Real-World Validation of Arkangel AI: A Conversational Agent for Real-Time, Evidence-Based Medical Question-Answering

Natalia Castano Villegas, Isabella Llano
Evidence Department
Arkangel AI
Bogotá, Colombia
natalia@arkangel.ai, isabella.llano@arkangel.ai

Maria Camila Villa, Jose Zea
Product Department
Arkangel AI
Bogotá, Colombia
camila.villa@arkangel.ai, jose@arkangel.ai

***Abstract*—The paper presents the first external validation of Arkangel AI (formerly MedSearch), a retrieval-augmented large language model-based conversational agent for evidence-based medical question-answering. *Problem:* Large language model-powered conversational agents are evaluated mainly on medical question-answering datasets with strong benchmark performance; however, multiple-choice formats do not assess complex, open-ended clinical reasoning or real-world search behavior. *Why it matters:* In healthcare, validity, safety, and currency of answers matter as much as speed; prior work has not fully addressed these in ecologically valid settings with mixed healthcare personnel. *Gap:* Existing evaluations rarely combine time-to-answer, number of searches, and expert-rated validity across multiple domains in a single blinded trial. *Solution:* The authors conducted a randomized, double-blind trial comparing Arkangel AI versus traditional non-AI search in healthcare personnel answering four clinical cases (four questions each). Validity was assessed with six domains (accuracy, consensus, bias, currency, safety) on a 3-point scale; specialists were blinded to group. *Main outcome:* Arkangel AI users achieved higher validity scores across all domains ($p < 0.01$), arrived at a final answer in less than half the time (three minutes faster) than the control group, with approximately 50% fewer searches per case; total average acceptability score 2.86 on a scale from 1 to 3. Most users found Arkangel AI helpful for daily practice and would recommend it. The main conclusion is that large language model-supported methods can improve clinical search efficiency without sacrificing, and even augmenting, answer quality in this setting; broader validations are needed.**

Keywords*—Large language model assessment; human evaluation; healthcare; real-world validation.

I. INTRODUCTION

Application of large language models (LLMs) as conversational agents (CAs) in healthcare has been evaluated using medical question-answering (QA) datasets, with excellent performance on international licensing-style exams [1, 2, 3]. Multiple-choice formats fall short when the goal is to assess open-ended clinical reasoning, source traceability, and real-world search behavior. To address this, we developed Arkangel AI (formerly MedSearch), an LLM-powered CA that performs real-time, evidence-based searches and provides curated references [4]. Our first manuscript reported internal validation (90.26% on MedQA). Here we report the first external validation using real-world healthcare personnel.

Research questions. (RQ1) Does Arkangel AI improve response validity versus traditional search? (RQ2) Does it reduce time and number of searches? (RQ3) How acceptable is it to users? (RQ4) How much does evaluator variability affect scores?

Limitations of this work. The sample was predominantly clinical-year medical students (>70%) and all cases were non-urgent, outpatient; one specialist per specialty scored answers, so classical inter-rater agreement is not reported for this dataset.

Paper structure. Section II describes methodology; Section III reports results; Section IV discusses findings; Section V concludes.

II. METHODOLOGY

Design and participants. Randomized, double-blind trial; over 100 healthcare personnel recruited in Colombia via social media, professional networks, and masterclasses. After informed consent, participants were randomly assigned to Group A (Arkangel AI) or Group B (traditional search: Google, PubMed, guidelines; no AI). Four outpatient clinical cases (orthopedics, psychiatry, pediatrics, gynecology), four questions per case (diagnostic, management, research, general), were designed by external specialists. Quizizz recorded time per question; Airtable captured specialist ratings. Researchers and evaluators were blinded to participant identity and group assignment.

Outcomes. (1) Validity: six domains (correctness, consensus alignment, demographic bias, treatment bias, currency, patient risk) on a 3-point ordinal scale; one specialist per

specialty scored all QA pairs. Total Average Validity Score = mean of the six domain scores. (2) Efficiency: time per case and number of searches per case (self-reported). (3) Acceptability: four items on a 3-point scale, assessed in Group A only.

