JESK J Ergon Soc Korea 2025; 44(4): 421-436
https://dx.doi.org/10.5143/JESK.2025.44.4.421
https://jesk.or.kr eISSN:2093-8462

Exploring User Experiences Across Information Retrieval Interfaces: Keyword Search, Rule-Based, and LLM-Based Chatbots

Heesoo Sa[1], Heeyoung Park[2], Yunha Choi[3], Daehyun An[2], Dongwook Hwang[2]

[1]Kwangwoon University, Department of Metaverse Convergence, Seoul, 01890
[2]Kwangwoon University, School of Media and Communication, Seoul, 01890
[3]EchoMarketing, Seoul, 05510

정보 탐색 방식에 따른 사용자 경험 비교: 키워드, 규칙 기반, LLM 기반 챗봇을 중심으로

사희수[1], 박희영[2], 최윤하[3], 안대현[2], 황동욱[2]

[1]광운대학교 메타버스융합학과
[2]광운대학교 미디어커뮤니케이션학부
[3]에코마케팅

Corresponding Author
Dongwook Hwang
Kwangwoon University, School of Media and Communication, Seoul, 01890
Email : dongwookkr@kw.ac.kr

Received : May 22, 2025
Revised : July 18, 2025
Accepted : August 04, 2025

Copyright@2025 by The Human Factors & Ergonomics Society of Korea. All right reserved.

Objective: This study aims to evaluate and compare the user experience of three search modalities: keyword-based search, rule-based chatbots, and large language model (LLM)-based chatbots, especially in the context of healthcare information retrieval. The goal is to identify the strengths, weaknesses, and potential opportunities to improve the user experience of each modality.

Background: With the rapid growth of digital transformation, efficient search technologies have become essential. Different search modalities offer unique advantages and constraints. Traditional keyword-based searches are most widely used but limited in handling complex or domain-specific queries. In contrast, rule-based chatbots offer structured, guided interactions, while LLM-based chatbots enable flexible, conversational interactions through advanced natural language processing. In specialized fields such as healthcare, users often struggle with information overload and a lack of specialized knowledge, highlighting the need for more intuitive retrieval systems.

Method: A total of 60 participants were asked to complete a series of cancer-related information retrieval tasks using three distinct search modalities. The experiment measured task completion time, error rates, workload and user-perceived usability. Subjective assessments were conducted using the NASA-TLX and Likert-scale questionnaires. Statistical comparisons among modalities were performed using ANOVA.

Results: The ANOVA results revealed significant differences among the three modalities. Notably, rule-based chatbots demonstrated the shortest task completion times and the lowest error rates, whereas LLM-based chatbots suffered from information overload that resulted in higher error rates. Keyword-based search required greater cognitive effort due to manual query input and navigation complexities.

Conclusion: Rule-based chatbots effectively reduce cognitive load and errors in

structured search environments, while LLM-based chatbots show promise in handling complex queries but require improvements to mitigate information overload. Although keyword-based search systems remain useful for simple queries, they are less efficient for specialized searches. The findings suggest that developing hybrid chatbots with integrating the structured strengths of rule-based approaches with the flexible natural language capabilities of LLM could optimize information retrieval in healthcare.

Application: The outcomes of this study contribute to designing enhanced information retrieval systems that cater to user needs, particularly in healthcare settings where accuracy and usability are critical. Future research should focus on hybrid chatbot systems that combine the advantages of structured and flexible search modalities.

Keywords: Cognitive load, Information retrieval, Keyword-Based search, Chatbot, Large language model

1. Introduction

현대의 디지털 환경에서는 인터넷과 모바일 기술의 발전으로 정보의 양이 폭발적으로 증가하고 있으며, 이에 따라 사용자가 필요한 정보를 신속하고 정확하게 탐색할 수 있는 검색 시스템의 중요성이 그 어느 때보다 부각되고 있다. 정보 탐색 방식은 단순한 자료 제공을 넘어 사용자의 탐색 효율성과 만족도에 직결되는 요소로 작용하며, 선택된 검색 방식에 따라 최종 결과의 품질 또한 크게 달라질 수 있다. 이러한 배경 속에서 인공지능 기술의 발전과 함께 다양한 검색 기술이 등장하였고, 특히 정보 탐색의 정확성과 효율성을 높이기 위한 알고리즘 성능 향상이 지속적으로 이루어지고 있다(Croft et al., 2010; Joachims, 2002).

최근에는 딥러닝 기반의 자연어 처리(Natural Language Processing, NLP) 기술을 활용한 정보 탐색 모델이 기존 키워드 검색 방식보다 더 높은 성능을 보일 수 있음이 여러 연구를 통해 입증되고 있다(Guo et al., 2016; Karpukhin et al., 2020; Vaswani et al., 2017). 특히 GPT와 같은 대규모 언어 모델(Large Language Model, LLM)의 등장은 전통적 검색 엔진의 한계를 극복하고자 하는 새로운 가능성을 열었으며, 이러한 LLM 기반 시스템은 사용자의 의도를 더 정확하게 파악하고 자연어 기반의 대화형 인터페이스를 통해 정보 탐색 경험을 혁신할 수 있는 잠재력을 지니고 있다(Brown et al., 2020). 예를 들어, ChatGPT, Bing, Claude와 같은 LLM 기반 시스템은 사용자가 자연어로 질문을 입력하면 그 의미를 해석해 완성도 높은 문장으로 조직된 답변을 제공함으로써, 보다 직관적이고 사용자 친화적인 검색 경험을 가능하게 한다. 그러나 이러한 대화형 시스템은 상대적으로 응답 속도가 느리며, 때때로 사실과 다른 정보를 생성하는 '할루시네이션(hallucination)' 문제를 수반할 수 있다는 단점이 존재한다. 반면, 전통적인 키워드 검색 방식은 빠른 응답 속도와 명확하게 구조화된 결과 목록을 제공하는 장점이 있지만, 사용자가 개별 검색 결과를 직접 탐색하고 필요한 정보를 선별해야 하는 불편함이 따른다. 또한, 키워드 일치에 기반한 검색 방식은 사용자의 구체적인 의도나 문맥을 충분히 반영하지 못하는 경우가 많으며, 이는 검색어를 반복적으로 수정하거나 여러 단계를 거쳐야 하는 비효율적인 탐색 경험으로 이어질 수 있다(Manning et al., 2008). 이러한 현상은 특히 짧은 문서나 비정형 데이터 탐색 시 더욱 두드러지며, 결과적으로 검색의 정확성과 전반적인 사용자 경험을 저하시킬 수 있다.

이처럼 각 검색 방식은 고유의 장단점을 지니고 있으며, 사용자의 요구와 사용 환경에 따라 그 효율성에 차이가 발생한다. 따라서 다양한 검색 시스템의 특성과 성능을 비교 분석함으로써, 보다 향상된 정보 탐색 경험을 제공할 수 있는 방안을 도출하는 것이 필요하다. 이러한 연구는 빠르게 변화하는 디지털 정보 환경 속에서 보다 직관적이고 효과적인 검색 시스템 설계에 있어 실질적인 기초 자료를 제공할 수 있을 것이다. 그러나 현재까지 검색 방식 간 차이를 체계적으로 분석한 연구는 제한적이며, 기존 연구들은 주로 새로운 검색 기술의 개발 및 적용에 초점을 맞추어 실제 사용자 경험에 기반한 검색 방식의 효율성 비교 사례는 미비한 실정이다. 이에 본 연구에서는 전통적인 키워드 검색 방식과 LLM 기반 대화형 챗봇 뿐만 아니라 규칙 기반 챗봇 등의 검색 방식에 대한 효용성을 비교 분석하고자 한다. 이를 통해 각 검색 방식의 장단점을 면밀히 규명하고, 향후 검색 시스템 발전 방향에 대한 실질적인 시사점을 도출하는 것을 목적으로 한다. 본 연구는 암 환자와 보호자들이 암 관련 정보를 보다 효과적으로 습득할 수 있도록 지원하는 '함께 나리'라는 정신건강 관리 앱을 활용하여, 다양한 검색 시스템이 제공하는 정보 탐색의 효율성과 사용자 경험을 평가하고자 한다. 암에 대한 사전 지식이 제한적인 일반 사용자들이 방대한 정보를 신속하게 검색할 수 있도록 하는 시스템의 설계 및 개선 방향에 관한 논의를 통해, 향후 의료 정보 탐색 시스템의 발전에 기여하고자 한다.

