Intelligence-Based Medicine 12 (2025) 100274

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Arkangel AI: A conversational agent for real-time, evidence-based medical question-answering

Maria Camila Villa, Natalia Castano-Villegas *, Isabella Llano, Julian Martinez, Maria Fernanda Guevara, Jose Zea, Laura Velásquez

Arkangel AI, Bogotá, Colombia

ARTICLE INFO

Keywords:
Artificial Intelligence
Conversational agent
Medical question-answering
Large language models

ABSTRACT

Introduction: Large Language Models (LLMs) have been trained and tested on several medical question-answering (QA) datasets built from medical licensing exams and natural interactions between doctors and patients to fine-tune them for specific health-related tasks.
Objective: We aimed to develop LLM-powered Conversational Agents (CAs) equipped to produce fast, accurate, and real-time responses to medical queries in different clinical and scientific scenarios. This paper presents Arkangel AI, our first conversational agent and research assistant.
Methods: The model is based on a system containing five LLMs; each is classified within a specific workflow with pre-defined instructions to produce the best search strategy and provide evidence-based answers. We assessed accuracy, intra/inter-class variability, and Cohen's Kappa using the question-answer (QA) dataset MedQA. Additionally, we used the PubMedQA dataset and assessed both databases using the RAGAS framework, including Context, Response Relevance, and Faithfulness. Traditional statistical analysis was performed with hypothesis tests and 95 % IC.
Results: Accuracy for MedQA (n: 1273) was 90.26 % and Cohen's kappa was 87 %, surpassing current SoTAs for other LLMs (GPT-4o, MedPaLM2). The model retrieved 80 % of the expected articles and provided relevant answers in 82 % of PubMedQA.
Conclusion: Arkangel AI showed proficient retrieval and reasoning abilities and unbiased responses. Evenly distributed medical QA datasets to train improved LLMs and external validation for the model with real-world physicians in clinical scenarios are needed. Clinical decision-making remains in the hands of trained health-care professionals.

1. Introduction

Large language models (LLMs) and LLM-powered conversational agents (CAs) can assist healthcare personnel by automating repetitive processes and reducing the pace needed for them to keep up with the flow of information and knowledge [1]. They can be trained in multiple tasks to support healthcare professionals, including real-time literature searches retrieved from trusted medical libraries, improving research efficiency and precision. LLMs can also help physicians revise specific information in clinical charts, resulting in faster, more informed decisions.

Healthcare-related LLMs and CAs have usually been trained and tested on medical question-answering (QA) datasets built from medical licensing exams and natural interactions between doctors and patients. Popular examples are MedQA [2], MedMCQA [3], and PubMedQA [4]. Their performance has surpassed human scores in all three datasets [5–7], and some have even surpassed experts' performance, such as GPT-4 using MedQA and PubMedQA [2]. CAs like ChatGPT have also been tested, attaining passing-level performance in the United States Medical Licensing Examinations (USMLE) [8].

State-of-the-art (SoTA) refers to the highest-performing LLMs using a specific QA dataset; Med-Gemini's accuracy using MedQA was 91.1 % [3]; with PubMedQA, GPT-4 obtained 81.6 %, using a prompting strategy named Medprompt [4]; Med-PaLM 2 had a 72.3 % accuracy using MedMCQA [9].

We aimed to develop an LLM-powered CA equipped to produce accurate, real-time responses to medical queries in different clinical and scientific fields. We present Arkangel AI, our first CA and research

* Corresponding author.
E-mail address: natalia@arkangel.ai (N. Castano-Villegas).

https://doi.org/10.1016/j.ibmed.2025.100274
Received 7 January 2025; Received in revised form 13 June 2025; Accepted 6 July 2025
Available online 11 July 2025

assistant, by describing its development and internal validation. We assess the model using MedQA as a reference standard [11]. Finally, we approach some limitations in LLM assessment by using different databases (PubMedQA) and frameworks that explore LLM beyond their accuracy.

2. Methodology

2.1. Development of the model

Arkangel AI is a CA powered by five LLMs that answer clinical questions using information from the internet, research articles, clinical guidelines, and medical libraries. Arkangel AI has a real-time Retrieval-Augmented Generation (RAG) system. This system consists of two modules: Information Retrieval and Text Generation. Retrieval extracts the relevant pieces (chunks) of information using two Application Programming Interfaces (APIs), computational strategies for communication between the software and the application [10], from Google and PubMed. Second, the Text Generation module consists of LLMs that use the retrieved information as context to generate responses.

The baseline LLMs used are in the Generative Pre-trained Transformer (GPT) family, chosen due to their SoTA performance in general language reasoning tasks; GPT-4o [11], with an accuracy of 88.7 % using the MMLU databases [12], 90.5 % using the MGSM [13], and 90.2 % with a Human Evaluation [14].

While developing the system, we incorporated improvements as we worked. The result was three versions of the model: V1, V2, and V3. The rationale for the different versions is elaborated in section *2.1.1.3* and Table 1. Unless specified otherwise, the term "Arkangel AI" throughout the manuscript refers to the final version (V3).

2.2. RAG and API systems

Arkangel AI uses Google and PubMed's APIs as sources of information. Google's API can generate search engines that create search strategies on particular pages. We used the functionality SafeSearch to avoid retrieving information from unsafe sources. The PubMed API allows for the retrieval of articles with free access. Data in PubMed can be sorted and filtered according to Best Match, Publication Date, Authors, and Journals. For Arkangel AI, searches are filtered by "Best Match".

The "Best Match" sorting feature is based on an algorithm that assigns different weights (relevance scores) to each citation resulting from a search query based on the number of search term frequency, length, use of MeSH (Medical Subject Headings) terms, date of publication, type of study and citation factor. On top of these, an ML algorithm re-ranks the papers using "*over 150 signals*" computed from combinations of the above factors to improve relevance [15].

Each search using either API chooses the ten most relevant articles to build a response. The number is a very conservative approach to the non-saturation of the model's baseline LLMs (GPT-4o); although there is no exact description of the maximum text length these LLMs can support, the literature describes them to drop in performance and consistency with contexts longer than 64.000 to 128.000 tokens [16], roughly equivalent to 48.000 to 96.000 words (using the rule-of-thumb equivalence of 1token = 0,75words [17]). This would result in between 9.6 and 19 manuscripts of 5.000 words each.

2.3. Workflows

We created four workflows to route the user's query to an optimal search strategy (Fig. 1). These correspond to the most frequent clinical question types in our experience with focus groups in healthcare personnel.

Table 1
Accuracy results for the three different versions of Arkangel AI.

Metric	Vitruvius				Improvement V1 to V2 (n = 288)		Improvement V2 to V3 (n = 1273)	
	Version 1	Version 2		Version 3	Percentage difference	p-value (z-test)	Percentage difference	p-value (z-test)
Accuracy	85.76 % (81.28; 89.46)	90.28 % (86.32; 93.28)	85.15 % (83.08; 87.02)	93.06 % (89.46; 95.54)	5 %	0.095	6 %	<0.05
# correct questions	247	260	1084	268				
Total # questions	288	288	1273	288				

90.26 % (88.50; 91.80)[a] 1149 1273

[a] All confidence intervals were calculated using statsmodels. stats.proportion.proportion_confint(.) in Python using the binomial test method.

