강 효 진 교수 지도

석사학위 청구논문

LLM 기반 교육용 챗봇의 사용성 평가 지표 개발 및 적용

– 영어 회화 앱 서비스를 중심으로 –

2025

성신여자대학교 대학원

미래융합기술공학과

김 지 효

LLM 기반 교육용 챗봇의 사용성 평가 지표 개발 및 적용

- 영어 회화 앱 서비스를 중심으로 -

강 효 진 교수 지도

이 논문을 석사 학위 논문으로 제출함

2024년 11월

성신여자대학교 대학원
미래융합기술공학과
김 지 효

인 준 서

김지효의 석사학위 논문으로 인준함

2025년 1월

심사위원장 이 여 름 (서명 또는 인)

심 사 위 원 강 효 진 (서명 또는 인)

심 사 위 원 김 환 (서명 또는 인)

성신여자대학교 대학원

논 문 개 요

본 연구는 대형 언어 모델(LLM)을 기반으로 한 교육용 챗봇 서비스의 사용성 평가 지표를 개발하고, 이를 검증하여 실무에 적용할 수 있는 디자인 가이드라인을 제안하는 데 목적이 있다. 에듀테크(EdTech) 시장이 급속히 성장하고 있으며, LLM 기반 챗봇은 학습자에게 맞춤형 학습 경험을 제공하며 기존 교육 방식에 혁신을 가져왔다. 그러나, LLM 기반 교육용 챗봇의 특성을 반영한 신뢰할 수 있는 사용성 평가 지표에 관한 연구는 미흡한 실정이다.

이에 본 연구는 네 가지 주요 연구 질문에 따라, 기존 AI 기반 챗봇의 사용성 평가 지표를 분석하고, LLM 및 교육 분야에 특화된 평가 요소를 통합한 새로운 사용성 평가 지표 체계를 제안하였다. 연구는 선행 연구 분석, 어피니티 다이어그램, 전문가 그룹 검토, 설문 조사 및 탐색적 요인분석을 통해 진행되었으며, 그 결과 LLM 기반 교육용 챗봇의 사용성 평가 지표는 기존 AI 기반 평가 지표와 차별화된 요소를 포함하였다.

제안된 사용성 평가 지표 체계는 LLM 기반 교육용 챗봇 서비스 설계 및 개발 단계에서 활용될 수 있도록 디자인 가이드라인 형태로 제시되었으며, 교육 서비스 품질 개선과 사용자 경험 향상에 기여할 수 있는 실질적인 도구를 제공한다. 이러한 연구 결과는 향후 LLM 기술 기반의 다양한 산업 분야로의 확장 가능성을 시사하며, 연구 및 실무 적용에 중요한 참고 자료로 활용될 것이다.

목 차

그 림 목 차

표 목 차

Ⅰ. 서 론

1. 연구 배경 및 목적

4차 산업 사회로 접어들면서, 교육 분야에서는 교육(Education)과 기술(Technology)을 결합한 '에듀테크(EdTech)'가 새로운 트랜드로 자리잡고 있다. 에듀테크는 시간과 공간의 제약을 극복하기 위해 정보 통신 기술을 활용하여 교육 콘텐츠를 전달하는 수준을 넘어, 4차 산업혁명의 핵심 기술인 인공지능(AI), 클라우드(Cloud),사물 인터넷(IoT), 모바일(Mobile),빅데이터(Big Data) 등을 교육에 적극적으로 활용하여 학생들에게 수준별 맞춤형 교육을 제공하는 것을 의미한다.[1)]

특히 에듀테크의 도입은 학습자의 적응형 학습 지원 및 지능형 교육행정 등 다양한 방식으로 교육 혁신을 이끌고 있다. 에듀테크는 지능정보기술을 활용해 기존 교육 방식에 혁신을 가져왔으며, 교육 서비스 및 출판 서비스 기업을 중심으로 빠르게 성장하고 있다.[2)] 코로나19 펜데믹 장기화로 교육 분야에서는 온라인 수업 활성화 등 비대면 서비스로의 패러다임 전환이 가속화되었으며, 이에 따라 에듀테크에 대한 민간 투자와 정책적 지원도 확대되었다.[3)] 교육부에 따르면 국내 에듀테크 시장 규모는 2020년 6조 5605억원에서 2023년에 8조 5140억원으로 성장했으며, 2026년 11조원에 이를 것이라고 전망된다.[4)] 에듀테

---

1) TTA정보통신용어사전. https://terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=193122-1
2) 인공지능 챗봇 트렌드 2021: 산업 별 전망 (교육). https://tonyaround.com/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5-%EC%B1%97%EB%B4%87-%ED%8A%B8%EB%A0%8C%EB%93%9C-2021-%EC%82%B0%EC%97%85-%EB%B3%84-%EC%A0%84%EB%A7%9D-%EA%B5%90%EC%9C%A1/
3) 동아일보(2020). 비대면 교육서비스 수요 증가 대학가에 AI 챗봇 '현명한 앤써니' 보급
4) 교육부(2023). 공교육과 기술이 함께 발전하는 '교육 정보 기술(에듀테크)' 시대 열린다.

크 시장 내에서는 학습자의 감성적 지도를 위한 수단으로 인공지능 챗봇의 활용이 두드러지고있다. 단순한 화면상의 설명이나 인터페이스만으로는 학생들의 성취도와 만족도를 높이기 어려운 한계가 있었으나, 대화 방식의 챗봇은 학습 능력 향상에 기여하고 있다.

챗봇은 인공지능과 자연어 처리 기술 등의 발전에 따라 다양한 분야에서 활용이 확장되고 있다. 2022년 11월, OpenAI는 대형 언어 모델(Large Language Model, 이하 LLM) GPT-3.5를 기반으로 한 ChatGPT를 공개하였으며, 공개 5일만에 사용자 100만 명을 달성하는 등 폭발적인 관심을 받았다.[5] LLM은 텍스트 뿐만 아니라 이미지, 오디오 등의 다양한 유형의 데이터를 동시에 처리할 수 있는 딥러닝 알고리즘으로[6], ChatGPT와 같은도구의 등장은 교육 연구에서 새로운 교육적 의미와 과제, 기회를 제시하여 대중과 학계의 주목을 받고 있다.[7]

특히 LLM과 같은 언어 모델 기반 챗봇 기술을 효과적이고 안전하게 사용하기 위해서는 인간의 면밀한 감독과 제어가 필요하며, 이를 뒷받침 하기위한 이론적 연구가 요구된다.[8] 이러한 배경에서 기존 사용성 평가 방법은 기술 발전과 새로운 사용 맥락에 맞춰 보완되고 대체되고 있으며,[9] 사용성 평가 지표에 대한 연구는 지속적으로 발전하고 있다. 그러나 LLM기반 교육용 챗봇에 특화된 사용성 평가에 대한 연구는 여전히 부족하다.

본 연구는 LLM이 도입된 교육용 챗봇이 기존 AI기반 챗봇과 차별화되는 사

---

5) Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S. C., Klein, J., & Bissyandé, T. F. (2023). Is ChatGPT the Ultimate Programming Assistant—How far is it?. arXiv preprint arXiv:2304.11938.

6) 김지현(2024). AI의 시작과 발전 과정, 미래의 전망.

7) Li, W., Zhang, X., Li, J., Yang, X., Li, D., & Liu, Y. (2024). An explanatory study of factors influencing engagement in AI education at the K-12 Level: an extension of the classic TAM model.Scientific Reports, 14.

8) Oermann, E. K., & Kondziolka, D. (2023). On Chatbots and Generative Artificial Intelligence.Neurosurgery,92(4), 665-666. https://doi.org/10.1227/neu.0000000000002415

9) Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research.International journal of human-computer studies,64(2), 79-102.

용성 평가 지표 특성을 분석하고, 이를 바탕으로 사용성 평가 지표 체계를 개발하는 것을 목적으로 한다. 이를 위해 우선 사용성 평가에 대한 이해를 바탕으로 LLM, 교육분야 특화 요소를 구분한 사용성 평가 지표 체계를 도출한다. 이후 사용성 평가 지표를 활용하여 LLM기반 교육용 챗봇 개발 단계에서 활용될 수 있는 디자인 가이드라인을 제안한다.

2. 연구 범위 및 절차

본 연구는 LLM이 도입된 영어 교육용 챗봇의 사용성 평가 지표를 개발하고 이를 검증하며, 해당 지표를 활용한 LLM이 도입된 영어 교육용 챗봇 서비스의 디자인 가이드라인을 제안하는 것을 최종 목표로 한다. 이러한 목적을 달성하기 위한 연구 질문은 다음과 같다.

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

연구질문 2. LLM 기반 챗봇 서비스의 사용성 평가 지표 중 'LLM'과 '교육 분야'에 특화된 사용성 평가 지표 요인은 어떤 것이 있는가?

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

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

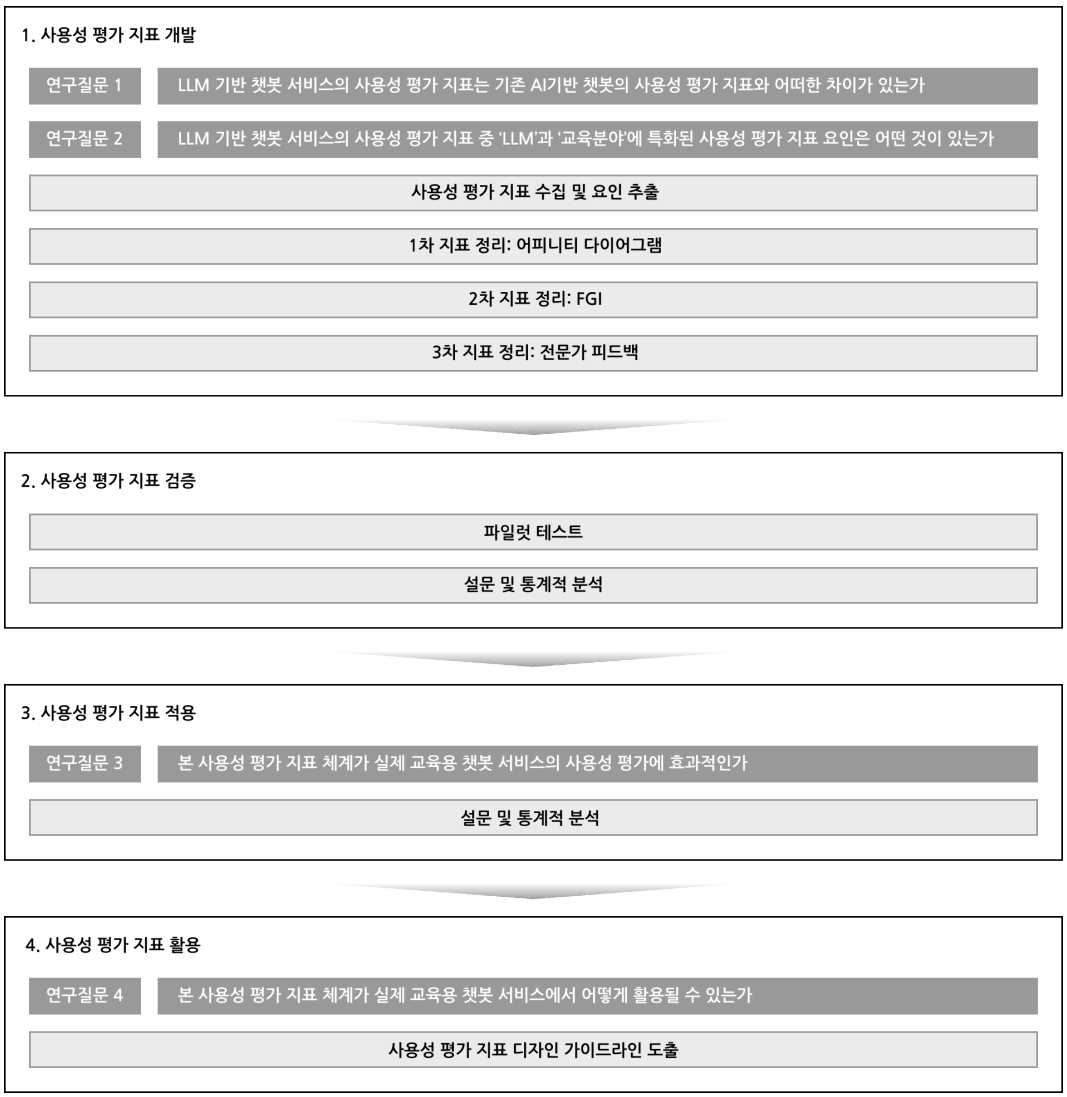
연구 질문을 구체화하기 위한 연구 범위 및 절차는 다음 【그림 1-1】과 같다. 첫째, 사용성 평가 지표 개발 단계에서는 LLM이 도입되기 전 기존 챗봇의 사용성 평가 지표와 LLM이 도입된 사용성 평가 지표에 대한 이론적 고찰을 위하여 국내외 문헌 조사를 수행한다. 수집된 결과를 바탕으로 AI 챗봇의 특성, LLM의 특성, 교육 분야의 특성을 각각 구분하여 비교 분석하고, 이를 통해 LLM 및 교육 분야에 특화된 사용성 평가 지표를 추출한다. 사용성 평가

지표를 체계적으로 확립하기 위해 (1) 연구자의 어피니티 다이어그래밍, (2) FGI를 활용한 그룹 어피니티 다이어그래밍 워크숍, (3) 전문가 정성평가를 포함한 3단계 절차를 거쳐 지표를 정리한다.

둘째, 사용성 평가 지표 검증 단계는 개발된 사용성 평가 지표의 타당성을 확인하는 단계이다. 먼저 본 설문 전 서비스디자인 분야 연구원 대상으로 평가 항목의 난이도와 적합성 등 설문지를 최종적으로 구성하기 위해 파일럿 테스트를 진행한다. 이후 설문 문항 별 적합한 상세 질문을 도출하여 구성된 최종 설문지를 바탕으로 본 설문을 진행한다. 본 설문은 LLM이 도입된 교육용 챗봇 중 영어 회화 챗봇에 한정하여 사용 경험이 있는 사용자를 대상으로 진행한다. 수집된 설문 데이터를 바탕으로 통계 소프트웨어(SPSS)를 활용한 탐색적 요인 분석을 수행하여 지표의 신뢰성과 타당성을 검증한다.

셋째, 사용성 평가 지표 활용 단계에서는 검증된 사용성 평가 지표의 유효성을 확인하기 위해 실제 사례 연구를 진행한다. 이를 위해 기존에 사용되고 있는 LLM 기반 영어 교육용 애플리케이션 3종을 선정하고, 서비스디자인 관련 분야 연구원들에게 스마트폰 애플리케이션을 활용한 태스크를 태스크를 수행하도록 요청한다. 이후 검증된 사용성 평가 지표가 담긴 설문지를 통해 각 애플리케이션의 사용성을 평가한다.

넷째, 사용성 평가 지표 활용 단계에서는 LLM 기반 교육용 사용성 평가 지표 개발, 검증, 적용 단계를 거친 사용성 평가 지표의 디자인 가이드라인을 제안한다. Nielsen(1993)에 따르면, 어떤 사용성 측정을 선택할 것인가에 대한 질문을 사용자 인터페이스 설계 및 개발에 대한 많은 접근방식에서 핵심이 된다. 이에 따라 사용성 평가 지표를 기반으로 LLM 기반 교육용 챗봇 디자인 가이드라인을 구체화한다. 이러한 디자인 가이드라인은 추후 교육용 챗봇 개발하는 단계에서 참고 및 활용될 수 있다.


1. 사용성 평가 지표 개발
연구질문 1
LLM 기반 챗봇 서비스의 사용성 평가 지표는 기존 AI기반 챗봇의 사용성 평가 지표와 어떠한 차이가 있는가
연구질문 2
LLM 기반 챗봇 서비스의 사용성 평가 지표 중 'LLM'과 '교육분야'에 특화된 사용성 평가 지표 요인은 어떤 것이 있는가
사용성 평가 지표 수집 및 요인 추출
1차 지표 정리: 어피니티 다이어그램
2차 지표 정리: FGI
3차 지표 정리: 전문가 피드백
2. 사용성 평가 지표 검증
파일럿 테스트
설문 및 통계적 분석
3. 사용성 평가 지표 적용
연구질문 3
본 사용성 평가 지표 체계가 실제 교육용 챗봇 서비스의 사용성 평가에 효과적인가
설문 및 통계적 분석
4. 사용성 평가 지표 활용
연구질문 4
본 사용성 평가 지표 체계가 실제 교육용 챗봇 서비스에서 어떻게 활용될 수 있는가
사용성 평가 지표 디자인 가이드라인 도출

【그림 1-1】연구 흐름도

Ⅱ. 이론적 배경

1. LLM 기술 및 챗봇 서비스에 관한 고찰

1) LLM 기술에 관한 고찰

OpenAI, Google, Meta와 같은 글로벌 선도 기업은 LLM 연구 및 상용화에 있어 주요한 역할을 하고 있으며, 이를 통해 글로벌 시장에서 LLM 기술의 표준을 구축하고 있다.[10] OpenAI의 ChatGPT는 사용자 친화적인 대화형 AI 모델로, 교육, 고객 지원, 창의적 콘텐츠 생성 등 다양한 응용 분야에서 활용되고 있다. Google의 Bard는 구글 검색 엔진과 통합되어 정보 검색을 고도화하는 데 중점을 두고 있으며, Meta는 Llama 시리즈를 통해 연구자와 개발자에게 오픈소스 LLM 툴을 제공하고 있다.[11]

국내에서도 이러한 글로벌 추세에 발맞춰 네이버, 카카오 등 주요 IT 기업들이 LLM 기술 개발에 앞장서고 있다.[12] 네이버의 HyperClova는 한국어에 최적화된 LLM으로, 한국어 데이터 기반 학습을 통해 언어적 맥락을 세밀하게 이해하는 데 초점을 맞추고 있다. 예를 들어, HyperClova는 한국의 교육 환경, 문화적 맥락, 특정 단어의 미묘한 뉘앙스까지 반영하여 학습 콘텐츠 생성, 질의응답 시스템, 기업 내부 시스템의 자동화를 지원한다.[13] 또한, 카카오는 자체

10) 김근용, 윤기하, 김량수, 류지형, & 김성창. (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
11) 조영임. (2023). 초거대 AI 와 생성형 인공지능.ICT Standard Weekly,1145, 1-9.
12) 이현주, 성창수, & 전병훈. (2023). 빅카인즈를 활용한 GenAI (생성형 인공지능) 기술 동향 분석: ChatGPT 등장과 스타트업 영향 평가.벤처창업연구,18(4), 65-76.
13) 김성희, & 이승민. (2024). 생성형 AI 의 기술적 특성과 사서의 개인적 특성이 생성형 AI 사용 의도에 미치는 영향.한국비블리아학회지,35(2), 109-133.

AI 플랫폼인 KoGPT를 통해 한국어 및 한글 데이터를 기반으로 한 대화형 AI와 생성형 AI 기술을 개발하고 있으며, 이를 다양한 서비스에 통합하고 있다. 특히, 비즈니스 고객을 대상으로 한 챗봇, 음성 인식 및 생성, 그리고 맞춤형 콘텐츠 생성 서비스가 주목받고 있다.[14] 이와 함께 국내 스타트업들도 LLM 기술을 활용한 다양한 서비스와 솔루션을 개발하고 있다. 예를 들어, 교육 스타트업은 LLM을 활용해 학생들에게 개인화된 학습 콘텐츠를 제공하고 있으며, 금융 및 법률 스타트업은 복잡한 법률 문서 요약, 자동화된 상담 서비스 등을 통해 업무 효율성을 높이고 있다. LLM은 챗봇을 넘어 다양한 산업 및 분야에서 활용되며, 인간의 작업을 보완하고 효율성을 높이는 데 기여하고 있다.[15]

다만, LLM 기술의 적용에는 몇 가지 한계와 과제가 존재한다. 우선, 데이터 편향, 윤리적 문제, 그리고 프라이버시 우려가 지속적으로 제기되고 있다.[16] 이를 해결하기 위해 LLM 연구자들은 신뢰성 있는 데이터 수집, 비용 효율적인 학습 방법, 사용자 데이터 보호를 위한 기술 개발에 집중하고 있다. 이러한 기술 발전과 함께, 끊임없이 변화하는 사용자 요구를 충족하고 기술의 효과성을 검증하기 위해 적합한 사용성 평가 지표 세트를 활용하는 것은 필수적이다. 사용성 평가는 웹 및 앱 서비스를 비롯한 다양한 제품이 특정 요소를 얼마나 효과적으로 충족하는지 측정하는 데 중점을 둔다. 그러나 각 제품의 특성에 따라 사용성 요소의 중요도가 다르게 나타나며, 다양한 환경과 지식을 가진 사용자들로부터 엇갈린 평가가 이루어질 수 있다. 따라서, 이러한 복잡성을 해결하기 위해 체계적이고 표준화된 사용성 평가 지표의 개발 및 적용이 요구된다.

---

14) 최수진. (2023). 챗 GPT 따라잡아라… 속도 내는 IT 거인들.한경 BUSINESS, (1420), 30-31.
15) Stade, E.C., Stirman, S.W., Ungar, L.H.et al.Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation.npj Mental Health Res3, 12 (2024).
16) Yang, J., Wang, Z., Lin, Y., & Zhao, Z. (2024). Global Data Constraints: Ethical and Effectiveness Challenges in Large Language Model.arXiv preprint arXiv:2406.11214.

2) LLM 챗봇과 교육 분야 챗봇

이처럼 대형 언어 모델(LLM)은 방대한 양의 데이터를 기반으로 훈련된 자연어 처리 모델로, 인간과 유사한 수준의 언어 이해와 생성 능력을 갖추고 있다. LLM 기반 챗봇은 기존의 단순한 질의응답(Q&A) 시스템을 넘어, 사용자의 의도를 파악하고 대화의 맥락에 적합한 답변을 생성하는 데 있어 뛰어난 성능을 발휘한다.[17] 특히, 사용자와의 지속적인 대화에서 일관성을 유지하며, 자연스럽고 유창한 언어 표현을 통해 보다 인간적인 소통 경험을 제공한다는 점에서 기존 챗봇의 한계를 효과적으로 극복하고 있다.[18]

이러한 특성은 교육, 헬스케어, 금융, 콘텐츠 생성 등 다양한 산업 분야에서 응용 가능성을 넓히고 있지만, 특히 교육 분야에서의 활용 가능성이 주목받고 있다. 예를 들어, LLM 챗봇은 학습자의 질문에 맞춤형 답변을 제공하거나[19], 대화형 학습을 통해 학생들이 보다 능동적으로 학습에 참여하도록 유도한다. 또한, 학습자의 수준과 학습 목표를 기반으로 개별화된 학습 경로를 제안하여 학습 효과를 극대화할 수 있다.

LLM 기반 챗봇은 사용자 데이터를 지속적으로 학습하면서 점점 더 개인화된 대화 서비스를 제공할 수 있다. 이러한 점에서 LLM 챗봇은 단순히 사용자의 요구를 충족시키는 도구를 넘어, 사용자 경험을 혁신적으로 개선하고, 새로운 서비스 품질의 표준을 정립할 수 있는 중요한 기술로 자리 잡고 있다. 결과적으로, LLM 기반 챗봇은 사용자와의 상호작용 효율성을 높이고, 서비스의 전반적인 가치를 향상시키는 데 핵심적인 역할을 하고 있다.

---

17) 박대민. (2023). 신뢰할 수 있는 인공지능 기반의 저널리즘 인공지능: 언론 신뢰와 인공지능 신뢰성 간 통약가능성을 바탕으로.언론과 사회,31(4), 5-47.
18) 이항, & 김준환. (2023). 통합기술수용모델이 챗 GPT 이용자의 디지털리터러시와 수용의도에 미치는 영향.융복합지식학회논문지,11(2), 33-43.
19) 황홍섭. (2021). 초등 사회과 마이크로러닝을 위한 챗봇의 개발. 사회과교육, 60(3), 81-104.

2. 사용성 평가 지표에 관한 고찰

사용성은 HCI(Human Computer Interection)의 핵심 용어이다. 대화형 시스템의 사용성을 어떻게 사용하고 개선할지는 HCI의 핵심 연구 질문이며, 연구 질문을 다루는 연구는 시스템 사용성을 개선하기 위한 지침(Smith and Mosier, 1986), 사용성을 측정하는 방법(ISO,1998), 사용성 문제 예측 방법에 대한 논의(Molich and Nielsen, 1990)로 이어졌다.

사용성은 밀러(Miller, 1971)[20]가 처음으로 사용 용이성(Ease of use)을 측정하기 위한 목적으로 사용된 개념이다.[21] 이후 여러 분야에서 논의를 거쳐 확장된 의미로 발전하며 다양하게 정의되었다. 섀클(Shackel, 1981)은 '인간이 어떠한 시스템에 대해 쉽고 효율성 있게 사용할 수 있는 능력'으로 정의하였으며[22], 닐슨(Nielson, 1993)은 '인터페이스의 사용하기 쉬운 정도에 관한 품질 속성'으로 정의하였다.[23]

사용성 평가는 제품의 복잡한 기능들을 사용자가 보다 쉽게 조작할 수 있도록 인터페이스 개발 또는 개선하기 위해 문제점을 찾아내고 개선 방향에 대한 아이디어를 발굴하는 과학적 조사 과정이다.[24] 사용성 평가의 적용 범위는 전자기기, 애플리케이션 및 웹 인터페이스, 헬스케어 등 제품 및 서비스 사용과 관련된 모든 범위에 해당한다. 따라서 사용성 평가는 어떤 대상의 성격에 초점을 맞추는가에 따라 다양한 항목과 요인으로 구성된다. 이에 따라 정확한 사용성 평가를 하기 위해서는 각 제품과 서비스의 성격이나 특징에 따라 사용성 평

---

20) Miller, R.B., 1971. Human ease of use criteria and their tradeoffs. IBM Report TR 00.2185, 12 April. IBM Corporation, Poughkeepsie, NY.
21) Shackel, B. (2009). Usability－Context, framework, definition, design and evaluation.Interacting with computers,21(5-6), 339-346.
22) Shackel, B. (1981).The concept of usability(pp. 1-30). Poughkeepsie, New York: Proceedings of IBM Software and Information Usability Symposium.
23) Nielsen, J. (1993).Usability Engineering.Boston: Morgan Kaufmann.
24) 이승희, 손원준. (2022). 시니어 세대를 위한 모바일 어플리케이션에 관한 사용성 평가 연구 - 국내 유통기업 사례를 중심으로. 상품문화디자인학연구,(68), 1-12.

가 기준이 정해져야 한다.

사용성에 대한 정의는 연구자의 목적에 따라 다양하다. 다음은 전통적으로 사용되어 오고 있는 전문가 및 전문 기관의 사용성 평가 원칙이다. 【표 2-1】은 제이콥 닐슨(Nielsen, J, 1993)의 휴리스틱 평가(Heuristic Evaluation) 지표이다. 총 10개의 사용성 평가 요인, 가시성(Visibility od System Statue), 정확성(Match Between Bystem and the Real World), 통제성(User Control and Freedom), 일관성(Consistency and Standard), 오류성(Error Prevention), 효율성(Recognition Rather than Recall), 신속성(Flexibility and Effciency of Use), 심미성(Aesthetic and Minimalist Design), 역조작(Help Users Recognizw, Diagnose, and Recover from Errors), 이해성(Help and Documentation)이 있으며, 전문가와 비전문가들이 시스템 인터페이스의 UI 디자인의 단점이나 문제점을 파악하기 위해 평가 목적으로 개념화시킨 방법론이다.[25)]

【표 2-1】 제이콥 닐슨의 10가지 휴리스틱 사용성 평가 지표[26)]

평가 지표	평가 내용
Visibility od System Statue	시스템은 적절한 시간과 피드백으로 사용자에게 진행 상황을 알려줘야 한다.
Match Between Bystem and the Real World	전문용어 또는 시스템 지향 언어를 자제하며 사용자에게 친숙한 단어로 말한다.
User Control and Freedom	사용자는 자신의 실수를 금세 복구할 수 있어야 하며 대표적으로 '이전'으로 돌아갈 수 있도록 도와주는 명령어 'control+z'와 같은 비상구가 있어야 한다.
Consistency and Standard	인터페이스의 일관성(내부/외부)을 제공하고 표준화시켜야 한다.
Error Prevention	오류가 발생하기 쉬운 조건을 제거하거나, 오류를 확인하고 사용자가 작업을 수행하기 전에 확인 옵션을 제시해야 한다.
Recognition Rather than Recall	사용자가 별도 학습 또는 기억 없이 해당 기능에 대해 쉽게 인식할 수 있어야 한다.

25) 서창희. "시니어를 위한 애플리케이션의 사용성 평가지표에 관한 연구." 국내석사학위논문 상명대학교 일반대학원, 2022. 서울

26) Nielsen, J. (1994). 10 Usability Heuristics for User Interface Design [On-Line]. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/ten-usability-heuristics/

Flexibility and Effciency of Use	자주 쓰는 메뉴 모음이나 순서 변경같이, 숙련된 사용자를 도울 방법을 연구해야 한다.
Aesthetic and Minimalist Design	불필요한 요소가 사용자에게 필요한 정보로부터 사용자의 주의를 분산시키지 않도록 한다.
Help Users Recognizw, Diagnose, and Recover from Errors	쉽고 명확한 언어로 에러 표시를 해야 하며, 동시에 빠른 해결책이 필요하다.
Help and Documentation	사용자가 어려움에 직면할 때 해당 기능에 대한 도움말 문서를 쉽고 빠르게 찾아볼 수 있도록 상황별로 제시해야 한다.

【표 2-2】는 피터 모빌(Peter Morville, 2004)의 사용자 경험 허니콤 모델(User Experience Honeycomb)이다. 사용자 경험을 측정하기 위한 총 7개의 사용성 평가 요인, 유용한(Useful), 사용하기 쉬운(Usable), 매력적인(Desirable), 발견 가능한(Findable), 접근 가능한(Accessible), 신뢰하는(Credible), 가치 있는(Valuable)이 있으며, 필요에 따라 재정의될 수 있다.[27)]

【표 2-2】 피터 모빌의 허니콤 모델[28)]

평가 지표	평가 내용
유용성 (useful)	사용자가 제품과 시스템을 사용할 때 진정으로 유용한가를 고려할 필요가 있다.
매력성 (desirable)	사용자가 시스템을 사용할 때 고민이나 어려움 없이 사용할 수 있다.
몰입성 (immersion)	감성적인 측면에서 오감을 만족시키고, 사용성에 있어 만족감을 얻는다.
사용성 (usable)	사용자는 자신이 필요로 하는 정보를 스스로 찾을 수 있다.
타당성 (validity)	장애를 가진 사용자도 이용할 수 있도록 특정 환경을 고려한다.
신뢰성 (credible)	제품이나 시스템이 안정적이어서 사용자의 신뢰를 받을 수 있다.

27) Kim, N.-H. (2020). User Experience Validation Using the Honeycomb Model in the Requirements Development Stage.International Journal of Advanced Smart Convergence,9(3), 227-231. https://doi.org/10.7236/IJASC.2020.9.3.227

28) Morville, P. (2004). Ambient Findability, Educational technology research and development, 54(6), 623-626.

가치성 (valuable)	사용자의 목표에 기여하고 가치를 포함한다.

【표 2-3】은 대표적인 소프트웨어 국제 품질특성과 척도에 대한 표준 지침인 국제표준화기구 ISO (International Organization for Standardization, 1991) 9126 소프트웨어 품질 평가 모델이다.[29] 소프트웨어 품질특성과 척도에 관한 지침으로 고객 관점에서 소프트웨어에 관한 품질특성과 하위 품질 평가 요소, 평가 기준을 정의하고 있다. 평가 요소로 기능성(Functionality), 신뢰성(Reliability), 사용성(Usability), 효율성(Efficiency), 유지 보수성(Maintainability), 이식성(Portability)이 있다.

【표 2-3】 ISO/IEC-9126, 6가지 품질특성[30]

평가 지표	평가 내용
기능성 (Functionality)	특정 조건에서 사용될 때 명시된 요구와 내재된 요구를 만족하는 기능을 제공하는 S/W 제품의 능력
신뢰성 (Reliability)	명시된 조건에서 사용될 대, 성능 수준을 유지할 수 있는 S/W 제품 능력
사용성 (Usability)	명시된 조건에서 사용될 경우, 사용자에 의해 이해되고, 학습되고 사용되고 선호될 수 있는 S/W 제품의 능력
효율성 (Efficiency)	명시된 조건에서 사용되는 자원의 양에 따라 요구된 성능을 제공하는 S/W 제품의 능력
유지 보수성 (Maintainability)	S/W 제품이 변경되는 능력, 환경, 요구사항 및 기능적 명세에 따른 수정, 개선, 혹은 개작 등 포함
이식성 (Portability)	어느 환경에서 다른 환경으로 전이될 수 있는 S/W 제품의 능력

29) https://www.iso.org/standard/16722.html

30) Lopez, C.M., Lopez, J.E., Buchely, A.B., & Lopez, D.F. (1998). Ergonomic requirements for office work with visual display terminals (VDTs).

1) AI 챗봇 관련 사용성 평가 지표

챗봇(ChatBot)은 채팅(Chatting)과 로봇(Robot)의 합성어로 음성이나 문자를 통한 인간의 대화를 최초의 챗봇은 1960년대 MIT에서 Joseph Weizenbaum에 의해 개발된 엘리자(ELIZA)[31]로 알려져 있다. 엘리자는 텍스트 기반 에이전트를 사용한 인간-컴퓨터 간의 상호작용을 위한 시스템으로, 초기 생각하는 기계라 불렸다.[32] ELIZA는 영어 입력에 타이핑된 응답으로 응답했고, DOCTOR라는 페르소나를 가지고 심리 치료사를 시뮬레이션하거나 "패러디"하는 것을 목표로 했다.[33] 정해진 응답만 가능하던 초기 챗봇을 시작으로 AI와 자연어 처리 기술이 발전하여 1990년대 들어서는 AI 머신러닝 알고리즘의 등장으로 스스로 규칙을 찾아 학습하는 챗봇 서비스로 발전하였다.[34] 최근의 사례, Siri와 ALEXA와 같은 현대 대화 에이전트, ChatGPT는 모두 이 최초의 챗봇에서 영감을 받았다고 할 수 있다.[35]

다음은 AI 챗봇의 사용성 평가 지표 관련 선행 연구를 조사한 내용이다. 【표 2-4】는 컴퓨터 기반 협력적 논증에서 개인 논증을 지원하기 위한 챗봇 개발(곽헌동, 2023)에서 개발한 도구에 대한 의견을 수집하기 위한 사용성 평가 설문지에서 사용된 평가 요소이다. 개인 논증 능력을 지원하는 챗봇 도구를 개발하고, 평가하기 위해 피터모빌의 허니콤 모델을 교육용 챗봇에 적합하도록 수정하여 '유용성', '사용성', '매력성', '신뢰성', '검색성' 요소를 채택하였다.

---

31) Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (Jan. 1966), 36-45. https://doi.org/10.1145/365153.365168

32) ELIZA Archaeology. ELIZA Archaeology Project. https://sites.google.com/view/elizaarchaeology/

33) Berry, DM (2023). 계산의 한계: Joseph Weizenbaum과 ELIZA 챗봇.Weizenbaum Journal of the Digital Society,3(3). https://doi.org/10.34669/WI.WJDS/3.3.2

34) SK hynix newsroom (2024). [All Around AI 1편] AI의 시작과 발전 과정, 미래 전망. https://news.skhynix.co.kr/post/all-around-ai-1

35) Electronic book review (2024). Reading ELIZA: Critical Code Studies in Action. https://electronicbookreview.com/essay/reading-eliza-critical-code-studies-in-action/

【표 2-4】 LLM 기반 챗봇 사용성 평가 지표: 컴퓨터 기반 협력적 논증에서 개인 논증을 지원하기 위한 챗봇 개발[36]

평가 지표	평가 내용	선행 연구
유용성	본 프로토타입은 논증에 대해 전혀 모르는 학습 자들(하 수준)에게 유용하게 활용될 수 있다.	Peter Morville (2004)
	본 프로토타입은 논증에 대해 조금 아는 학습자들(중 수준)에게 유용하게 활용될 수 있다.	
	본 프로토타입은 논증에 대해 많이 아는 학습자들(상 수준)에게 도 유용하게 활용될 수 있다.	
사용성	본 프로토타입은 온라인 상황에서 사용하기에 용이하다	
	본 프로토타입에 포함된 전반적인 대화의 흐름은 인간 대화의 규칙을 부합하여 적절하게 구성되어 있다.	
	본 프로토타입은 예상치 못한 에러가 없고 원활히 이루어진다.	
	본 프로토타입을 원활히 사용하는 데 걸린 시간은 적절하다.	
	본 프로토타입에 활용되는 인지 처리량은 적절하다.	
매력성	본 프로토타입은 이해하기 쉽다.	
	본 프로토타입의 내용 구성이 적절하다.	
	본 프로토타입의 전반적인 디자인은 적절하다.	
	본 프로토타입의 색감, 이미지, 도형 등은 적절하다.	
신뢰성	본 프로토타입은 개인적 논증 구축을 수행하는 데 신뢰성 있게 활용될 수 있다.	
	본 프로토타입이 제공하는 정보 중 오류가 없으며 높은 질의 정보를 제공한다.	
검색성	본 프로토타입은 협력 환경에서 쉽게 찾아 작동할 수 있다.	
	본 프로토타입을 통해 학습자가 원하는 혹은 필요한 정보를 쉽게 얻을 수 있다.	
	본 프로토타입을 사용하여 원하는 학습 단계나 정보로 쉽게 이동할 수 있다.	

【표 2-5】는 대화형 챗봇(Chat-Bot) 서비스디자인 - 패브릭 인테리어 컨설팅을 중심으로(양정아, 2023)에서 컨설팅 챗봇 서비스를 통해 사용자의 유형에

36) 곽헌동. "컴퓨터 기반 협력적 논증에서 개인 논증을 지원하기 위한 챗봇 개발." 국내석사학위논문 서울대학교 대학원, 2023. 서울

따른 맞춤 서비스 제공 가능성을 확인하기 위한 사용성 평가 설문지에서 사용된 평가 요소이다. 정성적 사용성 평가 방법인 John Brooke(1986)의 SUS(System Usability Scale)의 주요 측정 요소 '효과성', '효율성', '만족도'와 '이해 용이성', '유용성'을 추가하여 평가를 진행하였다.

【표 2-5】 LLM 기반 챗봇 사용성 평가 지표: 대화형 챗봇(Chat-Bot) 서비스디자인 - 패브릭 인테리어 컨설팅을 중심으로[37)]

평가 지표	평가 내용	선행 연구
효과성	사용자가 목표를 달성하는데, 챗봇이 얼마나 효과적이었는지	John Brooke (1986)
효율성	사용가가 원하는 서비스를 얻기 위해 얼마나 큰 노력과 시간이 필요한지	
만족도	챗봇이 사용자의 전반적인 경험을 향상하셨는지	
이해 용이성	챗봇이 사용자의 요구를 정확히 이해하여 올바른 대응을 할 수 있는지	
유용성	챗봇이 사용자의 요구와 목적을 이해하고 그에 맞는 도움을 제공하는지	

【표 2-6】은 의료서비스 로봇의 사용성 평가를 위한 사용성 평가 지표 개발(노지혜 외 2, 2023)에서 병원 내원객을 대상으로 한 '안내 로봇'과 '키즈 로봇'의 사용성 평가 지표다. 보편적 사용성 평가 항목, 의료서비스 로봇 사용성 평가 항목, 의료서비스 대상자의 특징과 사용 맥락을 수집하여 '효과성', '학습 용이성', '효율성', '만족도', '신뢰성', '기능성', '반응성', '조작 편의성', '오류', '접근성', '안전성', '직관성', '정확도', '친숙성' 지표를 추출하여 연구 목적에 맞게 사용성 평가 지표 체계를 수립하였다.

37) 양정아. (2023). 대화형 챗봇(Chat-Bot) 서비스디자인 : 패브릭 인테리어 컨설팅을 중심으로 [석사학위논문, 홍익대학교]. https://www.riss.kr/link?id=T16817464

【표 2-6】 LLM 기반 챗봇 사용성 평가 지표: 의료서비스 로봇의 사용성 평가를 위한 사용성 평가 지표 개발: 병원 안내 로봇, 키즈 로봇 중심으로

평가 지표	평가 내용	선행 연구
효과성	시스템을 통해 얼마나 정확하게 목적을 달성하는가	• Shackel, B • