Development and Application of a User Experience Evaluation Scale for Educational Chatbots Based on LLM: Focusing on English Conversation App Services

Ji-hyo Kim[1], Hyo-Jin Kang[2*]

[1]Department of Future Convergence Technology Engineering, Master Graduate, Graduate School of Sungshin Women's University, Seoul, Korea

[2]Department of Service Design Engineering, Associate Professor, Sungshin Women's University, Seoul, Korea

Abstract

Background With the rapid expansion of generative artificial intelligence (AI), large language model (LLM)-based chatbots are increasingly used in educational settings to support tasks such as answering questions, providing personalized guidance, and delivering automated feedback. However, existing usability evaluations have mainly focused on functional convenience and therefore do not sufficiently address interaction characteristics unique to LLM-based services, such as context retention, personalization, and feedback quality. Accordingly, this study aims to develop a user experience evaluation framework that reflects these distinctive attributes.

Methods This study was conducted through the phases of indicator development, validation, and application. First, a pool of candidate indicators was derived through literature review and referential sampling, and their structure was refined using affinity diagramming and expert review. Subsequently, exploratory factor analysis (EFA) was performed based on survey data to verify validity, and the indicators were applied to the evaluation of LLM-based educational applications to confirm practical applicability. Finally, the validated indicators were translated into design guidelines that can be utilized in service design and evaluation.

Results This study derives a user experience evaluation index system of LLM-based educational chatbot through literature research and expert review. The final system is composed of three categories: Usability, User Value, and User Acceptance. A total of 9 well-known and 28 detailed indicators were confirmed. In addition, the definition and core characteristics of each indicator were summarized and presented as criteria that can be used in the design and evaluation process.

Conclusions This study supplements the limitations of the existing function-oriented evaluation by proposing an indicator system to evaluate the user experience of the LLM-based educational chatbot. The proposed index identifies the difference in experience between services by reflecting on the learning context and confirms that it can be used in the design, development, and operation stages. In addition, the guidelines provide a practical basis for application to service quality improvement.

Keywords User Experience Scale, LLM, Generative AI, Educational Chatbot, Design Guideline, Conversational Interface

This research was expanded upon Kim, J. H. (2025). Development and Application of Usability Evaluation Scale for Educational Chatbots based on LLM: Focusing on an English Conversation App Services (Master dissertation), Sungshin Women's University, Seoul, Korea.

This paper was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A8076896).

*Corresponding author: Hyo-Jin Kang (hjkang@sungshin.ac.kr)

Citation: Kim, J. H., & Kang, H-J. (2026). Development and Application of a User Experience Evaluation Scale for Educational Chatbots Based on LLM: Focusing on English Conversation App Services. *Archives of Design Research, 39*(2), 265-288.

http://dx.doi.org/10.15187/adr.2026.05.39.2.265

Received : Aug. 20. 2025 ; **Reviewed :** Mar. 10. 2026 ; **Accepted :** Mar. 16. 2026
pISSN 1226-8046 **eISSN** 2288-2987

1. 서론

1. 1. 연구 배경 및 목적

정보통신기술(Information and Communication Technology, ICT)의 고도화는 4차 산업혁명의 기반이 되었으며, 이를 통해 다양한 산업 분야에서 디지털 전환(Digital Transformation, DX)이 확산되고 있다. 특히 인공지능(Artificial Intelligence, AI), 빅데이터 사물인터넷, 클라우드 컴퓨팅과 같은 핵심 ICT 기술은 산업 구조와 운영 방식을 변화시키는 동시에 새로운 서비스와 비즈니스 모델의 등장을 이끌고 있다. 이러한 변화는 교육 분야에서도 교수·학습의 전달 방식과 상호작용 구조, 학습자 경험 전반에 영향을 미치고 있다.

특히 AI 기술은 에듀테크 고도화의 핵심 요소로서, 학습자의 행동 데이터를 기반으로 한 개인 맞춤형 콘텐츠 제공, 적응형 학습 시스템, 지능형 튜터, 자동 피드백 기능 등을 통해 학습 경험의 정교화를 가능하게 한다. 이와 함께 학습 지원 챗봇(Chatbot)은 질의응답, 개념 설명, 퀴즈 제공, 학습 관리 등 다양한 기능을 수행하며 자기주도학습을 지원하고 교사의 반복적 업무를 보조하는 도구로 활용되고 있다.

이러한 흐름은 대규모 언어 모델(Large Language Model, LLM)의 등장 이후 더욱 가속화되었다. 2022년 OpenAI의 ChatGPT 공개 이후, LLM은 대규모 텍스트뿐만 아니라 이미지와 오디오 등 복합 데이터를 이해하고 생성할 수 있는 모델로 주목받으며 교육 분야에서도 새로운 가능성과 과제를 동시에 제시하고 있다. LLM 기반 챗봇은 자연스러운 언어 생성과 맥락 기반 응답을 통해 기존 AI 기반 챗봇보다 확장된 상호작용 경험을 제공한다. 그러나 이러한 특성을 교육적으로 효과적이고 안전하게 활용하기 위해서는 인간의 감독과 이를 뒷받침하는 체계적 연구가 필요하다.

기존의 사용성 평가는 기술 발전에 따라 확장되어 왔으나, LLM 기반 교육용 챗봇에 특화된 평가 체계는 아직 초기 단계에 머물러 있으며 교육적 맥락과 사용자 경험의 복합성을 충분히 반영하지 못하고 있다. 특히 디지털 서비스 환경에서는 기능 수행의 용이성뿐만 아니라, 사용 과정에서 형성되는 인지적 이해, 정서적 반응, 학습 동기와 같은 경험적 요소가 서비스에 대한 전반적인 평가와 학습 성과에 중요한 영향을 미친다. 이에 본 연구는 LLM이 도입된 교육용 챗봇의 특성을 반영한 사용자 경험 평가 지표 체계를 개발하고, 기존의 사용성 평가를 사용자 경험 평가의 관점으로 확장하여 적용하고자 한다.

1. 2. 연구 범위 및 절차

본 연구의 목적은 LLM이 도입된 영어 교육용 챗봇을 대상으로 사용자 경험 평가 지표를 개발·검증하고, 이를 실제 서비스 설계에 적용 가능한 디자인 가이드라인으로 확장하는 데 있다. 이러한 목적을 달성하기 위한 연구 질문은 다음과 같다.

연구질문 1. LLM 기반 교육용 챗봇 서비스의 사용자 경험 평가 지표는 기존 AI기반 챗봇의 사용성 평가 지표와 어떠한 차이가 있는가?

연구질문 2. LLM 기반 챗봇 서비스의 사용자 경험 평가 지표 중 'LLM'과 '교육분야'에 특화된 사용자 경험 평가 지표 요인은 어떤 것이 있는가?

연구질문 3. 본 사용자 경험 평가 지표 체계가 실제 교육용 챗봇 서비스의 사용자 경험 평가에 효과적인가?

연구질문 4. 본 사용자 경험 평가 지표 체계가 실제 교육용 챗봇 서비스에서 어떻게 활용될 수 있는가?

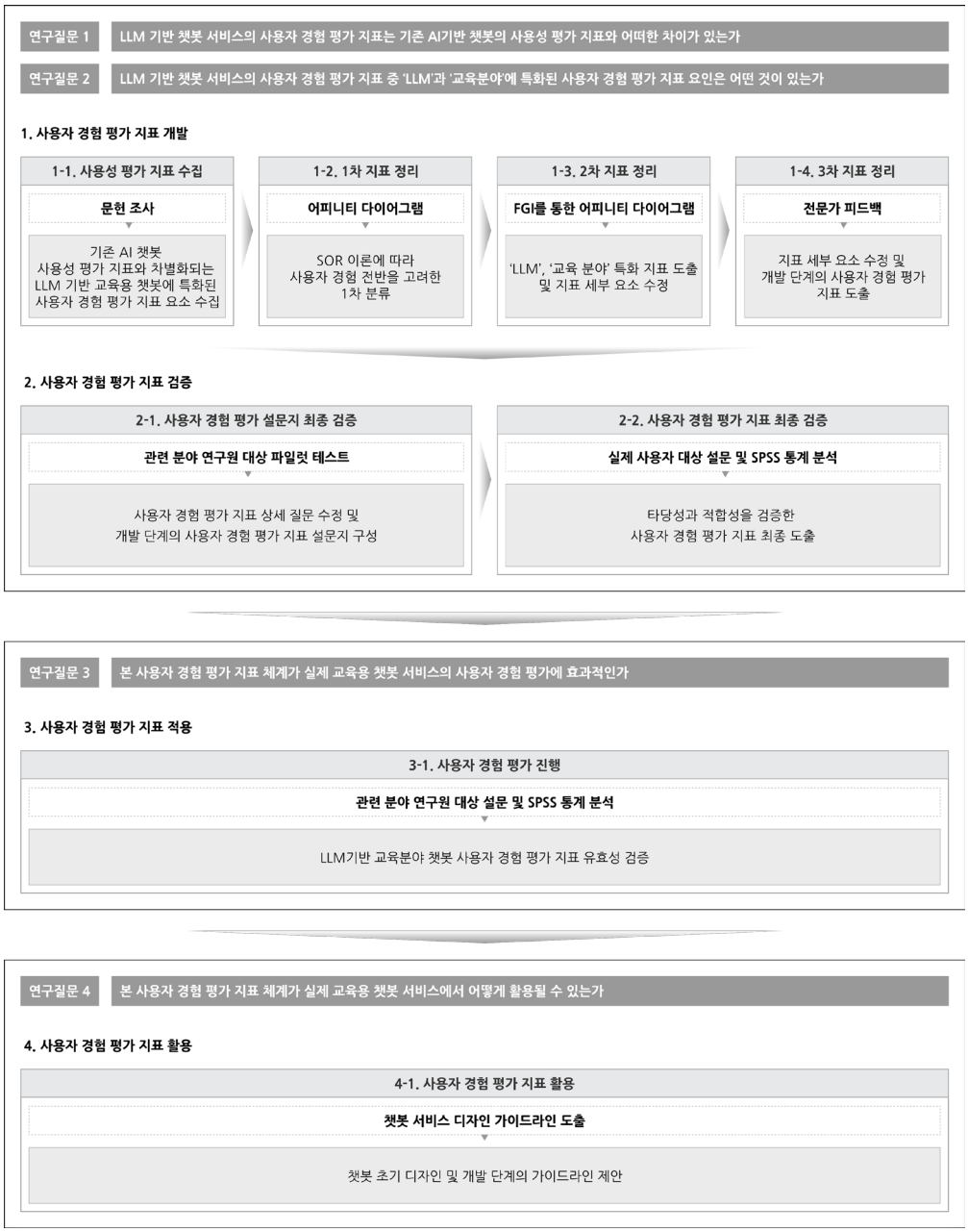

Figure 1 Research Flowchart

연구의 전체 절차는 〈Figure 1〉에 제시된 연구 흐름도와 같다.

첫째, 사용자 경험 평가 지표 개발 단계에서는 기술 및 교육 도메인의 특성을 반영하기 위해 선행연구 분석을 통해 초기 평가 항목을 수집한다. 이후 연구자의 어피니티 다이어그래밍, FGI(Focus Group Interview)를 통한 어피니티 다이어그래밍, 전문가 정성 평가의 3단계 절차를 통해 지표를 정제하고 초기 평가 지표 체계를 구축한다.