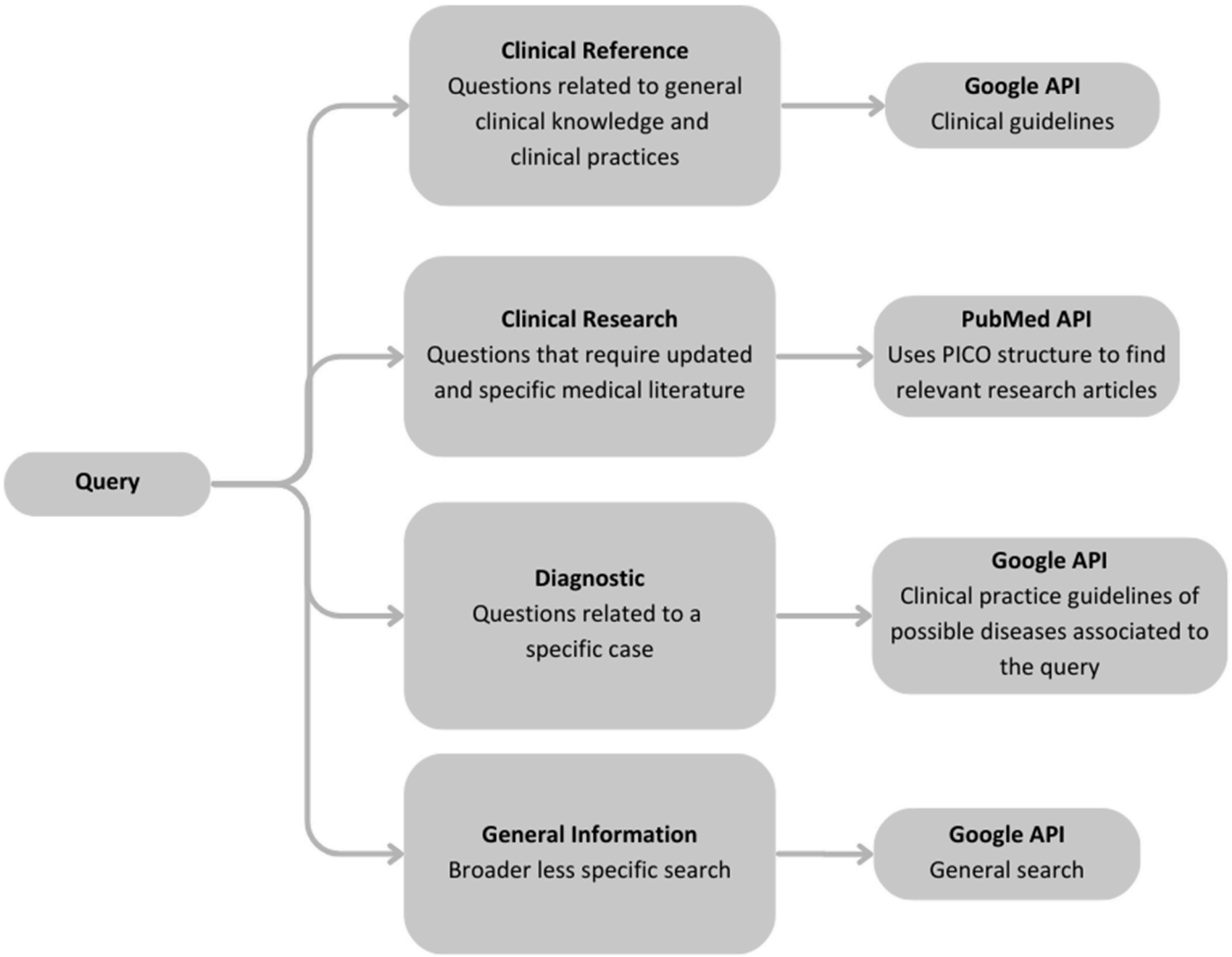


Fig. 1. Arkangel AI classifies questions into one of four different workflows.

1. Clinical Reference Questions: general medical knowledge. Their responses are based on official clinical practice guidelines (CPGs) and use the Google API.
2. Clinical Research Questions: require retrieving research articles to address medical and research queries. The model uses the PubMed API. The input question is automatically transformed into a PICO format (Patient, Intervention, Comparison, Outcome). The model transforms queries following a prompt strategy to focus on key terms to find relevant articles.
3. Diagnostic Questions: They are concerned with diagnostic information. Given a specific case, this workflow uses the Google API to search for CPGs and public textbooks.
4. General Information Questions: require a broader, less specific search. It uses the Google API.

The system works in English, Spanish, and Portuguese. Internal instructions, information search, and retrieval are in English. The final response is translated into the user's language. The model provides literary references in APA format.

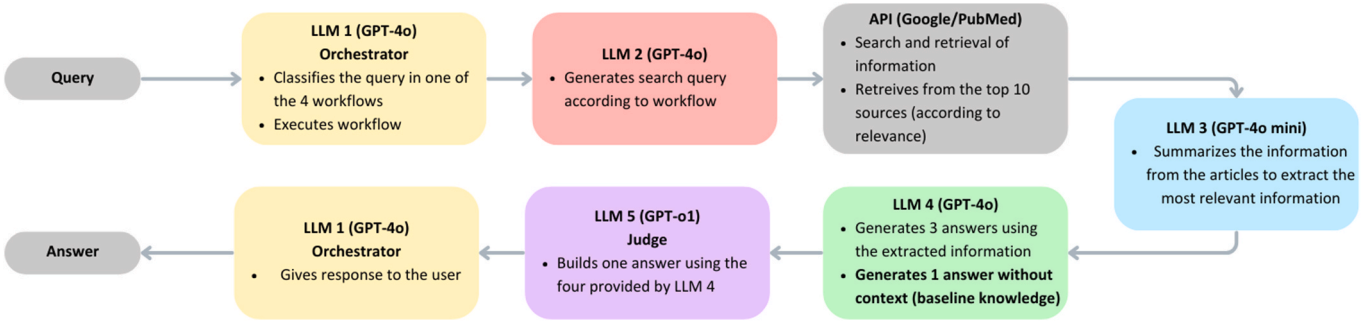


Fig. 2. Arkangel AI pipeline.

2.4. Individual LLMs and Answer Generation

LLM1 chooses the workflow and determines the search strategy. LLM2 improves it and passes it onto the corresponding API for the search. Answer Generation has three steps: LLM3 summarises the information from the first ten papers retrieved by the API. Next, LLM4 generates four answers to the initial query, three using the retrieved content and one using the LLM's baseline knowledge.

The rationale for the generation of four answers is.

a. The earliest versions of Arkangel AI, V1 and V2, comprised four LLMs and used 288 QA; for LLMs 1, 2, and 3, all three versions' characteristics and functionalities were maintained. However, LLM4 only produced one answer in V1. Its accuracy was 85.76 % (Table 1), and we tried to improve it by allowing LLM4 to produce two more answers from the same context source and adding a fifth LLM that would use those three answers as context to produce a more accurate response. Different answers are generated in LLMs due to the random effect, where several responses can be generated from one query [18].
b. Accuracy for Arkangel AI V2 improved by 5 % (90.28 %) on the 288 QA. We applied the same experiment to the entire dataset of 1273 QA; the accuracy obtained was 85.15 %. We considered adding a fourth "contextless" response to the previous three, meaning that LLM4 would generate the three answers using retrieved-context and a fourth using only the baseline LLMs' background training. Table 1 depicts how this approach helped to improve accuracy (Accuracy improved by 6% from the baseline).

LLM5 used OpenAI's o1 model [11]. It was instructed to produce one answer out of the four created by LLM4. This model acted as a judge, performing complex reasoning and creating a long chain of thought before generating a response. Finally, LLM1 presents the information to the user and closes the question-answer cycle (Fig. 2). If Arkangel AI cannot find reliable information to produce a response, it is instructed to declare it to lower the hallucination potential.

2.5. Dynamic user interface

Arkangel AI is available through a custom URL created by the Arkangel AI software platform. This platform is the backbone of Arkangel AI and enables LLM integration and core functionalities. Users input their queries into the platform through a conversational interface that allows follow-up questions and conversation-style interaction (Fig. 3).

2.6. Model assessment

2.6.1. Methods

We used the English dataset from MedQA, introduced by Jin et al., in 2020. It comprises 12723 multiple-choice QA from the medical licensing exams in the United States (USMLE) (Fig. 5.). Each set of questions is divided into development, training, and testing splits. Queries are presented as follow-up questions derived from a given clinical case, which requires multiple reasoning steps and is more challenging than simple QAs [19].

We summarized the general performance of Arkangel AI's different versions (Table 1). Given their independence and sample size, statistical

Table 2
Confusion matrix for the different categories results in Arkangel AI, compared to MedQA.

