Clinical Med students' validation of Arkangel AI: Are their responses any better when supported by the AI?

Natalia Castano-Villegas[1]*, Isabella Llano[2], María Camila Villa[3], Jose Zea[4] , Laura Velasquez[5].
Arkangel AI

[1] Physician and Epidemiologist, CID Specialist- Research Leader, Arkangel AI
[2] Biomedical Engineer- Master Student, The Pennsylvania State University
[3] Machine Learning Engineer- Arkangel AI
[4] CEO, Arkangel AI
[5] President, Arkangel AI

*Corresponding author:
Natalia Castaño-Villegas
Cra 17 #103–31, Bogotá, Colombia
Email: natalia@arkangel.ai
Telephone +57 313 7191729

ABSTRACT (295)

Introduction: Large Language Models (LLMs) in healthcare practice and education have been evaluated using medical question-answering (QA) datasets, with excellent performance. However, multiple-choice questions fall short when assessing more complex language interactions.
Objective: To evaluate the time invested and validity of medical students' responses to clinical questions using ArkangelAI, compared to traditional search methods.
Methods: Randomized, double-blind trial with clinical medical students assigned to two groups. Each group answered four clinical questions from each of four clinical cases, one using ArkangelAI, the other using traditional research methods: Google, PubMed, etc. Field specialists evaluated the responses using six pre-established criteria to define the answers' validity. Total average validity (the mean of individual scores) was compared by groups with hypothesis testing and 95% CI. The time to respond was also compared.
Results: Eighty-three medical students were randomized to groups A (43) and B (40). Average differences responded in half the time (three minutes faster) than the control group, with 98% fewer searches needed. The model's answers were valid (accurate, non-biased, aligned with consensus, and safe) with a total validity score of 2.84 (group A) and 2.69 (group B). Most Arkangel AI users found it helpful for daily practice and would recommend it to colleagues.
Conclusion: LLM-supported methods appear to have a positive influence on effective clinical search without sacrificing, and even augmenting, the quality of answers. This is applicable at the clinical medical student level and for non-critical clinical reasoning. Validations, including graduated physicians and specialists, are needed to further understand the effect of LLMs in education and clinical practice.

Key words: LLM assessment, Human Evaluation, Healthcare, Real-World Validation

INTRODUCTION

Medical pedagogy and practice are increasingly influenced by AI, particularly large language models (LLMs), which have demonstrated expert-level knowledge in clinical reasoning. However, the assessment of medical LLMs presents limitations. Evaluations frequently rely on single undergraduate medical exams or multiple-choice questions, which determine performance on static databases (1,2).

While LLMs have shown impressive performance on these tests (3–5), such evaluations may not fully capture the nuances of qualitative reasoning and practical clinical application (6). There is a recognized need for more sophisticated and qualitative approaches to assess the responses generated by LLMs, moving beyond accuracy compared to standardized tests (7).

This calls for a deeper exploration into the validity and quality of AI-generated content in medical and academic contexts. Ensuring the reliability and accuracy of AI outputs is paramount for safe and effective integration into medical education and practice (8). To address these issues, we developed ArkangelAI, an LLM-powered CA specialized in medical information. The app performs real-time internet searches based on the best available scientific evidence to answer medical questions, providing direct links to curated literature references. Our first manuscript presented its development and internal validation. It demonstrated state-of-the-art accuracy (90.26%) compared to other LLMs like GPT 4o and MedPaLM2 when using the same public medical question-answering (QA) database (the MedQA) and the exact sample sizes (2723 QA) (Table 1) (6).

In this paper, we intend to perform the first of a series of external validations of Arkangel in clinical medical students and physicians. We asked medical students to respond to clinical questions based on clinical outpatient scenarios, with and without using ArkangelAI or traditional search practices. This manuscript analyzes their knowledge outcomes.

Table 1. Reported SoTA LLMs on the MedQA (USMLE) test set, including Arkangel AI.

METHODOLOGY 1150

Study Design and Sample Size

This is a randomized trial, where eighty-three medical students are randomly allocated to two groups (A and B) to answer clinical questions using ArkangelAI (group A) vs traditional research methods, including Google search, medical libraries, books, guidelines, and colleagues, except another AI (group B) (Figure 1). The leading researchers were blinded to the subjects' identities and their allocation. Each group's answers were evaluated by expert physicians, external to Arkangel AI (from private practices or teaching hospitals). They were also blinded to the students' identities and the group from which answers came. The answers in which AI was used were indistinguishable from those that were not; they were mixed in the same matrix and were set to random distribution to avoid introducing bias due to observer fatigue (9).

Figure 1. Participant Flow chart

The students were recruited through social media, medical professional groups, the snowball method, and masterclasses in medicine faculties in Colombia, including Facultad de Medicina Universidad de Antioquia, Universidad de los Andes, Universidad del Bosque, and Universidad CES. Upon first contact, they completed a form in the Notion app (10) and received an automatically generated email containing the overview of the research and informed consent form (Supplementary Material 1). Once signed, they were randomly assigned to groups A or B, using Microsoft Excel 365 formulae.

The selection criteria were any medical student in their clinical practice years (fourth year and above), willing to participate, from a certified Colombian higher education medical institution. Subjects in both groups received a compensation of 150,000 COP (Colombian Pesos, equivalent to 38 USD) and a one-year subscription to ArkangelAI.

Selection criteria for specialist reviewers were a degree in the respective field, five or more years of experience, and at least three years of outpatient consultation in a tertiary care medical institution. The selected experts do not have any ties to Arkangel AI and signed a consent form indicating this. They were offered 1,200,000 COP for eight hours of revision, the estimated time for each specialist to revise the 1328 question-answer Q-A pairs (Figure 1). Each case was only evaluated by the expert in that field.

The experiment consisted of four pre-determined real-world-like clinical cases, created by specialists different from the reviewers, also external to Arkangel AI, who willingly provided them. The selected cases were in orthopedics, psychiatry, pediatrics, and gynecology (Supplementary Material 2). The cases were selected based on convenience: they represented fictitious, non-urgent, outpatient scenarios to test not the students' knowledge per se, but their answer performance when aided by traditional research methods or AI. The clinical cases followed the structure of an electronic health record (EHR) regulated by the Colombian Ministry of Health (11–15).

Materials and Methods 1524

We developed four types of questions for each case: diagnostic, initial approach, research, and general knowledge. Group A would study the cases and respond using ArkangelAI as a searching tool. Group B would answer the questions using traditional search strategies, excluding AI agents. Participants developed the test online, asynchronically, and did not have a time limit to answer. We used Quizizz (16) to create user-friendly questionnaires.

We also used Airtable (17), where the specialist's assessment was registered. We used Python 3.12.2, Cursor (18), Jamovi, and Microsoft Excel for the data analysis. The complete database with the students' responses and specialists' assessment is presented as Supplementary Material B.

Outcomes

To assess the impact of using Arkangel AI vs. traditional searching tools, we evaluated three aspects: 1. Model acceptability, 2. Answer efficiency, and 3. Validity of answers.

Validity of answers

This construct was evaluated using six Likert-like questions, developed by the team's epidemiologists and expert physicians (Table 2). These questions were based on essential aspects of Evidence-Based Medicine: correctness of response, agreement with current consensus, no introduction of sociodemographic biases, treatments aligned with the standard of care, updated information, and safe recommendations.

To assess answer validity, the specialists scored each QA pair answered by the participants. All four experts were blinded to each other's evaluations and the participant's group. They assigned a mark to each answer on a three-point scale.

Table 2. Questions to assess answer validity

The Total Average Validity Score was the average punctuation for the six criteria, which were compared by hypothesis testing according to the variables' distribution, using a p-value<0.05 for statistical differences. The 95% Confidence Intervals for averages are presented. The relative differences in frequencies were stated using the formula $(final\ value - initial\ value)/initial\ value) * 100$ (19), where the final value was the average for Group A and the initial value was the average for Group B. These analyses are also performed by medical specialty and type of question

Answer efficiency

We measured the average time and the average number of searches required to respond to each question. We used the times automatically recorded in the *Quizziz* platform to calculate the time per case. For the number of searches, we asked each participant to record the number of searches required per question in each case. Time to respond was compared between groups, using both means and medians to consider extreme values.

Model acceptability

We asked the participants allocated in group A to score their experience using ArkangelAI by answering four pre-defined questions developed by the research team, using a three-point scale evaluation (Table 3). We assessed their perceived usefulness of the model, confidence in its responses, the likelihood of daily use, and the likelihood of recommending it to a fellow student.

Table 3. Acceptability questions and possible values

Model Description

We described the development and internal validation of Arkangel AI in a previous paper (6). Improvements and iterations are ongoing, defining a fluid process in which efficiency, confidence, security, and satisfaction are priorities. Since then, the model has had two important updates:

The first one was for workflow classification (Figure 1, Table 4). Our previous system had five LLMs to classify the information into workflows according to the type of question (diagnostic, clinical management, research, common knowledge). Although it had an accuracy of 95% compared to the human standard, it presented an average response time of 2.63 minutes per question, a very long time in fast-paced clinical scenarios.

The second was for the retrieval. For our previous model, retrieval was made using Google and PubMed's
However, only 80% of the retrieved contexts were relevant

In the current model, the workflow is selected by the user, which prevents reprocessing, and we use the AI search engine Tavily (20) for retrieval. This system works like a traditional Search API; the difference is that it uses a proprietary AI to search, scrape, filter, rank, and extract information from online sources. These two modifications improved efficiency and response time by simplifying the steps and providing a faster retrieval.

Supplementary Material 3 provides a more detailed description of the system´s modifications, and Supplementary Material 4 explains the architecture and connectivity of the Tavily retriever and the Arkangel workflows, for reproducibility and transparency, together with a list of all Arkangel AI's sources. The specific prompts are available upon reasonable request.

Figure 2. Previous and current system architecture

Table 4. LLMs used by each of the current model workflows

RESULTS 433

Eighty-three medical students answered 1328 questions from 332 clinical case units (clinical case +4 QA pairs) (Figure 1). Using randomized allocation, 43 subjects were placed in group A (intervention with ArkangelAI) and 40 in group B (traditional search resources). Two participants assigned to group A were excluded due to a platform error, which resulted in them not answering any questions, and one participant from group B was excluded for not following instructions and using ChatGPT to answer the questions. One test in group A was incomplete and was not considered in the validity or acceptability analysis. One quiz was duplicated.

Validity of answers compared by groups

Table 5 shows a general comparison of the scores obtained in each group for the six validity questions. Using the Mann-Whitney U (not normal distribution), we compared median scores between groups and found statistical differences between all six answers, with Group A consistently scoring higher than Group B in all categories. Table 5 also displays the relative change in scores, showing the most significant increase in response accuracy (12.08%) and the lowest in treatment bias (0.70%).

Table 5. Average score for validity questions by group with 95% IC and hypothesis test

Validity score by specialty and type of question

Group A achieved significantly higher validity scores than Group B across all specialties and question types (both Mann–Whitney' U and Kruskal-Wallis' test p<0.01). Among specialties, the most significant relative improvement was in gynecology (11.48%), while the smallest was in psychiatry (1.75%) (Post Hoc Pairwise comparisons – DSCF method´s p-value was 0.005). Regarding question types, research questions showed the greatest improvement (9.47%), whereas general knowledge questions showed the least (2.95%) (DSCF´s p-value < 0.001). Full results are presented in Supplementary Material 5 (specialties) and Supplementary Material 6 (question types).

Efficiency by groups

Table 6 presents the efficiency results for Groups A and B. The time between cases was determined using the records from the Quizziz platform. The times between questions were calculated in Jamovi V2.6.45, where each observation had an automatically assigned weight. On average, Group A took almost four minutes less than Group B. Regarding the number of searches, Group B conducted more than twice as many searches as Group A. The average number of searches per case was 14 for Group A and 30 for Group B, confirming the previous statement.

Table 6. Efficiency results between groups: Average time and number of searches.

Tool acceptability

Figure 2 shows the average scores obtained for the acceptability questions in Table 3, as ranked by subjects in group A. Perceived tool utility, likelihood of daily use, and recommendation to a colleague were scored 2.98, 2.88, and 2.92, respectively. The lowest score was observed for confidence in the model's truthfulness, with 2.63. Total average tool acceptability is 2.85.

Figure 3. Average acceptability score for Arkangel AI (only for Group A)

DISCUSSION 1508

This is intended to be the first external validation of a series of studies assessing how ArkangelAI speeds up the process of finding clinically sound answers with the quality of a trained medical professional. Unlike traditional LLM evaluation approaches, which focus on standardized multiple-choice QA exams, this research takes a more analytical approach. It employs expert physicians' knowledge as the standard of reference against which the human-in-the-loop agent's answers are appraised regarding their quality and validity. It also compares the time taken using the different research methods.

The model's performance can translate into abilities such as effective clinical information search, analysis, and synthesis skills, cognitive support, and evidence-based decisions. These abilities help the learning process and could act as pedagogical support, with AI digital medical education used as a tool to support evidence-based learning and early clinical reasoning (21–23).

Our findings support these claims and are consistent with recent studies showing how AI-based conversational assistants can improve the quality of clinical reasoning in medical students, provided they are integrated critically and supervised within training (21), and how they save time by obtaining faster responses, maintaining or augmenting their quality.

Efficiency and Acceptability

The results demonstrated that students who used Arkangel AI solved clinical questions in significantly less time than the control group (median 2:54 vs. 5:28; relative reduction of 52%) and with fewer searches (3 vs. 6; relative decrease of 50%). Overall, the AI group required half the time and fewer than half the searches that the traditional search group. These differences were statistically significant (Mann–Whitney U test, $p < 0.01$). These suggest that Arkangel AI can enhance cognitive efficiency and enhance information seeking, essential skills for self-directed learning and evidence-based medical practice (24,25).

Overall acceptability was high (2.85 on a scale of 1 to 3), highlighting its usefulness and participants' willingness to recommend it to their peers. However, the dimension with the lowest score was confidence in the accuracy of the responses (2.63), a finding consistent with that reported by Sakamoto et al. (26), among others (27), who document skepticism towards artificial intelligence systems in clinical and educational contexts. Rather than a weakness, this result represents an opportunity to critically assess how the supervised use of AI can strengthen digital competence, encouraging source verification, responsible interpretation of evidence, and understanding of the limitations of language models.

To promote informed trust, strategies of explainability and transparency should be demanded from any clinical LLM, such as direct visualization of the sources consulted, indication of the level of evidence, explicit reference to clinical guidelines, and breakdown of the reasoning followed by the model. This type of transparency not only improves the perception of reliability but also enhances the pedagogical value of the tool by transforming each interaction into an active exercise in learning and critical reflection.

Validity

The validity of the responses, evaluated by independent experts in six dimensions (accuracy, consistency with guidelines, absence of demographic or therapeutic bias, timeliness, and safety), was consistently higher in the ArkangelAI group. The most significant increase in performance was observed in accuracy (10.78%, Mann–Whitney U p-value <0.01), followed by unbiased and updated information. These results indicate that AI can improve student responses by facilitating access to curated and up-to-date scientific literature, thereby reinforcing their training in

evidence-based medicine. Total Average Validity Score was 2.84 for group A and 2.69 for group B (Mann–Whitney U p-value <0.01)

Furthermore, these findings align with recent language model evaluation frameworks that prioritize dimensions such as accuracy, consistency, bias, safety, and clinical relevance (28–31), supporting the potential for supervised use of ArkangelAI to strengthen the quality, reliability, and educational value of responses produced in medical training settings.

The validity construct was also explored by specialty (Gynecology, Psychiatry, Orthopedics, and Pediatrics) and type of question (diagnostic, management, research, and general knowledge). Group A consistently scored higher than Group B. Absolute and relative differences were statistically significant, with no overlap in 95% confidence intervals. All hypothesis tests revealed p<0.05 (Supplementary Materials 5 and 6).

The most significant performance improvement was for gynecology (relative difference 11.4%) compared to psychiatry (2.7%, Mann–Whitney U p-value <0.01), which is consistent with analysis describing that specialties with more patient interaction must respond to constant changes and will thus be the least "impacted" by AI, while clinical and surgical components will be more influenced)(25). Research questions had the greatest improvement with 8.65% (Mann–Whitney U p-value <0.01), compared to general knowledge with 2.87% (Kruskal-Wallis DSFC post hoc correction's p-value <0.001, Supplementary Materials 8 and 9)

Although large language models (LLMs) have been widely described as clinical support tools, few studies have evaluated their impact in medical training contexts. Goh et al. assessed ChatGPT-4 in diagnostic reasoning using a randomized design with 50 physicians. They found no significant differences between groups, likely due to their limited sample size and focus exclusively on diagnostic performance(32).

In contrast, our study covered multiple domains of knowledge (diagnosis, management, research, and general knowledge) and showed significant differences in all dimensions evaluated. The differential performances observed in areas such as gynecology, pediatrics, or psychiatry are consistent with previous research showing variability in the accuracy of LLMs depending on the medical specialty (33).

Human evaluation frameworks have also been proposed to prioritize the accuracy, consistency, bias, and safety of responses generated by LLMs, seeking to ensure the validity and reliability of information in educational and clinical settings (34). In Latin America, although the adoption of AI in medical education is in its infancy, its potential for undergraduate learning contexts is particularly relevant (35). In this regard, the present study provides evidence on the pedagogical value of ArkangelAI as a tool to support evidence-based learning in medical students.

Strengths and limitations

This study is one of the first external validations of an AI-based tool (Arkangel AI) applied to undergraduate students in a medical education context. Its randomized, double-blind design, clinical cases developed by external specialists, and independent expert evaluation provide methodological rigor and bias control. In addition, applying a multidimensional validity construct that included accuracy, concordance with guidelines, bias, timeliness, and safety offers a more comprehensive approach than traditional studies based solely on precision or accuracy of responses.

One methodological limitation was that responses were evaluated by a single specialist in each clinical area. This could introduce bias as the standard of reference is subject to their clinical educated opinion, and could be influenced by many human factors, like the specialist's own experience with medical AI. In addition, it prevents the estimation of inter-rater variability, as each specialist only evaluated the case of their expertise. Therefore, even if the reviewers met experience criteria and were independent of Arkangel AI, some individual variability in scoring, inherent in human evaluations, cannot be ruled out. Nevertheless, human judgment remains the reference standard for LLMs

(36). We plan to include more than one evaluator per specialty to assess the consistency and reproducibility of scores in future validations among medical staff.

Furthermore, the clinical cases used were fictitious, field-limited, and constrained to outpatient and low-complexity scenarios. Consequently, the results can only be applied to similar contexts for medical students at the clinical level. The results also reflect academic reasoning rather than clinical decision-making.

Another aspect is the potential conflict of interest, as Arkangel AI employs all authors. Although this relationship was transparently declared and mitigated by external evaluators and blind analysis, the need for independent validation by universities or hospitals to ensure replicability and scientific credibility is underscored. To address this further, we provide the project's datasets for the replication of the analyses and results. We also describe the methodology and results as transparently as possible.

Since we did not use or divulge any patient data, we did not seek the approval of an ethics committee for this study, nor did we register it as a clinical trial (it does not fulfill the criteria). The cases were fictional and were not based on any patient. Every participant signed a consent form before entering the experiment.

Educational implications and future lines of research

The evidence suggests that ArkangelAI can support students in developing efficient search skills, scientific information analysis, and initial clinical judgment, provided that its use is accompanied by faculty guidance and ethical standards (37)**.** If properly framed, the incorporation of AI-based tools into medical education can become a catalyst for reflective learning and critical thinking (25)

Future research should explore the curricular integration of AI and evaluate its impact on students' cognitive, ethical, and attitudinal competencies, which were not the aim of this study. Similarly, it will be essential to conduct multicenter external validations, expand the sample size, and analyze the effect of the tool in actual clinical practice through studies focused on physicians and specialists that confirm the reproducibility of these findings and consolidate the scientific basis for the ethical and responsible integration of artificial intelligence into medical education and practice. Furthermore, comparison with other AI models will enable Arkangel AI to position itself within the international landscape of educational and clinical tools based on large language models (LLMs) (38–41).

Conclusion

LLM-supported methods appear to have a positive influence on effective clinical search without sacrificing, and even augmenting, the quality of answers. This is applicable at the clinical medical student level and for non-critical clinical reasoning. Validations, including graduated physicians and specialists, are needed to further understand the effect of LLMs in education and clinical practice.

Declarations

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

While preparing this work, the author(s) used Jenni AI (42) and Anara (43) to structure the introduction section. After using this tool/service, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the published article. We also used DeepL Translator (44) to translate the clinical cases into English and Grammarly (45) to help with the writing and editing processes. All scientific insights, data analysis, and conclusions were the sole responsibility of the authors.

Ethics approval and informed consent

This study involved medical professionals who evaluated fictitious clinical cases designed exclusively for research purposes. All participants provided written informed consent prior to participation. No real patient data, medical records, or identifiable personal health information were used or disclosed at any stage of the study.

According to Resolution 8430 of 1993 of the Ministry of Health and Social Protection of Colombia (Ministerio de Salud y Protección Social de Colombia), which establishes the ethical standards for health research in Colombia, this study was classified as minimal risk research. Under Article 11 of this resolution, studies involving exclusively fictitious cases and no use of real patient data are exempt from mandatory review by an institutional ethics committee. Therefore, formal ethical approval by an Institutional Review Board was waived.

The study was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki (2022) [28].

Funding

The project was funded Arkangel AI.

Authorship Statement

All authors met the ICMJE criteria for authorship and approved the final manuscript.

Acknowledgments

We would like to extend our sincere gratitude to the physicians who prepared the clinical cases and to those who participated in the evaluations, including Dr. Juliana Muñoz Restrepo, Dr. Laura Ovadia Cardona, Dr. Estefanía Bahamonde, Dr. Jorge Luis Molina Valderrama, Dr. Nataly Ávila, and Dr. Andrés David Álvarez. Finally, we appreciate the hundreds of physicians and medical students who responded to our invitation and completed their task with distinction, making this project possible.

Conflicts of interest

All authors are employees or founders of Arkangel-AI.

Data sharing

The deidentified datasets generated and analyzed during this study, including the questions, answers, and evaluation scores with their data dictionary, are provided in the supplementary material attached to this manuscript. The corresponding author can provide further details upon reasonable request.

Contributions

NCV: Conceptualization, Methodology, Data curation, Formal analysis, Project administration, Supervision, Writing – review & editing
MCV: Data curation, Software Writing – review & editing
IL: Data curation, Formal analysis, Investigation
JZ: Resources, Supervision, Writing – review & editing

LV: Resources, Writing – review & editing

REFERENCES

1. Wan N, Jin Q, Chan J, Xiong G, Applebaum S, Gilson A, et al. Humans and Large Language Models in Clinical Decision Support: A Study with Medical Calculators [Internet]. arXiv; 2025 [cited 2025 Oct 8]. Available from: http://arxiv.org/abs/2411.05897

2. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering [Internet]. arXiv; 2019 [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/1909.06146

3. Papers with Code - MedQA Benchmark (Question Answering) [Internet]. [cited 2024 Oct 3]. Available from: https://paperswithcode.com/sota/question-answering-on-medqa-usmle

4. Bardhan J, Colas A, Roberts K, Wang DZ. DrugEHRQA: A Question Answering Dataset on Structured and Unstructured Electronic Health Records For Medicine Related Queries [Internet]. PhysioNet; [cited 2024 Aug 28]. Available from: https://physionet.org/content/drugehrqa/1.0.0/

5. Kotschenreuther K. EHR-DS-QA: A Synthetic QA Dataset Derived from Medical Discharge Summaries for Enhanced Medical Information Retrieval Systems [Internet]. PhysioNet; [cited 2024 Aug 28]. Available from: https://physionet.org/content/ehr-ds-qa/1.0.0/

6. Villa MC, Castano-Villegas N, Llano I, Martinez J, Guevara MF, Zea J, et al. Arkangel AI: A conversational agent for real-time, evidence-based medical question-answering. Intell-Based Med. 2025 Jan 1;12:100274.

7. Kim Y, Wu J, Abdulle Y, Wu H. MedExQA: Medical Question Answering Benchmark with Multiple Explanations [Internet]. arXiv; 2024 [cited 2025 Oct 8]. Available from: http://arxiv.org/abs/2406.06331

8. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study. JMIR Med Educ. 2024 July 25;10:e56342.

9. Lardner B, Adams AAY, Knox AJ, Savidge JA, Reed R. Do observer fatigue and taxon-bias compromise visual encounter surveys for small vertebrates? Wildl Res. 2019;46(2):127–35.

10. Notion Labs, Inc. Notion APP [Internet]. Notion Labs, Inc.; Available from: https://www.notion.com/

11. Ministerio de Salud y Protección Social. RESOLUCiÓN NÚMERO 866 DE 2021 [Internet]. RESOLUCiÓN NÚMERO 866 DE 2021 June 25, 2021 p. 25. Available from: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-866-de-2021.pdf?utm_source=chatgpt.com

12. Ministerio de Salud y Proteccion Social. Ley 2015 de 2020 [Internet]. Ley 2015 de 2020 p. 4. Available from: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=105472

13. Ministerio de Salud y Proteccion social. LEY 1438 DE 2011 [Internet]. LEY 1438 DE 2011 p. 44. Available from: https://www.funcionpublica.gov.co/eva/gestornormativo/norma_pdf.php?i=41355

14. Minesterio de Salud y Proteción Social. RESOLUCION NUMERO 1995 DE 1999 (Julio 8) Por la cual se establecen normas para el manejo de la Historia Clínica [Internet]. 1995 DE 1999 p. 7. Available from: https://www.minsalud.gov.co/normatividad_nuevo/resoluci%C3%93n%201995%20de%201999.pdf

15. Ministerio de Salud y Protección Social. LEY 23 DE 1981 [Internet]. LEY 23 DE 1981 p. 12. Available from: https://www.mineducacion.gov.co/1621/articles-103905_archivo_pdf.pdf

16. 2025 Quizizz Inc. (DBA Wayground). Wayground, formerly Quizziz [Internet]. 2025 Quizizz Inc. (DBA Wayground); Available from: https://wayground.com/admin

17. Airtable [Internet]. Available from: https://www.airtable.com/

18. 2025 Anysphere, Inc. Cursor [Internet]. 2025 Anysphere, Inc.; Available from: https://cursor.com/

19. Reyes F, Alvarez Ochoa R, Cordeo G, Coronel A, Llanares M. Lo esencial en Bioestadística Médica. 2024.

20. 2025 Tavily Inc. Tavily [Internet]. 2025 Tavily Inc.; Available from: https://www.tavily.com/

21. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312.

22. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–40.

23. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models [Internet]. arXiv; 2023. Available from: http://arxiv.org/abs/2305.09617

24. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025 Mar 7;25:117.

25. The Medical Futurist. Mapping the Future of Healthcare: Six Practical Foresight Methods You Can Use Today [Internet]. [cited 2025 Apr 30]; 2025. Available from: https://medicalfuturist.com/mapping-the-future-of-healthcare-six-practical-foresight-methods-you-can-use-today/

26. Sakamoto T, Harada Y, Shimizu T. Facilitating Trust Calibration in Artificial Intelligence-Driven Diagnostic Decision Support Systems for Determining Physicians' Diagnostic Accuracy: Quasi-Experimental Study. JMIR Form Res. 2024;8:e58666.

27. Beger J. Not someone, but something: Rethinking trust in the age of medical AI. Eur J Radiol Artif Intell. 2025;3:100038.

28. Abeysinghe B, Circi R. The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches [Internet]. arXiv; 2024 Aug. (arXiv preprint). Available from: http://arxiv.org/abs/2406.03339

29. Ibrahim L, Huang S, Ahmad L, Anderljung M. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks [Internet]. 2024. (arXiv preprint). Available from: http://arxiv.org/abs/2405.10632

30. Hamzic D, Wurzenberger M, Skopik F, Landauer M, Rauber A. Evaluation and Comparison of Open-Source LLMs Using Natural Language Generation Quality Metrics. In IEEE; 2025. p. 5342–51. Available from: https://ieeexplore.ieee.org/abstract/document/10825576

31. Confident AI Blog [Internet]. 2024. LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide. Available from: https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

32. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Netw Open. 2024;7(10):e2440969.

33. Alkalbani AM, Alrawahi AS, Salah A, Haghighi V, Zhang Y, Alkindi S, et al. A Systematic Review of Large Language Models in Medical Specialties: Applications, Challenges and Future Directions. Information. 2025;16(6):489.

34. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024 Sept 28;7(1):258.

35. García Mogollón JM, Rojas Contreras WM, Sanabria M. El rol de la inteligencia artificial en la detección de tendencias emergentes en publicaciones científicas. Rev Científica Gen José María Córdova. 2025 Jan 1;23(49):1–17.

36. van der Lee C, Gatt A, van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. In: van Deemter K, Lin CY, Takamura H, editors. Tokyo, Japan: Association for Computational Linguistics; 2024. p. 355–68. Available from: https://aclanthology.org/W19-8643

37. Ogunyemi O, Rahman M, Aziz T, Singh G. Proposing a Principle-Based Approach for Teaching AI Ethics in Medical Education. JMIR Med Educ. 2024 Feb 15;10(1):e55368.

38. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. 2019 Sept 13; Available from: http://arxiv.org/abs/1909.06146

39. Arora RK, Wei J, Soskin Hicks R, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. ArXiv Prepr ArXiv2505XXXXX [Internet]. 2025; Available from: https://github.com/openai/simple-evals

40. Bedi S, Cui H, Fuentes M, Unell A, Wornow M, Banda JM, et al. MedHELM: Holistic Evaluation of Large Language Models for Medical Tasks. ArXiv Prepr ArXiv250523802 [Internet]. 2025; Available from: https://arxiv.org/abs/2505.23802

41. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. What Disease Does This Patient Have? A Large-Scale Open-Domain Question Answering Dataset from Medical Exams. ArXiv Prepr ArXiv200913081. 2020;

42. 2025 Jenni, Inc. Jenni AI [Internet]. 2025 Jenni, Inc.; Available from: https://jenni.ai/

43. Powered by Mintlify. Anara [Internet]. Powered by Mintlify; Available from: https://anara.com/new

44. Deepl. Deepl Translator [Internet]. Deepl; Available from: https://www.deepl.com/

45. Grammarly. Grammarly [Internet]. Grammarly; Available from: https://app.grammarly.com/

Tables

Table 1. Reported SoTA LLMs on the MedQA (USMLE) test set, including Arkangel AI.

Model	Test set questions used (n)	Accuracy
Arkangel AI v3	1273	90.26%
GPT-4o	1273	87.51%
GPT-4o mini	1273	73.61%
Med-PaLM 2	1273	85.40%
GPT-4 (Medprompt)	255	90.20%

Table 2. Questions to assess answer validity

Score	**1**	**2**	**3**
The information in the response is	Incorrect	Partially Correct	Correct
Does the response agree with the medical/scientific consensus?	There is no consensus	Opposed to consensus	Aligned with consensus
Does the response contain information that does NOT apply or is inaccurate for a specific demographic group?	Yes	Partially	No
Does it favor any particular intervention or medication not considered standard of care?	Yes	Partially	No
Does the response contain outdated information?	Yes	Partially	No
Does the response pose a risk of harm to the life or integrity of the patient?	High	Medium	Low

Table 3. Acceptability questions and possible values

Score	**1**	**2**	**3**
Were the responses helpful in answering the questions?	Did not help me	Told me exactly what I already knew	Did help me
How confident are you in the truthfulness of the tool's responses?	Not confident	Neither confident nor not confident	Confident
Would you use this tool in your daily	Probably not	Not sure	Probably yes

practice?			
Would you recommend this tool to your colleagues?	Probably not	Not sure	Probably yes

Table 4. LLMs used by each of the current model workflows

Workflow	Model	Description
Clinical	**GPT-4o**	For clinical questions, the search is directed towards indexed medical literature found in sources such as PubMed, Web of Science, Google Scholar, Research Gate, Scielo, governmental health agencies, Embase, Science Direct, and more.
Pharma	**GPT-4o**	For questions about pharmaceuticals. Direct the search to https://dailymed.nlm.nih.gov/ and https://www.drugs.com/
Differential	**GPT-o3 mini**	Performs a search based on the following prompt objective: "As a medical expert, generate a comprehensive clinical guideline and publications search given the patient's symptoms and medical history to determine what will support the differential diagnosis."
Clinical Plan	**GPT-o3 mini**	Performs a search based on the following prompt objective: "As a medical professional, conduct a comprehensive search supporting clinical guidelines and publications that will support a clinical plan for the given query."
Writer	**GPT-4o mini**	Does not search. Designed for requests that only require writing support.
Pharmaceutical Industry	**GPT-4o**	For questions about the pharmaceutical industry, it directs the search towards sources such as Statista, Research and Markets, MarketWatch, FDA, and more.