둘째, 사용자 경험 평가 지표 검증 단계에서는 파일럿 테스트를 통해 설문 문항의 적합성을 검토한 후 본 설문을 실시한다. 수집된 자료는 통계 소프트웨어(SPSS)를 활용한 탐색적 요인분석(EFA)을 통해 분석하여 지표의 신뢰성과 타당성을 검증한다.

셋째, 사용자 경험 평가 지표 적용 단계에서는 상호작용 특성이 상이한 LLM 기반 영어 교육용 애플리케이션 3종을 선정하여 사례 연구를 수행한다. 서비스디자인 분야 연구원을 대상으로 동일한 태스크를 수행하도록 하여 각 애플리케이션을 평가하고, 평가 결과를 비교·분석함으로써 지표의 실제 적용 가능성을 검증한다.

넷째, 사용자 경험 평가 지표 활용 단계에서는 최종 도출된 지표를 기반으로 LLM 기반 교육용 챗봇 설계에 적용 가능한 디자인 가이드라인을 제안한다. 이는 적절한 사용성 측정이 사용자 인터페이스 설계의 핵심 요소라는 선행연구(Nielsen, 1993)의 논의를 확장하여, 교육 도메인에서의 실질적 설계 기준을 제공하는 것을 목표로 한다.

2. 이론적 배경

본 연구는 LLM 기반 교육용 챗봇의 사용자 경험 평가 지표 체계를 개발하는 것을 목적으로 한다. 이에 따라 본 장에서는 사용자 경험 평가로의 확장을 위한 이론적 토대로서 기존 사용성 평가 연구를 중심으로 선행 이론과 지표 체계를 검토한다. 다만 선행연구에서는 관련 지표 체계를 주로 '사용성 평가'의 범주에서 논의해 왔으므로, 본 장에서는 선행연구의 맥락을 반영하여 해당 용어를 그대로 사용하고자 한다.

2. 1. 사용성 평가 지표 체계

사용성 평가는 제품이나 서비스가 사용자에게 얼마나 직관적으로 이해되고 효율적으로 사용될 수 있는지를 진단하는 개념으로, 전자기기, 모바일 애플리케이션, 웹 서비스 등 다양한 디지털 환경에서 활용되어 왔다. Nielsen(1993)의 휴리스틱 평가, Morville(2004)의 사용자 경험 허니콤 모델, ISO 9126 및 ISO 25010과 같은 대표적 지표 체계는 효율성, 유용성, 학습 용이성, 오류 회복력, 만족도 등을 중심으로 사용자와 시스템 간 상호작용 품질을 평가하는 기준을 제시하였다.

이러한 지표 체계는 사용성 평가를 다차원적으로 확장하는 데 기여하였으나, 최근 디지털 서비스 환경에서는 한계가 지적되고 있다. 특히 사용자와 시스템 간 상호작용 과정에서 형성되는 정서적 반응, 인지적 가치 인식, 지속적 사용 의도와 같은 경험적 요소가 서비스 평가에 중요한 영향을 미친다는 점이 강조되고 있다. 이에 따라 기존의 기능 중심 평가를 넘어 사용자 경험 전반을 고려하는 평가 관점의 확장 필요성이 제기되고 있다.

2. 2. LLM 챗봇의 사용성 평가 지표

대규모 언어 모델(LLM, Large Language Model)은 방대한 데이터 학습을 기반으로 인간과 유사한 수준의 언어 이해 및 생성 능력을 갖춘 자연어 처리 모델이다. LLM 기반 챗봇은 기존 규칙 기반 또는 단순 AI 챗봇과 달리, 대화 맥락을 유지하고 사용자 의도를 유연하게 해석하며 자연스러운 언어 표현을 생성하는 등 고도화된 상호작용 특성을 보인다. 이러한 능력은 기존 챗봇이 보였던 기계적 응답, 맥락 단절 등 한계를 완화해 보다 인간 친화적인 대화 경험을 제공한다.

그러나 LLM 기반 챗봇은 환각(hallucination), 편향(bias), 정보 최신성의 한계, 의인화 등 새로운 사용자 경험 문제를 동반한다. 이에 따라 기존 사용성 또는 UX 평가 지표만으로는 LLM 기반 상호작용의 특성을 충분히 평가하기 어렵다는 논의가 제기되고 있다. 안무정·강태임(2023)은 ChatGPT와 Bing 챗봇 분석을 통해 기존 휴리스틱 모델을 보완한 LLM 특화 UX 요소를 제안하며, 의인화와 사회적 실재감과 같은 요소의 중요성을 강조하였다.

이러한 연구들은 LLM 기반 챗봇의 평가가 기능적 정확성을 넘어 맥락 유지 능력, 신뢰성, 자연스러운 상호작용, 개인화 수준, 사회적 실재감 등 복합적인 경험 요소를 포함해야 함을 시사한다. 따라서 LLM 기반 챗봇을 평가하기 위해 기존 지표를 보완한 LLM 특화 사용자 경험 평가 지표 체계가 요구된다.

2. 3. 교육용 챗봇의 사용자 경험 평가 이해

교육용 챗봇은 단순한 정보 제공 도구를 넘어 학습자의 이해를 촉진하고 학습 과정을 안내하며 피드백을 제공하는 교육적 상호작용 학습환경을 형성한다. 특히 LLM 기반 교육용 챗봇은 교사와의 직접적 대면 수업이 아닌 시스템 중심 상호작용을 통해 학습이 진행된다는 점에서 전통적 교육 맥락과 구별된다. 이러한 환경에서는 학습자의 학습 조절 능력과 내재적 동기 유지가 사용자 경험 형성에 중요한 요소로 작용한다.

본 연구에서는 이러한 특성을 반영하여 교육용 챗봇의 사용자 경험 평가 체계를 수립하기 위한 핵심 교육학 이론으로 자기조절학습(Self-Regulated Learning; SRL)과 자기결정성이론(Self-Determination Theory; SDT)을 선택하였다. SRL은 학습자가 목표 설정, 수행 점검, 전략 조절을 통해 학습을 관리하는 과정을 설명하며(Zimmerman, 2002), 학습 조절 경험을 이해하는 개념적 기반을 제공한다. 이에 본 연구에서는 SRL 이론에서 제시하는 학습 조절 관련 개념을 바탕으로 personalized, 행동적 몰입, 학습 동기 등을 초기 지표 후보군으로 도출하였다(Appendix 1).

SDT는 자율성, 유능감, 관계성의 심리적 욕구 충족이 내재적 동기와 지속적 참여를 강화한다고 설명한다(Deci & Ryan, 2000). 이에 본 연구에서는 목표 설정 및 학습 진행 상황 모니터링, 피드백, self-directed, usefulness 등을 초기 지표 후보군으로 도출하였다(Appendix 1).

본 연구에서 관련 문헌을 통해 도출된 초기 지표 후보군은 자기조절학습과 자기결정성이론에서 제시하는 핵심 개념을 이론적 근거로 하여 구성되었다. 이후 통합·재구조화 과정을 거쳐 최종 사용자 경험 평가 지표 체계로 정립되었다. 이러한 접근은 LLM 기반 교육용 챗봇의 특성과 교육 맥락을 동시에 반영한 평가 체계를 구축하기 위한 이론적 기반을 제공한다.

3. 연구 방법

본 연구는 LLM 기반 교육용 챗봇의 사용자 경험 평가 지표 체계를 구축하기 위해 ①지표 개발, ②지표 검증, ③지표 적용, ④지표 활용의 네 단계를 따라 진행되었다. 각 단계는 연구 목적 달성을 위한 세부 절차로 구성된다.

3. 1. 사용자 경험 평가 지표의 개발

①-1단계: 지표 수집 및 후보군 도출

LLM 기반 교육용 챗봇의 사용자 경험 평가 지표를 도출하기 위해 2024년 4월 25일부터 6월 18일까지 총 55일간 문헌 조사를 실시하였다. Google Scholar, DBpia, RISS, arXiv 등 주요 학술 DB를 활용하여 문헌을 탐색하였으며, 〈일반적 사용성〉, 〈AI기반 챗봇〉, 〈LLM기반 챗봇〉, 〈교육용 챗봇〉을 핵심 탐색 키워드로 설정하였다. 초기 검색 결과에서 초록 검토를 통해 평가 지표를 포함하거나 LLM 및 교육 맥락과 관련성이 높은 문헌을 1차 선별하였다. 이후 연쇄 표집법(referential sampling)을 적용하여 참고문헌 목록을 추가 검토함으로써 탐색 누락 가능성을 보완하였다. 이를 통해 최종 분석 대상 문헌 29편을 선정하고 지표 후보군을 도출하였다.

①-2단계: 1차 지표 정리

도출된 지표 후보군은 인지심리학 기반의 SOR(Stimulus-Organism-Response) 이론을 활용하여 1차적으로 구조화하였다. SOR 이론은 외부 자극(S)이 사용자의 인지·정서적 상태(O)에 영향을 미치고, 그 결과 행동적 반응(R)으로 이어지는 과정을 설명하는 이론으로, 디지털 서비스 경험 연구에서 널리 활용된다(Kim et al., 2020).

본 연구는 서비스 이용 과정에서 사용자가 인지적으로 정보를 해석하고, 정서적으로 평가하며, 행동적으로 반응한다는 점에서 SOR 구조와 유사하다고 보았다. 이러한 관점은 사용성 평가를 단순한 기능적 사용성만 보는 데 한정하지 않고, 사용자가 느끼는 전반적인 사용자 경험을 총체적으로 이해하는 데 기여한다. 이에 따라 지표 후보군을 S-O-R 세 범주로 분류하고, 어피니티 다이어그램 기법을 활용하여 의미 유사 항목을 통합하여 주지표와 상세지표로 구성된 위계 구조를 도출하였다.

①-3단계: 2차 지표 정리

1차 정리된 지표 구조의 의미 적합성을 검토하기 위해 FGI를 실시하였다. FGI는 2024년 7월 23일 HCI 및 서비스디자인 연구원 3명과 HCI 분야 교수급 전문가 1명을 대상으로 진행되었다. 어피니티 다이어그램 방법론을 활용하여 지표의 정의, 범주 적합성, LLM 및 교육 맥락 반영 여부를 중심으로 검토하였다.

①-4단계: 3차 지표 정리

2차 지표 정리 내용을 바탕으로 지표 체계의 타당성 확보를 위해 2024년 8월 24일부터 29일까지 교수급 전문가 5명을 대상으로 평가를 실시하였다. 전문가별 분야 및 경력은 〈Table1〉과 같다. 지표 정의의 명확성, 분류 체계 적정성, 연구 목적 적합성을 중심으로 검토하였으며, 이를 통해 개발 단계의 사용자 경험 평가 지표를 확정하였다.

