Cureus

Open Access Original Article

DOI: 10.7759/cureus.43861

Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing

Review began 08/07/2023
Review ended 08/15/2023
Published 08/21/2023

Amita Kumari [1] , Anita Kumari [1] , Amita Singh [1] , Sanjeet K. Singh [2] , Ayesha Juhi [1] , Anup Kumar D. Dhanvijay [3] , Mohammed Jaffer Pinjar [1] , Himel Mondal [1]

1. Physiology, All India Institute of Medical Sciences, Deoghar, Deoghar, IND 2. Pathology, All India Institute of Medical Sciences, Deoghar, Deoghar, IND 3. Physiology, All India Institute of Medical Sciences Deoghar, Deoghar, IND

Corresponding author: Himel Mondal, himelmkcg@gmail.com

Abstract

Background

Large language models (LLMs), such as ChatGPT-3.5, Google Bard, and Microsoft Bing, have shown promising capabilities in various natural language processing (NLP) tasks. However, their performance and accuracy in solving domain-specific questions, particularly in the field of hematology, have not been extensively investigated.

Objective

This study aimed to explore the capability of LLMs, namely, ChatGPT-3.5, Google Bard, and Microsoft Bing (Precise), in solving hematology-related cases and comparing their performance.

Methods

This was a cross-sectional study conducted in the Department of Physiology and Pathology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India. We curated a set of 50 cases on hematology covering a range of topics and complexities. The dataset included queries related to blood disorders, hematologic malignancies, laboratory test parameters, calculations, and treatment options. Each case and related question was prepared with a set of correct answers to compare with. We utilized ChatGPT-3.5, Google Bard Experiment, and Microsoft Bing (Precise) for question-answering tasks. The answers were checked by two physiologists and one pathologist. They rated the answers on a rating scale from one to five. The average score of the three models was compared by Friedman's test with Dunn's post-hoc test. The performance of the LLMs was compared with a median of 2.5 by a one-sample median test as the curriculum from which the questions were curated has a 50% pass grade.

Results

The scores among the three LLMs were significantly different (p-value < 0.0001) with the highest score by ChatGPT (3.15±1.19), followed by Bard (2.23±1.17) and Bing (1.98±1.01). The score of ChatGPT was significantly higher than 50% (p-value = 0.0004), Bard's score was close to 50% (p-value = 0.38), and Bing's score was significantly lower than the pass score (p-value = 0.0015).

Conclusion

The LLMs reveal significant differences in solving case vignettes in hematology. ChatGPT exhibited the highest score, followed by Google Bard and Microsoft Bing. The observed performance trends suggest that ChatGPT holds promising potential in the medical domain. However, none of the models was capable of answering all questions accurately. Further research and optimization of language models can offer valuable contributions to healthcare and medical education applications.

Categories: Medical Education, Healthcare Technology, Hematology
Keywords: ai and robotics in healthcare, microsoft bing, google bard, chatgpt, pathology, hematology, hematologic diseases, natural language processing, search engine, pathologists

Introduction

Hematology, the specialized field of medicine focused on the study of blood and its associated disorders, holds paramount importance in the realm of healthcare. Precise diagnosis and effective management of hematologic conditions, such as anemia, leukemia, and coagulopathies, are imperative for optimizing patient outcomes and improving overall public health [1].

The advent of artificial intelligence (AI) and natural language processing (NLP) has led to the development of large language models (LLMs), which exhibit exceptional capabilities in processing and comprehending natural language data [2]. Prominent among these LLMs are ChatGPT, Google Bard, and Microsoft Bing, which have garnered substantial interest due to their capacity to comprehend textual information and generate contextually relevant responses [3]. These models have demonstrated various levels of accuracy in performing medical examinations, solving complex medical issues, or interpreting radiology reports [4,5,6,7,8].

