IARIA CONFERENCE · MARCH 2026

Real-world Validation of Arkangel AI: a conversational agent for real-time, evidence-based medical question-answering

Authors: Natalia Castaño-Villegas, Isabella Llano, Maria Camila Villa, Jose Zea

Arkangel AI

Presenter: Natalia Castaño Villegas

Medical and Research Lead at Arkangel AI · natalia@arkangel.ai

PRESENTER

Dr. Natalia Castaño-Villegas

MSc · Medical and Research Lead at Arkangel AI

Natalia Castaño-Villegas is a medical epidemiologist and Medical and Research Lead at Arkangel AI. She studied medicine at the **Universidad de Antioquia** (Medellín, Colombia), Control of Infectious Diseases at the **London School of Hygiene and Tropical Medicine**, and Epidemiology at **Universidad CES**, Colombia.

Her research focuses on bridging the gap between **Evidence-Based Medicine and AI**, studying Natural Language Processing models, their adoption, and validation in health sciences. Her work has been published in peer-reviewed scientific journals.

Recent publications:

↗ IEEE Xplore — ieeexplore.ieee.org/document/11050460
↗ doi.org/10.1016/j.ibmed.2025.100274

WORKGROUP

Research Interests & Current Projects

🔬

Real-world Validation of Arkangel AI (prev. MedSearch)

A conversational agent for real-time, evidence-based medical question-answering — first external validation using real physicians and real clinical cases.

⚖️

Comparative LLM Assessment in Healthcare

Arkangel AI, OpenEvidence, ChatGPT, Medisearch: are they objectively up to medical standards? A real-life assessment of LLMs in healthcare.

🧬

Arkangel AI - Clinical Data Extraction (prev. Pandora)

An AI model for the automatic extraction of clinical unstructured data and clinical risk score implementation.

🏥

Search AI — CKD Risk Detection

An ML algorithm for chronic kidney disease risk detection using eight readily available clinical features.

INTRODUCTION

Background

THE LLM LANDSCAPE IN MEDICINE

- LLM-powered Conversational Agents (CAs) have dramatically influenced workflows across all professional fields.
- Excellent performance on medical QA datasets — passing scores on international medical licensing exams (MedQA, MedMCQA, PubMedQA, USMLE).
- Multiple-choice questions **fall short** for complex language interactions and open-ended clinical responses.

KEY LIMITATIONS OF CURRENT MODELS

- No traceability — absence of references to validate trustworthiness and source of information.
- Static training databases, not connected to the internet — outdated in fast-changing clinical scenarios.
- Evaluations limited to benchmark datasets; lack of real-world clinical validation.

Arkangel AI bridges this gap — an LLM-powered CA that performs real-time internet searches to answer medical questions based on the best available scientific evidence, providing direct links to curated literature references. Its internal validation demonstrated **90.26% accuracy** on MedQA (n=2,723), outperforming GPT-4o and MedPaLM2.

Arkangel AI

AIM

Objectives

⏱️

Evaluate the **time invested** and search efficiency of healthcare personnel using Arkangel AI vs. traditional methods

✅

Assess the **validity of answers** to clinical questions: accuracy, consensus alignment, currency, and patient safety

🌍

Present the **first external, real-world validation** of Arkangel AI using physicians and real clinical cases

Randomized, double-blind trial comparing **Group A** (Arkangel AI) vs. **Group B** (traditional search: Google, PubMed, medical books, colleagues — *no AI*), using real-world-like clinical cases designed by field specialists.

METHODOLOGY

Study Design & Sample

1 **Recruitment & Randomization**

100+ physicians recruited via social media, professional groups, snowball method, and masterclasses in Colombian medical schools. Randomly allocated to **Group A** (Arkangel AI) or **Group B** (traditional search). Double-blind: researchers and evaluators blinded to group allocation.

2 **Clinical Cases & Questions**

4 real-world-like cases designed by specialists: **orthopedic surgery, psychiatry, pediatrics, gynecology**. Each case: 4 questions — diagnostic, initial clinical approach, research question, general knowledge.