Table 1 Professional Background of 5 Expert Evaluators

ID No.	전문분야	경력	직급
EXP-1	HCI, AI	27년	교수
EXP-2	IoT, AI	25년	조교수
EXP-3	HCI, AI	16년	조교수
EXP-4	AI, HCI	15년	조교수
EXP-5	IoT, AI	12년	조교수

3. 2. 사용자 경험 평가 지표 검증

②-1단계: 설문지 구성 및 파일럿 테스트

지표 평가를 위해 각 지표를 정량적으로 측정할 수 있는 설문 문항을 구성하고, 통계적 비교와 분석이 가능하도록 리커트 척도 기반 평가 방식을 적용하였다. 설문지의 적합성을 검토하기 위해 2024년 8월 19일 HCI 및 서비스디자인 연구원 7명을 대상으로 파일럿 테스트를 실시하였으며, 응답은 온라인 설문 시스템(Google Form)을 통해 수집하였다. 파일럿 테스트에서는 실제 사용자 경험 평가 상황을 가정하여 문항 난이도, 지표 정의와의 적합성, 문항 표현 방식(평서형/의문형), 척도 유형(5점/7점) 등에 대해 평가하였다. 수집된 피드백을 근거로 설문 문항을 수정·보완하여 본 설문에 사용할 최종 설문지를 확정하였다.

②-2단계: 본 설문 및 통계적 검증

개발된 사용자 경험 평가 지표의 타당성을 검증하기 위해 본 설문을 실시하였다. 설문은 2024년 9월 12일부터 10월 4일까지 온라인으로 진행되었으며, LLM 기반 영어 교육 챗봇 사용 경험자 204명을 대상으로 하였다. 설문지는 인구통계학적 특성 문항 6개와 사용자 경험 평가 문항 31개로 구성하였다. 지표 체계가 위계적 구조를 갖는 점을 고려하여 주지표 문항은 제외하고 상세 지표 문항만을 설문에 포함하였다.

수집된 자료는 SPSS를 활용한 탐색적 요인분석(EFA)을 통해 요인 구조 도출, 신뢰도 및 타당성 검증을 수행하였으며, 이를 통해 최종 사용자 경험 평가 지표 체계를 확정하였다.

3. 3. 사용자 경험 평가 지표 적용

③-1단계: 사용자 경험 평가 진행

확정된 지표의 적용 가능성을 검증하기 위해 사용자 경험 평가 실험을 실시하였다. 실험은 2024년 11월 4일부터 6일까지 HCI 및 서비스디자인 연구원 10명을 대상으로 진행하였다. 평가 대상은 LLM 기반 챗봇 기능을 제공하며 최근 3개월 이내 업데이트가 이루어진 영어 교육 애플리케이션 3종(Speak, Praktika, Memrise)으로 선정하였다. 각 서비스의 기능과 인터랙션 특성을 사전 분석하여 공통 수행이 가능한 동일 태스크를 설계하였다. 참가자는 세 애플리케이션에서 동일 태스크를 수행한 후, 개발된 사용자 경험 평가 지표를 기반으로 구성된 동일 설문지에 응답하였다. 이를 통해 서비스 간 인터랙션 특성이 지표별 평가 결과에 어떻게 반영되는지를 비교하였다.

수집된 자료는 SPSS를 활용하여 분석하였다. 세 서비스 간 점수 차이의 통계적 유의성을 검토하기 위해 일원분산분석(One-way ANOVA)을 우선 고려하였다. 그러나 정규성 검정 결과 모든 변수에서 유의확률이 0.050 이하로 나타나 정규성을 충족하지 않았으며, 표본 수가 30 미만인 점을 고려하여 비모수 검정인 Friedman 검정을 적용하였다.

3. 4. 사용자 경험 평가 지표 활용

④-1단계: 사용자 경험 평가 지표 디자인 가이드라인 도출

개발된 지표의 실무적 적용 가능성을 제시하기 위해 최종 사용자 경험 평가 지표를 기반으로 LLM 기반 교육용 챗봇 디자인 가이드라인을 도출하였다. 제안된 지표는 서비스 기획 단계에서는 요구사항 정의와 기능 설계의 기준으로, 개발 단계에서는 상호작용 품질 점검 기준으로, 운영 단계에서는 서비스 개선과 성능 평가 기준으로 활용될 수 있다.

본 연구는 지표가 실제 설계 과정에 적용될 수 있도록 주요 디자인 원칙과 구현 시 고려해야 할 요소를 정리하였다. 제안된 가이드라인은 LLM 기반 교육용 챗봇의 기획-설계-평가 전 과정에 활용 가능한 실천적 도구로서 의의를 갖는다.

4. 사용자 경험 평가 지표 개발 결과

4. 1. 사용자 경험 평가 지표 수집

LLM 기반 교육용 챗봇의 사용자 경험 평가 지표를 개발하기 위해 관련 문헌 조사를 실시한 결과 총 29편의 관련 연구가 최종 분석 대상으로 선정되었다. 문헌의 범주는 기존 AI 기반 챗봇(C), 교육용 챗봇(E), ChatGPT 기반 챗봇(G), LLM 기반 챗봇(L)의 네 유형으로 구분하여 관리하였다. 각 문헌으로부터 사용자 경험 평가와 관련된 항목을 추출한 결과, 총 274개의 초기 평가 항목이 도출되었으며 의미 중복, 연구 목적 부합성, 평가 가능성 등을 기준으로 1차 정제 과정을 거쳐 113개의 지표 후보군을 확정하였다.

문헌 범주별 분석 결과, (C), (G/L), (E) 모든 범주에서 공통적으로 포함된 핵심 평가 요소는 총 14개로 확인되었으며, 이는 기존 연구에서 제시된 전통적 사용성 평가 지표와 일치한다. 기존 AI 기반 챗봇(C) 문헌에서는 접근성, 프라이버시 보호 등 기능적 안정성과 기본적 사용성 요소가 강조된 반면, LLM 기반 챗봇(G/L) 문헌에서는 사회적 실재감, 의인화, 최신성, 환각, 편향 등 생성형 AI 특성에 기인한 상호작용 품질 요소가 두드러지게 나타났다. 교육용 챗봇(E) 문헌에서는 학습 동기, 개인 맞춤성, 학습 정보 제공, 피드백 등

학습 지원 요소가 핵심 지표로 제시되었다. 이러한 특성을 반영하여 구성된 초기 지표 후보군은 전통적 사용성 요소, LLM 특성 기반 요소, 교육 특화 요소가 통합된 구조를 가지며, 이는 본 연구의 지표 개발을 위한 기초를 형성하였다.

4. 2. 사용자 경험 평가 지표 체계 구성

4. 2. 1. 지표 체계 정립 과정

문헌 기반으로 도출된 113개의 초기 지표 후보군은 SOR 이론에 따라 분류한 뒤, 어피니티 다이어그램을 활용하여 1차 분석을 진행하였다. 그 결과 총 11개의 주지표와 40개의 상세 지표가 도출되었다. 이후 SOR 이론의 구성 요소를 본 연구 맥락에 맞게 재정의하였다. 자극(S)은 서비스 자체의 특성을 평가하는 요소로 판단하여 사용성(Usability)으로, 유기체(O)는 사용자의 기능적·감정적 가치 및 만족감과 관련된 요소로 보아 사용자 가치(User Value)로, 반응(R)은 서비스 수용 여부 및 지속 사용 의도와 관련되므로 사용자 수용도(User Acceptance)로 재구성하였다.

2차 분석은 FGI를 통해 수행되었으며, 지표의 의미 적합성과 범주를 재검토하여 11개의 주지표와 31개의 상세 지표로 수정·정리하였다. 이 과정에서 LLM 특화 지표와 교육 특화 지표를 구분하였다. 'LLM 특화 지표'는 주지표인 사회적 실재감과 상세 지표인 최신성, 환각, 의인화, 유연성, 공감, 인간다운 자연스러움 등 총 7개로 구성되었다. '교육 분야 특화 지표'는 주지표인 교육적 상호작용성과 상세 지표인 개인 맞춤성, 학습자 인지성, 학습 정보 제시성, 피드백, 이해 가능성, 몰입성, 자기주도성, 학습 동기 등 총 9개로 구성되었다.

마지막 3차 분석을 위해 전문가 의견을 종합한 결과, 다음과 같은 주요 수정 사항을 도출할 수 있었다. a. 의미 및 지표 특화에 따른 지표명 수정, b. 의미가 모호한 지표 정의 수정, c. 상세 지표 그룹 이동, d. 지표 추가 및 삭제 등의 주요 의견을 확인할 수 있었다.

a. 의미 및 지표 특화에 따른 지표명 수정

- LLM의 주요 특징인 맥락 일치를 강조해야 한다는 의견에 따라, 기존에 유사한 의미를 내포하고 있는 일관성을 '맥락 적합성'으로 수정하고 LLM 특화 지표로 변경하였다.(EXP-4)
- 챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용 하는 정도라는 정의가 유연성이라는 지표와 어울리지 않는다는 의견에 따라 지표명을 '적응성'으로 수정하였다.(EXP-2)
- 기능적 가치의 상세 지표인 효과성, 효율성, 신뢰성이 상세 지표보다 상위 개념의 지표로 느껴진다는 의견에 따라 지표명을 '학습 효과성', '학습 효율성', '학습 신뢰성'으로 수정하였다.(EXP-3)
- 다른 상세 지표명과 지표명의 수준에 차이가 느껴진다는 의견에 따라 학습 동기를 '학습 동기부여'로 변경하였다.(EXP-3)

b. 의미가 모호한 지표 정의 수정

- 주지표인 '접근성'의 의미가 교육적 상호작용성의 상세 지표인 이해 가능성과 유사하다고 느껴질 수 있다는 의견에 따라 챗봇 서비스가 사용자의 수준과 관계없이 용이하게 접근하여 사용할 수 있는 정도였던 정의를 '챗봇 서비스가 사용자의 신체적/환경적/인지적 수준의 제약 없이 시작할 수 있는 정도'로 수정하였다.(EXP-5)

c. 상세 지표 그룹 이동

- 안전성의 상세 지표였던 '환각 방지'가 주지표와 의미가 이질적이라는 의견에 따라 정보 전달성의 상세 지표로 이동하였다. '환각 방지'의 정의는 챗봇 서비스가 부정확하거나 신뢰할 수 없는 정보를 생성하지 않는 정도이다. 사용자가 챗봇이 생성하는 정보가 부정확하거나 신뢰할 수 없다고 해서 안전성이 떨어진다고 느끼기보다는 챗봇이 전달하는 내용이 의미상으로 명확하지 않다고 느낀다고 판단하였다.(EXP-1, EXP-2, EXP-5)

d. 지표 추가 및 삭제

- 접근성의 상세 지표인 물리적 접근성이 너무 많은 의미를 내포하고 있다는 의견에 따라, '신체적 접근성'과 '환경적 접근성'으로 지표를 분리하였다.(EXP-3)
- 챗봇 서비스가 생성한 부정확한 정보를 사용자가 잘못 활용하는 것을 방지하는 태도를 의미하는 오용 방지의 평가 내용이 모호하다는 의견에 따라 지표를 삭제하였다.(EXP-1)
- 안전성 평가 항목으로 챗봇이 프라이버시 침해 혹은 환각성 결과물을 생성한다고 느끼는지 평가하는 항목이 필요하다는 의견에 따라 챗봇 서비스가 제공하는 정보나 상호작용 과정에서 불공정한 편향과 차별이 없는지를 평가하는 '윤리성'을 추가하였다.(EXP-5)

이와 같은 과정을 통해 최종적으로 11개의 주지표와 31개의 상세 지표로 구성된 사용자 경험 평가 지표 체계를 확정하였다.