2. Literature Review

2.1 LLM-based chatbot

LLM을 기반으로 한 생성형 AI 챗봇은 방대한 양의 텍스트 데이터를 학습하여 자연스러운 대화와 문맥 이해를 수행하는 특징을 지닌다. 대표적인 사례로 ChatGPT, Claude, Bing 등이 있으며, 이들은 사용자의 질문에 대해 인간과 유사한 응답을 생성함으로써 정보 탐색 분야에서 새로운 접근 방식을 제시하고 있다(Brown et al., 2020). 최근 이러한 LLM의 발전으로 연구, 교육, 금융, 의료 등 다양한 분야에 혁신적인 영향을 미칠 것으로 기대되고 있으며, 많은 전문가들은 이러한 시스템이 기존의 패러다임을 변화시킬 잠재력을 가지고 있다고 평가하고 있다(Bender et al., 2021).

정보 탐색의 맥락에서, ChatGPT와 같은 LLM 기반 챗봇이 전통적인 검색 엔진과 비교하여 개인의 정보 탐색 행동에 미치는 영향을 조사한 연구는 아직 매우 미비한 실정이다. 현재까지 알려진 바로는, 의료 분야에서 ChatGPT와 구글 검색을 사용하여 건강 관련 질문에 대한 응답을 비교한 연구가 진행된 바 있으며, 그 결과 ChatGPT가 제공하는 응답이 구글 검색이 제공하는 정보만큼 혹은 그 이상으로 유용할 수 있음을 제시하고 있다(Fernández-Pichel et al., 2025; Sun et al., 2024). 그러나 이러한 연구들은 연구 범위가 제한적이며, 의료 전문가들의 의견을 충분히 반영하지 못하는 한계점이 존재한다.

또한 최근 정보과학 및 학술정보서비스 분야에서도 LLM 기반 검색 도구가 기존의 도서관 및 데이터베이스 기반 탐색 방식의 대체제의 역할을 할 것으로 내다보고 연구가 진행되고 있다. 기존 검색 환경은 사용자가 검색 목표를 세분화하고, 적절한 키워드를 선택하며, 다수의 결과를 직접 분석해야 하는 구조였다면, LLM 기반 검색은 자연어 기반 질문만으로도 요약된 정보를 실시간 제공함으로써 탐색 과정의 부담을 크게 줄이는 방식이다. Chowdhury와 Chowdhury (2024)의 연구에 따르면, 여러 LLM 기반 검색 도구들의 수행도를 비교 분석한 결과, 도구별 응답 품질이 상이하며, 프롬프트 설계 방식 등에 따라 결과의 질에 큰 영향을 주는 것으로 나타났다. 따라서, 사용자가 각 도구의 결과를 비판적으로 검토하고 검증하는 과정이 필수적임을 강조하고 있어, 이에 대한 추가 연구가 필요하다. 이처럼 LLM 기반 도구들 간의 상대적인 성능 차이를 비교하거나 프롬프트 설계 방식에 따른 결과 품질 변화에 초점을 맞춘 탐색적 연구들이 주를 이루고 있는 상황이고, 기존의 전통적 검색 방식과 LLM 기반 검색 시스템 간의 구조적 차이에 따른 사용자 경험에 대한 분석은 미비한 실정이다.

LLM 기반 챗봇은 복잡하거나 모호한 질의에 대해 보다 직관적이고 종합적인 응답을 제공할 수 있어, 전통적인 키워드 검색 방식보다 효율적인 정보 탐색 경험을 지원할 가능성이 있다. 최근 일부 연구에서는 검색 증강 생성(Retrieval-Augmented Generation, RAG) 구조를 도입하여 외부 데이터베이스나 인터넷의 최신 정보를 실시간으로 반영함으로써, 응답의 정확성과 최신성을 향상시키는 방법을 제안하고 있다(Lewis, 2020). 그러나 동시에, 할루시네이션, 허위 정보 생성, 저작권 침해 등 LLM 기반 챗봇이 내포할 수 있는 여러 잠재적 문제점들이 제기되고 있다. 최근 연구에서는 LLM 기반 대화형 검색이 사용자로 하여금 자신의 기존 신념에 부합하는 정보만을 탐색하게 만들며, 이로 인해 선택적 노출(selective exposure)과 의견 분극(opinion polarization)을 심화시킬 수 있음을 경고하기도 하였다. 이러한 현상은 단순히 기술적 한계를 넘어서, 정보 시스템과 사용자의 상호작용 방식에서 비롯된 사회적·인지적 위험으로 간주될 수 있어, 탐색 효율성 뿐만 아니라 정보의 신뢰성, 편향성 경감 등 LLM 기반 챗봇의 한계점을 극복하기 위한 추가적인 논의가 필요하다.

종합하면, LLM 기반 챗봇은 전통적인 검색 엔진과는 다른 방식으로 사용자에게 정보를 제공하며, 특히 복잡한 질의나 도메인 특화된 정보 탐색 상황에서 유용한 도구로 활용될 수 있다. 다만, 현재까지 이와 관련된 체계적인 비교 연구가 부족한 상황이므로, 본 연구에서는 이들 시스템이 실제 사용자의 정보 탐색 행동과 만족도에 미치는 영향을 면밀히 분석하고, 향후 보다 효율적인 정보 탐색 도구 개발에 기여하고자 한다.

2.2 Rule-based chatbot

규칙 기반 챗봇은 카카오톡과 같은 문자 기반 메신저에서 흔히 사용되며, 사용자가 미리 정의된 옵션 또는 시나리오를 선택하여 대화를 진행하는 방식을 말한다. 이러한 방식은 사용자가 복잡한 자연어 입력 없이도 손쉽게 서비스를 이용할 수 있게 하여, 은행의

고객 서비스나 온라인 쇼핑몰의 제품 추천 등 특정 업무에 최적화된 응답을 빠르게 제공하는 장점을 갖는다(Luger and Sellen, 2016; Weizenbaum, 1966). 실제로, 초기 챗봇 연구인 ELIZA (Weizenbaum, 1966)는 규칙 기반 시스템의 대표적인 예로, 제한된 대화 범위 내에서 사용자와 상호작용할 수 있음을 보여주었다.

그러나 규칙 기반 챗봇은 미리 설정된 옵션에 의존하기 때문에, 사용자가 예상치 못한 질문이나 옵션 외의 요청을 할 경우 적절한 대응이 어려워 전체적인 정보 접근성이 제한될 수 있다. 이러한 한계는 복잡한 정보 탐색 과정에서 사용자의 인지적 부담을 증가시키며, 결과적으로 서비스의 효율성을 저하시킬 우려가 있다(Luger and Sellen, 2016). 또한, 챗봇 서비스 이용 의도에 관한 연구에서는 접근성이 사용자의 서비스 수용에 결정적인 영향을 미친다는 결과가 도출되었는데, 이는 규칙 기반 챗봇이 제공하는 제한된 정보 접근성이 사용자의 만족도와 재이용 의도에 부정적인 영향을 미칠 수 있음을 시사한다(Shawar and Atwell, 2007). Gnewuch et al. (2017)은 챗봇의 개념적 이해와 분류를 통해, 규칙 기반 시스템의 한계와 함께 사용자 요구에 맞는 보다 지능적인 인터페이스 설계의 필요성을 강조하였다.

따라서, 규칙 기반 챗봇은 사용 편의성과 빠른 응답 제공이라는 강점을 지니고 있으나, 정보 접근성의 제약과 복잡한 정보 탐색 상황에서의 인지적 부담 증가 등의 한계를 보완하기 위해, 사용자의 다양한 요구에 유연하게 대응할 수 있는 지능형 대화 구조 또는 하이브리드 시스템의 개발이 요구된다.

2.3 Theories of information seeking

정보 탐색과 관련된 주요 이론으로는 인지 부하 이론(Cognitive Load Theory)과 정보 포락 이론(Information Foraging Theory)이 있다. 이들 이론은 사용자가 정보를 탐색하고 처리하는 과정을 설명하며, 효과적인 검색 시스템의 설계와 평가에 중요한 시사점을 제공한다(Pirolli and Card, 1999; Sweller, 1988).