Nevertheless, their efficacy and suitability for domain-specific applications, particularly in addressing medical inquiries pertaining to hematology, remain relatively unexplored. The intricate nature of hematology, characterized by a lexicon replete with specialized terminology and a wide array of conditions with nuanced diagnostic and therapeutic considerations, necessitates meticulous scrutiny of language models' performance in this context.

The present study aimed to address this gap in knowledge by exploring the capability of LLMs in solving hematology cases and conducting a comparative analysis of three LLMs, namely, ChatGPT, Google Bard, and Microsoft Bing.

Materials And Methods

Study type and settings

This was a cross-sectional study conducted in the Department of Physiology and Pathology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India. The data were collected from three LLMs, namely, Google Bard Experiment (https://bard.google.com/), Microsoft Bing Precise Conversation based on GPT-4 (https://www.bing.com/), and ChatGPT-3.5 Free Research version (https://chat.openai.com/), available free to any registered users. We used the free version of the LLMs. Henceforth in the article, the models are mentioned as Bard, Bing, and ChatGPT.

Hematology cases

A set of hematology-related cases covering various topics and complexities were carefully designed by subject matter experts. These questions encompassed laboratory calculations, disease interpretations, and other relevant aspects of hematology. A total of 50 cases and related questions were made, and content validity was checked by two experts of curriculum design. An example question and a part of the answer generated from Google Bard are shown in Figure *1*.

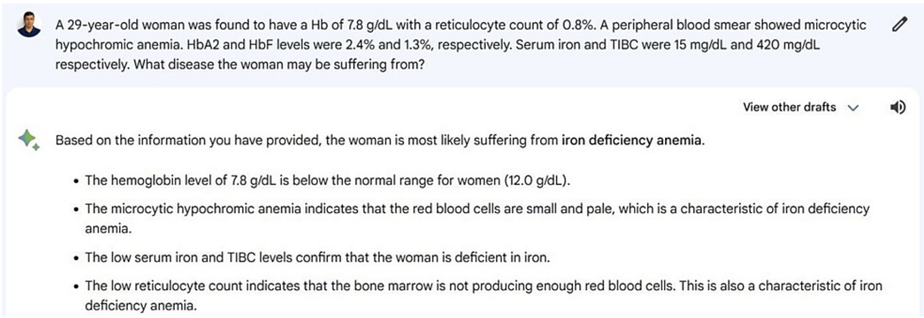

FIGURE 1: An example of a question and a part of the answer generated from Google Bard

Screenshot captured from Google Bard (https://bard.google.com). The image in the screenshot is that of the corresponding author.

Another example question and a part of the answer generated from Microsoft Bing are shown in Figure *2*.

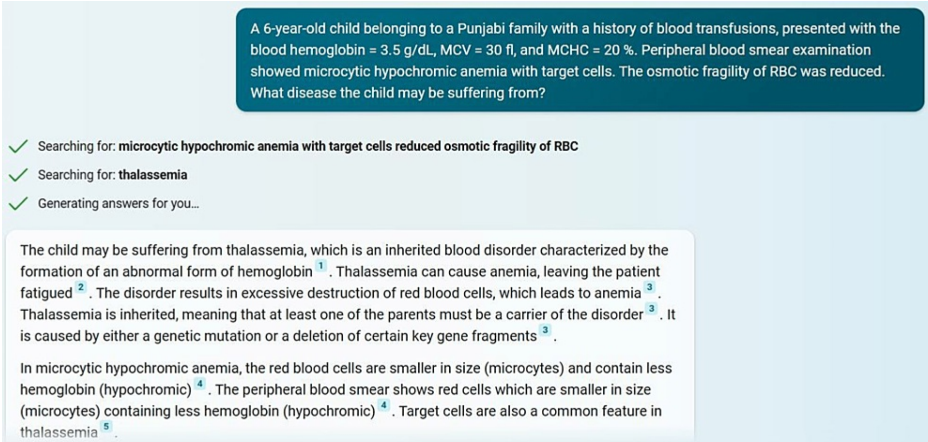

FIGURE 2: An example of a question and a part of the answer generated from Microsoft Bing Chat

Screenshot captured from Microsoft Bing (https://www.bing.com/).

