GPT-3.5, Gemini, and GPT-4 Performance on the Advanced Trauma Life Support Exam

Hilary Y. Liu[1*], Mario Alessandri Bonetti[1], Alain C. Corcos[2], Jenny A. Ziembicki[2], Francesco M. Egro[1,2]

[1] Department of Plastic Surgery, University of Pittsburgh Medical Center, 1350 Locust Street, G103, Pittsburgh, PA 15219, United States.
[2] Department of Surgery, University of Pittsburgh Medical Center, 1350 Locust Street, G103, Pittsburgh, PA 15219, United States.

* Corresponding author. Tel.: (404) 861-79444; email: liuh23@upmc.edu (H.Y.L.); m.alessandribonetti@gmail.com (M.A.B.); corcosac@upmc.edu (A.C.C.); ziembickija@upmc.edu (J.A.Z.); francescoegro@gmail.com (F.M.E.)
Manuscript submitted May 6, 2025; accepted May 20, 2025; published July 17, 2025.
doi: 10.18178/JAAI.2025.3.3.180-186

Abstract: The Advanced Trauma Life Support (ATLS) certification evaluates the ability of medical professionals to manage trauma patients effectively in emergency settings. With the rapid evolution of Large Language Models (LLMs), there is growing interest in exploring how these tools might integrate into clinical practice. This study assessed the performance of three LLMs—GPT-3.5, Gemini, and GPT-4—on the ATLS written examinations. Each model answered three different ATLS 10th edition exams. Their responses were compared to official answer keys, and average scores were calculated. Differences in performance among the LLMs were analyzed using chi-square testing. In addition, performance was examined based on question type: direct knowledge questions versus clinical scenario questions. GPT-3.5 achieved an average score of 65%, Gemini 61.7%, and GPT-4 83.3%. Among the three models, only GPT-4 surpassed the passing threshold of 75%. There was no statistically significant difference between the scores of GPT-3.5 and Gemini (p = 0.59). However, GPT-4 significantly outperformed both GPT-3.5 (p = 0.0012) and Gemini (p = 0.0002). No significant differences in performance were noted between direct and clinical scenario questions within each model. GPT-4 demonstrated the ability to successfully pass the ATLS examination, highlighting its advanced technical knowledge. Nonetheless, occasional inaccuracies or "hallucinations" were observed, particularly with more complex questions. With continued development and rigorous validation, LLMs like GPT-4 have the potential to serve as valuable adjuncts in clinical decision-making and trauma education.

Keywords: artificial intelligence, Advanced Trauma Life Support (ATLS), ChatGpt, Google, trauma

1. Introduction

The Large Language Model (LLM) ChatGPT (OpenAI, San Francisco, CA) has recently garnered attention for its potential to transform medical decision-making and education [1–8]. It has demonstrated strong reasoning abilities and a broad knowledge base across multiple medical fields. Indeed, it has passed the United States Medical Licensing Exam [[9], Plastic Surgery In-Service Exam [10], and board exams in radiology [11], neurosurgery [12], orthopedic surgery [13], ophthalmology [14], and otolaryngology [15].

Initially launched in November 2022 as GPT-3.5, ChatGPT later saw an upgraded release, GPT-4, by March 2023. GPT-4, similar to its predecessor, was trained using both supervised and unsupervised techniques on

a dataset encompassing around 100 trillion parameters [16]. GPT-4 underwent further refinement through Reinforcement Learning from Human Feedback (RLHF), a machine learning approach where user interactions inform a reward-based system to enhance model outputs. Around the same period, Google (Mountain View, CA) introduced its own large language model, Gemini. Unlike ChatGPT and GPT-4, which were trained on data available up to September 2021, Gemini has the capability to access live Internet information via Google. As artificial intelligence continues to evolve in healthcare, careful verification of AI-generated content and proactive management of its limitations will be critical for safe and effective clinical integration [17]. Indeed, a notable concern with AI-generated content is the phenomenon of "hallucinations", where inaccurate or outdated information is presented in a convincing way [18].