4. 2. 2. 지표 체계 정립 결과

본 연구는 사용자 경험 평가 지표 개발 단계를 거쳐 〈Table 2〉와 같이 개발 단계의 LLM기반 교육용 사용자 경험 평가 지표 체계를 구성하였다.

Table 2 User Experience scale and definitions: Development Phase

U. 사용성(Usability)		
U.1. 정보 전달성 (Information Delivery)		**챗봇 서비스가 제공하는 정보가 의미상으로 명확하게 전달되고 있는 정도**
	U.1.1. 명료성 (Clarity)	챗봇 서비스가 제공하는 정보가 명확하고 간결하게 의미를 전달하는 정도
	U.1.2. 투명성 (Transparency)	챗봇 서비스가 제공하는 정보가 시스템의 상태를 투명하게(사실대로) 전달하는 정도
	U.1.3. 맥락 적합성 (Contextual Conformity)	챗봇 서비스가 제공하는 정보가 특정 주제나 대화 맥락에 적합한 내용을 전달하는 정도
	U.1.4. 최신성 (Up-to-Dateness)	챗봇 서비스가 제공하는 정보가 최신 정보를 반영하고 있는 정도
	U.1.5. 환각 방지 (Hallucination Prevention)	챗봇 서비스가 부정확하거나 신뢰할 수 없는 정보를 생성하지 않는 정도
U.2. 시각적 전달성 (Visual Delivery)		**챗봇 서비스가 제공하는 정보가 시각적으로 명확하게 전달되고 있는 정도**
	U.2.1. 가시성 (Visibility)	챗봇 서비스가 제공하는 정보의 시각적 표현이 명확하게 전달되고 있는 정도
	U.2.2. 직관성 (Intuitiveness)	챗봇 서비스가 제공하는 정보의 시각적 표현이 직관적으로 이해하기 쉬운 정도
U.3. 접근성 (Visual Delivery)		**챗봇 서비스를 사용자의 신체적/환경적/인지적 수준의 제약 없이 시작할 수 있는 정도**
	U.3.1. 신체적 접근성 (Physical Accessibility)	챗봇 서비스를 신체적 조건 및 수준의 제약 없이 시작할 수 있는 정도
	U.3.2. 환경적 접근성 (Environmental Accessibility)	챗봇 서비스를 시간/장소 등 사용 환경의 제약 없이 시작할 수 있는 정도
	U.3.3. 인지적 접근성 (Cognitive Accessibility)	챗봇 서비스를 인지적 수준(연령/교육 수준 등)의 제약 없이 시작할 수 있는 정도
U.4. 안전성 (Safety)		**챗봇 서비스가 제공하는 정보나 상호작용 과정이 안전하다고 지각되는 정도**
	U.4.1. 프라이버시 보호 (Privacy Protection)	챗봇 서비스가 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하는 정도
	U.4.2. 오류 관리 (Error Management)	챗봇 서비스가 제공하는 정보가 시스템 사용 시 발생할 수 있는 오류를 예방하고, 오류발생 시 적절히 대처하는 정도
	U.4.3. 윤리성 (Ethicality)	챗봇 서비스가 제공하는 정보나 상호작용 과정에 불공정한 편향과 차별이 없는 정도

U.5. 사회적 실재감 (Social Presence)		**챗봇 서비스가 사용자에게 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용하는 정도**
	U.5.1. 의인화 (Personification)	챗봇 서비스에 상황과 태스크에 따라 적절한 인격이 투영되어 의인화된 정도
	U.5.2. 적응성 (Adaptiveness)	챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용하는 정도
	U.5.3. 공감 (Empathy)	챗봇 서비스가 사용자에게 인지적/감정적으로 공감하여 상호작용하는 정도
	U.5.4. 인간다운 자연스러움 (Human Naturalness)	챗봇 서비스와의 상호작용이 실제 인간같이 자연스럽고 이질감이 없는 정도
U.6. 교육적 상호작용성 (Educational Interaction)		**챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도**
	U.6.1. 개인 맞춤성 (Personalization)	챗봇 서비스가 사용자의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 학습 내용을 제공하는 정도
	U.6.2. 학습 정보 제시성 (Presentation of Learning Information)	챗봇 서비스가 교육(학습) 내용 및 진행 상황에 대한 정보를 명확하게 제시하는 정도
	U.6.3. 피드백 (Feedback)	챗봇 서비스가 교육(학습)에 필요한 피드백을 적시 적소에 제공하는 정도
	U.6.4. 몰입성 (Immersion)	챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있는 정도
	U.6.5. 학습 동기부여 (Learning Motivation)	챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하는 정도
	U.6.6.자기 주도성 (Self Directedness)	챗봇 서비스가 자기 주도적인 교육(학습) 경험을 제공하는 정도
	U.6.7. 이해 가능성 (understandability)	챗봇 서비스가 제공하는 교육(학습) 내용 및 정보를 잘 이해할 수 있는 정도
V. 사용자 가치(User Value)		
V.1. 기능적 가치 (Functional Value)		**챗봇 서비스 사용 시 기능적으로 얻는 혜택과 가치**
	V.1.1. 학습 효과성 (Learning Effectiveness)	챗봇 서비스를 통해 교육(학습) 목적 달성의 효과를 얻는 정도
	V.1.2. 학습 효율성 (Learning Efficiency)	챗봇 서비스를 통해 교육(학습)을 효율적으로 하는 정도
	V.1.3. 학습 신뢰성 (Learning Credibility)	챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있는 정도
V.2. 감정적 가치 (Emotional Value)		**챗봇 서비스 사용 시 감정적으로 얻는 혜택과 가치**
	V.2.1. 즐거움 (Enjoyment)	챗봇 서비스 사용 경험에서 즐거움과 흥미를 느끼는 정도
	V.2.2. 심미성 (Aesthetics)	챗봇 서비스 사용 경험에서 심미성을 느끼는 정도
	V.2.3. 친밀감 (Intimacy)	챗봇 서비스 사용 경험에서 심리적 친밀감을 느끼는 정도
	V.2.4. 자기효능감 (Self Efficacy)	챗봇 서비스 사용 경험에서 본인의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻는 정도
A. 사용자 수용도(User Acceptance)		
A.1. 만족도 (Satisfaction)		챗봇 서비스의 전반적인 경험에 대해 만족하는 정도
A.2. 태도 (Attitude)		챗봇 서비스 경험을 긍정적으로 생각하는 정도
A.3. 지속 사용 의도 (Continuous Intention)		챗봇 서비스를 지속적으로 사용하고자 하는 정도

5. 사용자 경험 평가 지표 검증 결과

5. 1. 사용자 경험 평가 파일럿 테스트 및 설문 구성

개발 단계에서 도출된 상세 지표를 정량적으로 측정하기 위해 설문 문항을 구성하고 파일럿 테스트를 실시하였다. 파일럿 테스트 결과, 의문형 문항이 응답하기 용이하다고 인식한 응답자의 비율이 57.1%였으며, 5점 리커트 척도가 적합하다는 응답이 85.7%로 나타났다. 이에 따라 최종 설문에서는 문항 표현 방식을 의문형으로 통일하고, 5점 리커트 척도를 적용하였다. 최종 설문지 구성은 〈Table 3〉과 같다.

Table 3 Final survey questions

사용성		
정보 전달성		**챗봇 서비스가 제공하는 정보가 의미상으로 명확하게 전달되고 있는 정도**
	1. 명료성	챗봇 서비스가 제공하는 정보가 명확하고 간결하게 의미를 전달하고 있나요?
	2. 투명성	챗봇 서비스가 제공하는 정보가 시스템의 상태를 투명하게(사실대로) 전달하고 있나요?
	3. 맥락 적합성	챗봇 서비스가 제공하는 정보가 특정 주제나 대화의 맥락에 적합한 내용을 전달하고 있나요?
	4. 최신성	챗봇 서비스가 제공하는 정보가 최신 정보를 반영하고 있나요?
	5. 환각 방지	챗봇 서비스가 부정확하거나 신뢰할 수 없는 정보를 생성하지 않고 있나요?
시각적 전달성		**챗봇 서비스가 제공하는 정보가 시각적으로 명확하게 전달되고 있는 정도**
	6. 가시성	챗봇 서비스가 제공하는 정보의 시각적 표현이 명확하게 전달되고 있나요?
	7. 직관성	챗봇 서비스가 제공하는 정보의 시각적 표현이 직관적으로 이해하기 쉽나요?
접근성		**챗봇 서비스를 사용자의 신체적/환경적/인지적 수준의 제약 없이 시작할 수 있는 정도**
	8. 신체적 접근성	챗봇 서비스를 신체적 조건 및 수준의 제약 없이 시작할 수 있나요?
	9. 환경적 접근성	챗봇 서비스를 시간/장소 등 사용 환경의 제약 없이 시작할 수 있나요?
	10. 인지적 접근성	챗봇 서비스를 인지적 수준(연령/교육 수준 등)의 제약 없이 시작할 수 있나요?
안전성		**챗봇 서비스가 제공하는 정보나 상호작용 과정이 안전하다고 지각되는 정도**
	11. 프라이버시 보호	챗봇 서비스가 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하고 있나요?
	12. 오류 관리	챗봇 서비스가 제공하는 정보가 시스템 사용 시 발생할 수 있는 오류를 예방하고, 발생 시 적절히 대처하고 있나요?
	13. 윤리성	챗봇 서비스가 제공하는 정보나 상호작용 과정에 불공정한 편향과 차별이 없나요?
사회적 실재감		**챗봇 서비스가 사용자에게 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용하는 정도**
	14. 의인화	챗봇 서비스에 상황과 태스크에 따라 적절한 인격이 투영되어 의인화되었나요?
	15. 적응성	챗봇 서비스가 나의 반응이나 변화를 수용하고 적용하여 상호작용하고 있나요?
	16. 공감	챗봇 서비스가 나에게 인지적/감정적으로 공감하여 상호작용하고 있나요?
	17. 인간다운 자연스러움	챗봇 서비스와의 상호작용이 실제 인간같이 자연스럽고 이질감이 없나요?
교육적 상호작용성		**챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도**
	18. 개인 맞춤성	챗봇 서비스가 나의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 교육(학습) 내용을 제공하고 있나요?
	19. 학습 정보 제시성	챗봇 서비스가 교육(학습) 내용 및 진행 상황에 대한 정보를 명확하게 제시하고 있나요?
	20. 피드백	챗봇 서비스가 교육(학습)에 필요한 피드백을 적시 적소에 제공하고 있나요?
	21. 몰입성	챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있나요?
	22. 학습 동기부여	챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하고 있나요?
	23. 자기 주도성	챗봇 서비스가 자기 주도적인 교육(학습) 경험을 제공하고 있나요?
	24. 이해 가능성	챗봇 서비스가 제공하는 교육(학습) 과정과 내용을 신뢰할 수 있나요?

사용자 가치	
기능적 가치	**챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도**
25. 학습 효과성	챗봇 서비스를 통해 교육(학습) 목적 달성의 효과를 얻고 있나요?
26. 학습 효율성	챗봇 서비스를 통해 교육(학습)을 효율적으로 하고 있나요?
27. 학습 신뢰성	챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있나요?
감정적 가치	**챗봇 서비스 사용 시 감정적으로 얻는 혜택과 가치**
28. 즐거움	챗봇 서비스 사용 경험에서 즐거움과 흥미를 느끼나요?
29. 심미성	챗봇 서비스 사용 경험에서 심미성을 느끼나요?
30. 친밀감	챗봇 서비스 사용 경험에서 심리적 친밀감을 느끼나요?
31. 자기효능감	챗봇 서비스 사용 경험에서 본인의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻고 있다고 느끼나요?

