Annika Meyer*, Ari Soleman, Janik Riese and Thomas Streichert

Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum

https://doi.org/10.1515/cclm-2024-0246
Received February 24, 2024; accepted May 13, 2024; published online May 29, 2024

Abstract

Objectives: Laboratory medical reports are often not intuitively comprehensible to non-medical professionals. Given their recent advancements, easier accessibility and remarkable performance on medical licensing exams, patients are therefore likely to turn to artificial intelligence-based chatbots to understand their laboratory results. However, empirical studies assessing the efficacy of these chatbots in responding to real-life patient queries regarding laboratory medicine are scarce.

Methods: Thus, this investigation included 100 patient inquiries from an online health forum, specifically addressing Complete Blood Count interpretation. The aim was to evaluate the proficiency of three artificial intelligence-based chatbots (ChatGPT, Gemini and Le Chat) against the online responses from certified physicians.

Results: The findings revealed that the chatbots’ interpretations of laboratory results were inferior to those from online medical professionals. While the chatbots exhibited a higher degree of empathetic communication, they frequently produced erroneous or overly generalized responses to complex patient questions. The appropriateness of chatbot responses ranged from 51 to 64 %, with 22 to 33 % of responses overestimating patient conditions. A notable positive aspect was the chatbots’ consistent inclusion of disclaimers regarding its non-medical nature and recommendations to seek professional medical advice.

Conclusions: The chatbots’ interpretations of laboratory results from real patient queries highlight a dangerous dichotomy – a perceived trustworthiness potentially obscuring factual inaccuracies. Given the growing inclination towards self-diagnosis using AI platforms, further research and improvement of these chatbots is imperative to increase patients’ awareness and avoid future burdens on the healthcare system.

Keywords: ChatGPT; chatbot; AI; laboratory results; health forum

Introduction

Laboratory medical reports are crucial in guiding clinical decision-making. Nonetheless, their technical nature often poses comprehension challenges for individuals without medical training [1]. Consequently, many seek clarification online [1], increasingly trusting artificial intelligence (AI)-powered chatbots over conventional search engines for medical advice [2].

This shift has been notably influenced by the launch of the AI chatbot “ChatGPT” (Chat Generative Pre-trained Transformer) in late 2022, which has not only provided the general public with access to an advanced AI [3] but also achieved unprecedented user growth [4]. In the medical domain, research indicates that 78 % of its users are inclined to employ ChatGPT for self-diagnosis purposes [5]. The emergence of other AI chatbots such as Gemini and Le Chat has further broadened user options [6].

While these chatbots demonstrate proficiency in handling binary medical multiple-choice questions [7–13], their performances in academic tests may not adequately reflect the complex and nuanced reality of clinical practice [14]. This discrepancy highlights the importance of evaluating these chatbots in real-world scenarios, particularly as they were not specifically designed for medical inquiries [15–17].

In this context, an initial study by Cadamuro et al., which employed 10 fictive laboratory scenarios, revealed ChatGPT’s ability to identify laboratory tests, categorize values within given reference intervals and provide

*Corresponding author: Annika Meyer, Institute of Clinical Chemistry, Faculty of Medicine and University Hospital, University Hospital Cologne, Kerpener Str. 62, 50937 Cologne, Germany, E-mail: ameyer18@smail.uk-koeln.de. https://orcid.org/0000-0002-8411-8799
Ari Soleman, Faculty of Medicine and University Hospital, University Hospital Cologne, Cologne, Germany
Janik Riese, Institute of Pathology, Faculty of Medicine, RWTH Aachen University, Aachen, Germany
Thomas Streichert, Institute of Clinical Chemistry, Faculty of Medicine and University Hospital, University Hospital Cologne, Cologne, Germany.

https://orcid.org/0000-0002-6588-720X

Open Access. © 2024 the author(s), published by De Gruyter.
superficial interpretations. However, this study also highlighted the need for more extensive research involving a wider range of medical laboratory reports [1] to reduce misinterpretations in the post-analytical phase [18].

Building on this foundation, our study evaluates the capabilities of three chatbots (ChatGPT [GPT-4], Gemini [Gemini Pro], and Le Chat [Mistral Large]) using 100 patient inquiries focused on laboratory medical reports from an online health forum. This approach seeks to bridge the gap between theoretical data and real-life applications. Thus, the objective of this research is to explore the practical applicability and reliability of these chatbots in the field of laboratory medicine, offering insights into their potential utility and limitations in a genuine medical context.