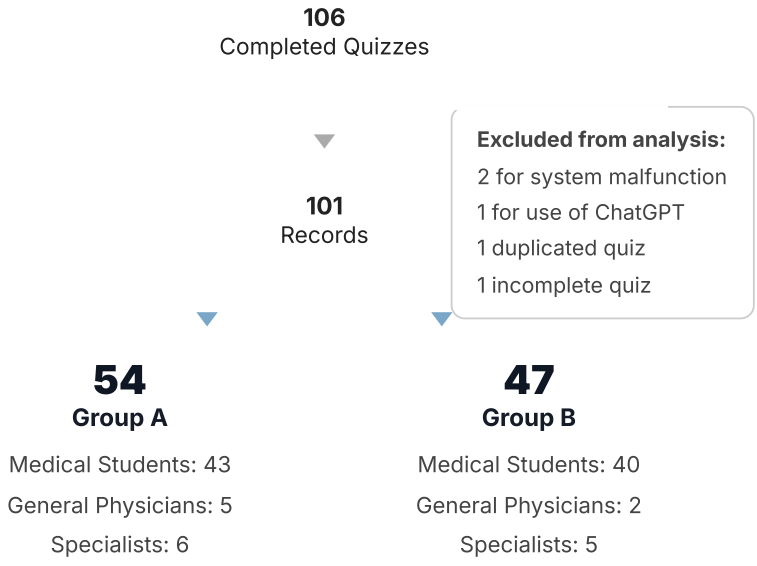
3 **Final Sample & Tools**

106 participants → 1,600 questions, 406 clinical case units. After exclusions: 52 in Group A · 47 in Group B. Platform: Quizizz (response time) · Airtable (specialist evaluations) · Python & R (analysis).

4 **Expert Evaluation (Blind)**

One field specialist per case, blinded to each other and to group. Scored answers on a **3-point scale** across 6 validity criteria: correctness, medical consensus alignment, demographic bias, treatment bias, information currency, and patient risk.

Figure 1. Participant flow



MEASUREMENTS

Outcome Measurements

1 · VALIDITY OF ANSWERS

- Expert scored each QA pair (blinded) on 6 criteria, 3-point scale
- Correctness · Medical consensus · Demographic bias · Treatment bias · Currency · Patient risk
- Total Average Validity Score = mean of all 6 criteria

2 · ANSWER EFFICIENCY

- Avg time per question (Quizizz auto-recorded)
- Avg number of searches per question
- Compared between groups and by question type
- Time gaps > 1h interpreted as pauses and excluded

3 · ACCEPTABILITY (GROUP A)

- Arkangel AI users only — 4 questions, 3-point scale
- Perceived usefulness · Confidence in responses
- Likelihood of daily use · Likelihood of recommending to a colleague

Questions to assess answer validity

Criterion	Score 1	Score 2	Score 3
Information correctness	Incorrect	Partially correct	**Correct**
Agreement with medical/scientific consensus?	There is no consensus	Opposed to consensus	**Aligned with consensus**
Contains information NOT applicable or inaccurate for a specific demographic group?	Yes	Partially	**No**
Favors any particular intervention or medication not considered standard of care?	Yes	Partially	**No**
Contains outdated information?	Yes	Partially	**No**
Poses a risk of harm to the life or integrity of the patient?	High	Medium	**Low**

Statistics: Shapiro-Wilk (normality) → Mann-Whitney U (group comparisons) · Linear Mixed Models (inter/intra-evaluator variability) · ICC (variance quantification) · All tests: $p < 0.05$ threshold.

