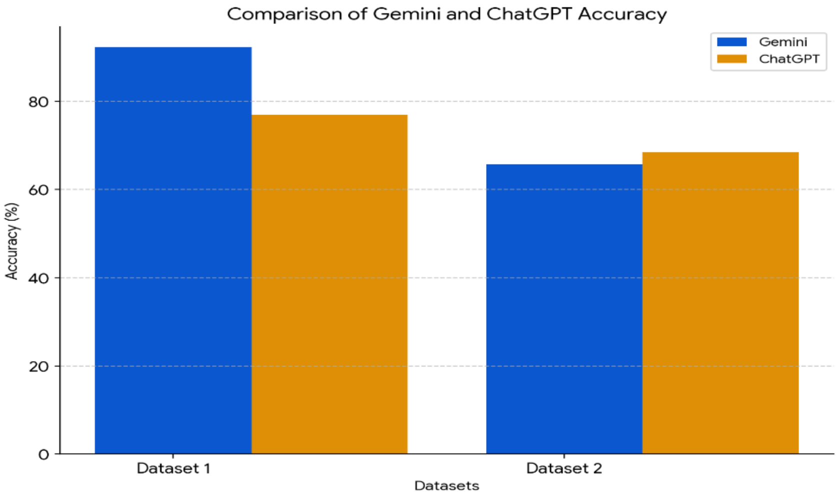


Figure 3: Accuracy comparison of Gemini and ChatGPT models

Although Gemini outperformed another Chatbot in the correct answering of questions in dataset 1, this trend was reversed in dataset 2, where ChatGPT scored slightly better in the correct answering of questions.

3.1. Question Difficulties

When answering questions about our chosen datasets, both LLMs did not encounter significant difficulties learning the questions. For factual questions, our chosen LLMs presented information already available in the training data. For example, answering questions like "Who was his father?" and "When did Sher Shah Suri rule?" involved retrieving established information and presenting it clearly. We instructed both models to provide short answers to the questions. In response, ChatGPT provided concise answers, while Gemini offered more concise information, occasionally providing deeper insights for clarity.

3.2. Correctness and Relevance of the Answers

Generally, the answers provided by the LLMs were scientific and relevant to the queries posed to the Chatbots, and at times, they were very precise and outstanding. ChatGPT provided concise information up to the required standard format (see Appendix A). On the other hand, Gemini provided additional information for answers but did not stray outside the scope of the posed question (see Appendix B).

The additional content was useful for learning and the clarity of information for the learners. We received mostly similar responses from both models in the majority. For dataset 2, we also received consistent answers from two models. In a few cases, both models did not provide the same answers (see Appendix C). The limitations of the models in answering certain questions were identified.

3.3. Quality of the Content (Clarity and Comprehensiveness of the Data Generated by LLMs)

The quality of the responses has been evaluated for clarity, correctness, and comprehensiveness, highlighting the appropriate structure and organization of the texts. For example, when asked about the nationality or origin of Sher Shah Suri, ChatGPT provided a clear response: "Sher Shah Suri's nationality or origin is Afghan."

4. Discussion

These results demonstrate the capabilities and limitations of Chatbots such as ChatGPT and Gemini to reveal information about historical personality. Both Chatbots showed greater access to factual information about historical figures like Sher Shah Suri, Mughal emperors, and Sikhs in the subcontinent. ChatGPT provided precise and short answers to the questions asked compared to Gemini. The Gemini response remained very close to the ChatGPT application. However, the Gemini application showed a limited understanding of the context and nuance. For example, when answering the factual information about Sher Shah Suri, Gemini failed to fully grasp the historical context, complexities, and motivation about historical personalities. Our posed questions were closed-ended and MCQs in the majority. However, we observed that answering an open question requiring critical thinking, such as 'What was Sher Shah Suri's greatest impact on India?' may create difficulty for Gemini in providing insightful responses. In general, the quality of the responses to the posed questions depended on the accuracy of the information on which they were trained. Despite these limitations, both Chatbots are valuable tools for history learners. Chatbots provide efficient and quick ways to collect information about historical personalities, saving time and effort compared to traditional learning methods. The concise information provided by ChatGPT can serve as a starting point for further exploration and research on historical topics.

The literature shows that search engines on the Internet and YouTube provide incorrect answers in certain critical situations to a large number of their users [10, 11] However, this study revealed that ChatGPT and Gemini have correct data at hand, unlike many other Internet sites. It is suggested that it can be used in various other fields. In our study, ChatGPT reached a 72.75% accuracy, and Gemini received a 79.00% accuracy proportion in public exams, where it was most successful, Still, ChatGPT has an error proportion of around 27.25%, and Gemini with 21.00% error proportion. Given that, historians can be more successful in asking questions about their own fields. While ChatGPT is undergoing further improvements in upcoming versions, we can consider it a more appropriate model to reduce the workload on historians to read books for longer times in preparation for their exams.