Thus far, the performance of none of these LLMs has been evaluated on the Advanced Trauma Life Support (ATLS) exam. The Advanced Trauma Life Support (ATLS) program, created by the American College of Surgeons (ACS) Committee on Trauma, aims to train healthcare providers in the immediate evaluation and management of trauma patients [19]. The objective of the ATLS course is to teach the knowledge and skills necessary to effectively evaluate, stabilize and treat trauma patients during the "golden hour" immediately following injury, during which effective medical treatment can prevent death.

This study aimed to compare the performance of GPT-3.5, Gemini, and GPT-4 on the ATLS exam with a particular focus on accuracy and the occurrence of hallucinations based on question type.

2. Methods

The ATLS exam consists of a written multiple-choice exam and a practical exam evaluating the participant's skills in a simulated trauma scenario. For this study, only the written portion of the exam was assessed. The 10th Edition of the ATLS exam prepared by the ACS (updated October 2020) was used. There were three different ATLS exams with 40 multiple-choice questions each (120 questions in total). Each question had five answer choices, with a single most correct answer. There is no penalty for incorrect answers. The passing threshold for each exam is 75% (30 out of 40 questions correct). Questions were also categorized by question type: direct questions that tested basic knowledge in a straightforward manner or clinical scenarios that tested applied knowledge in the context of a trauma situation. Question categorization was independently performed by two authors, with disagreements adjudicated by the senior author.

GPT-3.5, Gemini, and GPT-4 were queried with the 120 exam questions on July 18th, 2023. Each answer provided by an LLM was compared to the exam answer key provided by the ACS. In the case of a wrong answer, the question was asked again, and in the case of a correct answer on the second attempt, the response to a third attempt was deemed to be the conclusive answer. Questions that an LLM opted not to respond to were considered incorrect. To minimize potential bias related to memory retention within the LLMs, new user accounts were created for this study, and chat histories were cleared between each question prompt.

All statistical analyses and visualizations were conducted using Prism 9.0 (GraphPad Software Inc., CA) and Microsoft Excel (Redmond, WA). Categorical variables were summarized as frequencies and percentages, while continuous variables were reported as means and standard deviations. Differences in LLM performance were calculated using a chi-square test. Chi-square tests were also performed to evaluate differences in LLM performance according to question type. A p-value < 0.05 was considered statistically significant.

3. Results

On the ATLS exam, GPT-3.5 received an average score of 62.5 ± 6.6% (75 out of 120), Gemini 61.7 ±8.8%

(74 out of 120), and GPT-4 83.3 ±1.4% (100 out of 120). The average scores received by the three LLMs, as well as scores for each individual exam, are shown in Table 1. If GPT-3.5, Gemini, and GPT-4 were candidates in the ATLS course, GPT-4 would have passed on all three exams since it scored more than 75%, whereas GPT-3.5 and Gemini both would have failed all three exams. There was no statistically significant difference between the average scores of GPT-3.5 and Gemini (65.0% vs. 61.7%, $p = 0.59$). Meanwhile, GPT-4 achieved significantly higher scores compared to both GPT-3.5 (83.3% vs. 65.0%, $p = 0.0012$) and Gemini (83.3% vs. 61.7%, $p = 0.0002$).

Table 1. GPT-3.5, Gemini, and GPT-4 Performance on Three ATLS Exams

LLM	Exam 1, % (no. correct)	Exam 2, % (no. correct)	Exam 3, % (no. correct)	Mean ± SD (total no. correct)
GPT-3.5	67.5 (27)	55 (22)	65 (26)	62.5 ± 6.6 (75)
Gemini	62.5 (25)	52.5 (21)	70 (28)	61.7 ± 8.8 (74)
GPT-4	82.5 (33)	82.5 (33)	85 (34)	83.3 ± 1.4 (100)

Note: LLM: Large Language Model; no.: Number; SD: Standard Deviation

Among the 120 total questions, 52 (43.3%) were direct questions, while 68 (56.7%) were presented as clinical scenarios. On the direct questions, GPT-3.5 scored 65.4% (34 out of 52), Gemini scored 61.5% (32 out of 52), and GPT-4 scored 86.5% (45 out of 52). On the clinical scenarios, GPT-3.5 scored 64.7% (44 out of 68), Gemini scored 61.8% (42 out of 68), and GPT-4 scored 80.9% (55 out of 68). No difference in performance was found based on the type of question (direct vs. clinical scenario) for GPT-3.5 (65.4% vs. 64.7%, $p = 0.9384$), Gemini (61.5% vs. 61.8%, $p = 9798$), or GPT-4 (86.5% vs. 80.9%, $p = 0.4100$).

Of note, the trend in LLM performance remained consistent when questions were divided by type. On direct questions, on direct questions, there was no significant difference in performance between GPT-3.5 and Gemini (65.4% vs. 61.5%, $p = 0.6838$), whereas GPT-4 outperformed GPT-3.5 (86.5% vs. 65.4%, $p = 0.0116$) and Gemini (86.5% vs. 61.5%, $p = 0.0036$). On clinical scenarios, there was no significant difference in performance between GPT-3.5 and Gemini (64.7% vs. 61.8%, $p = 0.7221$), whereas GPT-4 outperformed GPT-3.5 (80.9% vs. 64.7%, $p = 0.0340$) and Gemini (80.9% vs. 61.8%, $p = 0.0137$). The average performance of the three LLMs overall, and on direct and clinical scenario question types, are shown in Fig. 1.

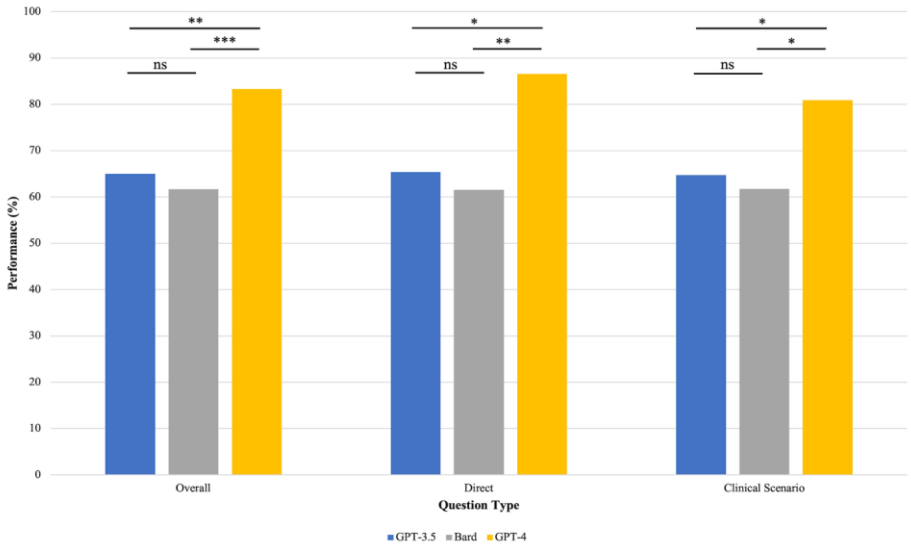

Fig. 1. GPT-3.5, Gemini, and GPT-4 performance by question type. Results of comparative chi-square tests are also shown. *$p < 0.05$; **$p < 0.01$; ***$p < 0.0001$; *ns*, not significant.

4. Discussion

The ATLS certification aims to train healthcare professionals to immediately and accurately evaluate, prioritize, and manage patients with life-threatening injuries [19]. While AI's knowledge of basic medical topics has been explored, comprehensive investigations into specific areas like burns and trauma are lacking [8, 20, 21]. This study is the first to evaluate the knowledge of multiple LLMs on trauma.

If GPT-3.5, Gemini, and GPT-4 were participants in the ATLS course, only GPT-4, which achieved a score of 83.3%, would have passed. Both GPT-3.5 and Gemini received scores below the 75% passing threshold. Thus, GPT-4 significantly outperforms its GPT-3.5 predecessor and Google competitor, on both direct and clinical scenario questions. The disparity in LLM performance may be due, in part, to the different datasets on which LLMs were trained. Whereas Gemini was trained on Google's Infiniset, a dataset of internet content such as public forums that are deliberately curated to enhance Gemini's conversational tone, GPT-4 may have been trained on a larger dataset that focused on technical and applied knowledge ("Google Bard AI—What Sites Were Used To Train It?") [22]. Information on ATLS test questions, which are likely based on updated and specialized knowledge that is not widely available in open-access internet sources and "grey literature," is thus more likely to be found in the latter dataset. This serves as a potential explanation for GPT-4's superior performance on the ATLS exam. However, little is known about the exact parameters and algorithm employed by GPT-4, as OpenAI has not released the technical details of this product. Moreover, it is important to note that GPT-4 is a paid subscription service with a limited number of questions that the user can ask every 4 h. Thus, although GPT-4 is technically more advanced than GPT-3.5 and Gemini, it is currently not as accessible as the latter LLMs. However, the issue of the limited accessibility of more advanced LLMs may become less limiting as tech companies release newer models, rendering the latest ones outdated and, consequently, free or easier to access.

One of the most concerning findings of this study was that when GPT-3.5, Gemini, or GPT-4 were incorrect, they often provided erroneous explanations that, at first glance, seem very credible. These incorrect explanations, which have been identified in other LLM studies as well, are described as "hallucinations" [18]. While incorrect information was generally not observed for basic medical knowledge, it is important to recognize that the risk of disseminating misinformation becomes more pronounced when querying LLMs for more intricate technical knowledge. For example, hallucinations are particularly alarming in the context of high-risk trauma situations, where incorrect information presented by LLMs with confidence could potentially result in life-threatening mistakes. Thus, scrutiny is needed when using LLMs to avoid the repercussions of relying on misleading or inaccurate medical knowledge presented as fact. Nevertheless, it is important to note that LLMs are advancing at a remarkable pace, as demonstrated by the significantly better performance of GPT-4 on the ATLS exam compared to GPT-3.5. Thus, as improved training sets and Internet access capabilities are integrated into updated LLMs, the incidence of hallucinations will likely decrease, contributing to an overall improvement in their performance.

LLMs have tremendous potential to revolutionize healthcare, particularly in the time-dependent and high-risk context of trauma care. Through their interactive interface and rapidly accessible knowledge repositories, LLMs offer the ability to supplement medical decision-making processes and bolster the education of healthcare practitioners and patients alike. This is particularly valuable in regions that are distant from trauma centers. In such areas, LLMs can serve as an invaluable resource, providing insights and guidance that might otherwise be inaccessible in critical situations. Indeed, many people may be using LLMs as a "curbside consult", including medical students, residents, and non-specialists. Moreover, there are ongoing efforts to integrate AI into medicine, with Epic and Microsoft working to integrate this technology into the electronic health record to improve productivity and patient communication ("Epic and Microsoft Bring GPT-4 to EHRs") [23]. The adoption of artificial intelligence in medicine is inevitable, whether or not it

is universally embraced.

It is therefore essential for researchers, clinicians, and healthcare professionals to develop a thorough understanding of the current capabilities of LLMs, especially in areas where they can independently validate the information. This understanding will empower them to make informed decisions about the suitability and limitations of LLMs, particularly when used as curbside consults. Through continued research efforts, the medical community can establish the groundwork for the careful integration of LLMs into emergency medicine, ultimately improving outcomes for both patients and providers. For example, with the increasing development of advanced LLMs, another prospective avenue of research involves leveraging the success rates of LLMs to gauge the complexity of test questions—whether they lean towards specialized or fundamental knowledge. This methodology could offer significant insights into question composition, assisting in striking a balance between specialized and foundational understanding. Moreover, investigating AI's potential as a supplementary "back-check" mechanism rather than a substitute for human judgment presents promising opportunities to refine LLM testing procedures and enhance results. Importantly, this study is an early contribution to the ethical and practical discourse on LLM performance, shedding light on the quantitative differences in LLM performance between burns and trauma, using data trained on open access internet sources and "grey literature" rather than exclusively on peer-reviewed papers.

Despite the excitement and potential benefits of the integration of AI in medicine, it is important to recognize the potential drawbacks. One particularly concerning drawback highlighted by this study is the proliferation of misinformation through "hallucinations", which has the potential to adversely affect both patient and physician decision-making. For this reason, it is essential to validate LLM outputs. The authors would like to emphasize that this study does not necessarily validate the use of LLMs as a replacement for clinical judgment in trauma situations, as the results revealed that only the most advanced LLM, GPT-4, which is not yet accessible to most users, passed the ATLS exam. However, as AI technology improves, more updated LLMs may be validated. Therefore, it is crucial to balance the adoption of validated AI technologies with the preservation of human clinical judgment. This is particularly relevant in trauma situations such as mass casualty events with hostile intervention, where LLMs may not be available. While LLMs are a useful tool, they cannot replace the clinical judgment, hands-on experience, and human connection required of emergency healthcare providers. Ongoing research and ethical discussion will pave the way for LLM integration in emergency medicine, in a way that benefits both patients and healthcare providers without undermining the invaluable capabilities of human cognition.

5. Conclusion

This study serves as an initial benchmark in assessing LLM performance in trauma scenarios. GPT-4 passed the ATLS exam, outperforming GPT-3.5 and Gemini. This is promising for the future of LLM application in trauma care. However, due to hallucinations and continuing ethical considerations, LLMs are best suited as a complement to human cognition. Future research should focus on validating LLM performance across a broader range of clinical scenarios, exploring strategies to minimize hallucinations, and developing best practices for safe and effective integration of AI into emergency medical workflows.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

H.Y.L. performed literature search, study design, data collection, data analysis, data interpretation, writing.

M.A.B. performed literature search, study design, data collection, data analysis, data interpretation, writing. A.C.C. performed critical revision. J.A.Z. performed critical revision. F.M.E. performed study design, critical revision, and supervision of the project.

References

[1] Alessandri-Bonetti, M., Giorgino, R., Naegeli, M., Liu, H. Y., & Egro, F. M. (2023). Assessing the soft tissue infection expertise of ChatGPT and Bard compared to IDSA recommendations. *Annals of Biomedical Engineering*.

[2] Alessandri-Bonetti, M., Liu, H. Y., Giorgino, R., Nguyen, V. T., & Egro, F. M. (2023). The first months of life of ChatGPT and its impact in healthcare: A bibliometric analysis of the current literature. *Annals of Biomedical Engineering*.

[3] Alessandri-Bonetti, M., Liu, H. Y., Palmesano, M., Nguyen, V. T., & Egro, F. M. (2023). Online patient education in body contouring: A comparison between Google and ChatGPT. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 87*, 390–402.

[4] Carrarini, M. J., Liu, H. Y., Perez, C. K., & Egro, F. M. (2025). Evaluating Large Language Model's accuracy in current procedural terminology coding given operative note templates across various plastic surgery sub-specialties. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *106*, 50–52.

[5] Jeong, T., Liu, H., Alessandri-Bonetti, M., Pandya, S., Nguyen, V. T., & Egro, F. M. (2023). Revolutionizing patient education: ChatGPT outperforms Google in answering patient queries on free flap reconstruction. *Microsurgery*, *43(7)*, 752–761.

[6] Liu, H. Y., Alessandri-Bonetti, M., Arellano, J. A., & Egro, F. M. (2024). Can ChatGpt be the plastic surgeon's new digital assistant? A bibliometric analysis and scoping review of ChatGpt in plastic surgery literature. *Aesthetic Plastic Surgery*, *48(8)*, 1644–1652.

[7] Liu, H. Y., Alessandri Bonetti, M., De Lorenzi, F., Gimbel, M. L., Nguyen, V. T., & Egro, F. M. (2024). Consulting the digital doctor: Google versus ChatGPT as sources of information on breast implant-associated anaplastic large cell lymphoma and breast implant illness. *Aesthetic Plastic Surgery*, *48(4)*, 590–607.

[8] Liu, H. Y., Alessandri Bonetti, M., Jeong, T., Pandya, S., Nguyen, V. T., & Egro, F. M. (2023). Dr. ChatGPT will see you now: How do Google and ChatGPT compare in answering patient questions on breast reconstruction? *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *85*, 488–497.

[9] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., *et al.* (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2(2)*, e0000198.

[10] Bhayana, R., Krishna, S., & Bleakney, R. R. (2023). Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. *Radiology*, *307(5)*, e230582.

[11] Hopkins, B. S., Nguyen, V. N., Dallas, J., Texakalidis, P., Yang, M., Renn, A., Guerra, G., *et al.* (2023). ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *Journal of Neurosurgery*, 1–8.

[12] Lum, Z. C. (2023). Can artificial intelligence pass the american board of orthopaedic surgery examination? orthopaedic residents versus chatgpt. *Clinical Orthopaedics and Related Research*, *481(8)*, 1623–1630.

[13] 13 Mihalache, A., Popovic, M. M., & Muni, R. H. (2023). Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmology*, *141(6)*, 589–597.

[14] Hoch, C. C., Wollenberg, B., Lüers, J.-C., Knoedler, S., Knoedler, L., Frank, K., Cotofana, S., *et al.* (2023).

ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *European Archives of Oto-Rhino-Laryngology*, *280(9)*, 4271–4278.

[15] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., *et al.* (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*.

[16] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *The New England Journal of Medicine*, *388(13)*, 1233–1239.

[17] 17 Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGpt: Implications in scientific writing. *Cureus*, *15(2)*, e35179.

[18] Gascon, G. M., Steinberg, S., Kovach, S., & Falcone, R. E. (2022). Mixed methods approach to understanding the influence of changes in successive versions of the advanced trauma life support training program on student performance. *Surgery*, *171(3)*, 584–589.

[19] Alessandri-Bonetti, M., Liu, H. Y., Donovan, J. M., Ziembicki, J. A., & Egro, F. M. (2024). A comparative analysis of ChatGPT, ChatGPT-4, and Google bard performances at the advanced burn life support exam. *Journal of Burn Care & Research*, *45(4)*, 945–948.

[20] Alessandri-Bonetti, M., Liu, H. Y., Donovan, J. M., Ziembicki, J. A., & Egro, F. M. (2024). A comparative analysis of ChatGPT, ChatGPT-4, and Google bard performances at the advanced burn life support exam. *Journal of Burn Care & Research*, *45(4),* 945–948.

[21] Alessandri Bonetti, M., Giorgino, R., Gallo Afflitto, G., De Lorenzi, F., & Egro, F. M. (2023). How does chatgpt perform on the italian residency admission national exam compared to 15,869 medical graduates? *Annals of Biomedical Engineering*.

[22] Google Bard AI—What Sites Were Used To Train It? Retrieved from https://www.searchenginejournal.com/google-bard-training-data/478941/

[23] Epic and Microsoft Bring GPT-4 to EHRs. Retrieved from https://www.epic.com/epic/post/epic-and-microsoft-bring-gpt-4-to-ehrs