5. 2. 사용자 경험 평가 및 통계 결과 분석

본 연구는 LLM 기반 영어 교육용 챗봇 서비스의 사용자 경험 평가 지표 체계를 검증하기 위해 설문 조사를 실시하였으며, 영어 회화 기반 LLM 챗봇 사용 경험이 있는 응답자 204명의 자료를 확보하였다. 응답자의 인구통계학적 특성은 〈Table 4〉에 제시하였으며, 표본은 20~30대가 다수를 차지하고(91.1%), 대졸 이상 비율이 높았으며(85.8%), 디지털 기기 친숙도는 '보통 이상'이 87.2%로 나타났다. 또한 영어 교육용 챗봇을 주 1~2회 이상 이용하는 응답자가 75.4%로 확인되어, 본 연구 목적에 부합하는 사용자 표본이 확보되었음을 확인하였다.

Table 4 Demographic characteristics of survey participants

성별	**남성**				**여성**			
	101명 (49.2%)				103명 (50.2%)			
연령	**10대**	**20대**	**30대**	**40대**	**50대 이상**			
	2명 (0.9%)	128명 (62.4%)	59명 (28.7%)	13명 (6.3%)	2명 (0.9%)			
디지털 기기 친숙도	**전혀 친숙하지 않다**	**친숙하지 않다**	**보통이다**	**친숙하다**	**매우 친숙하다**			
	9명 (4.3%)	16명 (7.3%)	32명 (15.6%)	84명 (40.9%)	63명 (30.7%)			
교육 수준	**초등학교 졸업**	**고등학교 졸업**	**대학교 졸업**	**대학원 이상**				
	1명 (0.4%)	27명 (13.1%)	154명 (75.1%)	22명 (10.7%)				
영어 교육용 챗봇 평균 이용 빈도	**비지속적 사용**	**월 4회 미만**	**주 1~2회**	**주 3~4회**	**주 5~6회**	**매일 사용**		
	49명 (23.9%)	27명 (13.1%)	81명 (39.5%)	31명 (15.1%)	6명 (2.9%)	10명 (4.8%)		

지표 체계의 통계적 타당성을 검증하기 위해 상세 지표 31개 항목을 대상으로 탐색적 요인분석(EFA)을 실시하였다. 이때 '사용자 가치' 영역은 '사용성'보다 개념적 범위가 넓고, 일부 항목이 사용 경험을 기반으로 형성될 수 있으므로 하나의 구조로 분석할 경우 요인 해석이 혼재될 가능성이 있다. 이에 분석의 개념적 명확성을 확보하기 위해 사용성 영역과 사용자 가치 영역을 분리하여 EFA를 수행하였다.

아래 〈Table 5〉, 〈Table 6〉은 EFA를 통해 상세 지표 31개의 항목에 대해 항목 구조를 파악한 결과이다. 주성분 분석은 VARIMAX 직교 회전으로 수행되었으며, 요인 추출 기준은 고윳값 1 이상으로 설정하였다. 사용성 카테고리의 총 6차례의 반복된 EFA 단계를 거쳤다. 우선 사용성 카테고리의 상세 지표인 변수 1 ~ 변수 31을 대상으로 변수 제거 없이, 고윳값 기준으로 1차 EFA를 진행하였다. 분석 결과 개발 단계에서 정의한 상세 지표 그룹과 상이한 공통 구조가 나타났으며, 이를 고려해 고윳값 기준으로 파악된 요인 수(4개)를 고정해 4차례 추가 EFA를 실시하였다. 그 결과 변수 5(환각 방지), 변수 8(신체적 접근성), 변수 14(의인화)가 지속적으로

불규칙적으로 구분되었으며, 기존 상세 지표 그룹과 다른 속성을 띠고 있음을 확인할 수 있었다. 이에 세 변수를 제외한 28개 변수를 대상으로 재분석을 수행하였고, 최종적으로 21개의 상세 지표가 사용성 영역에 적합한 구조를 형성하는 것으로 나타났다.

Table 5 Final EFA Results: Usability Category

KMO의 표본 적합도(MSA) m 검정		0.926				
Bartlett의 구형성 검정		근사 카이제곱		2667.393		
		자유도		210		
		유의확률		〈.001		
요인	요소	성분				공통성
		1	2	3	4	
요인 1	6. 가시성	**0.763**	0.193	0.175	0.253	0.714
	3. 맥락 적합성	**0.753**	0.285	0.181	0.222	0.731
	4. 최신성	**0.746**	0.252	0.181	0.149	0.676
	7. 직관성	**0.702**	0.227	0.163	0.245	0.630
	1. 명료성	**0.702**	0.239	0.116	0.392	0.716
	2. 투명성	**0.686**	0.261	0.156	0.184	0.597
요인 2	17. 인간다운 자연스러움	0.310	**0.776**	0.183	0.058	0.735
	16. 공감	0.238	**0.765**	0.162	0.093	0.677
	15. 적응성	0.344	**0.650**	0.214	0.265	0.657
	18. 개인 맞춤성	0.185	**0.618**	0.160	0.411	0.610
	20. 피드백	0.290	**0.608**	0.117	0.448	0.668
	19. 학습 정보 제시성	0.254	**0.568**	0.152	0.419	0.606
요인 3	10. 인지적 접근성	0.044	0.041	**0.816**	0.210	0.714
	11. 프라이버시 보호	0.189	0.110	**0.815**	0.162	0.738
	12. 오류 관리	0.344	0.234	**0.706**	0.119	0.686
	13. 윤리성	0.258	0.110	**0.696**	0.163	0.590
	9. 환경적 접근성	0.031	0.251	**0.627**	0.069	0.462
요인 4	24. 이해 가능성	0.295	0.009	0.293	**0.761**	0.753
	20. 몰입성	0.229	0.280	0.177	**0.708**	0.664
	22. 학습 동기부여	0.288	0.315	0.191	**0.703**	0.714
	23. 자기 주도성	0.328	0.337	0.153	**0.687**	0.716

사용성 영역의 요인분석 결과, 표본 적합도(KMO)는 0.926으로 나타났으며, Bartlett의 구형성 검정에서도 통계적으로 유의한 결과($p<.001$)를 보여 요인분석 수행의 적합성이 확인되었다(Table 5). 이러한 기초 검증을 바탕으로 총 네 개의 사용성 요인이 최종적으로 도출되었다.

Table 6 Final EFA Results: User Value Category

KMO의 표본 적합도(MSA) m 검정		0.871		
Bartlett의 구형성 검정		근사 카이제곱	675.952	
		자유도	210	
		유의확률	〈.001	
요인	요소	성분		공통성
		1	2	
요인 5	25. 학습 효과성	**0.843**	0.190	0.746
	26. 학습 효율성	**0.842**	0.219	0.756
	27. 학습 신뢰성	**0.707**	0.427	0.600
	28. 즐거움	**0.629**	0.452	0.682
요인 6	29. 심미성	0.174	**0.879**	0.716
	30. 친밀감	0.289	**0.770**	0.803
	31. 자기효능감	0.404	**0.744**	0.676

사용자 가치 영역에서도 KMO 값은 0.871로 적정 수준을 충족하였으며, Bartlett 검정 결과 역시 통계적으로 유의하게 나타났다(p〈.001)(Table 6). 이러한 기초 검증을 바탕으로 총 두 개의 사용성 요인이 최종적으로 도출되었다. 특히 '즐거움'은 초기 분류에서는 감정적 가치 영역에 포함되었으나, EFA 결과 기능적 가치 요인으로 분류되었다. 이는 교육 서비스 맥락에서 즐거움이 엔터테인먼트적 감정적 만족이라기보다, 학습 과정에서의 흥미·참여를 높이는 기능적 요소로 인식되는 경향이 반영된 결과로 해석할 수 있다.

5. 3. LLM기반 교육용 챗봇 사용자 경험 평가 지표 최종 도출

통계 분석 결과를 바탕으로 사용자 경험 평가 지표 체계를 재구성하였다. 개발 단계에서 도출된 11개의 주지표와 31개의 상세 지표는 검증 단계를 거치며 9개의 주지표와 28개의 상세 지표로 축소·정제되었다. 기존 문헌 연구와 통계 분석 결과를 종합하여 주지표가 포함되는 각 요인(요인 1~6)의 명칭을 재정의하였다. 사용성(Usability) 영역에서는 요인 1을 '효과적인 전달성', 요인 2를 '사회적 실재감', 요인 3을 '신뢰 가능한 상호작용', 요인 4를 '교육적 상호작용성'으로 명명하였다. 사용자 가치(User Value) 영역의 요인 5는 '감정적 가치', 요인 6은 '기능적 가치'로 재정의하였다.

한편, 본 지표체계의 상세지표인 '개인 맞춤성', '학습정보 제시성', '피드백' 세 지표는 학습·교육 분야 선행연구에서 추출된 항목으로, 교육학 및 학습 분야에서 널리 활용되는 분류에 기반한다. 문헌에서도 해당 지표는 학습 상호작용 기능과 관련해 핵심적으로 다루어진다. 탐색적 요인분석(EFA)에서는 이 항목들이 통계적으로 LLM 요인인 '사회적 실재감'과 군집되었으나, 이는 교육적 의미와 지표 본래 의도를 충분히 반영하지 못한다는 한계가 있다. 선행연구에 따르면, 요인 구조가 이론적 체계와 상충할 경우 명확한 이론적 근거가 존재하거나 전문가 합의가 확보된 경우에는 통계적 결과를 그대로 수용하지 않고 개념적 정합성을 기준으로 최종 구조를 확정한 사례가 보고된 바 있다(김방희, 2015; 이대용, 2012). 이러한 기준에 따라 본 연구에서도 통계적 결과는 보조적 근거로 활용하되, 지표가 교육적 상호작용 기능을 수행한다는 문헌적 근거를 우선 고려하였다. 또한 1차 연구 과정에서 참여했던 교수급 전문가들에게 추가 검토를 요청한 결과, 세 지표는 학습 과정에서 독립적 기능을 수행하므로 하나의 영역으로 분리하는 것이 타당하다는 의견이 제시되었다. 이에 본 연구에서는 세 지표를 교육 특화 영역인 '교육적 상호작용성'으로 재분류하였으며, 이를 통해 지표 간 의미적 일관성과 이론적 정합성을 확보하였다.

〈Table 7〉은 LLM 기반 교육용 챗봇의 최종 사용자 경험 평가 지표 체계를 제시한다.