먼저, 인지 부하 이론은 학습자나 사용자가 정보를 처리하는 데 소요되는 정신적 노력, 즉 인지 부하의 구조를 설명하며, Sweller (1988)는 이를 세 가지 유형으로 구분하였다. 먼저, 내재적 부하(intrinsic load)는 과제 자체의 복잡성과 정보의 양에 따라 결정되는 인지적 요구로, 사용자가 본질적으로 이해하고 처리해야 할 내용의 복잡도에 대한 부하를 의미한다. 반면, 외재적 부하(extraneous load)는 정보 제시 방식, 인터페이스 디자인, 불필요한 탐색 과정 등 과제 외적인 요소로 인해 발생하는 인지적 부하이다. 마지막으로 본질적 부하(germane load)는 학습이나 문제 해결을 위한 의미 있는 정보 처리 활동에 사용되는 정신적 자원을 의미하며, 사용자가 정보를 구조화하고 내면화하는 과정에서 발생하는 긍정적인 인지 부하로 간주된다(Paas et al., 2003).

이러한 세 가지 인지 부하 유형은 검색 시스템 설계 시 중요한 요소로 고려되어야 한다. 예를 들어, 복잡한 검색어 입력이나 다단계 탐색은 내재적 및 외재적 부하를 증가시켜 사용자의 정신적 피로도를 높이는 반면, 직관적인 인터페이스와 요약 중심의 결과 제공은 외재적 부하를 줄이고 본질적 부하를 촉진하여 학습과 문제 해결에 긍정적인 영향을 줄 수 있다. Chandler와 Sweller (1991)는 이러한 맥락에서 정보의 제시 방식이 사용자의 인지적 부담에 미치는 영향을 실험적으로 입증하였다.

반면, 정보 포락 이론은 사용자가 최소한의 노력으로 최대한의 정보를 획득하려는 경향을 설명한다. Pirolli와 Card (1999)는 사용자가 정보 탐색 과정에서 예상되는 정보 가치와 탐색 비용을 비교하여 최적의 탐색 전략을 선택한다고 보았으며, 이러한 특성을 반영한 검색 시스템이 보다 높은 사용자 만족도를 제공할 수 있음을 제시하였다. Fu와 Pirolli (2007)는 웹 내 사용자 내비게이션 모델을 통해, 사용자가 정보를 탐색할 때 발생하는 인지적 비용과 효율성을 분석하였다. 또한, Marchionini (2006)는 탐색 과정에서 사용자가 경험하는 불확실성과 탐색 비용이 검색 결과의 만족도에 미치는 영향을 논의하였다.

이처럼 인지 부하 이론은 사용자가 정보를 얼마나 쉽게 처리하고 학습할 수 있는지를, 정보 포락 이론은 사용자가 정보에 얼마나 효율적으로 접근할 수 있는지를 설명함으로써, 검색 시스템의 설계와 사용자 경험 개선을 위한 이론적 기반을 제공한다. 따라서, 본 연구에서는 이러한 이론에 근거하여 LLM 기반 챗봇, 규칙 기반 챗봇, 그리고 전통적인 키워드 검색형 방식이 각각 상호작용의 유형과 정보 접근 방식에서 어떠한 차이를 보이는지 비교 분석하고자 한다. 이들 방식은 사용자 만족도, 편리함, 신뢰도 등 다양한 측면에서 서로 다른 경험을 제공하며, 이를 종합적으로 평가함으로써, 어떠한 상황에서 어느 접근 방식이 더 적합한지를 비교하고자 한다.

3. Method

본 연구에서는 모바일 애플리케이션을 활용한 태스크 수행을 통해 세 가지 검색 방식(키워드 검색형, 규칙 기반 챗봇, LLM 기반 챗봇)을 비교하는 피험자 간 설계(between-subjects design)를 적용하였다. 참가자들은 무작위로 세 가지 조건 중 하나에 배정되었으며, 각 조건에서 동일한 검색 태스크를 수행하였다.

3.1 Participants

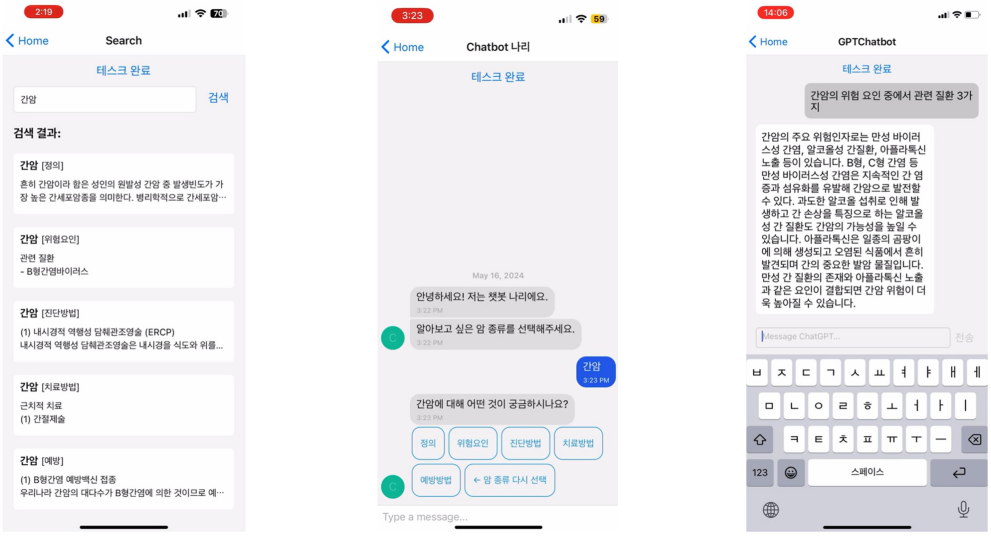
본 연구는 서울의 한 4년제 대학 학부생을 대상으로 한 참여로 진행됐다. 일주일간 교내 게시판에 연구 참여를 위한 공고를 게시해 실험 참여자를 모집했다. 실험 참여를 신청한 총 60명의 학생들 중 남성은 56%인 34명, 여성은 43.3%인 26명이었다. 평균 연령은 21.7세(SD=2.22)였다. 모든 참여자는 실험에 앞서 연구 참여동의서에 서명했다.

3.2 Experimental materials

본 연구에서는 보다 현실적인 검색 환경을 구현하기 위해 실제 암 정보 검색 플랫폼인 함께나리(HamkkeNari) 앱의 데이터베이스를 기반으로 세 가지 검색 방식을 모바일 애플리케이션 챗봇으로 개발했다. 함께나리 앱은 암 환자 및 생존자들이 신뢰할 수 있는 의료 정보를 보다 쉽게 탐색할 수 있도록 설계된 정보 제공 플랫폼으로, '함께'와 '개나리'의 상징적 의미를 결합하여 암 치료 여정에 동반자적 역할을 수행하는 것을 목표로 한다. 해당 앱은 질병 관련 정보 뿐만 아니라 명상 및 글쓰기 기능을 통해 심리적 안정을 지원하는 기능도 포함하고 있다.

본 연구에서는 함께나리 앱의 구조를 기반으로 구현된 세 가지 검색 방식을 비교하였다. Figure 1은 함께나리 앱에서 각각의 검색 방식을 적용한 샘플 이미지이다. 검색 방식에 대한 설명은 다음과 같다:

- 키워드 검색형 : 사용자가 입력한 키워드를 기반으로 관련 문서를 검색하고 목록 형태로 제공하는 방식
- 규칙 기반 챗봇: 미리 정의된 규칙과 시나리오에 따라 구조화된 검색 응답을 제공하는 챗봇
- LLM 기반 챗봇 : 대규모 언어 모델을 활용하여 자연어 질의 응답이 가능하도록 설계된 검색 시스템



(a) Keyword-based search (b) Rule-based chatbot (c) LLM-based chatbot

Figure 1. Examples of app search results by each search method: (a) Keyword-based search, (b) Rule-based chatbot, (c) LLM-based chatbot

각 검색 방식이 사용자의 탐색 행동과 검색 결과에 미치는 영향을 평가하기 위해서 본 연구에서는 여러 검색 시나리오를 모의 실험하는 실험적 방법을 적용하였다. 이는 통제된 실험 환경에서 참가자의 검색 행동을 체계적으로 관찰하고 분석할 수 있도록 하기 위함이다. 연구 참가자들은 각기 다른 검색 방식을 사용하여 정해진 태스크를 수행하였다. 참가자들이 수행한 태스크는 암 환자 및 일반 사용자가 실제로 자주 찾는 정보 유형을 반영하여 총 12가지로 구성되었다. 각 태스크는 함께나리 앱의 데이터베이스에 기반하여, 특정 암에 대한 정의, 진단 방법, 위험 요인, 치료법, 예방책 등 다양한 주제를 포괄한다. 예를 들어, "간암의 위험 요인 중 관련 질환 3가지", "갑상선암 진단에서 미세침흡인세포검사(FNAC)의 시행 조건", "위암의 가족력에 따른 위험도 증가 배수" 등과 같이, 참가자들은 실질적인 건강 정보 탐색 상황과 유사한 맥락에서 정보 탐색을 수행하였다.