Analysis. Distributions were non-normal (Shapiro–Wilk); Mann–Whitney U test for group comparisons; medians and IQR as primary estimates; means (SD) and 95% CI also reported. Subgroup analyses by specialty and question type are exploratory.

III. RESULTS

106 participants answered 1600 questions (406 case-units). After exclusions (platform error $n = 2$, protocol violation $n = 1$), 55 remained in Group A and 48 in Group B; more than 70% were medical students.

Validity. Group A (Arkangel AI) scored higher than Group B on all six validity domains (all $p < 0.01$). Largest relative improvement: response accuracy (13.12%); smallest: medical consensus alignment (3.25%). Table I summarizes the key outcomes across both groups.

TABLE I. KEY OUTCOMES: ARKANGEL AI (GROUP A) VS. TRADITIONAL SEARCH (GROUP B)

OUTCOME	GROUP A	GROUP B	P-VALUE
Validity (1–3)	Higher in all 6 domains	Lower	< 0.01
Time per case	~3 min faster (69%)	—	< 0.001
Searches per case	~50% fewer	—	< 0.001

Efficiency. Group A reached a final answer in less than half the time (approximately three minutes faster per case) than Group B, with approximately 50% fewer searches per case. Mann–Whitney U, $p < 0.001$.

Acceptability. In Group A, total average acceptability score was 2.86 on a scale from 1 to 3. Dimension scores: utility 2.98, daily use 2.87, recommendation 2.93, truthfulness confidence 2.65 (lowest).

Evaluator variability. Linear mixed models (group as fixed effect, evaluator as random effect) showed significant group effects across all six validity domains (all $p < 0.01$). ICC values indicated that most variance was within-evaluator (intra-evaluator). The single-rater-per-specialty design is a stated limitation.

IV. DISCUSSION

Results support that Arkangel AI improves efficiency (time and searches) and validity scores versus traditional search in this predominantly student sample. Strengths include the randomized, double-blinded design, expert-designed cases, and multidimensional validity framework. Key limitations are the overrepresentation of medical students (>70%), single specialist per specialty, and acceptability assessed only in the intervention group. All authors are affiliated with Arkangel AI; evaluators and case designers were external. Findings align with recent work showing that retrieval-augmented LLM workflows can reduce clinical search burden [5], while persistent physician skepticism toward AI-driven tools remains a barrier [6].

V. CONCLUSION AND FUTURE WORK

This extended abstract presents the first external validation of Arkangel AI, demonstrating superior validity, efficiency, and acceptability versus traditional search in an elective clinical setting. Large language model-supported, evidence-based search can enhance physician workflows in this context. Future work includes multicenter recruitment, two independent raters per answer with adjudication and inter-rater reliability, pre-registered protocols, and validation in general practitioners, specialists, and higher-acuity scenarios.

ACKNOWLEDGMENT

The authors thank the healthcare personnel and medical students who participated, and the specialist physicians who designed the cases and evaluated responses.

REFERENCES

[1] A. Gilson et al., “How Does ChatGPT Perform on the USMLE?” JMIR Med Educ., vol. 9, p. e45312, 2023.

[2] A. J. Thirunavukarasu et al., “Large language models in medicine,” Nat. Med., vol. 29, pp. 1930–1940, 2023.

[3] K. Singhal et al., “Towards expert-level medical question answering with large language models,” arXiv:2305.09617, 2023.

[4] I. Llano et al., “MedSearch: A conversational agent for real-time, evidence-based medical question-answering,” SSRN 5092702, 2025.

[5] S. Shool et al., “A systematic review of LLM evaluations in clinical medicine,” BMC Med. Inform. Decis. Mak., vol. 25, p. 117, 2025.

[6] T. Sakamoto, Y. Harada, and T. Shimizu, “Facilitating trust calibration in AI-driven diagnostic support,” JMIR Form. Res., vol. 8, p. e58666, 2024.