Table 7 Final User Experience scale and definitions

U. 사용성(Usability)	
U.1. 효과적인 전달성 (Effective Delivery)	**챗봇 서비스가 제공하는 정보가 의미적/시각적으로 명확하게 전달되고 있는 정도**
U.1.1. 명료성 (Clarity)	챗봇 서비스가 제공하는 정보가 명확하고 간결하게 의미를 전달하는 정도
U.1.2. 투명성 (Transparency)	챗봇 서비스가 제공하는 정보가 시스템의 상태를 투명하게(사실대로) 전달하는 정도
U.1.3. 맥락 적합성 (Contextual Conformity)	챗봇 서비스가 제공하는 정보가 특정 주제나 대화 맥락에 적합한 내용을 전달하는 정도
U.1.4. 최신성 (Up to Dateness)	챗봇 서비스가 제공하는 정보가 최신 정보를 반영하고 있는 정도
U.1.5. 가시성 (Visibility)	챗봇 서비스가 제공하는 정보의 시각적 표현이 명확하게 전달되고 있는 정도
U.1.6. 직관성 (Intuitiveness)	챗봇 서비스가 제공하는 정보의 시각적 표현이 직관적으로 이해하기 쉬운 정도
U.2. 신뢰 가능한 상호작용 (Reliable Interaction)	**챗봇 서비스가 안전하고 신뢰할 만한 상호작용을 제공하는 정도**
U.2.1. 환경적 접근성 (Environmental Accessibility)	챗봇 서비스를 시간/장소 등 사용 환경의 제약 없이 시작할 수 있는 정도
U.2.2. 인지적 접근성 (Cognitive Accessibility)	챗봇 서비스를 인지적 수준(연령/교육 수준 등)의 제약 없이 시작할 수 있는 정도
U.2.3. 프라이버시 보호 (Privacy Protection)	챗봇 서비스가 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하는 정도
U.2.4. 오류 관리 (Error Management)	챗봇 서비스가 시스템 사용 시 발생할 수 있는 오류를 예방하고, 오류 발생 시 적절히 대처하는 정도
U.2.5 윤리성 (Ethicality)	챗봇 서비스가 제공하는 정보나 상호작용 과정에 불공정한 편향과 차별이 없는 정도
U.3. 사회적 실재감 (Social Presence)	**챗봇 서비스가 사용자에게 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용하는 정도**
U.3.1. 적응성 (Adaptiveness)	챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용하는 정도
U.3.2. 공감 (Empathy)	챗봇 서비스가 사용자에게 인지적/감정적으로 공감하여 상호작용하는 정도
U.3.3. 인간다운 자연스러움 (Human Naturalness)	챗봇 서비스에 상황과 태스크에 따라 적절한 인격이 투영되어 실제 인간같이 자연스럽고 이질감이 없는 정도
U.4. 교육적 상호작용성 (Educational Interaction)	**챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도**
U.4.1. 개인 맞춤성 (Personalization)	챗봇 서비스가 사용자의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 내용을 제공하는 정도
U.4.2. 학습 정보 제시성 (Presentation of Learning Information)	챗봇 서비스가 교육(학습) 내용 및 진행 상황에 대한 정보를 명확하게 제시하는 정도
U.4.3. 피드백 (Feedback)	챗봇 서비스가 사용자에게 필요한 피드백을 적시 적소에 실시간으로 제공하는 정도
U.4.4. 몰입성 (Immersion)	챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있는 정도
U.4.5. 학습 동기부여 (Learning Motivation)	챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하는 정도
U.4.6. 자기 주도성 (Self Directedness)	챗봇 서비스가 자기 주도적인 교육(학습) 경험을 제공하는 정도
U.4.7. 이해 가능성 (understandability)	챗봇 서비스가 제공하는 교육(학습) 내용 및 정보를 잘 이해할 수 있는 정도

V. 사용자 가치(User Value)	
V.1. 기능적 가치 (Functional Value)	**사용자가 서비스 사용 시 기능적으로 얻는 혜택과 가치**
V.1.1. 학습 효과성 (Learning Effectiveness)	챗봇 서비스를 통해 교육(학습) 목적 달성의 효과를 얻는 정도
V.1.2. 학습 효율성 (Learning Efficiency)	챗봇 서비스를 통해 교육(학습)을 효율적으로 하는 정도
V.1.3. 학습 신뢰성 (Learning Credibility)	챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있는 정도
V.1.4. 학습 흥미성 (Learning Interest)	챗봇 서비스의 교육(학습) 과정에서 즐거움과 흥미를 느끼는 정도
V.2. 감정적 가치 (Emotional Value)	**사용자가 서비스 사용 시 감정적으로 얻는 혜택과 가치**
V.2.1. 심미성 (Aesthetics)	챗봇 서비스 사용 경험에서 심미성을 느끼는 정도
V.2.2. 친밀감 (Intimacy)	챗봇 서비스 사용 경험에서 심리적 친밀감을 느끼는 정도
V.2.3. 자기효능감 (Self Efficacy)	챗봇 서비스 사용 경험에서 본인의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻는 정도

A. 사용자 수용도(User Acceptance)	
A.1. 만족도 (Satisfaction)	**챗봇 서비스의 전반적인 경험에 대해 만족하는 정도**
A.2. 태도 (Attitude)	**챗봇 서비스 경험을 긍정적으로 생각하는 정도**
A.3. 지속 사용 의도 (Continuous Intention)	**챗봇 서비스를 지속적으로 사용하고자 하는 정도**

6. 사용자 경험 평가 지표 적용 결과

6. 1. 실제 서비스 환경 분석 결과

사용자 경험 평가를 위해 선정한 세 개의 애플리케이션을 분석한 결과, 모두 LLM 기반 대화형 인터페이스를 제공하였으나 인터랙션 방식은 다르게 나타났다. 스픽은 음성 기반 대화와 즉각적 문법·어휘 피드백을 중심으로 난이도 조절 기능을 제공하였고, 프랙티카는 아바타 기반 1:1 대화를 지원하며 실시간 교정 기능을 제공하였으나 대화 중 난이도 변경은 제한적이었다. 멤라이즈는 GPT 기반 AI 파트너를 통해 텍스트 중심의 자유 대화를 지원하였으나 추가 설명 기능은 제한적으로 제공되었다. 이러한 인터랙션 차이는 이후 지표별 사용자 경험 평가 결과를 해석하는 기준으로 활용하였다.

6. 2. 개발된 사용자 경험 평가 지표 체계 적용 결과

〈Appendix 2〉는 세 가지 실제 서비스에 대한 Friedman 검정 결과를 제시한다. 분석 결과, LLM 특화 지표인 맥락 적합성, 최신성, 개인 맞춤성, 적응성, 공감, 인간다운 자연스러움, 친밀감에서 서비스 간 통계적으로 유의미한 차이가 나타났으며, 교육 특화 지표에서도 학습 신뢰성을 제외한 대부분 항목에서 유의한 차이가 확인되었다. 반면 일반 지표의 대다수 항목에서는 유의미한 차이가 나타나지 않았다.

이는 세 서비스가 LLM 기반 챗봇으로서 기본 기능 수준은 유사하나, 상호작용 방식 및 학습 지원 전략에 따라 사용자 경험이 차별적으로 형성될 수 있음을 시사한다. 따라서 본 연구에서 개발된 사용자 경험 평가 지표 체계는 인터랙션 구조가 상이한 LLM 기반 영어 학습 서비스 간 사용자 경험 차이를 효과적으로 구분하고 해석하는 데 활용 가능함을 확인하였다.

7. 사용자 경험 평가 지표 활용

7. 1. 지표를 활용한 효과적인 사용자 경험 평가 방안 제안

본 연구에서는 일반 AI 사용자 경험 평가 지표, LLM 특화 지표, 교육 분야 특화 지표를 통합적으로 반영한 사용자 경험 평가 지표 세트를 개발하였다. 해당 지표는 서비스의 기획·개발·운영 전 단계에서 활용 가능하다.

서비스 기획 단계에서는 지표를 기반으로 요구사항과 기능 설계를 명확히 정의함으로써, 챗봇의 대화 품질, 피드백 방식, 맞춤형 학습 지원 기능 등 사용자 경험 요소를 초기 설계에 반영하고 설계 방향의 타당성을 사전에 점검할 수 있다. 개발 단계에서는 프로토타입 또는 베타 버전의 챗봇을 평가하여 서비스 완성도에 영향을 미치는 상호작용 품질, 사용자 편의성, 피드백 적합성 등을 정량적으로 검토하고 개선 방향을 도출할 수 있다. 운영 단계에서는 상용화된 챗봇 서비스를 비교 평가함으로써 사용자 경험 기반의 서비스 품질을 분석하고, 경쟁 서비스 대비 차별화 요소를 도출하는 데 활용될 수 있다.

한편, 교육용 챗봇의 사용자 경험 평가는 학습자의 언어 수준, 학습 목적 등 사용자 특성에 따라 상호작용 방식과 기대 수준이 달라질 수 있으므로, 충분한 대화 시나리오 구성과 사용자 특성 반영이 요구된다. 이는 평가 결과의 신뢰도를 확보하고 지표의 적용 범위를 확장하는 데 중요하다.

7. 2. 사용자 경험 평가 지표 활용: 디자인 가이드라인 수립

본 연구는 앞서 개발된 사용자 경험 평가 지표를 활용하여, 다양한 교육 분야에서 적용 가능한 LLM 기반 챗봇의 디자인 가이드라인을 수립하였다. 〈Table 8〉은 사용성(usability) 카테고리에 포함되는 주지표 4개와 상세지표 21개를 기반으로 한 디자인 솔루션을 제시한다. 본 연구에서 제시한 가이드라인은, 서비스 설계 과정에서 실제 구현이 가능한 요소와 밀접하게 연관된 사용성(usability) 지표를 중심으로 작성되었다.

Table 8 Design Guidelines Based on User Experience scale for Educational Chatbots based on LLM

사용성	
효과적인 전달성	**교육용 챗봇 서비스가 제공하는 정보가 의미적/시각적으로 명확하게 전달되어야 함**
명료성	• 챗봇이 문장을 간결하게 구사하고 문단의 구분을 명확하게 해야 함
투명성	• 챗봇이 사실에 기반한 내용을 전달할 때 정확한 출처를 제공함 • 나의 학습 정보가 어떻게 쓰이고 있는지 설명함
맥락 적합성	• 챗봇이 사용자와 나눴던 이전의 답변을 기억하고 맥락에 맞는 답변을 제공함
최신성	• 챗봇이 정치, 경제, 사회, 문화, 라이프, 스포츠 등의 최신 정보를 반영한 답변을 제공함
가시성	• 챗봇의 음성 제어 버튼, 해석 제공 버튼 등 사용자가 그 의미를 명확하게 알고 사용할 수 있는 디자인을 제공함
직관성	• 챗봇이 응답을 생성하는 것을 사용자가 즉각적으로 인지할 수 있도록 모션을 제공함 • 챗봇이 사용자가 음성으로 대답하는 경우 말의 시작과 마침을 인식할 수 있는 모션을 제공함
신뢰 가능한 상호작용	**교육용 챗봇 서비스가 안전하고 신뢰할 만한 상호작용을 제공해야 함**
환경적 접근성	• 챗봇이 사용자와의 대화 내용을 수집하고 있다면 사전에 고지하고 동의를 받아야 함
인지적 접근성	• 대화 중 사용자가 대화 종료를 시도했을 때, 대화 종료/저장 여부를 확인하는 메시지를 보냄
프라이버시 보호	• 챗봇이 윤리에 어긋나는 발언을 하는 경우를 대비하여 발화 전 자체 검열 시스템을 제공함 • 챗봇이 윤리에 어긋나는 발언을 하는 경우 사용자가 신고할 수 있는 기능을 제공함
오류 관리	• 챗봇이 사용자가 이동 중이거나 공공장소에 있을 때 주변 소음으로 인해 음성으로 대답하는 데 어려움이 없도록 대화 방식에 다양한 옵션을 제공해야 함
윤리성	• 대화 시작 전 사용자의 교육 수준에 대한 테스트를 진행함 • 대화 중 사용자가 직접 챗봇 수준을 조정할 수 있는 기능을 제공함 • 대화 중 챗봇이 사용자의 수준을 고려하여 직접 난이도를 조정하는 기능을 제공함