전체 태스크는 간암, 갑상선암, 담낭담도암, 대장암, 위암, 유방암, 신장암, 전립선암, 자궁경부암, 췌장암, 폐암, 난소암 등 12종의 암에 대해 하나씩 할당되었으며, 각 문항은 암종별로 균형 있게 배분되었다. 또한, 특정 암종이나 질문 유형에서 과도하게 복잡하거나 난해한 정보 탐색이 요구되지 않도록, 정보 유형(정의, 원인, 진단, 치료, 예방 등)의 다양성과 난이도 간 균형을 고려하여 태스크를 설계하였다.

본 연구에서 개발된 실험 처치물의 구현을 위해 사용된 기술 스택은 다음과 같다. 프론트엔드 개발에는 JavaScript (Node.js v20.10.0, npm 10.2.3)를 언어로 사용하였으며, React Native와 Expo를 프레임워크로 활용하였다. 백엔드 개발은 Java 17을 언어로, Spring Boot 3.1.7을 프레임워크로 사용하여 구현하였다. 통합 개발 환경(IDE)으로는 IntelliJ IDEA Community Edition 2023.3.2와 Visual Studio Code를 사용하였으며, 데이터베이스(DB)로는 MySQL Community Edition 8.0을 채택하였다. 또한, Google Cloud Translation API와 OpenAI의 Fine-Tuning API를 활용하였다. 본 연구에서 사용된 각 검색 방식의 구체적인 동작 메커니즘과 설계 과정은 다음과 같다.

키워드 검색형(Figure 1(a))는 사용자가 입력한 검색어를 기반으로 관련 문서를 찾고, 문서의 우선순위를 부여하여 사용자에게 결과를 제공하는 방식으로 설계되었다. 입력된 검색어는 Lucene의 KoreanAnalyzer 라이브러리를 활용해 형태소 단위로 토큰화하였으며, 검색어의 앞에 위치한 단어에 더 높은 가중치를 부여하여 중요도를 평가하였다. 이 방식은 사용자가 "유방암 치료 방법"과 같은 문구를 입력했을 때, "유방암"과 관련된 문서가 우선적으로 노출되도록 하였다. 결과적으로 사용자가 문서를 선택하고 세부 내용을 확인할 수 있도록 구현하였다.

규칙 기반 챗봇(Figure 1(b))는 사용자가 사전에 정의된 선택지를 선택함으로써 정보를 탐색하는 구조로 설계되었다. 사용자가 특정 암의 종류를 선택하면 해당 암에 대한 정의, 위험요인, 진단 방법, 치료 방법, 및 예방법 등 세부 항목이 나타나도록 구성하였으며, 3개의 레벨로 설계하였다. 선택된 항목에 따라 관련 텍스트 데이터를 화면에 표시하며, 추가적인 선택지가 필요한 경우 추가 레벨로 진행할 수 있도록 설계하였다. 이전 레벨로 돌아갈 수 있는 기능을 포함하여 사용자의 탐색 과정을 보다 유연하게 지원하였다.

LLM 기반 챗봇(Figure 1(c))는 자연어 처리 기술을 활용하여 사용자의 자유로운 질문에 응답할 수 있도록 설계되었다. OpenAI의 GPT-3.5-turbo 모델을 암 관련 정보로 파인튜닝하여 학습시켰으며, 학습 데이터는 동일한 질문을 다양한 표현 방식으로 구성하여 모델이 높은 정확도로 답변을 생성할 수 있도록 구성하였다. 사용자가 한글로 질문을 입력하면 Google Cloud Translation API를 통해 영어로 번역하여 모델에 전달하고, 모델의 답변을 다시 한글로 번역하여 사용자에게 제공하였다.

3.3 Experimental procedure

실험은 참가자 간 설계로 구성하여, 키워드 검색형, 규칙 기반 챗봇, LLM 기반 챗봇 중 무작위로 하나에 노출됐다. 실험 참여자들은 그룹별로 무작위 할당됐다. 실험은 온전히 실험에 집중할 수 있도록 외부로부터의 빛과 소음을 완벽히 차단한 실험실에서 진행했다. 연구자는 간략한 실험 절차를 설명했고 실험 참여자는 연구 참여 및 개인 정보 제공 동의서, 그리고 사전 설문지를 작성했다. 이후, 참여자에게 본인 핸드폰을 사용하여 Expo Go 앱을 설치하도록 하고 QR 코드를 통해 챗봇 앱에 접속할 수 있도록 했다. 모든 실험 참여자는 앱을 통해 미리 정해진 태스크 12문항의 답을 찾도록 했다. 다만, 키워드 검색형의 경우 한 단어만 검색할 수 있도록 제한했다. 태스크를 모두 종료하고 사후 설문지를 작성했으며, 전체 실험 소요 시간은 약 20분이었다.

3.4 Evaluation metrics

본 연구에서는 검색 성과 및 사용자 경험을 평가하기 위해 일반적으로 활용되는 다양한 측정 지표를 적용한다. 이에 포함되는 평가 지표로는 검색 작업 수행에 소요된 시간, 오답률, 사용성 평가 지표, 작업부하 등이 포함된다. 검색 작업 수행에 소요된 시간을 측정하기 위해 각 태스크를 시작할 때 앱 상단의 시작 버튼을 누르고, 태스크의 답안을 입력하기 전에 앱 상단의 완료 버튼을 누르게 함으로써 걸린 시간을 측정했다. 또한, 주어진 태스크에서 적절한 답변을 제시하였는지를 판단하여 전체 응답에서 오답 비율이 어느 정도인지를 측정하는 오답률(%)도 확인하였다. 실험에 사용된 앱의 사용 편리성을 알아보기 위해 SUS (System Usability Scale) 평가를 실행했다. 1점(전혀 동의하지 않음)에서 5점(매우 동의함)까지의 5점 리커트(Likert) 척도를 기반으로 구성되었으며, 총 10문항을 통해 측정했다. 또한, 실험 참여자의 작업부하를 평가하기 위해서 NASA-TLX 지표를 활용하였다. NASA-TLX는 인간이 특정 작업을 수행할 때 경험하는 인지적 및 신체적 부담을 정량적으로 평가하기 위한 대표적인 주관적 작업부하 측정 도구이다(Hart and Staveland, 1988). 해당 지표는 다양한 실험 및 실제 작업 환경에서 작업 난이도, 피로도, 스트레스 등 작업 수행과 관련된 다차원적 부하를 평가하는 데 널리 사용된다. 본 연구에서는 Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration 등 6가지 하위 차원을 10점 리커트 척도로 측정했다.