		MedQA				
		A	B	C	D	
Arkangel AI	**A**	313	10	5	10	**338**
	B	18	281	12	8	**319**
	C	13	11	318	10	**352**
	D	9	7	11	237	**264**
		353	**309**	**346**	**265**	**1273**

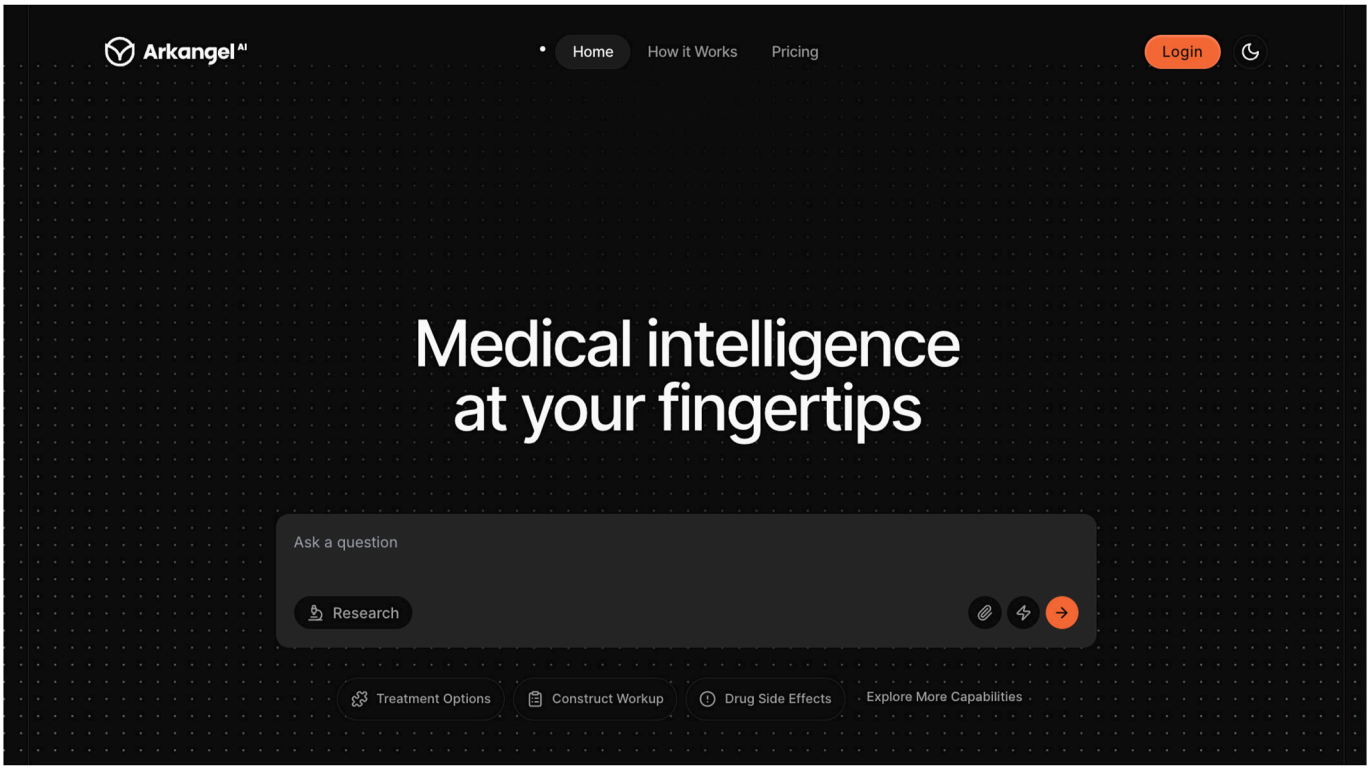


Fig. 3. Arkangel AI user interface.

Table 3
Values for sensitivity, precision and F1-score by class.

a)

Metric	Class				Kruskall-Wallis[a] p-value
	A	B	C	D	
Sensitivity/ Recall[c]	88.67 %	90.94 %	91.91 %	89.43 %	0.392
Precision[d]	92.60 %	88.09 %	90.34 %	89.77 %	0.392
F1-score[e]	90.59 %	89.49 %	91.12 %	89.60 %	0.392

b)

Metric	Value
Accuracy[b]	90.26 %
Cohen's Kappa	86.96 %
Observed (Po)	90.26 %
Expected (Pe)	25.28 %

[a] Both Cochrane's Q and Kruskal-Wallis tests resulted in the same p-value.
[b] Accuracy calculated as: (TP + TN)/(TP + TN + FP + FN).
[c] Sensitivity (Recall) calculated as: TP/(TP + FN).
[d] Specificity (Precision) calculated as: TP/(TP + FP).
[e] F1 score calculated as: (2*Precision*Recall)/(Precision + Recall).

differences between V1, V2, and V3 were explored using a z-test. We used absolute and relative frequencies to describe the distribution of medical specialties and subspecialties among the QAs (Fig. 6). For the model's internal variability, we described its operational characteristics divided by class (A, B, C, D); we performed Cochrane's Q and Kruskal-Wallis test to evaluate possible classification bias. For workflow classification, we described relative frequencies and used the Kruskall-Wallis test to asses statistical differences in performance per workflow. Statistical significance was a P-value <0.05, and 95 % confidence intervals (CI 95 %) were reported. We used Python and R for tests, coding, and the Arkangel AI app for Arkangel AI.

2.6.2. Main outcomes

MedQA (Supplementary Material 1) was evaluated using: 1. Accuracy and Consistency. Accuracy measures the favorable agreement with the standard of reference and is the metric reported in all benchmarks and SoTAs [5]. The model's consistency was evaluated by assessing internal and external variability. We calculated operative characteristics per class (Table 3a). By default, classes were defined as A, B, C, or D; 2. We used Cohen's Kappa (Table 3b) 3 for external variability. Model efficiency was defined as the average response time.

2.6.3. Secondary outcomes

1. Workflow Classification

Workflow Classification assessment was performed using the MedQA database and Human Evaluation as the reference standard. Two researchers individually revised the classification in a subset of the original database (n = 127, 10 %). The subsample size was defined based on convenience, considering the costs and human resources needed. Researchers manually classified 127 questions into four workflows, blinded to the model and each other's results. Workflow distribution is described in Fig. 4. Human classification was guided by the prompt's definition of each workflow. Prompts are available upon reasonable request. Workflow accuracy is presented in Table 4 and Supplementary Material 4A.

2.6.4. RAGAS framework

Automatic evaluation of RAG systems is an ongoing topic of investigation [20]. In addition to accuracy, the ability to find relevant contexts to answer questions or the model's use of those contexts should also be considered. However, these assessments often require extensive human annotation to produce reference values, and there are limited options for robust alternatives to a human standard of reference [21].