Another example is shown in Figure 3, where the response was generated from ChatGPT-3.5.

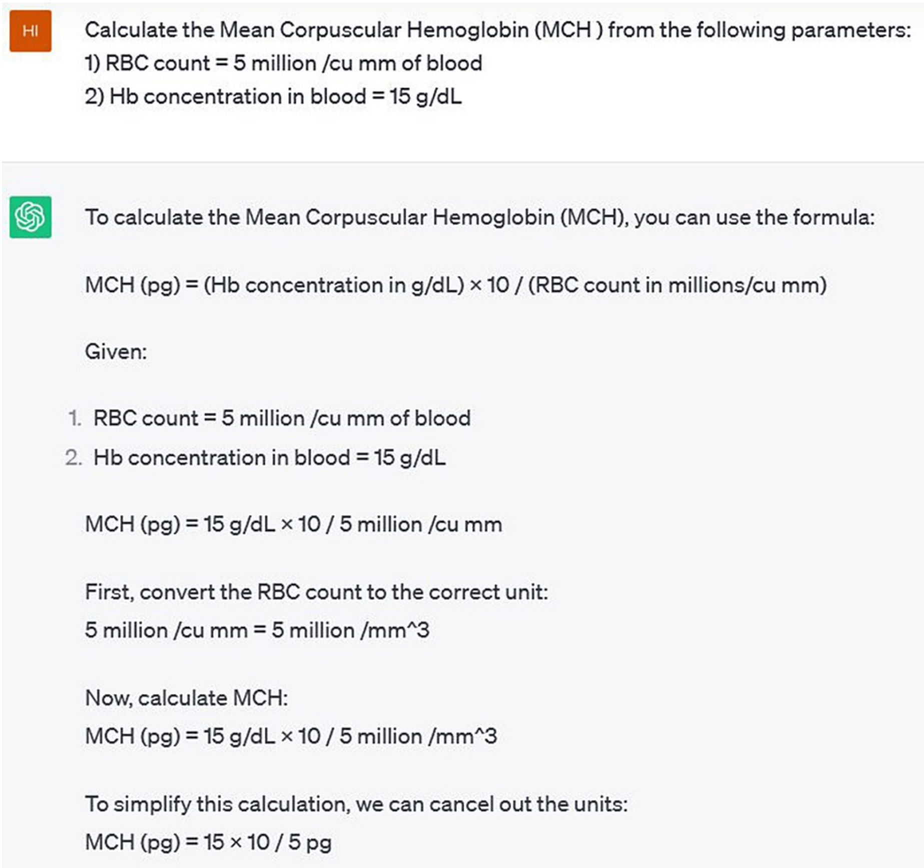

FIGURE 3: An example of a question and a part of the answer generated from ChatGPT-3.5

Screenshot captured from ChatGPT (https://chat.openai.com/).

Data collection from LLMs

The questions were asked to three LLMs, namely, Bard, Bing, and ChatGPT, on July 30, 2023 to get answers.

Generated answers were stored for further analysis. The answers were coded and blinded to the raters to reduce bias.

Assessing accuracy by three raters

To assess the accuracy of the LLMs' responses, three independent raters, with expertise in hematology and medical education, were recruited. The raters evaluated each LLM-generated answer and scored them based on their correctness with an accuracy score ranging from 1 to 5. The detailed scoring method was as follows:

5 - Highly accurate: The answer provided by the AI is thoroughly accurate, aligning perfectly with clinical knowledge and best practices.
4 - Moderately accurate: The answer provided by the AI is mostly accurate, with only minor discrepancies that do not significantly impact its clinical reliability.
3 - Somewhat accurate: The answer provided by the AI contains several inaccuracies that may require clarification or verification by a medical professional.
2 - Slightly accurate: The answer provided by the AI has noticeable inaccuracies, and its clinical reliability is questionable without substantial correction.
1 - Inaccurate: The answer provided by the AI is fundamentally incorrect and could pose serious risks to patient care if relied upon without thorough review and correction.