Min and Kwak (2023)의 연구에 따르면, 기존의 사용성 평가 관련 선행연구들을 분석한 결과, 많은 연구에서 유용성, 신뢰성, 편리성을 주요 평가 지표로 활용하고 있음이 확인되었다. 본 연구에서는 이러한 기존 지표에 더해 가독성과 만족도 항목을 추가하였으며, '편리성' 항목은 보다 세분화하여 '편의성'과 '사용성'으로 구분하여 설문 문항을 구성하였다. 이에 따라 본 연구는 기존 선행연구(You et al., 2019; Bubaš et al., 2023)에서 타당성이 검증된 문항들을 기반으로 일부 수정하여, 유용성, 신뢰성, 사용성, 편의성, 가독성, 만족도로 총 6가지 항목으로 구성된 UX 피라미드 기반의 공통 설문지를 설계하였다. 구체적인 항목 설명과 문항 수는 Table 1과 같다. 유용성은 사용자가 챗봇 시스템을 통해 실질적인 도움을 얻는 정도를 평가하며, 시스템이 사용자에게 유익하고 의미있는 경험을 제공하는지 여부를 측정한다. 총 4개의 문항을 통해 측정했다. 신뢰성은 챗봇이 일관적이고 신뢰할 수 있는 성능을 제공하는지를 평가한다. 즉, 사용자가 시스템의 반응을 예측 가능하고 안정적이라고 인식하는 정도를 말하기도 하며, 총 2개 문항으로 측정했다. 사용성은 챗봇의 사용이 얼마나 용이한지를 평가하는 변인으로, 사용자가 직관적으로 시스템을 이해하고 쉽게 조작할 수 있는 지로 해석될 수 있다. 총 7개의 문항을 통해 측정했다. 편의성은 사용자가 챗봇을 얼마나 편리하게 사용할 수 있는지를 나타내는 항목으로, 특히 챗봇이 사용자의 시간과 노력을 얼마나 절약해주는지를 중심으로 평가했다. 총 4개 문항으로 측정했다. 가독성은 챗봇 대화 내용이 사용자에게 쉽게 읽히고 이해되는 정도를 평가하며, 정보가 명확하게 이해 가능하게 제공되는지를 측정한다. 총 3개 문항을 통해 측정했다. 만족도는 사용자가 챗봇을 사용 후 느끼는 전반적인 만족감을 말하며, 총 4개 문항으로 측정했다. 각 항목은 5점 리커트 척도로 측정했다. 신뢰도는 크론바하 알파 계수를 통해 검토하였으며, 유용성(α=.791), 신뢰성(α=.830), 사용성(α=.834), 편의성(α=.830), 가독성(α=.850), 만족도(α=.901) 모두 0.7 이상의 값을 보여 측정 도구로서의 내적 일관성이 확보되었음을 확인할 수 있었다.

Table 1. Survey Items (adapted from the UX Pyramid)

Category	Item No.	Survey item
Usefulness	U1	I was able to easily find the information I needed while using this app.
	U2	The information provided matched my questions or needs accurately.
	U3	The information provided was easy to understand.
	U4	The information fulfilled my needs for information.
Reliability	R1	The information provided is trustworthy
	R2	The search results are consistently accurate.
Usability	US1	This app is easy to use.
	US2	The interface of this app is intuitive.
	US3	I can quickly find the information I need.

Table 1. Survey Items (adapted from the UX Pyramid) (Continued)

Category	Item No.	Survey item
Usability	US4	I have no difficulty in understanding or using the search function.
	US5	This app responds to my queries as expected.
	US6	This app is user-friendly.
	US7	This app can be used without prior instruction or a manual.
Convenience	C1	This app allows quick access to necessary features and information.
	C2	The app's search function meets my needs.
	C3	It takes many steps to find information.
	C4	The process of using this app is simple and efficient.
Readability	Re1	The font size and color contrast make the information easy to read.
	Re2	I can easily identify important information.
	Re3	The amount of information provided is appropriate.
Satisfaction	S1	I am satisfied with this app.
	S2	I would recommend this app to a friend.
	S3	This app is fun to use.
	S4	I feel it is necessary to use this app

수집된 데이터는 집단 간 평균 차이를 분석하기 위해 일원분산분석(ANOVA)을 사용했다. 분석은 jamovi 통계 소프트웨어를 사용했고, 유의수준은 0.05로 설정했다. 등분산성이 만족된 경우에는 Tukey HSD 사후분석을, 등분산성이 만족되지 않은 경우에는 Games-Howell 사후분석을 실시했다. 등분산성 여부는 Levene의 등분산 검정으로 확인했다.

4. Results

NASA-TLX의 하위 항목별 평균값을 조건별로 살펴본 결과, 유의확률은 모두 .05를 상회하여 통계적으로 유의미한 차이는 없었으나, 조건에 따라 인지적 경험의 경향성이 다르게 나타났다(Table 2). 정신적 요구에서는 규칙 기반 챗봇(4.00)이 가장 높은 점수를 보였고, 키워드 검색형(3.80), LLM 기반 챗봇(3.75) 순으로 나타났다. 이는 규칙 기반 조건에서 사용자가 정해진 흐름에 맞춰야 한다는 압박감을 상대적으로 더 크게 느꼈을 가능성을 시사한다. 신체적 요구는 세 조건 모두 전반적으로 낮은 점수를 보였으며, 세 조건 간 평균값 차이는 매우 미미했다. 특히 LLM 기반 챗봇(2.60), 규칙 기반 챗봇(2.65), 키워드 검색형(2.79) 모두 큰 차이를 보이지 않았으며, 전반적으로 물리적 조작에 대한 부담은 낮았던 것으로 보인다. 시간적 요구에서는 키워드 검색형(4.25)과 규칙 기반 챗봇(4.05)가 상대적으로 높은 점수를 기록하여 시간적 압박을 더 많이 경험했음을 시사하며, 반면 LLM 기반 챗봇(2.60)은 상대적으로 여유 있는 상호작용을 가능하게 한 것으로 보인다. 수행도 항목에서는 규칙 기반 챗봇이 가장 낮은 점수(1.85)를 보였으며, 이는 참가자들이 해당 조건에서 자신의 수행을 가장 성공적으로 인식했음을 의미한다. 키워드 검색형(2.85)과 LLM 기반 챗봇(2.70)은 상대적으로 자신의 수행에 대해 낮은 만족을 보였다. 노력 항목의 경우 세 조건 모두 유사한 수준이었으나, 규칙 기반 챗봇(3.15)이 가장 높아 사용자가 작업을 완료하는 데 더 많은 노력이 들었다고 인식한 경향이 있었다. 좌절감 수준에서는 규칙 기반 챗봇(2.60)이 가장 낮은 점수를 기록하였으며, 이는 해당 조건이 상대적으로 스트레스를 덜 유발한 환경이었다는 점을 시사한다. 키워드 검색형과 규칙 기반 챗봇은 각각 3.00, 3.10으로 큰 차이는 없었다. 전반적인 6개 하위 항목의 평균 점수(Raw NASA-TLX)에서는 LLM 기반 챗봇이 2.96점으로 가장 낮았으며, 키워드 검색형이 3.23점으로 가장 높았다.

소요 시간은 키워드 검색형은 6분 14초, 규칙 기반 챗봇은 3분 52초, LLM 기반 챗봇이 7분 7초를 기록했다. 진행한 태스크의 오답률

은 평균적으로 키워드 검색형이 0.2개, 규칙 기반 챗봇이 0.15개, LLM 기반 챗봇이 1.2개로 나타났다. SUS에서는 규칙 기반 챗봇이 가장 높은 점수를 기록하였다. 키워드 검색형의 SUS 평균 점수는 80.9, 규칙 기반 챗봇은 82.4, LLM 기반 챗봇은 75.4로 나타났으며, 통계적으로 유의미한 차이는 나타나지 않았다(Table 2).

Table 2. Results of Nasa-tlx and SUS ANOVA

Variable	Group	N	Mean	SD	*F*	*p*-value
Nasa-tlx Mental demand	Keyword Search (a)	20	3.80	2.42	0.04	0.95
	Rule-Based (b)	20	4.00	2.75		
	LLM-Based (c)	20	3.75	2.79		
Nasa-tlx Physical demand	Keyword Search (a)	20	2.79	2.64	0.00	0.99
	Rule-Based (b)	20	2.65	2.98		
	LLM-Based (c)	20	2.60	2.46		
Nasa-tlx Temporal demand	Keyword Search (a)	20	4.25	3.24	2.37	0.10
	Rule-Based (b)	20	4.05	2.91		
	LLM-Based (c)	20	2.60	2.28		
Nasa-tlx performance	Keyword Search (a)	20	2.85	1.84	2.60	0.08
	Rule-Based (b)	20	1.85	1.14		
	LLM-Based (c)	20	2.70	2.32		
Nasa-tlx effort	Keyword Search (a)	20	2.80	1.79	0.10	0.89
	Rule-Based (b)	20	3.15	2.54		
	LLM-Based (c)	20	3.00	2.68		
Nasa-tlx frustration level	Keyword Search (a)	20	3.00	2.69	0.24	0.78
	Rule-Based (b)	20	2.60	1.98		
	LLM-Based (c)	20	3.10	2.59		
Nasa-tlx means of 6 subscales	Keyword Search (a)	20	3.23	1.54	0.11	0.89
	Rule-Based (b)	20	3.05	1.88		
	LLM-Based (c)	20	2.96	2.07		
SUS	Keyword Search (a)	20	80.9	8.63	1.07	0.35
	Rule-Based (b)	20	82.4	13.22		
	LLM-Based (c)	20	75.4	17.36		