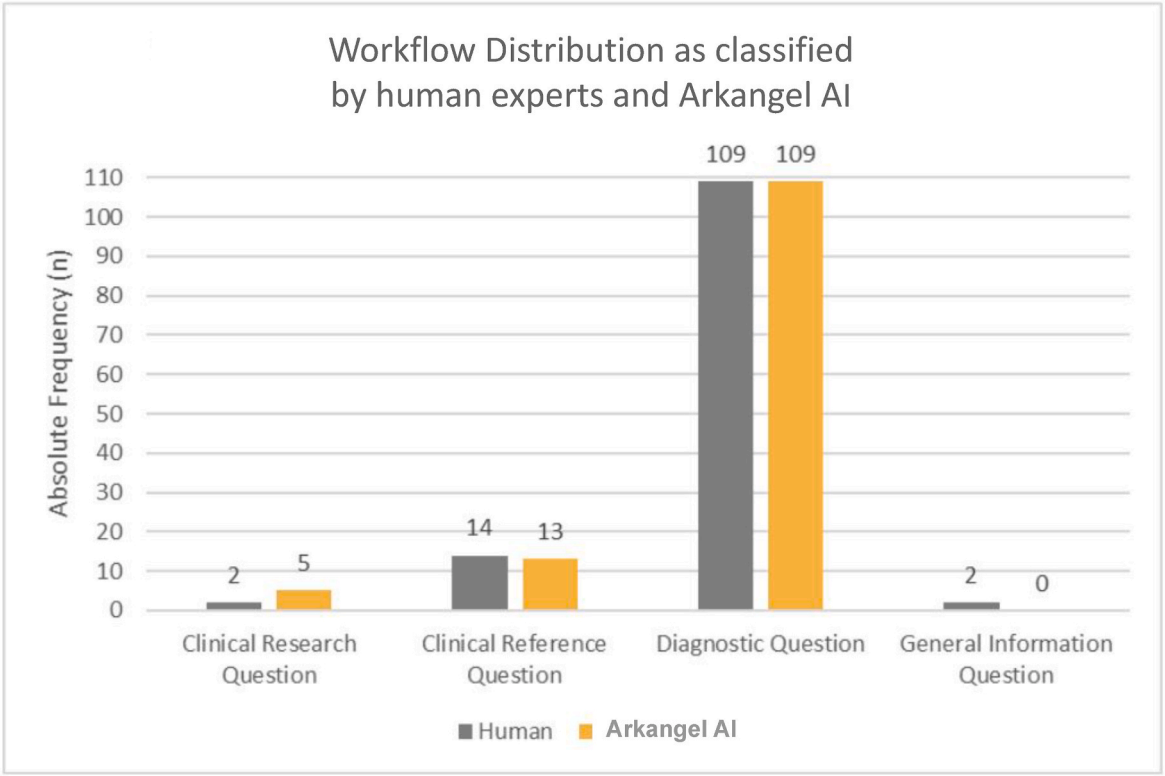


Fig. 4. Workflow distribution as classified by Human experts and Arkangel AI.

2. Retrieval and Validity

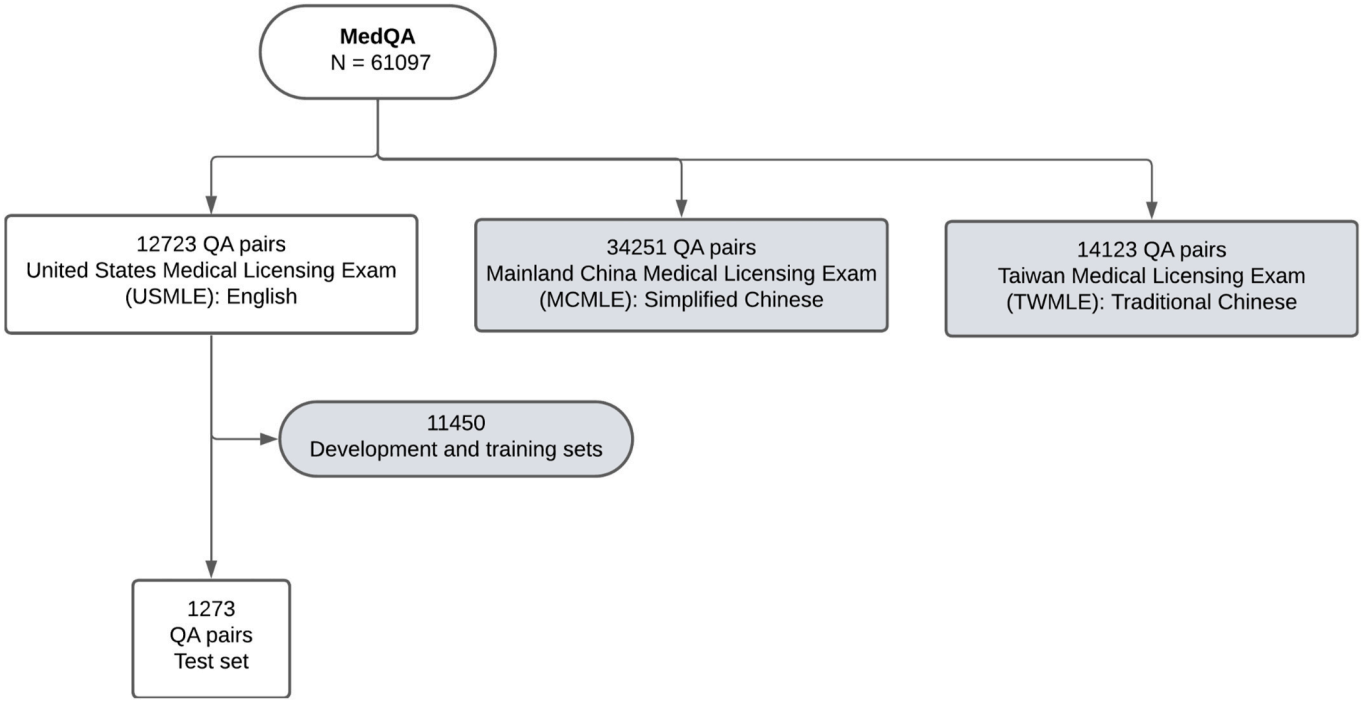


Fig. 5. MedQA distribution.

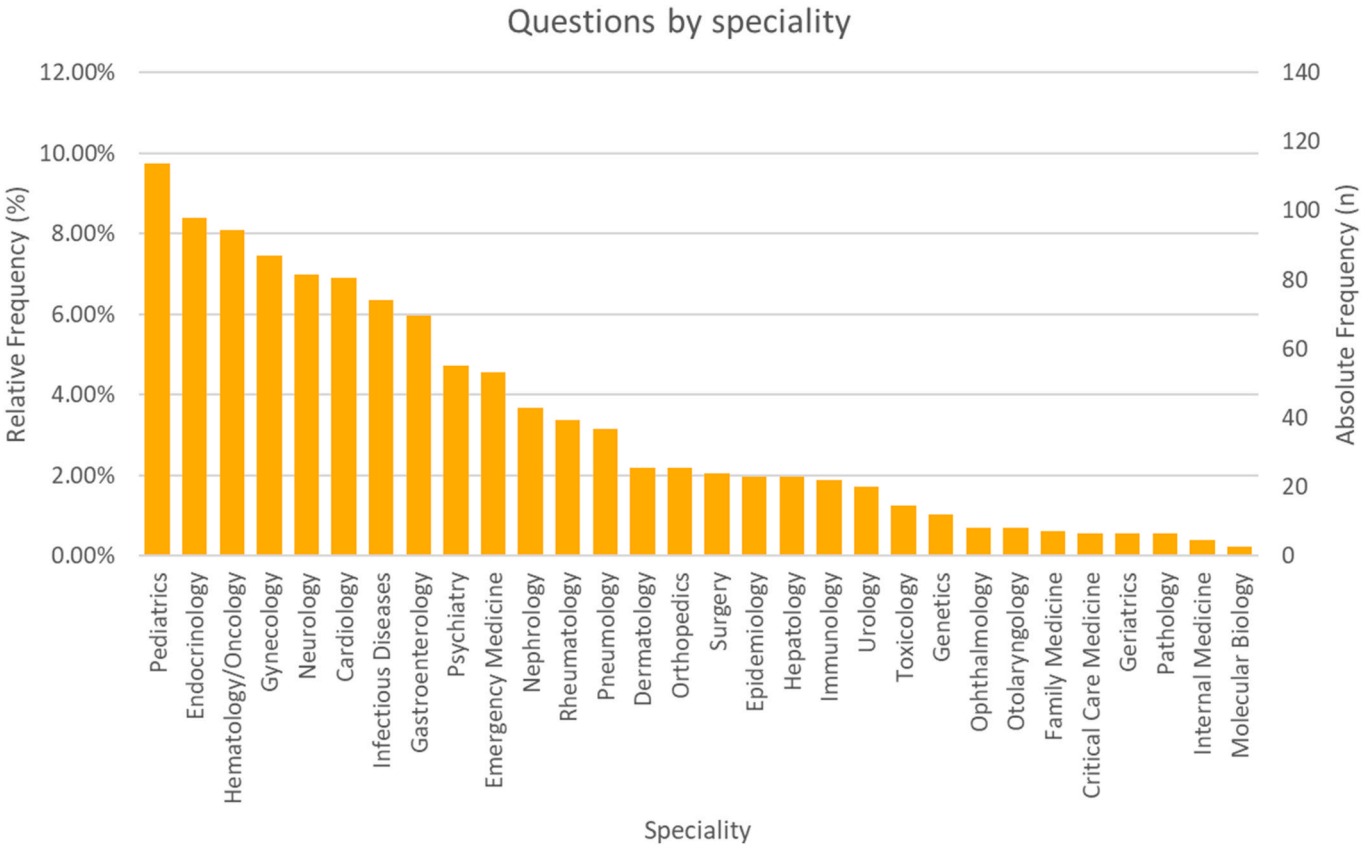


Fig. 6. MedQA questions by medical specialty.