It can be assumed that the same results can be reproduced by using the approximate measurement instrument and unchanged objectives. It is supported by the data sets used in this study. Due to this, we can trust the results generated by the two models. Moreover, the correct result frequencies of the two datasets evidence that the results are reliable.

Researchers from the history domain can be motivated to conduct more research by delving deeper into the history topics. Both Chatbots are easier to access and provide user-friendly interfaces to learn history and answer questions in natural language, which is accessible to a large number of users, regardless of their prior knowledge of history topics.

A large body of literature suggests that Chatbots should be used carefully to avoid any possibility of inaccuracies and decrease critical thinking skills [2] [12]. Moreover, incorrect referencing is not comprehensively addressed by these Chatbots at hand. It is imperative to make further effective developments employing the supervising and moderating strategies.

The research methods used in this study were subject to selection bias because the manual collection of questions was conducted from two sources. Despite using the random selection of questions, we deliberately omitted the image-based questions as well as questions in the table. Reliance on the specific question format could have reduced the relevance across the diverse domains and particularly focused on history education. Therefore, it could potentially affect the generalizability of outcomes to other MCQ-based examinations. We were unable to assess the images and table questions; they cannot be supplied as input to ChatGPT-3.5 and ChatGPT-4.0. The evaluation of the current study was performed on two datasets from the history domain. However, the number of questions selected from both data sets was limited.

5. Conclusions

This study explored the performance of two Chatbots such as ChatGPT and Gemini on historical data of Sher Shah Suri and Mughal emperors and Sikhs in the subcontinent. Both Chatbots showed the ability to present factual information. ChatGPT provided very concise and short information on historical figures such as Sher Shah Suri and the Mughal emperors. Gemini revealed very exact information for the ChatGPT application. ChatGPT and Gemini presented valuable tools to help history learning in understanding the historical data. In the future, large datasets on diverse domains can be used to assess the performance of a number of Chatbots.

Conflict of Interest

There are no conflicts of interest to declare regarding this manuscript.

Data Availability

Data supporting the findings of this study are provided in this manuscript.

Funding Statement

This research study did not receive any funding.

6. References

[1] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido and J. Maningo, "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models," *PLoS Digital Health,* vol. 2(2), pp.e0000198, 2023.

[2] C. Kooli, "Chatbots in education and research: A critical examination of ethical implications and solutions," *Sustainability,* vol. 15(7), p.5614, 2023.

[3] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz and K. Kanclerz, "ChatGPT: Jack of all trades, master of none," *Information Fusion,* vol. 99, p.101861, 2023.

[4] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving language understanding by generative pre-training," 2018, https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell, "Language models are few-shot learners," *Advances in Neural Information Processing Systems,* vol. 33, pp. 1877–1901, 2020.

[6] H. Einarsson, S. H. Lund and A. H. Jónsdóttir, "Application of ChatGPT for automated problem reframing across academic domains," *Computers and Education: Artificial Intelligence,* vol. 6, p.100194, 2024.

[7] A. Herrmann-Werner, T. Festl-Wietek, F. Holderried, L. Herschbach, J. Griewatz, K. Masters, S. Zipfel and M. Mahling, "Assessing ChatGPT's Mastery of Bloom's Taxonomy using psychosomatic medicine exam questions: mixed-methods study," *Journal of Medical Internet Research,* vol. 26, pp. e52113, 2024.

[8] T. E. S. P. L. (2024). https://testbook.com/objective-questions/mcq-on-sher-shah%E2%80%99s-administrative-reforms--5fc42346a1bc541cc2ffbc6b.

[9] CSS Times. https://www.csstimes.pk/pakistan-affairs-solved-mcqs-political-history-of-mughals-and-sikhs-for-competitive-exams/.

[10] A. Oztermeli and N. Karahan, "Evaluation of YouTube video content about developmental dysplasia of the hip," *Cureus,* vol. 12(8), 2020.

[11] K. C. Madathil, A. J. Rivera-Rodriguez, J. S. Greenstein and A. K. Gramopadhye, "Healthcare information on YouTube: a systematic review," *Health Informatics Journal,* vol. 21(3), pp. 173–194, 2015.

[12] W. Choi, "Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs," *BMC Medical Education,* vol. 23(1), p.864, 2023.