사용성 평가 설문의 ANOVA 결과는 Table 3과 같다. 유의미한 차이가 발견된 설문지 문항은 유용성 3번과 가독성 3번 문항, 만족도 1번 문항이었다. 유용성 3번 문항 "제공된 정보는 이해하기 쉬웠다"에 대해 일원분산분석 결과, 세 챗봇 간 유의미한 차이가 나타났으며, 사후분석 결과 키워드 검색형과 LLM 기반 챗봇 간의 차이가 통계적으로 유의미한 것으로 확인되었다. 키워드 검색형의 평균 점수는 4.5였으며, LLM 기반 챗봇의 평균 점수는 3.4로 나타났다. 가독성 3번 문항 "제공된 정보의 양은 적절하다" 역시 유의미한 차이는 보였으며, 사후분석 결과 키워드 검색형과 LLM 기반 챗봇 간, 그리고 규칙 기반 챗봇과 LLM 기반 챗봇 간의 차이가 모두 통계적으로 유의하게 나타났다. 키워드 검색형의 평균 점수는 4.3, 규칙 기반 챗봇은 4.1, LLM 기반 챗봇은 3.05로 나타났다. 반면, 만족도 1번 문

항 "이 앱에 만족한다"에 대해 세 가지 챗봇 간 유의미한 차이가 나타났으나, 사후분석에서는 통계적으로 유의한 집단 간 차이는 확인되지 않았다.

Table 3. Results of between-subject ANOVA

Variable	Group	N	Mean	SD	*F*	*p*-value	Post-hoc
U1	Keyword Search (a)	20	4.40	0.75	0.98	0.38	
	Rule-Based (b)	20	4.35	1.08			
	LLM-Based (c)	20	3.95	1.23			
U2	Keyword Search (a)	20	4.50	0.94	1.40	0.25	
	Rule-Based (b)	20	4.00	1.12			
	LLM-Based (c)	20	4.10	0.91			
U3	Keyword Search (a)	20	4.50	0.76	5.35	0.00	a > c
	Rule-Based (b)	20	4.00	0.91			
	LLM-Based (c)	20	3.40	1.35			
U4	Keyword Search (a)	20	4.40	0.75	1.41	0.25	
	Rule-Based (b)	20	4.30	0.65			
	LLM-Based (c)	20	4.00	0.91			
R1	Keyword Search (a)	20	4.10	1.02	2.59	0.08	
	Rule-Based (b)	20	4.05	0.94			
	LLM-Based (c)	20	3.45	0.99			
R2	Keyword Search (a)	20	4.35	0.87	1.30	0.28	
	Rule-Based (b)	20	4.35	0.87			
	LLM-Based (c)	20	3.90	1.07			
US1	Keyword Search (a)	20	4.65	0.67	0.85	0.43	
	Rule-Based (b)	20	4.40	0.94			
	LLM-Based (c)	20	4.30	0.97			
US2	Keyword Search (a)	20	4.50	0.82	0.67	0.51	
	Rule-Based (b)	20	4.15	1.22			
	LLM-Based (c)	20	4.25	0.85			
US3	Keyword Search (a)	20	4.45	0.68	0.61	0.54	
	Rule-Based (b)	20	4.15	1.04			
	LLM-Based (c)	20	4.25	0.85			
US4	Keyword Search (a)	20	4.65	0.48	1.67	0.20	
	Rule-Based (b)	20	4.50	0.51			
	LLM-Based (c)	20	4.20	1.00			
US5	Keyword Search (a)	20	4.40	0.82	2.61	0.08	
	Rule-Based (b)	20	4.00	1.02			

Table 3. Results of between-subject ANOVA (Continued)

Variable	Group	N	Mean	SD	*F*	*p*-value	Post-hoc
US5	LLM-Based (c)	20	3.65	1.22	2.61	0.08	
US6	Keyword Search (a)	20	4.05	1.14	1.173	0.31	
	Rule-Based (b)	20	4.15	1.04			
	LLM-Based (c)	20	3.65	1.08			
US7	Keyword Search (a)	20	4.70	0.65	1.31	0.28	
	Rule-Based (b)	20	4.40	0.82			
	LLM-Based (c)	20	4.30	1.12			
C1	Keyword Search (a)	20	4.50	0.68	2.77	0.07	
	Rule-Based (b)	20	4.40	0.88			
	LLM-Based (c)	20	3.85	1.18			
C2	Keyword Search (a)	20	4.30	0.65	2.86	0.06	
	Rule-Based (b)	20	4.30	0.81			
	LLM-Based (c)	20	3.75	1.02			
C3	Keyword Search (a)	20	3.70	1.30	0.69	0.50	
	Rule-Based (b)	20	3.95	1.14			
	LLM-Based (c)	20	4.15	1.18			
C4	Keyword Search (a)	20	4.45	0.68	1.00	0.37	
	Rule-Based (b)	20	4.50	0.76			
	LLM-Based (c)	20	4.15	1.04			
Re1	Keyword Search (a)	20	4.35	1.13	1.91	0.15	
	Rule-Based (b)	20	4.15	1.22			
	LLM-Based (c)	20	3.65	1.13			
Re2	Keyword Search (a)	20	4.25	0.96	2.91	0.06	
	Rule-Based (b)	20	3.95	1.09			
	LLM-Based (c)	20	3.30	1.45			
Re3	Keyword Search (a)	20	4.30	0.86	6.10	0.00	a > c b > c
	Rule-Based (b)	20	4.10	1.02			
	LLM-Based (c)	20	3.05	1.35			
S1	Keyword Search (a)	20	4.40	0.75	3.71	0.03	
	Rule-Based (b)	20	4.35	0.87			
	LLM-Based (c)	20	3.65	1.22			
S2	Keyword Search (a)	20	4.10	1.07	1.967	0.15	
	Rule-Based (b)	20	3.85	1.04			
	LLM-Based (c)	20	3.40	1.27			

Table 3. Results of between-subject ANOVA (Continued)

Variable	Group	N	Mean	SD	*F*	*p*-value	Post-hoc
S3	Keyword Search (a)	20	3.25	1.20	0.31	0.73	
	Rule-Based (b)	20	3.45	1.31			
	LLM-Based (c)	20	3.15	1.13			
S4	Keyword Search (a)	20	3.80	1.24	0.77	0.46	
	Rule-Based (b)	20	3.85	0.93			
	LLM-Based (c)	20	3.40	1.35			

5. Discussion

본 연구는 모바일 애플리케이션 환경에서 키워드 검색형, 규칙 기반 챗봇, LLM 기반 챗봇 간의 정보 탐색 행동과 사용자 경험을 비교 분석하였다. 실험 결과, 인지적 부담, 탐색 효율성, 오류 발생률, 주관적 사용성 평가 등 다양한 측면에서 세 방식 간 유의미한 차이가 나타났다.

먼저 인지 부하 측면에서, NASA-TLX 설문 결과 키워드 검색형(3.23)은 LLM 기반 챗봇(2.96)과 규칙 기반 챗봇(3.05)에 비해 더 높은 인지적 부담을 유발한 것으로 나타났다. 이는 Sweller (1988), Paas et al. (2003)이 주장한 바와 같이, 복잡한 검색어 입력과 다단계 탐색이 내재적 부하를 증가시키고 정신적 스트레스를 유발했음을 시사한다. 특히 시간적 요구 측면에서는 키워드 검색형(4.25), 규칙 기반 챗봇(4.05), LLM 기반 챗봇(2.6)의 순서로 높았으며, LLM 기반 챗봇은 사용자가 관련 질문만 작성하여 자연어 기반 대화를 통해 빠르게 응답함으로써 복잡한 검색어를 찾아 입력하거나 카테고리별 탐색 없이도 정보를 얻을 수 있었다. 이에 따라 정보를 탐색하는 과정에서 불필요한 과정을 배제하고 있어 인지적 부담과 스트레스가 상대적으로 낮았고 외재적 부하를 효과적으로 낮추는 것을 확인할 수 있었다.