Materials and methods

Chatbot selection

For this retrospective study, we selected chatbots based on Cascella et al.'s publication [6]. We excluded chatbots without web-based user interface, and those that were based on or performed below the level of the large language model GPT-3.5 and its predecessors. This refined our selection to three advanced chatbots, namely ChatGPT (GPT-4), Gemini (Gemini Pro), and Le Chat (Mistral Large).

Data collection

To assess the efficacy of these chatbots in interpreting laboratory reports, we sourced real-life patient queries from the 'AskDocs' subreddit on Reddit. This platform allows users to engage anonymously within specialized communities [19]. In 'AskDocs', users anonymously post medical questions which are then answered by verified physicians [20]. A comprehensive description of this forum is provided in the publication by Nobles et al. [20]. Our research methodology was designed to only observe, avoiding any direct interaction, thereby preserving the community's integrity and ensuring compliance with its guidelines [21].

In compliance to Reddit's Data API [22] as well as Developer Terms [23], we utilized the search term “CBC” (Complete Blood Count) to identify relevant posts. This term was selected for its prominence in general medical practice, as outlined by the European Federation of Clinical Chemistry and Laboratory Medicine Working Group on Artificial Intelligence (WG-AI) [1].

Our initial search yielded 635 posts. After applying exclusion criteria such as posts dated before ChatGPT’s knowledge cut-off date, absence of physician responses, presence of images, and deviation from the topic, we narrowed this selection down to 135 relevant posts. To determine an adequate sample size for robust statistical analysis, we conducted a Monte Carlo simulation. We assumed moderate factor loadings (λ=0.5) and aimed for a power of 0.8 [24]. This simulation indicated a minimum required sample size of 77. To enhance the stability of our results, we chose to include the 100 posts in our analysis, exceeding the minimum requirement to increase the study’s statistical power and potential generalizability. Thus, the posts were organized using Reddit’s internal sorting mechanisms, prioritizing relevance first, then categorizing them as ‘hot’, ‘top’, ‘new’, and ‘comments’. The first 100 posts were selected to ensure a balanced representation that captures both relevance and current trends in the discussions [25].

For each selected post, we gathered data including the title, text, number of upvotes, and comments, along with the most upvoted physician response. Upvotes serve as an internal Reddit rating system, allowing users to express approval of specific content [19]. Due to the possibility of users upvoting their own posts and the limited variability in upvotes across the chosen posts – as evidenced by a narrow interquartile range – we decided to exclude upvote data from the primary analysis. Instead, upvotes were utilized solely for descriptive statistics.

This approach mitigates potential biases in user engagement metrics related to self-upvoting. Moreover, some of the Reddit’s users provided reference ranges and some did not, addressing a potential confounder previously noted in the literature on ChatGPT [18]. Every post, along with its title, was then presented to all the three chatbots each in a new chat session to obtain their responses.

Evaluation by medical experts

Two medical experts, from an early-career physician to a professor, independently assessed the responses from the online physicians and the chatbots. Discrepancies were resolved through structured discussion using a consensus approach, reducing biases related to training level [1]. In instances of ambiguity, relevant literature databases such as NCBI, PubMed, and Amboss, or relevant medical textbooks [26], were consulted.

To prevent identification bias, all responses were anonymized and edited to remove any language that might reveal the non-human nature of the chatbots. Phrases such as “I am not a doctor,” “not a medical professional,” “AI,” “artificial,” “language model,” and “not a physician” were systematically searched for and removed.

The responses were then ranked relative to each other on a scale from 1 (best) to 4 (last), to allow a comparative perspective on the effectiveness of each respondent, chatbots and physicians. The ranking process for the chatbots involved assessing several criteria across the chatbots’ outputs: potential dangerous content, medical errors, technical errors (supplying faulty links, changing the output language), failure to answer the question, content-wise and technically correct answer to the question, empathetic and correct answer to the question.

Each response was also evaluated on a scale from 1 (excellent) to 6 (inadequate), focusing on the criteria of quality, clarity, medical accuracy and empathy. The overall quality assessment encompassed all response dimensions, while medical accuracy was specifically measured against medical content alone. Responses with incorrect information were rated lower than those with omissions or incomplete information.