Data analysis

The obtained raters' evaluations were tested for normality by the Shapiro-Wilk test. The data were found not to follow normal distributions. Hence, we used nonparametric tests. The data were presented in mean, standard deviation (SD), median, first quartile (Q1), and third quartile (Q3) [9]. The average (average of the three raters) score awarded to the three LLMs was compared by Friedman's test with Dunn's post-hoc analysis. The average score was also compared by a one-sample Wilcoxon signed-rank test with a hypothetical value of 2.5 (50% score is the passing score in the curriculum of hematology from where the cases and questions were framed). Intraclass correlation coefficient (ICC) was calculated to determine the agreement level among the raters in assessing the accuracy of the LLMs' responses [10]. We used IBM SPSS Statistics for Windows, version 20 (released 2011; IBM Corp., Armonk, New York, United States) for statistical analysis. A p-value < 0.05 was considered statistically significant.

Results

A total of 50 cases were analyzed by the three raters. The scores among the three LLMs were significantly different (p-value < 0.0001), with the highest score by ChatGPT (3.15±1.19), followed by Bard (2.23±1.17) and Bing (1.98±1.01).

The median score with 95% confidence interval (CI) is shown in Figure *4*. In the post-hoc analysis, Bard versus Bing did not show any significant difference (p-value = 0.33). However, Bard versus ChatGPT (p-value < 0.0001) and Bing versus ChatGPT (p-value < 0.0001) showed significant differences.

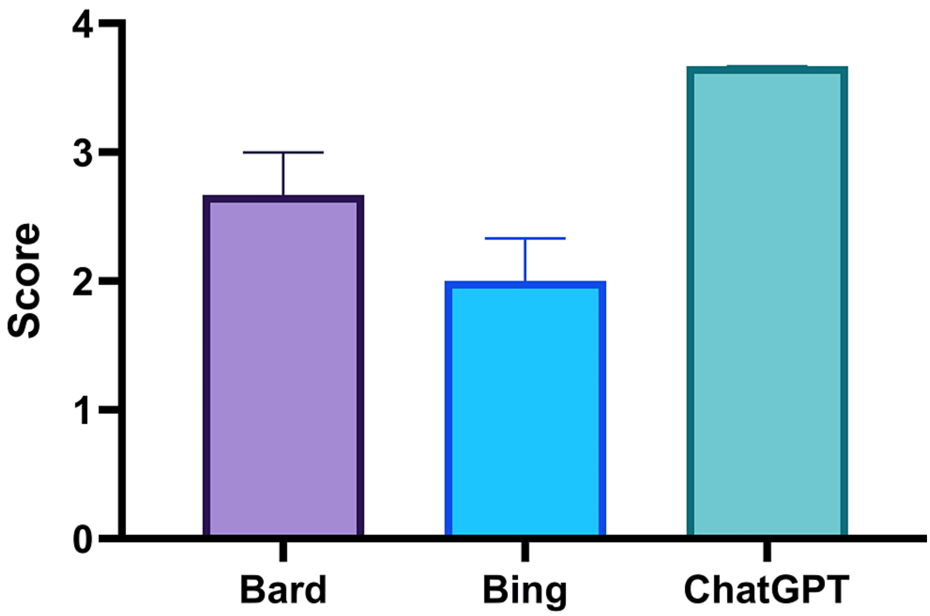

FIGURE 4: Average scores of Bard, Bing, and ChatGPT in the answers to the hematology questions

The bar and whisker indicate a median score with a 95% confidence interval. The comparison was done by Friedman's test (non-parametric analysis of variance (ANOVA)) with Dunn's post-hoc analysis.

The median and first quartiles/third quartiles were as follows: Bard 2.67 (1.67/3), Bing 2 (1.33/2.67), and ChatGPT 3.67 (3/4).

The scores of the three raters are shown in Table *1*. The scores among the raters showed a good to excellent level of reliability.

LLM	Central tendency	Rater 1	Rater 2	Rater 3	ICC, p-value
Google Bard	Mean±SD	1.88±1.19	2.12±1.41	1.94±1.11	0.83, <0.0001*
	Median (Q1-Q3)	3 (1-4)	2 (0.25 - 3)	2 (1-3)	
Microsoft Bing	Mean±SD	2.52±1.42	2.02±1.41	2.16±1.18	0.78, <0.0001*
	Median (Q1-Q3)	2 (1-3)	2 (1-3)	2 (1-3)	
ChatGPT-3.5	Mean±SD	3.4±1.26	3.02±1.25	3.02±1.27	0.98, <0.0001*
	Median (Q1-Q3)	4 (3.2-4)	3 (3-4)	3 (3-4)	

TABLE 1: Scores of the answers by the three LLMs as rated by the three raters

*Statistically significant p-value of the intraclass correlation coefficient (ICC)[10]

LLM: large language model, SD: standard deviation

Interpretation of ICC: <0.5 = poor reliability, 0.5-0.75 = moderate reliability, 0.75-0.9 = good reliability, >0.90 = excellent reliability[9]

When we conducted the Wilcoxon signed-rank test with a hypothetical value of 2.5 (the course from where the questions were prepared needs a 50% score to pass), we found that ChatGPT's score was significantly higher than the passing score (p-value = 0.0004). The score of Bard was close to 50% (p-value = 0.38). However, Bing's score was significantly lower than the passing score (p-value = 0.0015).

Discussion

We found that ChatGPT performed better than the other two LLMs in problem-solving abilities in hematology. These findings emphasize the importance of selecting the most appropriate language model based on its performance for specific tasks, enabling more effective problem-solving in various scenarios. The LLMs are evolving day by day. Hence, further studies are required in the future to fully explore their capability [11]. When we tested the models' performance with a minimum passing score (50%) of the curriculum, we found ChatGPT to pass it with the highest margin, indicating that it consistently outperformed the threshold. Meanwhile, Bard's score showed no significant difference from the passing score, suggesting that it performed at a level close to the passing requirement. By contrast, Bing's score was significantly lower than the passing score, indicating that it fell short of meeting the passing standard. These results highlight the varying proficiency levels of the language models in achieving accuracy with the same questions at the same time.

Based on the finding that laboratory calculations or interpretations for hematology diseases are weak in many cases, caution should be exercised when considering the use of LLMs in medical education. While LLMs can provide valuable information and learning resources, their limitations in accurately handling laboratory data for hematology diseases warrant careful supervision and validation by qualified medical educators. In the context of healthcare, where precision and accuracy are crucial, the use of LLMs for making clinical decisions or providing direct patient care should be approached with caution. LLMs can serve as helpful tools for information retrieval and initial insights, but they should not replace the expertise and clinical judgment of healthcare professionals. As technology advances and LLMs undergo further refinement, they may hold more promise for enhancing medical education and supporting healthcare, but their current limitations necessitate prudent utilization. Some of the potential use of LLMs are summarized in Table *2* [12,13,14,15,16].

Category	Brief
Students	LLMs can act as virtual tutors, helping medical students access supplementary information, explanations, and case studies to reinforce their learning.
Healthcare professionals	Doctors, nurses, and other healthcare professionals can use LLMs to stay updated with the latest medical research, treatment guidelines, and evidence-based practices.
Patients	LLMs can be used in health applications or chatbots to provide basic medical information and answer common health-related queries for patients.
Researchers	LLMs can assist researchers in searching for relevant literature, summarizing papers, and extracting key insights from a vast amount of medical literature.
Remote health assistants	LLMs can be particularly valuable in areas with limited access to healthcare resources, where individuals can seek initial medical advice and information.