탐색 성과 측면에서는 규칙 기반 챗봇이 가장 우수한 성과를 보였다. 수행 시간은 평균 3분 52초로 가장 짧았고, 오답률 역시 0.15로 가장 낮았다. 이는 규칙 기반 챗봇이 미리 정의된 탐색 경로와 선택지를 제공함으로써 정보 탐색 흐름을 구조화하고 사용자의 판단을 유도했기 때문으로 해석된다. 반면 키워드 검색 방식은 평균 6분 14초의 수행 시간과 0.2의 오답률을 보였으며, 적절한 검색어 입력과 검색 결과 페이지 탐색을 반복해야 하는 점에서 탐색 비용이 높았다. LLM 기반 챗봇은 유연한 대화 방식에도 불구하고 평균 7분 7초의 탐색 시간과 1.2의 오답률로 가장 낮은 탐색 성과를 기록하였다. 이는 과도한 정보 제공으로 인해 사용자가 핵심 정보를 신속히 파악하지 못한 결과로 해석된다. 예컨대, 특정 검사의 조건을 묻는 질문에 대해 LLM 기반 챗봇이 관련된 모든 진단법을 한꺼번에 제공함으로써 정보 식별을 어렵게 만든 사례가 있었다. 이는 정보량 증가가 반드시 탐색 효율로 이어지지 않음을 보여준다.

이러한 결과는 정보 포락 이론의 관점에서도 설명 가능하다. 사용자는 탐색 비용 대비 최대 정보를 획득하려는 경향이 있으며, 규칙 기반 챗봇은 탐색 경로를 단계적으로 제한함으로써 '정보 단서'를 명확히 제공하고, 예측 가능한 정보 패치를 통해 효율적인 탐색을 지원했다. 이에 반해, LLM 기반 챗봇은 정보 탐색 중 불필요한 정보 패치 이동을 유도하며, 오히려 포락 비용을 증가시킨 경향이 확인되었다. 특히 최근 연구에 따르면, 이러한 정보 포락 이론의 개념이 LLM 기반 챗봇 환경에도 적용 가능하다는 점을 제안하였으며, 챗봇과의 대화 중 정보 단서 약화와 과잉 패치 이동이 탐색 효율을 떨어뜨릴 수 있음을 강조하고 있다(Ragavan and Alipour, 2024).

사용자 경험 측면에서도 차이가 뚜렷했다. SUS 평가에서 규칙 기반 챗봇은 82.375점으로 가장 높은 점수를 기록했고, 키워드 검색형은 80.875점, LLM 기반 챗봇은 75.375점으로 가장 낮았다. 이는 규칙 기반 챗봇이 미리 정의된 탐색 흐름을 제공함으로써 사용자에게 명확하고 직관적인 사용 경험을 제공했기 때문으로 해석된다. 반면, LLM 기반 챗봇은 자유도가 높은 대신 정보 과잉과 구조화 부족으로 인해 사용성을 저해한 것으로 보인다.

또한, 사용성 설문 결과에 따르면 LLM 기반 챗봇은 '정보의 양이 적절하다'는 가독성 항목에서 평균 3.05점으로, 다른 조건에 비해 유의미하게 낮은 점수를 기록하였다. '제공된 정보가 이해하기 쉬웠다'는 항목에서도 키워드 검색형보다 낮은 평가를 받았으며, 이는 통계적으로 유의미한 차이를 보였다($p < .05$). 이러한 결과는 Sweller et al. (2011)의 인지 부하 이론에 근거해 해석할 수 있으며, 특히 사용자가 기존에 관련 지식을 충분히 갖추지 못한 상황에서는 방대한 정보가 오히려 이해를 어렵게 만들고 인지적 부담을 증가시킬 수 있음을 시사한다. 본 연구의 검색 태스크는 암과 같은 의료 전문 정보를 다루고 있으며, 이는 일반 사용자에게 생소하고 복잡한 개념이 많기 때문에, 응답 내용이 구조화되어 있지 않거나 핵심 정보가 명확히 부각되지 않을 경우 정보 수용에 어려움을 줄 수 있다. LLM 기반 챗봇은 자유로운 자연어 응답의 장점에도 불구하고, 이러한 상황에서는 핵심 정보와 보조 정보의 경계를 구분하기 어려운 비구조화된 응답을 제공함으로써, 오히려 정보의 과잉을 초래하고 이해도를 낮출 수 있다. 따라서, 사전 지식이 충분하지 않은 사용자에게는 LLM 기반 챗봇보다 키워드 기반 혹은 규칙 기반의 구조화된 검색 방식이 더 효과적일 수 있다.

한편, LLM 기반 챗봇은 만족도 항목에서도 상대적으로 낮은 평가를 받았다. 이는 해당 챗봇이 정보 제공 시 사용자의 요구를 넘어서 다양한 세부 정보를 한꺼번에 제공하는 특성, 정보의 출처가 명확하지 않거나 신뢰도를 판단하기 어려운 응답 방식 등 여러 요인들이 복합적으로 작용했을 가능성이 있다. 과도한 정보 제공으로 인해 사용자가 필요한 정보를 빠르게 찾기 어렵게 만들며, 판단과 선택 과정에서 인지적 부담을 가중시킬 수 있고, 간결하고 핵심적인 답변을 기대하지만, 실제 응답은 복잡하고 방대한 정보를 포함해 오히려 탐색을 어렵게 만들 수 있어 사용자 불안을 유발했을 가능성이 있다. 결과적으로, 정보 탐색 효율성과 사용자 만족도 모두에 부정적 영향을 미치는 것으로 나타났다. 이러한 사용자 경험 저하는 단순한 정보량의 문제를 넘어서, LLM의 고유한 한계인 할루시네이션과 응답 신뢰도 부족에서도 기인할 수 있다. 겉보기에 자연스럽지만 실제로는 부정확하거나 맥락에 맞지 않는 정보, 그리고 출처가 명확하지 않은 응답 방식은 사용자 판단 부담을 높이고, 정보 수용 과정에 혼란을 초래할 수 있다.

요약하자면, 규칙 기반 챗봇은 구조화된 정보 흐름과 명확한 탐색 경로 제공을 통해 인지 부담을 줄이고 탐색 성과를 높이는 데 효과적이었고, 키워드 검색형은 사용자의 기존 지식이 있을 경우 비교적 효율적인 탐색이 가능했지만, 복잡한 질의나 정보량이 많은 환경에서는 반복 탐색으로 인해 인지 부하가 커졌다. LLM 기반 챗봇은 자연어 질의 응답이라는 유연성을 갖추었지만, 정보 과잉, 응답의 구조화 부족, 핵심 정보 부각 실패 등으로 인해 탐색 성과와 사용자 만족도에서 제한점을 보였다.

6. Conclusion

본 연구는 모바일 애플리케이션 환경에서 키워드 검색형, 규칙 기반 챗봇, LLM 기반 챗봇을 비교 분석하여, 각 방식이 정보 탐색 행동, 인지적 부하, 탐색 효율성, 주관적 사용성에 미치는 영향을 평가하였다. 실험 결과, 규칙 기반 챗봇은 구조화된 대화 흐름과 단계적 탐색을 통해 가장 짧은 태스크 수행 시간, 낮은 오답률, 높은 SUS 점수를 기록하며 가장 효율적인 탐색 경험을 제공했다. 반면, 키워드 검색형은 높은 인지적 부담과 긴 수행 시간, 잦은 재검색 행동 등을 보였다. LLM 기반 챗봇은 자연어 대화의 유연성 덕분에 인지 부담은 낮았지만, 정보 과잉으로 인한 탐색 비효율과 높은 오답률, 낮은 만족도로 이어졌다. 이는 각 방식이 사용자에게 제공하는 정보 탐색 경험이 인지 부하와 탐색 비용 측면에서 뚜렷하게 다름을 보여준다.