Furthermore, appropriateness was scrutinized. In this context, overestimations were defined as evaluations that incorrectly identified healthy conditions as pathological, exaggerated the severity of a pathology, or recommended unnecessarily severe or invasive diagnostic steps or interventions.

Statistical analysis

All assessments were then statistically analyzed using the R programming language, specifically utilizing the packages ‘rio’ [27], ‘tidyverse’ [28], ‘gtsummary’ [29], ‘labelled’ [30], ‘ggpubr’ [31], ‘fastDummies’ [32].
For the descriptive statistics, categorical variables were presented as frequency and percentage, and continuous variables as median and interquartile range. The normal distribution of continuous data was refuted using the Kolmogorov–Smirnov and the Shapiro–Wilk test (Supplementary Material, Appendix 1). The Shapiro–Wilk and Kolmogorov–Smirnov tests were selected to assess the conformity of the data with the assumption of normality in its underlying distribution. This assumption served as a foundational requirement for subsequent parametric statistical analyses [33]. The analysis of categorical variables was conducted using the McNemar test, whereas the paired Wilcoxon signed rank test was applied for ordinal data. For models with continuous predictor variables, random intercept logistic regression was employed. A p-value of less than 0.05 was considered statistically significant. The Bonferroni correction was implemented to mitigate the heightened likelihood of a type I error arising from multiple comparisons. This adjustment was made to the p-values, utilizing the R package ‘gtsummary’ [29]. While alternative statistical methodologies such as the ‘Hochberg’ method were available, the adoption of Bonferroni’s approach was favored due to its conservative nature and capability to regulate the anticipated count of type I errors per family [34].

Results

General characteristics

The study analyzed posts (respective patient questions) with a mean word count of 146 (IQ: 82, 231) and their titles containing 7 (IQ: 5, 9) words. On average, posts received 1 (IQ: 1, 2) upvotes and 4.5 (IQ: 3.0, 6.0) comments. The mean age of respondents was 27 (IQ: 22, 31) and 52 % (49/94) identified as female. On average, online physicians used the fewest words (median: 20; IQ: 10, 35), while ChatGPT used the most (median: 357; IQ: 281, 379; p<0.001 for all comparisons). Le Chat and Gemini did not differ in their word count (p=0.9), with 261 (IQ: 216, 305) and 257 (IQ: 208, 301), respectively (Table 1, Figure 1A, Supplementary Material, Appendices 1–3).

Assessment by medical professionals

The evaluation highlighted significant differences in rankings between chatbots and physicians (<0.001 for all comparisons). Online physicians were frequently ranked highest, leading in 60 % (60/100) of cases, while Gemini was ranked last 39 % (39/100) of the time. Although there was no significant difference between online doctors and chatbots regarding the absence of estimations or underestimations, ChatGPT alone matched the online physicians in both quality (p=0.3) and accuracy (p=0.057).

In comparisons among the chatbots, ChatGPT and Gemini varied significantly in rank (p<0.001), quality (p<0.001) and accuracy (p=0.007). Le Chat, however, showed lower levels of empathy compared to ChatGPT (p<0.001) and Gemini (p=0.007), the latter two exhibiting similar empathy levels (p>0.9). Moreover, chatbots tended to overestimate conditions, with overestimation rates ranging from 22 to 33 % contrasting the 1.0 % (1/100) rate observed among online physicians (p<0.001 for all comparisons) (see Supplementary Material, Appendices 2–7, Figures 1B and C, and 2).

Furthermore, medical professionals observed several challenges in the chatbots’ interpretation of laboratory results. Notably, the chatbots exhibited difficulties in maintaining consistency, particularly in interpreting complex contexts, distinguishing between abnormal and critical laboratory values, and providing diagnostic recommendations. Inconsistencies were particularly evident in their application of reference intervals, where they applied differing standards to patients of identical sex and similar age without citing the underlying sources for such varying reference ranges. This absence of standardized reference values also led to divergent interpretations of the same laboratory data for a single patient between the chatbots (e.g., ChatGPT: “Your ALP value is 36 which is slightly low.”; Gemini: “Alkaline Phosphatase: Your level is 36 U/L, which is slightly high.”, Le Chat: “Alkaline Phosphatase (ALP) 36 U/L: This is a liver enzyme, and your result is within the normal range, which is typically 20–140 U/L.”). Furthermore, issues with the reliability of cited sources were frequently noted in Gemini and occasionally in ChatGPT, where the sources provided were often invalid, unsuitable, or misleading. Despite these limitations, several aspects of the chatbots responses were positively acknowledged. These included its consistent recommendations for professional medical consultation, interpersonal capabilities, adept use of verbal imagery, and the incorporation of gender-sensitive language. Gemini was particularly noted for its structured approach and ongoing recommendations for preparing for medical appointments (Table 2).