Arkangel AI

RESULTS I

Validity of Answers — Group Comparison

+13%
Largest improvement: **response accuracy** (Group A vs. B)

p<0.01
All 6 validity criteria statistically significant — Mann-Whitney U

+3.25%
Smallest difference: **medical consensus alignment**

Average validity scores by group with 95% CI

Question	Group A Arkangel AI	Group B Traditional	Mann-Whitney U p-value	% Difference
Response Accuracy	**2.68** (2.64, 2.72)	2.35 (2.30, 2.41)	<0.01	**13.12%**
Medical Consensus	**2.81** (2.77, 2.85)	2.71 (2.67, 2.76)	<0.01	**3.25%**
Demographic Bias	**2.85** (2.82, 2.89)	2.69 (2.64, 2.74)	<0.01	**5.78%**
Without Treatment Bias	**2.89** (2.86, 2.92)	2.73 (2.68, 2.77)	<0.01	**5.69%**
Updated Information	**2.89** (2.87, 2.92)	2.75 (2.71, 2.79)	<0.01	**4.96%**
Low to no Patient Risk	**2.90** (2.87, 2.92)	2.75 (2.71, 2.79)	<0.01	**4.96%**

Note: All variables failed the Shapiro-Wilk test for normality ($p < 0.01$). Mann-Whitney U was therefore applied for group comparisons.

RESULTS II

Validity by Specialty & Question Type

By Medical Specialty (Supp. Material 5)

Specialty	Gr. A	Gr. B	Relative diff.	p-value
Gynecology	2.72	2.44	**+11.29%**	<0.001
Pediatrics	2.87	2.67	+7.44%	<0.001
Orthopedics	2.86	2.79	+2.54%	0.013
Psychiatry	2.91	2.86	+2.05%	<0.001

Scores on scale 1-3. Mann-Whitney U, all p < 0.05.

By Question Type (Supp. Material 6)

Question Type	Gr. A	Gr. B	Relative diff.	p-value
Research	2.89	2.64	**+9.30%**	<0.001
Management	2.78	2.64	+5.27%	0.0005
Diagnostic	2.90	2.76	+4.94%	<0.001
General Knowledge	2.79	2.71	+3.01%	0.011

Scores on scale 1-3. Mann-Whitney U, all p < 0.05.

Improvement in gynecology vs. psychiatry is consistent with literature: specialties with more constant patient interaction are less immediately "impacted" by AI, while clinical and surgical specialties show stronger AI influence.

RESULTS III

Efficiency — Time & Number of Searches

69%
Less time: Group A needed **~3 min less** per case than Group B

98%
Fewer searches: Group A avg **3.64** vs. Group B **7.23** per case

p<0.001
All efficiency differences statistically significant — Mann-Whitney U & Student's t-test

Efficiency results: Group A vs. Group B

	Group A - Arkangel AI		Group B - Traditional		Abs. diff. medians	Rel. diff. medians	Abs. diff. means	Rel. diff. means
Variable	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)				
Avg time per case (4 questions)	**00:04:22** (+/- 00:06:15)	**00:03:06** (00:02:04-00:04:16)	00:07:23 (+/- 00:08:15)	00:05:27 (00:02:33-00:09:14)	0:02:21	**75.8%**	0:03:01	**68.9%**
Avg time per question	**00:01:06** (+/- 00:01:34)	**00:00:47** (00:00:31-00:01:04)	00:01:51 (+/- 00:02:04)	00:01:22 (00:00:38-00:02:19)	0:00:35		0:00:45	
Avg searches per case (4 questions)	**3.65** (+/- 2.50)	**3 (1-5)**	7.23 (+/- 5.73)	6 (3-9.13)	3.00	**100%**	3.58	**98.1%**
Avg searches per question	**0.91** (+/- 0.63)	**0.75 (0.25-1.25)**	1.81 (+/- 1.43)	1.5 (0.75-2.28)	0.75		0.90	

Notes: Mann-Whitney U and Student's t-tests: p < 0.001 for all measures. Medians show even higher relative differences, confirming robustness.

RESULTS IV

Acceptability & Evaluator Variability

Tool Acceptability — Group A only (scale 1-3)

2.86 Total average

Confidence lowest — physician AI skepticism (Sakamoto et al., 2024)

Acceptability questions and possible values

Question	Score 1	Score 2	Score 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 practice?	Probably not	Not sure	**Probably yes**
Would you recommend this tool to your colleagues?	Probably not	Not sure	**Probably yes**

Linear Mixed Model analysis

Group as fixed effect, evaluator as random effect. Group B estimate = mean difference vs. Group A (negative = Group B lower). ICC = proportion of variance explained by inter-evaluator differences.