사회적 실재감		**교육용 챗봇 서비스가 사용자에게 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용해야 함**
	적응성	• 챗봇이 사용자의 대화 흐름에 적절한 대답을 하지 못하는 경우 해당 답변을 평가할 수 있는 기능을 제공함
	공감	• 챗봇이 사용자의 기분을 파악하여 사용자 질문에 대답 전, 적시 적소에 공감하는 답변을 제공함
	인간다운 자연스러움	• 챗봇이 사용자가 원하는 상황에 맞게 적절한 페르소나를 제공함 • 대화 시작 전 사용자가 원하는 페르소나를 직접 설정할 수 있는 기능을 제공함
교육적 상호작용성		**교육용 챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공해야 함**
	개인 맞춤성	• 사용자의 교육 수준에 맞는 과정을 적절하게 제공함 • 사용자가 직접 자신의 교육 수준에 맞는 과정을 선택할 수 있는 기능을 제공함
	학습 정보 제시성	• 챗봇과의 대화 전·중·후에 대한 절차를 사전에 고지함 • 사용자가 대화 진행 상황을 실시간으로 확인할 수 있도록 명시하는 기능을 제공함 • 챗봇이 어떤 프로세스를 통해서 나의 학습을 분석하고 있는지 고지함
	피드백	• 챗봇이 사용자 응답의 의미적/문법적 오류 등을 파악하여 적절한 피드백을 제공함
	몰입성	• 사용자가 교육(학습) 상황에 몰입할 수 있도록 대화 중 실제 사람 같은 아바타 화면을 제공함
	학습 동기부여	• 사용자가 챗봇을 통한 학습 과정에서 의욕을 느낄 수 있도록 출석 도장 혹은 학습 완료 배지와 같은 교육(학습) 동기부여 요소를 제공함
	자기 주도성	• 챗봇이 지정된 시간에 교육(학습)할 수 있도록 알림을 제공함 • 사용자가 교육(학습)을 진행한 내용과 시간을 스스로 관리할 수 있는 페이지를 마련함
	이해 가능성	• 챗봇이 교육(학습) 목적을 명확하게 인식하고 정보를 제공할 수 있도록 구성해야 함 • 교육 콘텐츠 자체나 학습 방식을 사용자가 잘 이해할 수 있도록 챗봇이 잘 설명해 주도록 함 • 매뉴얼을 숙지하지 않아도 잘 이해할 수 있는 익숙한 인터페이스를 제공함

〈Table 8〉의 주요 지침 중 교육 특화 지표에 대한 구체적인 설계 예시는 다음과 같다.

- **개인 맞춤성**: 초보자에게는 단문 중심 예시 제공, 상급자에게는 고급어휘 추천 및 긴 문장 구성 힌트를 제공하는 방식.
- **학습 정보 제시성**: 화면 상단에 진도 바를 표시하거나, "지금은 2단계 '표현 확장' 단계입니다"와 같이 현재 수행 단계를 명시하는 방식.
- **피드백**: 정답 제시뿐 아니라 수정된 문장, 오류 설명, 대체 표현까지 함께 제시.
- **몰입성**: AI 아바타가 눈맞춤·표정 변화 등을 제공하고, 실시간 반응과 함께 발화 속도·억양을 평가하는 방식.
- **학습 동기부여**: 출석 스탬프, 학습 배지, 누적 점수 기록, 목표 달성 시 축하 메시지 제공.
- **자기 주도성**: 사용자가 정한 시간에 학습 알림 제공, "지난 학습에서 어려웠던 부분 다시 볼까요?" 등의 적시 안내.
- **이해 가능성**: 대화 시작 시 "오늘은 ordering food 상황을 연습합니다"와 같이 목표 명확화.

이와 같은 디자인 가이드라인과 솔루션은, 사용자 경험 평가 지표를 기반으로 LLM 기반 교육용 챗봇 설계 과정에서 구체적인 설계 원칙과 실천적 전략을 제공함으로써, 실제 서비스 개발에 적용 가능성을 높이는 데 의의가 있다.

8. 결론

8. 1. 연구 요약

본 연구는 LLM 기술의 확산에 따라 교육용 챗봇의 사용자 경험을 평가하기 위한 특화된 지표 체계를 개발하는 것을 목적으로 수행되었다. 이를 위해 기존 AI 기반 평가 지표와 함께 LLM 특화 지표 및 교육 분야 특화 지표를 수집·통합하고, 통계적 검증을 통해 최종 사용자 경험 평가 지표 체계를 도출하였다. 도출된 지표는 실제 영어 회화 앱 서비스(스픽, 프랙티카, 멤라이즈)에 적용되었으며, 서비스 간 사용자 경험 차이를 효과적으로 판별할

수 있음을 확인하였다. 또한 본 연구는 평가 지표를 기반으로 디자인 가이드라인을 제안함으로써, 서비스 기획 및 설계 과정에서 사용자 경험 요소를 반영할 수 있는 실질적인 적용이 가능하도록 하였다.

8. 2. 연구 가치 및 기여

본 연구는 다음과 같은 학문적 및 실무적 기여를 갖는다. 먼저 학문적 측면에서 첫째, 최신 LLM 기술을 반영한 사용자 경험 평가 지표의 학문적 경향과 흐름을 체계적으로 반영하였다. 둘째, 통계적으로 검증된 지표 체계를 기반으로, 일반 UX 관점, 일반 AI 지표, LLM 특화 지표, 교육 분야 특화 지표를 구분하고 구조화하였다. 셋째, SOR 이론을 적용하여 인지적·정서적·행동적 차원의 위계를 설정함으로써, 사용성(Usability)을 넘어 사용자 가치(User Value)와 사용자 수용도(User Acceptance)를 포함하는 다차원적 평가 체계를 제시하였다.

실무적 측면에서는 첫째, 실제 영어 회화 앱 서비스 사례를 통해 지표 체계의 유의미성을 검증하고, 서비스 간 실질적 차이를 확인함으로써 현장 적용 가능성을 입증하였다. 둘째, 개발된 지표 체계는 기존 출시 서비스의 UX 품질 개선뿐만 아니라, 서비스 출시 전 테스트 단계에서 진단 도구로 활용할 수 있다. 셋째, 디자인 가이드라인 형태로 제시되어 서비스 기획 및 개발 초기 단계에서 디자이너가 참고할 수 있는 실무적 자료로 활용 가능하다.

8. 3. 연구 한계 및 향후 연구 방안

본 연구는 영어 회화 중심의 교육 서비스 사례에 한정하여 지표 적용 결과를 검증하였으며, 제시된 디자인 가이드라인 또한 해당 도메인을 기준으로 도출되었다는 점에서 적용 범위에 한계가 있다. 또한 사용자 경험 평가 참여자의 표본 규모가 제한적(10명)으로 구성되었으므로, 결과 해석에 있어 신뢰도와 일반성을 확보하기 위해서는 표본 확장이 요구된다.

향후 연구에서는 평가 대상 서비스의 범위를 확대하고, 표본 규모 및 참여자 특성을 다양화하여 지표 체계의 일반성과 신뢰성을 추가적으로 검증할 필요가 있다. 이를 통해 본 연구에서 제안한 사용자 경험 평가 지표 체계가 다양한 LLM 기반 교육 서비스 환경에서 실증적 평가 도구로 활용될 수 있을 것으로 기대된다. 나아가 실제 서비스 설계·개선 사례에 적용하여 산업적 확장성과 실무적 활용성을 검증한다면, 지표 체계의 적용 가능성과 기여도를 더욱 강화할 수 있을 것이다.

References

1. Choi, S. (2021). Artificial Intelligence in Education: A Literature Review on Education Using Artificial Intelligence. *The Journal of Korean Association of Computer Education, 24*(3), 11-21. https://doi.org/10.32431/kace.2021.24.3.002
2. Choi, S. (2023). Chat GPT Catch Up… Fast IT giants [all things on ChatGPT]. *hankyungBUSINESS*. https://magazine.hankyung.com/business/article/202302082148b
3. Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79-102. https://doi.org/10.1016/j.ijhcs.2005.06.002
4. Hwang, H. (2021). Development of Chatbot for Elementary Social Studies Micro-learning. *Journal of social studies education, 60*(3), 81-104. https://doi.org/10.37561/sse.2021.09.60.3.81
5. Jo, Y. (2023). Super AI and Generative Artificial Intelligence. *ICT Standard Weekly, 1145*, 1-9. http://weekly.tta.or.kr/weekly/files/20232901012950_weekly.pdf
6. Kim, B. (2015). *Development and Validation of the Evaluation Indicators for Teaching Competency of STEAM Education in the Secondary School* (Doctoral Dissertation). Korea National University of Education. http://www.riss.kr/link?id=T13863664
7. Kim, G., Yoon, K., Kim, Y., Ryu, J., & Kim, S. (2024). Technical Trends in On-device Small Language Model Technology Development. *Electronics and Telecommunications Trends, 39*(4), 82-92. https://doi.org/10.22648/ETRI.2024.J.390409

8. Kim, J. (2024). The Beginning, Development Process, and Future Prospects of AI. SK Management and Economic Research Institute. https://news.skhynix.co.kr/all-around-ai-1/
9. Kim, S., & Lee, S. (2024). The Impact of Generative AI's Technical Characteristics and Librarians' Personal Traits on Intention to Use Generative AI. *Korean Biblia Society for Library and Information Science, 35*(2), 109-133. https://doi.org/10.14699/kbiblia.2024.35.2.109
10. Lee, D., & Kim, S. (2012). The Development and Validation of the Evaluation Indicators for Teaching Competency of Elementary Teachers. *A Study on the Evaluation of Education, 25*(4), 581-604. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001723092
11. Lee, H., Sung, C., & Jeon, B. (2023). GenAI(Generative Artificial Intelligence) Technology Trend Analysis Using Bigkinds: ChatGPT Emergence and Startup Impact Assessment. *Asia-Pacific Journal of Business Venturing and Entrepreneurship, 18*(4), 65-76. https://doi.org/10.16972/apjbve.18.4.202308.65
12. Lee, J. (2020). Crisis and Opportunities in Higher Education Stimulated by Edutech. *Korea Business Review, 24*(New Year's special issue), 151-171. https://doi.org/10.17287/kbr.2020.24.0.151
13. Lee, J. (2021). Trends in Artificial Intelligence (AI) and Data Utilization in Education. *Proceeding of KISDI AI Outlook, 21*(5), 39-51. https://library.kisdi.re.kr/%24/10160/contents/4333446?checkinId=1401349&articleId=626257
14. Lee, H., & Kim, J. (2023). Effects of UTAUT on the Digital Literacy and Acceptance Intention of ChatGPT Users. *The Society of Convergence Knowledge Transactions, 11*(2), 33-43. https://doi.org/10.22716/sckt.2023.11.2.014
15. Oermann, E., & Kondziolka, D. (2023). On Chatbots and Generative Artificial Intelligence. *Neurosurgery, 92*(4), 665-666. https://doi.org/10.1227/neu.0000000000002415
16. Park, D. (2023). Journalism Artificial Intelligence Based on Trustworthy Artificial Intelligence: Toward a Commensurability between Media Trust and Trustworthiness of Artificial Intelligence System. *Journal of mediasociety, 31*(4), 5-47. https://doi.org/10.52874/medsoc.2023.11.31.4.5
17. Park, J., & Gil, J. (2020). Edutech in the Era of the 4th Industrial Revolution. *Journal of KTSDE, 9*(11), 329-331. https://doi.org/10.3745/KTSDE.2020.9.11.329
18. Yang, J., Wang, Z., Lin, Y., & Zhao, Z. (2024). Global Data Constraints: Ethical and Effectiveness Challenges in Large Language Model. *arXiv e-prints*, arXiv-2406. https://arxiv.org/html/2406.11214v1
19. Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *Npj mental health research, 3*(1), 12. https://doi.org/10.1038/s44184-024-00056-z

Appendix 1. The process and results of deriving education-specific indicators

최종 지표	초기 지표 후보군	논문명	년도	교육학 이론
개인맞춤성	personalized	Effects of Generative Chatbots in Higher Education	2023	SDT
	Personalized Experience	OpineBot: Class Feedback Reimagined Using a Conversational LLM	2023	
	Learning materials	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	
학습정보 제시성	목표 설정 및 학습 진행 상황 모니터링	개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석	2024	SRL
	도움말 및 문서화	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	
피드백	시스템의 가독성	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	SRL
	피드백, Q&A 제공	개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석	2024	
	Course-Specific Feedback	OpineBot: Class Feedback Reimagined Using a Conversational LLM	2023	
몰입성	행동적 몰입	대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석	2023	SDT
	정서적 몰입	대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석	2023	
	인지적 몰입	대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석	2023	
학습 동기 부여	학습 동기	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	SDT
자기 주도성	self-directed	Effects of Generative Chatbots in Higher Education	2023	SRL
이해 가능성	상황 이해	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	SRL
	Can detect meaning and intent	개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석	2024	
학습 효과성	usefulness	AI literacy for ethical use of chatbot: Will students accept AI ethics?	2024	SRL
학습 효율성	사용의 유연성과 효율성	사용자 친화적인 챗봇 튜터 설계 지침 개발 연구	2023	SRL
학습 신뢰성	(LLM지표 후보군에서 도출되었지만 FGI 및 교수급 전문가와 협의 후 교육학적 지표로 지표명 수정함)			
자기효능감	수행 기대감	대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석	2023	SDT

Appendix 2. Friedman U Test Results by Application Service

지표	통계항목	대상 애플리케이션		
		스픽 (Speak)	프랙티카 (Praktika)	멤라이즈 (Memrise)
효과적 전달성		4.4(0.84)	4.4(0.51)	3.9(0.87)
명료성	평균(표준편차)	4.7(0.48)	4.3(0.48)	3.8(1.03)
	평균 순위	2.50	2.00	1.50
	유의확률		0.013*	
투명성	평균(표준편차)	4.6(0.51)	4.5(0.52)	4.1(0.87)
	평균 순위	2.25	2.15	1.60
	유의확률		0.023*	
맥락 적합성	평균(표준편차)	4.6(0.51)	4.5(0.97)	4.0(0.94)
	평균 순위	2.25	2.30	1.45
	유의확률		0.019*	
최신성	평균(표준편차)	3.7(1.05)	4.4(0.51)	3.4(1.07)
	평균 순위	1.85	2.55	1.60
	유의확률		0.018*	
가시성	평균(표준편차)	3.9(1.10)	4.6(0.69)	3.9(1.10)
	평균 순위	1.80	2.50	1.70
	유의확률		0.032*	
직관성	평균(표준편차)	3.9(1.10)	4.4(0.84)	4.1(1.10)
	평균 순위	1.80	2.25	1.95
	유의확률		0.401	
신뢰 가능한 상호작용		4.3(0.67)	4.4(0.51)	3.9(0.99)
환경적 접근성	평균(표준편차)	3.8(0.91)	4.1(1.19)	3.9(1.10)
	평균 순위	1.90	2.15	1.95
	유의확률		0.764	
인지적 접근성	평균(표준편차)	3.9(0.87)	4.1(1.19)	4.0(0.81)
	평균 순위	1.85	2.20	1.95
	유의확률		0.444	
프라이버시 보호	평균(표준편차)	3.8(0.78)	4.0(0.81)	3.6(0.96)
	평균 순위	1.95	2.30	1.75
	유의확률		0.144	
오류 관리	평균(표준편차)	3.8(0.91)	3.9(0.73)	3.2(1.22)
	평균 순위	2.15	2.25	1.60
	유의확률		0.163	
윤리성	평균(표준편차)	4.3(0.94)	4.3(0.67)	3.8(1.03)
	평균 순위	2.15	2.15	1.70
	유의확률		0.259	
사회적 실재감		4.0(0.94)	4.6(0.69)	3.1(0.99)
적응성	평균(표준편차)	4.6(0.69)	4.7(0.48)	3.9(1.19)
	평균 순위	2.30	2.20	1.50
	유의확률		0.022*	
공감	평균(표준편차)	4.1(0.87)	4.6(0.51)	3.1(1.28)
	평균 순위	2.15	2.55	1.30
	유의확률		0.008*	
인간다운 자연스러움	평균(표준편차)	4.3(0.48)	4.3(0.94)	2.8(1.13)
	평균 순위	2.40	2.40	1.20
	유의확률		0.004*	

교육적 상호작용성		4.0(0.81)	4.6(0.69)	3.5(1.17)
개인 맞춤성	평균(표준편차)	4.0(0.66)	4.8(0.63)	3.1(1.19)
	평균 순위	2.00	2.75	1.25
	유의확률		〈 0.001	
학습 정보 제시성	평균(표준편차)	3.5(1.43)	4.1(0.99)	3.7(1.05)
	평균 순위	1.75	2.30	1.95
	유의확률		0.331	
피드백	평균(표준편차)	4.2(1.03)	4.3(0.94)	3.2(1.54)
	평균 순위	2.25	2.25	1.50
	유의확률		0.069	
몰입성	평균(표준편차)	4.0(1.15)	4.7(0.67)	3.2(1.54)
	평균 순위	2.00	2.50	1.50
	유의확률		0.032*	
학습 동기부여	평균(표준편차)	3.3(1.05)	4.5(0.70)	3.2(1.31)
	평균 순위	1.85	2.60	1.55
	유의확률		0.026*	
자기 주도성	평균(표준편차)	3.9(1.10)	4.5(0.70)	3.3(1.33)
	평균 순위	2.00	2.50	1.50
	유의확률		0.028*	
이해 가능성	평균(표준편차)	4.4(0.51)	4.6(0.51)	3.8(1.22)
	평균 순위	2.05	2.35	1.60
	유의확률		0.022*	
기능적 가치		4.4(0.69)	4.5(0.52)	3.4(1.17)
학습 효과성	평균(표준편차)	4.3(0.94)	4.6(0.69)	3.2(1.31)
	평균 순위	2.25	2.45	1.30
	유의확률		0.002*	
학습 효율성	평균(표준편차)	4.2(1.03)	4.6(0.69)	3.3(1.33)
	평균 순위	2.2.	2.45	1.35
	유의확률		0.007*	
학습 신뢰성	평균(표준편차)	4.3(0.82)	4.5(0.70)	3.7(1.48)
	평균 순위	2.05	2.25	1.70
	유의확률		0.161	
학습 흥미성	평균(표준편차)	4.1(0.99)	4.6(0.69)	3.1(1.59)
	평균 순위	2.05	2.50	1.45
	유의확률		0.016*	
감정적 가치		3.4(1.26)	4.1(0.73)	2.4(1.07)
심미성	평균(표준편차)	3.3(1.25)	4.0(0.81)	2.5(1.26)
	평균 순위	2.10	2.60	1.30
	유의확률		0.004*	
친밀감	평균(표준편차)	3.1(1.28)	4.2(1.03)	2.1(1.10)
	평균 순위	2.00	2.70	1.30
	유의확률		0.002*	
자기효능감	평균(표준편차)	3.6(1.34)	4.3(0.82)	3.1(1.19)
	평균 순위	1.90	2.55	1.55
	유의확률		0.029*	
만족도	평균(표준편차)	4.0(0.81)	4.3(0.48)	3.0(1.05)
태도	평균(표준편차)	4.5(0.70)	4.7(0.48)	3.2(1.13)
지속 사용 의도	평균(표준편차)	4.2(1.03)	4.4(0.84)	2.7(1.33)

LLM기반 교육용 챗봇의 사용자 경험 평가 지표 개발 및 적용: 영어 회화 앱 서비스를 중심으로

김지효[1], 강효진[2*]

[1]성신여자대학교 미래융합기술공학과, 석사졸업생, 서울, 대한민국

[2]성신여자대학교 서비스디자인공학과, 부교수, 서울, 대한민국

초록

연구배경 생성형 AI 확산으로 LLM 기반 챗봇은 교육 환경에서 질의응답, 맞춤형 안내, 자동 피드백 등 학습 지원 기능을 수행하고 있다. 그러나 기존 사용성 평가는 기능적 편의성에 치중되어, 맥락 유지, 개인화, 피드백 품질 등 LLM 기반 상호작용 특성을 충분히 반영하지 못한다. 이에 본 연구는 이러한 특성을 고려한 사용자 경험 평가 지표 체계를 개발하고자 한다.

연구방법 본 연구는 지표 개발-검증-적용-활용 네 단계로 수행되었다. 먼저 문헌 조사와 연쇄 표집을 통해 지표 후보군을 도출하고, 어피니티 다이어그램과 전문가 검토를 통해 구조를 확정하였다. 이후 설문 기반 EFA로 타당성을 검증하였으며, LLM 기반 교육용 애플리케이션 평가에 적용해 활용 가능성을 확인하였다. 마지막으로 확정된 지표는 서비스 설계와 평가에 적용할 수 있는 가이드라인으로 제시하였다.

연구결과 본 연구는 문헌 조사와 전문가 검토를 통해 LLM 기반 교육용 챗봇의 사용자 경험 평가 지표 체계를 도출하였다. 최종 체계는 사용성(Usability), 사용자 가치(User Value), 사용자 수용도(User Acceptance)의 3개 범주로 구성되며, 총 9개의 주지표와 28개의 상세지표가 확정되었다. 또한 각 지표별 정의와 핵심 특징을 정리하여 설계 및 평가 과정에서 활용 가능한 기준으로 제시하였다.

결론 본 연구는 LLM 기반 교육용 챗봇의 사용자 경험을 평가하기 위한 지표 체계를 제안함으로써 기존 기능 중심 평가의 한계를 보완하였다. 제안된 지표는 학습 맥락을 반영하여 서비스 간 경험 차이를 식별하였으며, 설계·개발·운영 단계에서 활용 가능함을 확인하였다. 또한 가이드라인을 통해 서비스 품질 개선에 적용할 수 있는 실무적 근거를 제공하였다.

주제어 사용자 경험 평가 지표, LLM, 생성형AI, 교육용 챗봇, 디자인 가이드라인, 대화형 인터페이스

*교신저자 : 강효진 (hjkang@sungshin.ac.kr)