Table 1: Summary statistics for included posts and answers.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n=100*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of asker, years</td>
<td>27 (22, 31)</td>
</tr>
<tr>
<td>(Missing)</td>
<td>7</td>
</tr>
<tr>
<td>Gender of asker</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>49 (52 %)</td>
</tr>
<tr>
<td>Male</td>
<td>45 (48 %)</td>
</tr>
<tr>
<td>(Missing)</td>
<td>6</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
</tr>
<tr>
<td>Upvotes</td>
<td>1 (1, 2)</td>
</tr>
<tr>
<td>Comments</td>
<td>4.5 (3.0, 6.0)</td>
</tr>
<tr>
<td>Word count</td>
<td></td>
</tr>
<tr>
<td>Title’s word count</td>
<td>7 (5, 9)</td>
</tr>
<tr>
<td>Post’s word count</td>
<td>146 (82, 231)</td>
</tr>
</tbody>
</table>

*Median (interquartile range) or frequency (%).
Comprehending laboratory results presents a significant challenge for those outside the medical profession, largely due to the detailed and data-heavy nature of these results [1]. Consequently, patients may turn to AI-based chatbots for medical advice [1, 2, 36], a trend accelerated by digital advancements that enable patients to access their laboratory reports directly, bypassing initial consultations with physicians [36, 37]. However, several issues arise from the reliance on chatbots for interpreting laboratory reports.

Despite their user-friendly interfaces and their rapid, personalized, and seemingly expert responses [1, 6, 38], none of the three chatbots matched the proficiency of online physicians regarding laboratory report interpretations. This finding contradicts earlier research by Ayers et al., who preferred ChatGPT’s responses to general patient inquiries in the same forum [15]. As ChatGPT’s proficiency differs by specialty [13], the discrepancy may be due to this study’s focus on laboratory reports, thereby indicating that their interpretation poses a particular challenge for those chatbots.

In this context, all three chatbots occasionally treated laboratory results as “standalone information”, especially...
when clinical information was limited. This approach is broadly recognized as problematic [18] and likely contributed to their overestimation tendency. This pattern may be further enhanced by all three chatbots’ inability to distinguish between critical and abnormal laboratory values, aligning with previous literature on ChatGPT [1]. For ChatGPT, this overestimation tendency is also evident in fictive scenarios across various medical contexts, such as plastic surgery [39] and multidisciplinary patient vignettes [40]. Overestimations, deemed less harmful than underestimations, might be strategically employed by their developers to avoid legal repercussions [39]. Instead of mitigating unfound patient concerns, however, such overestimations might inadvertently amplify them, thereby increasing logistical and financial burdens on the healthcare system [39, 40].

Figure 2: Assessment of laboratory medicine report interpretation by chatbots and physicians in regard to quality, clarity, empathy and accuracy. Violin plots represent the distribution of performance metrics – accuracy, clarity, empathy, and overall quality – among ChatGPT, Gemini, Le Chat, and physicians in an online health forum.
Table 2: Narrative review of benefits and drawbacks in the laboratory result interpretation by chatbots.