Table 4
Model's Classification of workflows; global and class performance metrics.

Metric	Accuracy[a]	Sensitivity	Specificity	F1 Score
Global	0.945	0.945	0.942	0.941
Clinical Reference Question (n = 14)	0.786	0.786	0.846	0.815
Clinical Research Question (n = 2)	1.0	1.0	0.400	0.571
Diagnostic Question (n = 109)	0.982	0.982	0.982	0.982
General Information[b] (n = 2)	0.0	0.0	0.0	0.0

[a] Kruskall-Wallis test for accuracy p-value = 0.392.
[b] Both general information questions were misclassified as diagnostic by the model (Supplementary Material E).
2. Retrieval assessment

Table 5
Retrieval assessment for MedQA: Context precision.
3.Quality Assessment

Context Precision	Essential to answer the question	Absolute Frequency (n)	Relative Frequency (n)
	Yes (Essential)	70	55.12 %
	No (Non-Essential)	56	44.09 %
	Empty	1	0.79 %

Table 6
Quality assessment for MedQA: Response relevancy.

Response Relevance	Did the response address the initial query?	Absolute Frequency (n)	Relative Frequency (n)	Average score
	Yes (Relevant)	123	96.85 %	0.54
	No (Not Relevant)	3	2.36 %	0.0
	Empty	1	0.79 %	–

The Retrieval Augmented Generation Assessment (RAGAS) is an automated framework for RAG evaluation described by Shahul et al., in 2023 that doesn't depend on human annotations for ground truths [22]; Context Precision measures retrieval by determining if the retrieved context is essential for a response; Response Relevance determines if the model's response addresses the user's question (if the response is actually what was asked); Response Faithfulness quantifies if the answer comes from the context retrieved.

We performed RAGAS on a subset (10 %, n:127) of MedQA (Supplementary Material 2). Additionally, we used the PubMedQA dataset. This decision was made to also evaluate the RAGAS framework assessment of a dataset with a human-revised reference standard for the context retrieved, not just for the answers.

PubMedQA compiles 273500 QA. It was published in 2019 to evaluate the reasoning capacity of LLMs like GPT-4 and MedPaLM2 [9,23]. We used 500 QA Human Evaluated (Supplementary Material 3). The database contains QA pairs with questions structured from an abstract's title, an answer (Yes, No, Maybe), and the abstract as a reference context (Supplementary Material 4B). These characteristics make it helpful in evaluating retrieval, which was our main objective when using it. All the abstracts come from PubMed and are included in the original database, leaving out the Conclusion segment since the actual answer could be displayed there [4]. SoTA using this database is GPT-4 (plus Medprompt strategy), with an accuracy of 82 % [23].

2.6.5. Retrieval assessment

In 127 (10 %) of MedQA, retrieval was evaluated using Context Precision. We used 500 (50 %) of Human-Evaluated PubMedQA to calculate Context Recall, a variation of Context Precision used when there is a reference standard for the retrieved context. It quantifies the proportion of contexts adequately retrieved compared to the standard (Tables 5–10).

2.6.6. Validity/Quality Assessment

Quality assessment was performed for both subsamples. We calculated Response Relevancy and Response Faithfulness.

2.6.7. Data policy

The latest actualization of our data policy is available at https://medsearch.arkangel.ai/privacy. We recommend not providing private, confidential, or identifying information when entering a query. The website states, *"While we strive to protect your personal information, no method of transmission over the Internet or method of electronic storage is completely secure."*

3. Results

3.1. Dataset distribution and model improvement

MedQA's sample in English was 12723 (Fig. 5), 11450 QA for training, and 1273 for testing. As our model did not require fine-tuning, we used the test set directly. When described by medical specialty, QA pairs were unevenly distributed (Fig. 6). Pediatrics, endocrinology, and hematology/oncology were the most frequent. Molecular biology and general internal medicine (when subspecialty is not mentioned) were the least present.

Table 1 summarises performance and accuracy improvements between versions. Improvement from V2 to V3 (n = 1273) represented an increase of 6 % in absolute accuracy. The z-test p value demonstrates a statistical difference, with no overlap of IC95 %. On the contrary, improvement from V1 to V2 (n = 288) showed a similar relative increase with no statistical difference, while the CI 95 % for their accuracies overlapped.

Table 8
Retrieval assessment for PubMedQA; context recall.

Context Recall	Was context retrieval accurate by the standard of reference?	Absolute Frequency (n)	Relative Frequency (n)
	Yes (accurate retrieval)	401	80.20 %
	No (not accurate retrieval)	76	15.20 %
	Empty	23	4.60 %

Table 9
Quality assessment for PubMedQA: response relevance.

Response Relevance	Did the response address the initial query?	Absolute Frequency (n)	Relative Frequency (n)	Average score
	Yes (Relevant)	414	82.80 %	0.17
	No (Not Relevant)	86	17.20 %	0.0
	Empty	23	4.60 %	–

Table 7
Quality assessment for MedQA: Response faithfulness.

Faithfulness	Was the response taken from at least one of the contexts?	Absolute Frequency (n)	Relative Frequency (n)	Average Score	Answered Correctly (n, %)
	Yes (Faithful)	73	57.48 %	0.11	68 93.15 %
	No (Not Faithful)	53	41.73 %	0	47 88.68 %
	Empty	1	0.79 %	–	1 100.00 %

Table 10
Quality assessment for PubMedQA; response faithfulness.

Faithfulness	Was the response taken from at least one of the contexts	Absolute Frequency (n)	Relative Frequency (n)	Average Score	Answered Correctly (n, %)
	Yes (Faithful)	272	54.40 %	0.5	184 67.65 %
	No (Not Faithful)	205	41.00 %	0	100 48.78 %
	Empty	23	4.60 %	–	6 26.09 %