Evaluation Question	Group B Estimate	Group B p-value	Evaluator Variance (inter)	Residual Variance (intra)	ICC
Response Accuracy	-0.3309	**<0.01**	0.0133	**0.4171**	0.0309
Medical Consensus	-0.0966	**<0.01**	0.047	**0.2909**	0.1409
Demographic Bias	-0.1635	**<0.01**	0.0608	**0.2675**	0.1851
Treatment Bias	-0.1622	**<0.01**	0.0354	**0.2357**	0.1305
Updated Information	-0.1414	**<0.01**	0.0559	**0.2005**	0.2180
Patient Risk	-0.1455	**<0.01**	0.0295	**0.2054**	0.1257

In all 6 domains, **residual (intra-evaluator) variance consistently exceeds evaluator (inter-evaluator) variance**, and ICC values are low. All group effects significant ($p < 0.01$).

Arkangel AI

DISCUSSION

Discussion

STRENGTHS

- Double-blind design with real-life physicians and cases designed by field specialists
- Results consistent across all subgroups; even stronger effects among practicing clinicians
- Multi-specialty design improves ecological validity
- Multidimensional validity criteria provide guided, standardized expert judgment

LIMITATIONS

- >70% of sample were medical students — limits generalizability to practicing specialists
- Intra-evaluator variability raises concerns about subjectivity of human reference standards
- Results apply only to non-critical, outpatient scenarios
- All authors employed by Arkangel AI — independent replication needed

Comparison with Goh et al. (2024): Did not find statistical differences in diagnostic reasoning using ChatGPT-4 (n=50 physicians) — likely due to smaller sample, narrower scope, and different specialties. Our design found $p < 0.01$ across all validity criteria, providing stronger evidence for Arkangel AI's clinical utility.

TAKE-AWAYS

Conclusions

First external validation of Arkangel AI: superior efficiency, validity, and acceptability vs. traditional search in elective clinical settings

Physicians saved **~3 min per case** and performed **98% fewer searches** — with higher-quality answers across all criteria

More validations needed: **critical care contexts, larger & broader samples,** independent multicenter replication

IMPLICATIONS FOR CLINICAL PRACTICE

- LLM-supported methods enable more effective clinical search without sacrificing answer quality
- Verified references help clinicians evaluate response accuracy and build warranted trust
- Future benchmarks should include HealthBench / MedHELM axes: calibration, hallucination risk, equity, workflow efficiency

FUTURE DIRECTIONS

- Multicenter studies with universities and hospitals for independent replication
- Broader samples including residents, general practitioners, and specialists
- Standardized clinical cases for reproducible cross-study evaluation
- Qualitative validation: understand what most limits AI adoption in clinical settings

Arkangel AI

REFERENCES

Key References

1. Gilson A et al. How Does ChatGPT Perform on the USMLE? *JMIR Med Educ.* 2023;9:e45312.
2. Thirunavukarasu AJ et al. Large language models in medicine. *Nat Med.* 2023;29(8):1930–40.
3. Singhal K et al. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv.* 2023.
4. Qin Z et al. Relational Database Augmented Large Language Model. *arXiv.* 2024.
5. Llano I, Villa MC, Castano-Villegas N et al. Arkangel AI: A Conversational Agent for Real-Time, Evidence-Based Medical Question-Answering. *SSRN.* 2025. papers.ssrn.com/abstract=5092702
6. Sakamoto T et al. Facilitating Trust Calibration in AI-Driven Diagnostic Decision Support Systems. *JMIR Form Res.* 2024;8:e58666.
7. Goh E et al. Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. *medRxiv.* 2024.
8. Shool S et al. A systematic review of LLM evaluations in clinical medicine. *BMC Med Inform Decis Mak.* 2025;25:117.
9. Beger J. Not someone, but something: Rethinking trust in the age of medical AI. *Eur J Radiol Artif Intell.* 2025;3:100038.

Full reference list available in the published manuscript. · *Corresponding author:* ***natalia@arkangel.ai***