<table>
<thead>
<tr>
<th>Evaluated aspects</th>
<th>Description of performance</th>
<th>Examples/notes – ChatGPT</th>
<th>Examples/notes – Gemini</th>
<th>Examples/notes – Le Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference intervals</td>
<td>Inconsistency in reference interval</td>
<td>(27 year old male): ‘red blood cells (RBC): the normal range for men is generally 4.7 to 6.1 million cells per microliter (µL).’ (36 year old male): ‘RBC (red blood cell count); these cells carry oxygen throughout your body. A typical range for men is roughly between 4.5 and 5.5 million cells/µL’</td>
<td>(19 year old female): ‘these are borderline low (normal range for hemoglobin is 12.3–15.2 g/dL and hematocrit is 36.1–46.4 %).’ (17 year old female): ‘a hemoglobin of 13 might be high for you based on your previous results, but it’s generally within the normal range for most females (around 12–16 g/dL)’</td>
<td>(28 year old female): ‘white blood cells (WBC): your WBC count is slightly below the normal range, which is typically 4.5–11.0 thou/cu mm. This could be due to various reasons, such as a viral infection, nutritional deficiencies, or stress.’ (19 year old female): ‘WBC (white blood cell count) – 4.3 (normal range: 4.0–11.0 × 10³/µL)’</td>
</tr>
<tr>
<td></td>
<td>Inconsistency between given and used reference intervals</td>
<td>None found</td>
<td>Gemini applies different reference ranges than the laboratory given pH of 5.0–8.0: ‘normal findings except for a slightly high urine pH (7.5). Ideally, urine pH should be between 6.0 and 7.0.’</td>
<td>Le Chat applies different upper limit for CRP than the 0.9 given by the laboratory: ‘a normal CRP level is typically less than 1.0 mg/dL’</td>
</tr>
<tr>
<td></td>
<td>Value not correctly categorized within reference limits</td>
<td>While one reddit user’s RBC is at 5.38 µL ChatGPT states: ‘RBC is slightly high but not alarmingly so. Normal range for females is typically around 4.2 to 5.4 million cells per microliter (µL), although this can slightly vary between laboratories’</td>
<td>‘Hemoglobin: yours is at 17.4, again within the normal range (13.0–17.0)’</td>
<td>‘Hemoglobin: your result is 17.4 g/dL, which is within the normal range for men (13.0–17.0 g/dL)’</td>
</tr>
<tr>
<td>Others</td>
<td>ChatGPT applies differing units within one response: ‘white blood cells (WBC): a normal WBC count ranges from 4,500 to 11,000 white blood cells per microliter of blood. Your numbers have gone from 5.1 to 4.2, which is a slight decrease but still within the normal range’</td>
<td>None found</td>
<td>Gemini makes the following statement: ‘2.11 % – this is a very small percentage’</td>
<td>None found</td>
</tr>
<tr>
<td><strong>Contexts</strong></td>
<td>Citing faulty sources</td>
<td>According to the cancer society, Hodgkin’s lymphomas account for 12 % of childhood cancers [35], but ChatGPT describes 3 % (‘in fact, according to the American cancer society, only about 3 % of childhood cancers are lymphomas’))</td>
<td>Links often invalid (‘[invalid URL removed]’), with misleading content or in a language other than the input language (e.g., ‘www.toutiao.com/article/6482582255993093006/’)</td>
<td>Neither referring to sources nor supplying links</td>
</tr>
<tr>
<td></td>
<td>Misinterpretation of modality and its diagnostic value</td>
<td>ChatGPT misinterprets falling neutrophils and rising lymphocytes without additional blood smear as ‘your slight shift in neutrophil and lymphocyte percentages is known as a ‘shift to the left’’</td>
<td>Instead of ferritin, Gemini attributes the highest diagnostic value to iron in iron deficiency anaemia: ‘iron saturation: Low at 7 % (normal starts at 15 %). This is the key indicator of iron deficiency’</td>
<td>Le Chat overestimates the diagnostic value of CBC results when deciding on antibiotic therapy: ‘it is possible that they determined that the CBC results were not conclusive enough to warrant antibiotic treatment’</td>
</tr>
<tr>
<td></td>
<td>Limited understanding of complex interrelationships</td>
<td>While metabolic alkalosis due to the loss of stomach acid would be compensated by hyperventilation and thus the increase in pCO₂ [35], ChatGPT misinterprets: ‘the</td>
<td>Gemini fails to establish a connection between non-fasting and glucose: ‘glucose (121, normal range 65–99) – while mildly elevated, a single reading isn’t conclusive for diabetes. However, it</td>
<td>Le Chat fails to establish a connection between non-fasting and glucose: ‘glucose 121 (65–99): this is higher than the normal range. It could indicate</td>
</tr>
</tbody>
</table>


Table 2: (continued)

<table>
<thead>
<tr>
<th>Evaluated aspects</th>
<th>Description of performance</th>
<th>Examples/notes – ChatGPT</th>
<th>Examples/notes – Gemini</th>
<th>Examples/notes – Le Chat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
<td>low potassium and CO₂ in the second visit could also be related to the vomiting (losing stomach acid can lead to metabolic alkalosis, which can lower blood potassium)’</td>
<td>warrants further investigation considering the other symptoms’</td>
<td>hyperglycemia, which is high blood sugar. This could be due to many reasons, including diet, stress, or an underlying condition like diabetes. However, a single high reading does not necessarily mean diabetes, and further testing would be needed to confirm this’</td>
</tr>
<tr>
<td>Promoting medical professionals</td>
<td>Consistent recommendation for professional medical consultations</td>
<td>‘It’s important that she continues to work with her healthcare providers to manage her symptoms and monitor her health. You might also consider seeking a second opinion if you feel it would be beneficial’</td>
<td>‘Don’t self-diagnose: avoid searching the internet for diagnoses based on your results. Let your doctor guide you’</td>
<td>‘I would recommend reaching out to your healthcare provider to discuss these results. They have a full understanding of your medical history and can provide you with the most accurate advice’</td>
</tr>
<tr>
<td>Inclusion of prior medical assessments</td>
<td></td>
<td>‘They don’t necessarily indicate something more serious like lymphoma, especially considering your doctor’s lack of concern’</td>
<td>‘Doctor’s opinion: it’s positive that your doctor isn’t concerned, especially since your BUN and GFR are normal’</td>
<td>‘Your doctor, who has access to your complete medical history and test results, has mentioned that they are not concerned. This should be reassuring, as they have a better understanding of your overall health’</td>
</tr>
<tr>
<td>Advice on the preparation for a physician’s appointments</td>
<td></td>
<td>‘In the meantime, continue to advocate for yourself. Record your symptoms, make notes about anything that might be relevant (including diet, sleep, stress, etc.), and don’t hesitate to ask your doctors for clarification or more information’</td>
<td>‘Gather past CBC results: bring copies of previous blood tests to the hematology appointment. This will give the doctor a clearer picture. Prepare a list of questions: write down any questions you have about your son’s low MPV and his medications. I Hope this information is helpful! Again, it’s best to wait for your son’s pediatrician to give specific advice based on his individual case’</td>
<td>None found</td>
</tr>
<tr>
<td>Linguistic capabilities</td>
<td>Recognition and usage of figurative language</td>
<td>ChatGPT responds to ‘Doctor Google has officially scared me’ with: ‘don’t let Dr. Google scare you’</td>
<td>Gemini responds to ‘feeling like death’ with: ‘feeling like death: your CBC results alone likely wouldn’t explain feeling that bad. There could be other factors at play, and it’s important to discuss this with your doctor’</td>
<td>Le Chat does not employ linguistic imagery, but utilizes it: ‘medical interpretation requires a holistic view of your health status, and a single test result is just one piece of the puzzle’</td>
</tr>
<tr>
<td></td>
<td>Utilization of gender-conscious language</td>
<td>‘However, it’s important to note that while these could be potential explanations, it’s best to continue the conversation with your doctor. He or she knows your medical history best and will be able to guide you through the next steps, which might involve further testing or treatment to manage any ongoing issues that could be contributing to the elevated CRP’</td>
<td>‘If you’re still feeling anxious, talking to your doctor about your health anxiety might be helpful. They can provide resources and strategies for managing it’</td>
<td>‘I would recommend reaching out to your healthcare provider to discuss these results. They have a full understanding of your medical history and can provide you with the most accurate advice’</td>
</tr>
</tbody>
</table>
Furthermore, reference ranges posed another notable challenge to all three chatbots. Despite literature suggesting that ChatGPT and other chatbots (CopyAI and Writesonic) can navigate within provided reference intervals [1, 38], ChatGPT, Le Chat and Gemini struggled in their absence, often showing inconsistencies in ranges and units, as well as misclassifying laboratory values within them. Confirming previous research [38], the use of different reference ranges by each chatbot led to varying interpretations of identical blood values – a situation reminiscent of the saying “two doctors, three opinions”. Notably, even when reference ranges were specified, ChatGPT occasionally failed to classify laboratory values accurately, while Le Chat and Gemini struggled more frequently. Despite ChatGPT’s superior performance compared to its competitors [16, 17, 38], none of the three chatbots proved consistently reliable. This inconsistency extends to other medical specialties, such as risk stratification in non-traumatic chest pain scenarios, further compromising their clinical reliability [41]. This emphasizes the need for substantial medical expertise to accurately interpret and validate their outputs, thus limiting their standalone use in interpreting laboratory reports.