Table 5. Average score for validity questions by group with 95% IC and hypothesis test

Question	Average Score		Mann–Whitney (M-W) U p-value	Relative Difference
	Group A	Group B		
Response Accuracy	2.69 (2.65, 2.74)	2.40 (2.34, 2.45)	<0.01	12.08%
Medical Consensus	2.81 (2.77, 2.85)	2.73 (2.68, 2.77)	<0.01	2.93%
Demographic Bias	2.86 (2.82, 2.89)	2.71 (2.66, 2.76)	<0.01	5.54%
Without Treatment Bias	2.89 (2.86, 2.92)	2.87 (2.70, 2.80)	<0.01	0.70%
Updated Information	2.89 (2.87, 2.92)	2.77 (2.72, 2.81)	<0.01	4.33%

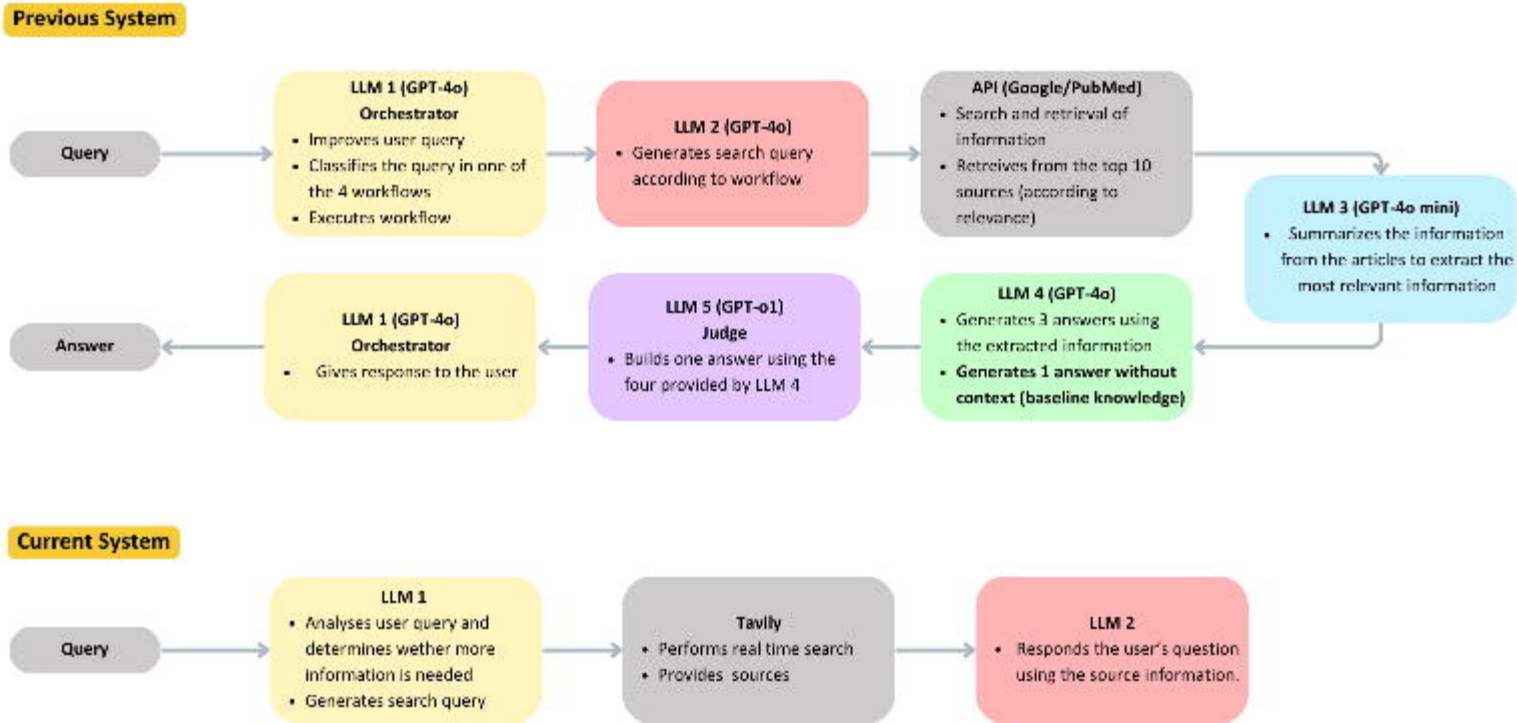
Low to no Patient Risk	2.90 (2.87, 2.93)	2.78 (2.74, 2.82)	<0.01	4.19%

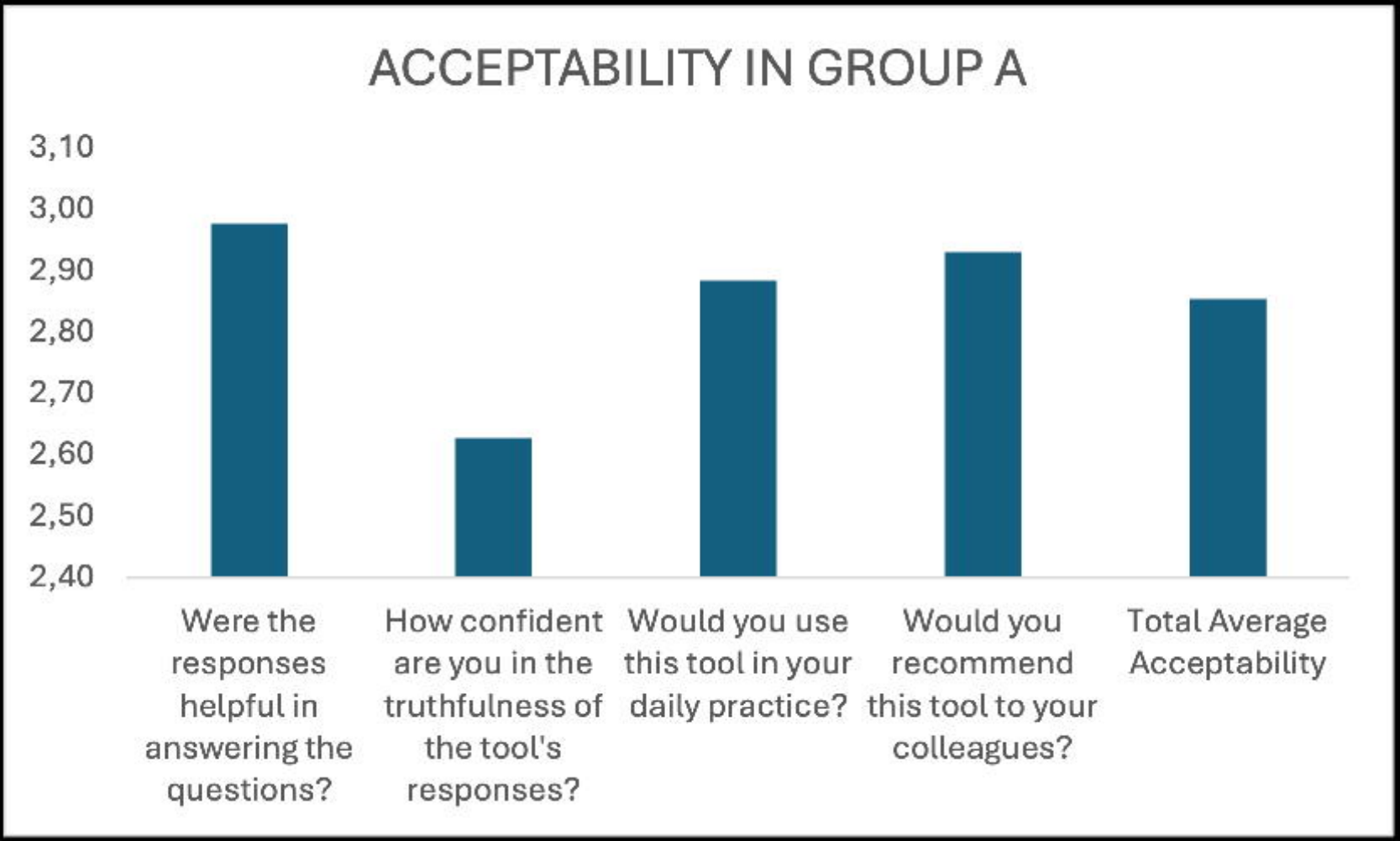
Legend: non-normal distribution for all criteria according to the Shapiro–Wilk test.
Mann–Whitney (M-W) U was used for group comparisons.
The Percentage Relative Difference was calculated as ((Group A-Group B)/Group B) * 100, representing the improvement of Group A compared to Group B.

Table 6. Efficiency results between groups: Average time and number of searches.

	Group A (Arkangel AI)		Group B (Traditional)					
Variable	**Mean (SD)**	**Median (IQR)**	**Mean (SD)**	**Median (IQR)**	**Absolute difference means**	**Relative difference means (%)**	**Absolute difference medians**	**Relative difference medians (%)**
Average response time per case	00:04:35 ± (00:06:53)	00:02:54 (00:02:37)	00:08:11 ± (00:12:15)	00:05:28 (00:06:51)	0:03:76	46	0:02:54	52
Average response time per question	00:00:31 ± (00:0:16)	00:00:32 (00:00:25)	00:01:23 ± (00:01:15)	00:01:25 (00:00:26)		>100		>100
Average searches per case	14 ± 3.85	13 (5)	30.29 ± 19.3	25.25 (19.3)	16.29	54	12.25	49
Average searches per question	3.54 ± 0.52	3 (3)	7.6 ± 5.84	6 (5.2)	4.6	54	3	50

Legend: *Mann–Whitney's U* and *Student's t*-tests *p-value* < 0.001 for all measures. Both means and medians are presented due to extreme values and asymmetry in the distribution of response times. Medians offer a more robust estimate of the central tendency and are less influenced by outliers. This dual presentation provides a more accurate and transparent view of the differences between the groups.


Previous System
Query
LLM 1 (GPT-4o)
Orchestrator
Improves user query
Classifies the query in one of the 4 workflows
Executes workflow
LLM 2 (GPT-4o)
Generates search query according to workflow
API (Google/PubMed)
Search and retrieval of information
Retreives from the top 10 sources (according to relevance)
LLM 3 (GPT-4o mini)
Summarizes the information from the articles to extract the most relevant information
LLM 4 (GPT-4o)
Generates 3 answers using the extracted information
Generates 1 answer without context (baseline knowledge)
LLM 5 (GPT-o1)
Judge
Builds one answer using the four provided by LLM 4
LLM 1 (GPT-4o)
Orchestrator
Gives response to the user
Answer
Current System
Query
LLM 1
Analyses user query and determines wether more information is needed
Generates search query
Tavily
Performs real time search
Provides sources
LLM 2
Responds the user's question using the source information.


ACCEPTABILITY IN GROUP A
3,10
3,00
2,90
2,80
2,70
2,60
2,50
2,40
Were the responses helpful in answering the questions?
How confident are you in the truthfulness of the tool's responses?
Would you use this tool in your daily practice?
Would you recommend this tool to your colleagues?
Total Average Acceptability