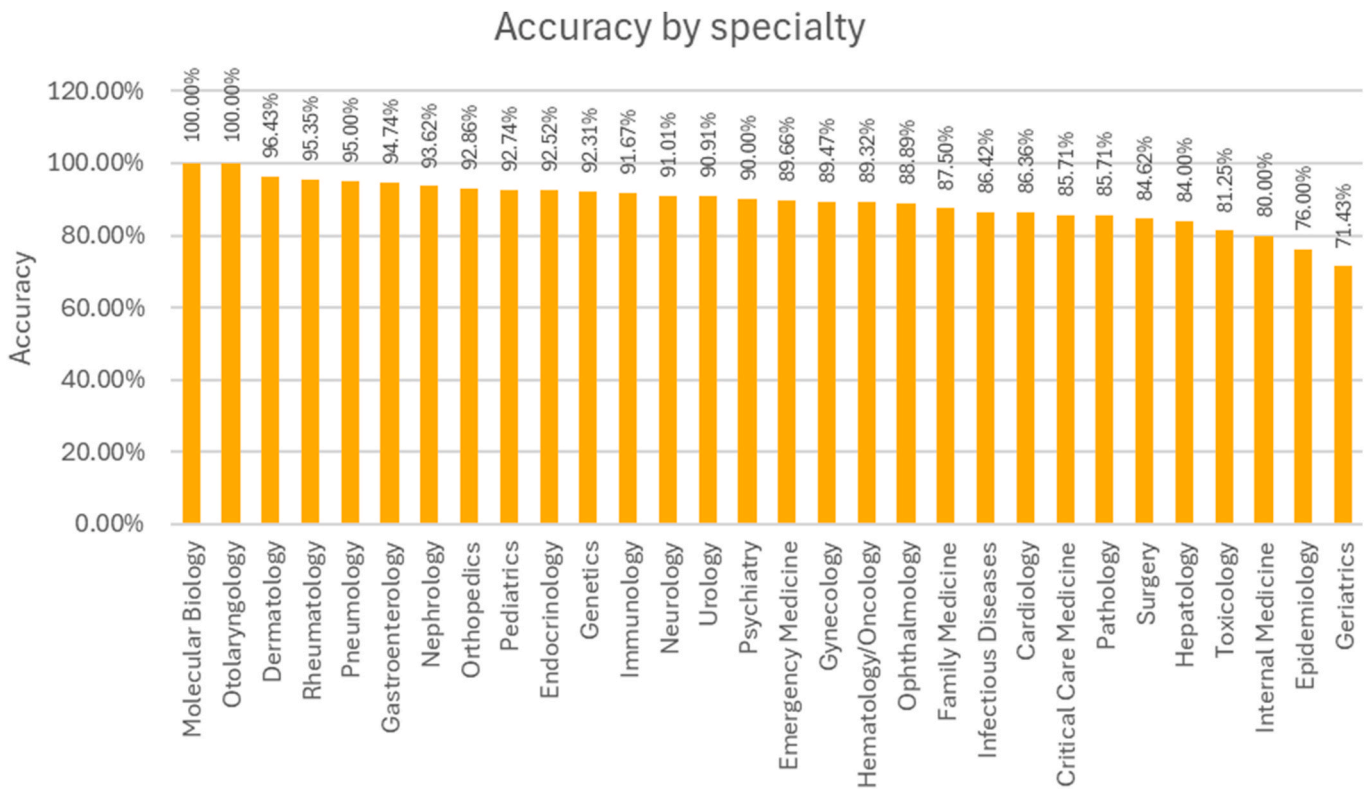


Fig. 7. Classification accuracy by medical specialty.

Please refer to the summary above to answer this question
The authors of the study have decided to conduct a follow-up analysis on their data. They decide to stratify their results by CD4+T-lymphocyte count at the time of diagnosis. Among patients with CD4+ cell counts below 200/mm3, cART adherence was a significant predictor of DLBCL risk (RR = 0.52, p = 0.01). However, among patients with CD4+ cell counts above 200/mm3, no relationship was found between DLBCL risk and cART adherence (RR = 0.96, p = 0.36). Which of the following explains for the difference observed between the two strata?"

"A": "Poor generalizability",
"B": "Stratified sampling",
"C": "Random error",
"D": "Effect modification"

MedQA Correct Answer: D
MedSearch Answer: D

Reasoning: The question pertains to analyzing study data with stratification, indicating an inquiry about research methodology and data interpretation, which makes it a Clinical Research Question. The ark_search did not return specific information related to CD4+ T-lymphocyte count, cART adherence, or DLBCL risk regarding the difference in risk ratios, but the concept of effect modification fits this scenario. Effect modification occurs when the effect of the main exposure (cART adherence) on an outcome (DLBCL risk) varies across levels of another variable (CD4+ T-lymphocyte count). Here, cART adherence affects DLBCL risk differently when stratified by CD4 count.

Fig. 8. QA pair marked as "empty" in all three RAGAS metrics.

3.2. Main outcome

1. Variability

We assessed internal and external variability based on each category's observed and expected values. The following matrix (Table 2) presents the observed values for the four options in MedQA.

Sensitivity, precision, and F1 Scores for all four classes were over 89 % and showed less than 5 % variability (Table 3a). Kruskall-Wallis test for performance by class demonstrated no statistical difference in the model's classification (A, B, C, or D), with p-values >0.05.

2. Accuracy

Arkangel AI achieved an overall accuracy of 90.26 % and a Cohen's Kappa coefficient of 86.96 % (see Table 3b), displaying near-perfect agreement [24] with the MedQA reference standard.

Fig. 7 presents accuracy by medical specialty. Geriatrics and epidemiology achieved under 80 % accuracy, while molecular biology and otolaryngology achieved 100 % accuracy. Other specialties presented accuracies ranging from 80 % to 96 % without remarkable absolute variation.

3. Efficiency

For MedQA's 10 % subset, the total response time was 334.14 min, with an average of 2.63 min per question, a maximum of 4.3 min, and a minimum of 29 s.

3.3. Secondary outcomes

1. Workflow classification

Accuracy for general workflow classification was 94.5 %. Accuracy per class was: Clinical Reference 78.6 %, Clinical Research 100 %, Diagnostic 98.2 % and General Information 0 % (Table 4). The confusion matrix is in Supplementary Material 4A.

3.3.1. Context Precision

Each QA used an average of eight out of ten contexts retrieved (articles, guidelines, websites). Seventy QA (55.12 %) had at least one retrieved-context qualified as "essential"(Table 5.). One QA pair did not have a score, probably indicating that this question was answered using the baseline LLM knowledge (Fig. 7.)

3.3.2. Response Relevancy

The assessment determined that 92.64 % of QA pairs addressed the user's query, with an average score of 0.54 on a scale from zero to one (Table 6). The same QA pair was answered using baseline LLM knowledge.

3.4. Response Faithfulness

If at least one of the texts retrieved for a QA pair had a score different than zero, it was marked as faithful. Seventy-two QA pairs (56.69 %) were deemed faithful, while the remaining 42.52 % were not. Correct and incorrect answers, following the MedQA standard, were calculated for each group (Faithful, Non-Faithful) (Table 7). Correct answers represented 94 % of those classified as Faithful. Nonetheless, for those QA responses marked as Non-Faithful, 88.89 % were accurate. Only one of the QA pairs was consistently shown as answered without using any context. This answer was correct. The average faithfulness score for the 72 faithful answers was 0.12.

3.5. RAGAS in PubMedQA

Arkangel AI retrieved 80.2 % (n = 401/500) of the standard abstracts (Table 8), with an accuracy of 59.5 % and an average response time of 2.2 min per question.

We explored how the transformation to a PICO research question influenced accuracy; Supplementary material 4C displays the confusion matrix for Arkangel AI using PubMedQA. When Arkangel AI classifies a query on the Research Question workflow, it transforms it to refine key terms. Given that most QA in this scenario were classified as Clinical Research Questions (77.6 %), Supplementary Material 4D explores accuracy based on whether the input question was initially classified as such. Accuracy in this subgroup was 60.05 %, and 45.54 % when classified in any other workflow. The z-test p-value was <0.05, (Supplemntary material 4E) evidencing statistical differences in performance depending on the workflow. An example of PICO transformation is found in Supplementary Material 4F.

3.5.1. Context Recall

PubMedQA had 4.6 % of QA answered with background LLM knowledge. It adequately retrieved 80 % of abstracts.

3.5.2. Response Relevance

Arkangel AI's responses were classified as relevant in 82.8 % of QA. In contrast, 86 (17.2 %) were not. The average score was 0.17 (Table 9).

3.5.3. Faithfulness

Arkangel AI's responses were deemed faithful 49.8 %. 25.7 % of the faithful responses were correct and had an average score of 0.31. Not faithful answers were correct in 22.8 % of QA, while "contextless" answers were correct in 26.09 % (Table 10).

3.6. Classification errors

We manually analyzed a sample of incorrect responses to better identify Arkangel AI's limitations. Fig. 9 depicts four of those cases. Supplementary Material 4G describes and discusses these examples.

Table 11 presents the accuracy of Arkangel AI Version 3 and other state-of-the-art models.

4. Discussion

We introduce the development and internal evaluation of Arkangel AI, a five-LLM system designed to answer medical questions in different fields. Through PubMed and Google's APIs, it finds, summarises, and selects relevant content from CPGs and peer-reviewed scientific articles.

4.1. Accuracy

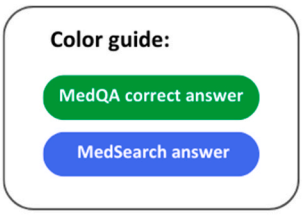
We aimed to evaluate the model's accuracy in answering clinical questions using MedQA. The exam's threshold for humans is 60 % (87 % for expert physicians) [25]. After experimenting with different approaches, we achieved 90.26 % accuracy. We concluded that Arkangel AI surpasses the currently described SoTA for MedQA; for Med-PaLM2, accuracy was 85.4 % [9]; for GPT-4o, 87.51 %, and 73.61 % for GPT-4o mini. GPT-4 obtained an accuracy of 90.2 % using 255 QA pairs [23], and 93.06 % using 72 QA [26]. Different sample sizes were not justified for the latter.

4.2. Consistency

Accuracy was improved through Arkangel AI's evolution by leveraging multiple answers using LLMs' random effect, background knowledge, and information retrieved. LLMs 1, 2, and 3 classified the query into workflows, generated improved search strategies, and retrieved contexts brought through the APIs, respectively. They remained unchanged. LLM4 generated one answer from the retrieved context. It was instructed to create two more from context and one answer from baseline LLM knowledge after encountering QAs in which reasoning steps required analyses based on unwritten rules, like questions concerning "common sense" or the analysis of human behavior, since baseline training could bring better insights.

Evidence from models using LLMs as judges has demonstrated improved accuracy and reliability [27]. We added LLM5 to function as a judge and unified the information into one answer. The result was a statistical improvement of 6 % in accuracy, confirmed by 95 % CI, which did not overlap. (V2 IC95 % (83.08; 87.02); V3 IC95 % (88.50; 91.80)).

LLMs' evaluation of MedQA has reported a bias where the model can disproportionately select or disregard a label [25]. We evaluated performance and accuracy for classes A, B, C, and D. We found minimal



A) Ethical question

A junior orthopaedic surgery resident is completing a carpal tunnel repair with the department chairman as the attending physician. During the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication. The attending tells the resident that the patient will do fine, and there is no need to report this minor complication that will not harm the patient, as he does not want to make the patient worry unnecessarily. He tells the resident to leave this complication out of the operative report. Which of the following is the correct next action for the resident to take?

"A": "Disclose the error to the patient and put it in the operative report",
"B": "Tell the attending that he cannot fail to disclose this mistake",
"C": "Report the physician to the ethics committee",
"D": "Refuse to dictate the operative report"

Reasoning: Guidelines on the ethical obligation to disclose medical errors to patients were searched. The correct answer is based on the professional and ethical responsibility to be honest and transparent with patients. The patient has the right to be informed about any errors or complications that occur during their treatment.

B) Behaviour

A 39-year-old man presents to the emergency department because of progressively worsening chest pain and nausea that started at a local bar 30 minutes prior. The pain radiates to the epigastric area. He has a 5-year history of untreated hypertension. He has smoked 1 pack of cigarettes daily for the past 5 years and started abusing cocaine 2 weeks before his emergency room visit. The patient is diaphoretic and in marked distress. What should be the first step in management?

"A": "Diltiazem",
"B": "Labetalol",
"C": "Propranolol",
"D": "Reassurance and continuous monitoring"

Reasoning: For this patient with chest pain possibly related to cocaine use, diltiazem, a calcium channel blocker, is the appropriate management to relieve coronary spasm without risk of uncontrolled alpha-adrenergic stimulation, unlike beta-blockers which are contraindicated.

C) Requires an image (dataset error)

A 5-year-old boy of African descent is presented to the emergency department by his parents. The child is clutching his abdomen and crying about pain in his belly. His mother reports intermittent episodes of jaundice with joint and bone pain since he was 5 months old. At presentation, the patient's vital signs are within normal limits. Physical examination reveals slight jaundice and pale conjunctiva. The spleen is tender and easily palpable. After a complete blood count with differential and an abdominal ultrasound, the patient is found to have sickle-cell disease with splenic infarct. A partial splenectomy is performed. After the operation, the physician provides vaccines against Neisseria meningitidis, Haemophilus influenzae, and Streptococcus pneumoniae. The picture shows a slide obtained from the resected portion of the patient's spleen. Dysfunction of the zone marked with which number predisposes the patient to the aforementioned infections?

"A": "2 only",
"B": "1 only",
"C": "1 and 2",
"D": "1, 2, and 3"

Reasoning: Based on the information obtained from the search, spleen dysfunction in sickle cell disease predisposes to infections by encapsulated bacteria such as Streptococcus pneumoniae, Haemophilus influenzae and Neisseria meningitidis. This dysfunction is related to damage to the areas of the spleen responsible for filtering these pathogens.

D) Reasoning not associated to one of the answers

A 65-year-old man with a history of hypertension, obesity, and alcoholic cirrhosis is seen in clinic for follow-up. He feels well and currently drinks 5 glasses of wine each night. Medications include atenolol and lisinopril. On physical exam, temperature is 98.1 deg F (36.7 deg C), blood pressure is 151/82 mmHg, pulse is 71/min, and respirations are 14/min. He has spider angiomata on his chest; no asterixis, jaundice, ascites, or peripheral edema is noted. Screening ultrasound reveals a new liver nodule, and follow up CT demonstrates a 2 cm right hepatic lobe lesion with enhancement in the arterial phase. No hypodense attenuation is seen on the venous or delayed phase. What is the next step in management?

"A": "Proceed with liver biopsy",
"B": "Refer for surgical resection",
"C": "Refer for radiofrequency ablation",
"D": "Observe and get follow-up imaging in 3 months"

Reasoning: The search revealed that the next step in the management of a patient with alcoholic cirrhosis presenting with a liver nodule with arterial phase enhancement on CT is to obtain a liver-specific contrast-enhanced MRI. This is because MRI can provide additional detail and confirm features typical of hepatocellular carcinoma (HCC), avoiding the need for biopsy. In addition, it is recommended to measure alpha-fetoprotein (AFP) levels and refer the patient to a hepatology or oncology specialist for appropriate treatment planning. Finally, the patient should be counselled on alcohol cessation.

Fig. 9. Four examples of incorrect responses
A) a mistake in an ethical question; B) a mistake in a question about behaviour; C) a question that required an image to be answered; D) the Model's reasoning does not match any of the possible answers.

Table 11
Reported SoTA LLMs using MedQA, 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 [4]	1273	85.40 %
GPT-4 (Medprompt) [10]	255	90.20 %

Source: https://paperswithcode.com/sota/question-answering-on-medqa-usmle

internal variability by class and evidenced a non-biased distribution in models' labeling (K-W p-value >0.05).

The model's workflow classification was also accurate. We human-evaluated 127 QA pairs and achieved 94.5 % general accuracy. Performance assessment was more challenging; using the Kruskal-Wallis test, we found no statistical difference (K-W p-value >0.05), which was ideal. Nevertheless, considering small sample sizes and zero accuracies in some groups, we included the assessment of the workflow classification on PubMedQA as well.

PubMedQA's accuracies demonstrated statistically better performance (z-test p-value <0,05) for Clinical Research Question (60 %) compared to other workflows (46 %), supporting a different conclusion. This comparison was made by recategorizing all the other workflows into one. Results of the K-W test for accuracies among ungroupped workflows are found in Supplementary Material 4H.

Moreover, workflow misclassifications evidence potential limitations in categorizing more complex queries. This questions the added value of workflow classification in these models, underlines the relevance of the quality of users' prompts, and emphasizes their influence on the accuracy of responses. Ideal prompting structures, their association with response quality, and resources for improving one's competence in building them are available elsewhere. The Arkangel AI website offers a free prompting course for physicians at (https://intro-ia-en-salud.thinkific.com/courses/curso-de-prompting-en-salud) other courses like Google's prompting essentials (https://grow.google/prompting-essentials/) are also available online. Additionally, it points to the unsuitability of multiple-choice QAs for assessing LLMs because static QAs do not depict real-world clinical questions from practitioners.

Despite consistent performance across specialties, MedQA shows an uneven distribution of medical specialties. Perfect accuracies in less-represented areas compared to lower accuracy in more frequent specialties alert to the introduction of a bias toward specific fields, potentially limiting the generalizability of findings to medical contexts different from those prevalent in licensing exams.

4.3. Retrieval

We used the RAGAS framework to assess beyond accuracy. It offers mathematical options when no human-evaluated standard exists [22]. We also used PubMedQA, specifically to evaluate retrieval; we expected it to perform better due to the availability of PubMed abstracts through the API. We used MedQA (10 %) and PubMedQA (50 %) samples. MedQA reflected the contexts classified as essential for answering (55.12 %), while for PubMedQA, it was 80.20 %. These results indicate proficiency in Arkangel AI's retrieval.