TABLE 2: Potential audience and brief where large language models (LLMs) can help

The substantial agreement among the raters may be attributed to several potential reasons. The raters were provided with clear and well-defined evaluation criteria, ensuring a consistent understanding of the scoring method. Their familiarity and expertise in assessing hematology-related questions could have contributed to more aligned judgments. However, in-depth analysis and examination of the specific data and rater feedback would be necessary to fully understand the underlying factors influencing the agreement among the raters.

Limitations

The study has several limitations that should be considered when interpreting the findings. The sample size of hematology questions and LLM responses was relatively small, potentially impacting the generalizability of the results. Rater bias and subjectivity in evaluating LLM responses might have introduced variability. Only three models were tested. The cross-sectional design offers only a snapshot of LLM performance, and real-world applications may present additional challenges. Despite these limitations, this study provides valuable insights into LLM accuracy in hematology questions and underscores the need for further research and exploration of ethical considerations in utilizing LLMs in medical education and healthcare.

Conclusions

There were variations in the LLM performance, with ChatGPT demonstrating the highest accuracy, Bard exhibiting moderate accuracy, and Bing showing comparatively lower accuracy in answering questions of hematology. While LLMs hold promise as valuable tools for medical education, caution is warranted in their use, considering their limitations in handling complex medical nuances and potential inaccuracies. This study emphasizes the need for continuous refinement and validation of LLMs for reliable healthcare applications.

Additional Information

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The corresponding author would like to thank Sarika Mondal and Ahana Aarshi for their moral support by sacrificing their family time during data analysis, visualization, and drafting of the manuscript. The language and grammar were edited with the help of ChatGPT-3.5 free research version (August 3, 2023 version) and Grammarly free version browser extension.

References

1. Obstfeld AE: Hematology and machine learning. J Appl Lab Med. 2023, 8:129-44. 10.1093/jalm/jfac108
2. Wang J, Deng H, Liu B, et al.: Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed. J Med Internet Res. 2020, 22:e16816. 10.2196/16816
3. Agarwal M, Sharma P, Goswami A: Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus. 2023, 15:e40977. 10.7759/cureus.40977
4. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU: ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification?. J Am Med Inform Assoc. 2023, 10.1093/jamia/ocad104
5. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023, 2:e0000198. 10.1371/journal.pdig.0000198
6. Bhayana R, Krishna S, Bleakney RR: Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology. 2023, 307:e230582. 10.1148/radiol.230582
7. Dhanvijay A D, Pinjar M, Dhokane N, et al.: Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. Cureus. 2023, 15:e42972. 10.7759/cureus.42972
8. Ghosh A, Bir A: Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus. 2023, 15:e37023. 10.7759/cureus.37023
9. Mondal H, Mondal S, Majumder R, De R: Conduct common statistical tests online. Indian Dermatol Online J. 2022, 13:539-42. 10.4103/idoj.idoj_605_21
10. Koo TK, Li MY: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016, 15:155-63. 10.1016/j.jcm.2016.02.012
11. Friederichs H, Friederichs WJ, März M: ChatGPT in medical school: how successful is AI in progress testing? . Med Educ Online. 2023, 28:2220920. 10.1080/10872981.2023.2220920
12. Dave T, Athaluri SA, Singh S: ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023, 6:1169595. 10.3389/frai.2023.1169595
13. Ruksakulpiwat S, Kumar A, Ajibade A: Using ChatGPT in medical research: current status and future directions. J Multidiscip Healthc. 2023, 16:1513-20. 10.2147/JMDH.S413470
14. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M: The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform. 2023, 305:644-7. 10.3233/SHTI230580
15. Shahsavar Y, Choudhury A: User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors. 2023, 10:e47564. 10.2196/47564
16. Mondal H, Mondal S, Podder I: Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J. 2023, 14:482-6. 10.4103/idoj.idoj_72_23