특히, 전통적인 키워드 검색 방식이 높은 정보 정확도를 기록한 점은, 여전히 해당 방식이 효율성과 정확성 면에서 강점을 지님을 시사한다. 그러나 키워드 검색형은 사용자 입력에 크게 의존하며, 복잡한 질의나 맥락을 파악하기 어렵다는 한계도 존재한다. 이에 따라, 규칙 기반 챗봇의 구조화된 정보 흐름과 LLM 기반 챗봇의 자연어 이해 및 유연한 응답을 결합한 하이브리드 챗봇이 유망한 대안으로 제시될 수 있다. 예를 들어, 사용자가 모호하거나 복잡한 질문을 입력했을 때, 하이브리드 챗봇은 규칙 기반 시스템을 통해 명확한 탐색 경로를 제공하고, 동시에 LLM 기반 언어 모델을 활용해 맥락을 이해하고 맞춤형 응답을 생성할 수 있다. 이러한 접근은 정보의 양과 질을 균형 있게 조절하여 정보 과잉 문제를 완화하고, 사용자가 원하는 정보를 더욱 효율적으로 탐색할 수 있도록 돕는다.

본 연구는 세 가지 챗봇 유형이 모바일 환경에서 사용자 정보 탐색 경험에 미치는 영향을 비교함으로써, 사용자 중심 챗봇 설계에 실질적인 시사점을 제공한다. 특히 키워드 검색형, 규칙 기반 챗봇, LLM 기반 챗봇을 동일 조건에서 정량적으로 비교하고, 그 결과를 인지 부하 이론과 정보 포락 이론의 틀로 해석함으로써 기존 이론을 실제 응용 환경에 확장 적용했다는 점에서 학문적 기여가 있다. 또한, 챗봇 설계 방식에 따른 사용자 인지 부담과 탐색 효율이 어떻게 달라지는지를 실험적으로 확인함으로써 실무적으로도 유용한 설계 방향을 제안한다. 다만, 의료 정보 탐색이라는 특정 도메인과 제한된 사용자 집단을 대상으로 했다는 점에서 결과의 일반화에는

한계가 존재한다. 특히 LLM 기반 챗봇의 응답은 실시간으로 생성되기 때문에, 응답의 일관성이나 신뢰성, 사용자의 감정적 반응 등 정성적 요소는 충분히 분석되지 못했다. 또한, 정보 탐색 능력, 사전 지식, 챗봇 사용 숙련도와 같은 사용자 특성에 따라 탐색 성과나 사용자 경험이 달라질 수 있다. 따라서 향후 연구에서는 이러한 개인차를 반영한 실험 설계와 분석을 통해 챗봇 유형에 따른 효과를 보다 정교하게 검증할 필요가 있다. 나아가, LLM 기반 챗봇의 탐색 실패 원인을 보다 다각도로 분석하기 위해, 응답의 신뢰도 평가 및 할루시네이션 발생 여부에 대한 정성적 분석도 함께 이루어질 필요가 있다. 이를 통해 챗봇 설계 전략에 대한 보다 실질적이고 보편적인 기준을 마련할 수 있을 것이다.

References

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., On the dangers of stochastic parrots: Can language models be too big?, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623, 2021. Association for Computing Machinery.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D., Language models are few-shot learners, *Advances in Neural Information Processing Systems*, 33, 1877-1901, 2020.

Bubaš, G., Cižmešija, A. and Kovačić, A., Development of an assessment scale for measurement of usability and user experience characteristics of Bing Chat conversational AI, *Future Internet*, 16(1), 2023. doi:10.3390/fi16010004

Chandler, P. and Sweller, J., Cognitive load theory and the format of instruction, *Cognition and Instruction*, 8(4), 293-332, 1991.

Chowdhury, G. and Chowdhury, S., Towards an agenda for information education and research for sustainable development, *Journal of Information Science*, 1-19, 2024. doi:10.1177/01655515241267156

Croft, W.B., Metzler, D. and Strohman, T., *Search Engines: Information Retrieval in Practice*, 2nd ed., Addison-Wesley, 2010.

Fernández-Pichel, M., Pichel, J.C. and Losada, D.E., Evaluating search engines and large language models for answering health questions, *npj Digital Medicine*, 8, 2025. doi:10.1038/s41746-025-01546-w

Fu, W.T. and Pirolli, P., SNIF-ACT: A cognitive model of user navigation on the World Wide Web, *Human-Computer Interaction*, 22(4), 355-412, 2007.

Gnewuch, U., Morana, S. and Maedche, A., Towards Designing Cooperative and Social Conversational Agents for Customer Service, *ICIS Proceedings*, 1-13, 2017.

Guo, J., Fan, Y., Ai, Q. and Chen, T., A deep relevance matching model for ad-hoc retrieval, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 55-64, 2016. ACM.

Hart, S.G. and Staveland, L.E., Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Advances in Psychology*, 52, 139-183, 1988. North-Holland.

Joachims, T., Optimizing search engines using clickthrough data, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133-142, 2002. ACM.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W., Dense passage retrieval for open-domain question answering, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769-6781, 2020. Association for Computational Linguistics.

Lewis, M., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *Advances in Neural Information Processing Systems*, 33, 9459-9474, 2020.

Luger, E. and Sellen, A., Like having a really bad PA: The gulf between user expectation and experience of conversational agents, *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5287-5296, 2016. ACM.

Manning, C.D., Raghavan, P. and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

Marchionini, G., Exploratory search: From finding to understanding, *Communications of the ACM*, 49(4), 41-46, 2006.

Min, Y. and Kwak, S., A comparative study on the usability by the platform of artificial intelligence chatbot service in library (Master's thesis), *Journal of the Korean Biblia Society for Library and Information Science*, 34(2), 2023.

Paas, F., Renkl, A. and Sweller, J., Cognitive load theory and instructional design: Recent developments, *Educational Psychologist*, 38(1), 1-4, 2003.

Pirolli, P. and Card, S., Information foraging, *Psychological Review*, 106(4), 643-675, 1999.

Ragavan, S.S. and Alipour, M.A., Revisiting Human Information Foraging: Adaptations for LLM-based Chatbots, *arXiv*, 2024. doi:10.48550/arXiv.2406.04452

Shawar, B.A. and Atwell, E., Chatbots: Are they really useful?, *Journal for Language Technology and Computational Linguistics*, 22(1), 29-49, 2007.

Sun, X., Ma, R., Zhao, X., Li, Z., Lindqvist, J., Ali, A.E. and Bosch, J.A., Trusting the search: Unraveling human trust in health information from Google and ChatGPT, *arXiv preprint*, arXiv:2403.09987, 2024.

Sweller, J., Cognitive load during problem solving: Effects on learning, *Cognitive Science*, 12(2), 257-285, 1988.

Sweller, J., Ayres, P. and Kalyuga, S., Altering Element Interactivity and Intrinsic Cognitive load, *Cognitive Load Theory*, Springer, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., Attention is all you need, *Advances in Neural Information Processing Systems*, 5998-6008, 2017.

Weizenbaum, J., ELIZA—A computer program for the study of natural language communication between man and machine, *Communications of the ACM*, 9(1), 36-45, 1966.

You, S.C., Choi, J. and Sim, M., UX design evaluation and its approach to mobile applications for smart appliances, *Smart Media Journal*, 8(3), 70-79, 2019.

Author listings

Heesoo Sa: heesoosa011228@gmail.com
Highest degree: B.A., School of Media and Communication, Kwangwoon University
Position title: M.A. Student, Department of Metaverse Convergence, Kwangwoon University
Areas of interest: Human-computer Interaction, Usability Evaluation, Interaction Design, VR, AI

Heeyoung Park: altntrkfn8282@naver.com
Highest degree: High School Diploma, Deokjeong High School
Position title: B.A. Student, Department of Media and Communication, Kwangwoon University
Areas of interest: Human-computer Interaction, Usability Evaluation, Interaction Design, UX/UI, AI

Yunha Choi: num22yh@naver.com
Highest degree: B.A., School of Media and Communication, Kwangwoon University
Position title: Staff, EchoMarketing
Areas of interest: Human-computer Interaction, Usability Evaluation, UX/UI, AI

Daehyun An: andh1006@naver.com
Highest degree: High School Diploma, Seoryeong High School
Position title: B.A. Student, School of Media and Communication, Kwangwoon University
Areas of interest: Human-computer Interaction, UX/UI, AI

Dongwook Hwang: dongwookkr@kw.ac.kr
Highest degree: PhD, Department of Industrial Engineering, Seoul National University
Position title: Assistant Professor, School of Media and Communication, Kwangwoon University
Areas of interest: Human Factors, XR, Human-computer Interaction, UX/UI, AI