Furthermore, the challenge of detecting inaccuracies is amplified by the sophisticated linguistic capabilities and empathetic responses of ChatGPT [15, 42], Gemini [43], and Le Chat [16, 44]. These features, while enhancing user engagement, may obscure the detection of errors [45]. Although all chatbots employ a cautionary tone, frequent disclaimers about their non-medical status, and recommendations for professional medical consultations similar to those described for ChatGPT in the literature [1, 15], it remains to be seen whether they can compensate for the misplaced trust in the chatbots’ interpretive abilities of laboratory reports [46, 47].

Despite these challenges, the chatbots’ rapid, personalized, and clear responses demonstrate significant potential for future applications in patient-centric communication. To optimize the utility and safety of chatbots in interpreting laboratory values, a strategic integration of the distinct strengths of various chatbots could prove beneficial. For instance, combining ChatGPT’s adept handling of reference ranges with Gemini’s structured advice and practical tips for medical appointment preparation could potentially streamline the logistical aspects of healthcare interactions. In fact, the implementation of AI in laboratory medicine could lead to significant financial savings, estimated at around 883.5 billion euros annually [48], and a reduction in labor hours by approximately 53.4 million [49]. The potential of AI in shifting healthcare efficiency and cost management is also reflected by the high amount of health start-ups focusing on screening and diagnostics [50]. This underscores the economic and safety imperatives of further developing AI-driven chatbots, emphasizing their role in enhancing patient safety and healthcare efficiency.

However, it’s essential to recognize that laboratory results are medical reports intended for interpretation by trained professionals. Given the current limitations of AI-based chatbots in accurately interpreting such results, while they may serve as tools for medical professionals, their unsupervised use by patients is not advisable.

**Limitations**

Overall, this study has limitations. The inherent variability of AI output, despite the high consistency observed in ChatGPT’s responses to simulated laboratory reports [1], poses a challenge to reproducibility. Furthermore, as AI...
models continue to evolve, the results of this study may not be applicable to future iterations of these chatbots.

Although the medical experts evaluated each physician and chatbot’s response to the best of their knowledge, subjective variation cannot be fully accounted for.

Moreover, the study’s focus on English language laboratory results from a single online health forum limits its generalizability across different specialties, forums, or languages, particularly given the English language bias in multilingual large language models like ChatGPT.

Conclusions

Overall, the chatbots’ perceived high trustworthiness, coupled with content inaccuracies and a tendency to overestimate medical laboratory reports, creates a risky scenario for medical non-professionals. Instead of alleviating unfounded patient concerns and thereby relieving the burden on the healthcare system, the chatbots may inadvertently promote over- and misdiagnosis. Thus, given the substantial patient inclination to self-diagnose using chatbots, it is crucial to enhance the chatbot’s development to safeguard patients and prevent future healthcare system burdens.

Acknowledgments: We acknowledge the support of DeepL, DeepL Write and ChatGPT (GPT-4) in the linguistic improvement and translation of the manuscript. Furthermore, ChatGPT was used to facilitate and accelerate the programming of the statistics. Any output from these artificial intelligences were critically reviewed by the authors.

Research ethics: This project was reviewed and approved by the local Ethics Committee on 11. August 2023 (23-1288-retro).

Informed consent: Not applicable.

Author contributions: TS and AM designed the study. AM collected the patient questions as well as the physician and ChatGPT responses. TS, AM, AS evaluated the responses regarding the questions. AS analyzed and interpreted the data statistically as well as wrote the manuscript. TS and AS critically reviewed and improved the manuscript. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest.

Research funding: None declared.

Data availability: In compliance with data protection regulations and to maintain confidentiality and privacy information, the authors are unable to share the raw data from this study. We are committed to upholding ethical standards and safeguarding sensitive data, and as such, cannot provide access to the dataset.

References

32. Kaplan J. fastDummies: fast creation of dummy (binary) columns and rows from categorical variables; 2023.
34. Andrade C. Multiple testing and protection against a type 1 (false positive) error using the Bonferroni and Hochberg corrections. Indian J Psychol Med 2019;41:99–100.
44. Lee YK, Suh J, Zhan H, Li JJ, Ong DC. Large language models produce responses perceived to be empathic. ArXiv 2022;10:e32578.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2024-0246).