4.4. Validity and quality

Response Relevance determines whether the answer addresses the question; the 97 % in MedQA and 82 % in PubMedQA reinforce confidence in the model's results. Scores are more challenging to interpret as they quantify the cosine proximity of embeddings [23]. Results might decrease if queries are too verbose, as in MedQA, or too succinct. As a mathematical process, it does not necessarily capture the nuances of responses, which is one of the common critiques of the measuring framework [28].

Response Faithfulness scores indicate responses extracted from the source contexts. For MedQA, 57 % of responses were considered faithful and accurate (94 %), and 43 % were non-faithful yet accurate (89 %). This discrepancy could indicate that responses provided with baseline LLM knowledge produce correct answers even if not adherent to the retrieved sources. Nevertheless, only one (<1 %) QA pair in MedQA did not use contexts.

For PubMedQA, 49.80 % were faithful, and the influence of background knowledge was more noticeable. Twenty-three (5 %) responses were produced without context, although only six were correct (accuracy 26 % for both). This is also attributed to the availability of predefined, public abstracts, but still raises questions about the model's use of untraceable evidence, potentially containing unverified information, regardless of accuracy. We are working to reduce the variability of responses by augmenting access to verified information sources and the models' prioritization of them.

Studies evaluating RAGAS highlight that faithfulness scores can vary due to the definition of "claims," the basic unit for its calculation, where simple statements might be paraphrased into multiple ones and complex statements might not be separated, affecting results [28]. The interpretability of scores is also questioned since similarity can be high between embeddings of random sentences [29].

5. Conclusions

Arkangel AI showed proficient retrieval and reasoning. We evidenced that adding baseline knowledge answers and using judge LLMs could improve performance. Furthermore, Arkangel AI proved to be unbiased in its multiple-choice QA answering. Prompt improvement of workflow classification will be explored. There is a need for evenly distributed QA datasets to train LLMs, and it is crucial to develop more robust frameworks for assessing LLMs' quality, involving clinicians to understand better their benefits, limitations, and areas of improvement. The model's external validation will allow us to explore these issues, perfecting its performance and assessment through real-world interaction with physicians.

CRediT authorship contribution statement

Maria Camila Villa: Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Natalia Castano-Villegas:** Writing – review & editing, Writing – original draft, Data curation, Methodology, Supervision, Investigation, Formal analysis, Conceptualization. **Isabella Llano:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Julian Martinez:** Validation, Supervision, Software, Project administration, Methodology, Conceptualization. **Maria Fernanda Guevara:** Visualization, Validation, Project administration, Conceptualization. **Jose Zea:** Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Laura Velásquez:** Supervision, Resources, Project administration, Funding acquisition.

Ethics statement

Our research does not require an ethics statement as it did not involve human participants or any patient or sensitive medical information. All questions were extracted from MedQA and PubMedQA which are widely used public medical question answering datasets.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Natalia Castano-Villegas reports a relationship with Arkangel AI that

includes: employment. Maria Camila Villa reports a relationship with Arkangel AI that includes: employment. Isabella Llano reports a relationship with Arkangel AI that includes: employment. Maria Fernanda Guevara reports a relationship with Arkangel AI that includes: employment. Julian Martinez reports a relationship with Arkangel AI that includes: employment. Jose Zea reports a relationship with Arkangel AI that includes: employment and equity or stocks. Laura Velasquez reports a relationship with Arkangel AI that includes: employment and equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2025.100274.

References

[1] Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare 2024 Jan 5;12(2):125.

[2] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the ustates medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023 Feb 8;9:e45312.

[3] Pal A, Umapathi LK, Sankarasubbu M. MedMCQA: a large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the conference on health, inference, and learning [internet]. PMLR; 2022. p. 248–60. https://proceedings.mlr.press/v174/pal22a.html.

[4] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. arXiv 2019. http://arxiv.org/abs/1909.06146.

[5] Papers with code - MedQA benchmark (question answering). https://paperswithcode.com/sota/question-answering-on-medqa-usmle.

[6] MedMCQA homepage https://medmcqa.github.io/.

[7] PubMedQA homepage https://pubmedqa.github.io/.

[8] 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.

[9] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. arXiv 2023. http://arxiv.org/abs/2305.09617.

[10] Biehl M. API architecture. API-University Press; 2015. p. 190.

[11] OpenAI o1 system card. https://openai.com/index/openai-o1-system-card/.

[12] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. arXiv 2021 [Internet], [cited 2024 Aug 15]. Available from: http://arxiv.org/abs/2009.03300.

[13] Shi F, Suzgun M, Freitag M, Wang X, Srivats S, Vosoughi S, et al. Language models are multilingual chain-of-thought reasoners. arXiv 2022. http://arxiv.org/abs/2210.03057.

[14] Chen M, Tworek J, Jun H, Yuan Q, Pinto HP, de O, Kaplan J, et al. Evaluating large language models trained on code. arXiv 2021. http://arxiv.org/abs/2107.03374.

[15] How does the PubMed best match feature work? NLM customer support center https://support.nlm.nih.gov/kbArticle/?pn=KA-03719.

[16] Lee Y, Kim J. Evaluating consistencies in LLM responses through a semantic clustering of question answering. arXiv 2024. http://arxiv.org/abs/2410.15440.

[17] What are tokens and how to count them? OpenAI help center. https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them.

[18] Wang RE, Durmus E, Goodman N, Hashimoto T. Language modeling via stochastic processes. arXiv 2023 [Internet] [cited 2025 Feb 18]. Available from: http://arxiv.org/abs/2203.11370.

[19] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. arXiv 2020 [Internet], [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2009.13081.

[20] Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv 2024 [Internet] [cited 2024 Aug 21]. Available from: http://arxiv.org/abs/2312.10997.

[21] 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 C, Takamura H, editors. Proceedings of the 12th international conference on natural language generation. Tokyo, Japan: Association for Computational Linguistics; 2019. p. 355–68 [Internet]. [cited 2024 Aug 10], Available from: https://aclanthology.org/W19-8643.

[22] Es S, James J, Espinosa Anke L, Schockaert S. RAGAs: automated evaluation of retrieval augmented generation. In: Aletras N, De Clercq O, editors. Proceedings of the 18th conference of the European chapter of the association for computational linguistics: system demonstrations. St. Julians, Malta: Association for Computational Linguistics; 2024. p. 150–8 [Internet]. [cited 2025 Mar 4], Available from: https://aclanthology.org/2024.eacl-demo.16/.

[23] Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv 2023 [Internet], [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2311.16452.

[24] Cerda LJ, Villarroel Del PL. Evaluación de la concordancia inter-observador en investigación pediátrica: coeficiente de Kappa. Rev Chil Pediatr 2008 Feb;79(1) [Internet] [cited 2025 Mar 8], Available from: http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0370-41062008000100008&lng=en&nrm=iso&tlng=en.

[25] Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns 2024 Mar 8;5(3) [Internet], [cited 2024 Jul 15], Available from: https://www.cell.com/patterns/abstract/S2666-3899(24)00042-4.

[26] Li J, Wang S, Zhang M, Li W, Lai Y, Kang X, et al. Agent hospital: a simulacrum of hospital with evolvable medical agents. arXiv 2024 [Internet], [cited 2024 Aug 13]. Available from: http://arxiv.org/abs/2405.02957.

[27] Badshah S, Sajjad H. Reference-guided verdict: llms-as-judges in automatic evaluation of free-form text. arXiv 2024 [Internet], [cited 2024 Sep 26]. Available from: http://arxiv.org/abs/2408.09235.

[28] Roychowdhury S, Soman S, Ranjani HG, Gunda N, Chhabra V, Bala SK. Evaluation of RAG metrics for question answering in the telecom domain. arXiv 2024 [Internet], [cited 2025 Mar 13]. Available from: http://arxiv.org/abs/2407.12873.

[29] Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv 2019 [Internet], [cited 2025 Mar 13]. Available from: http://arxiv.org/abs/1909.00512.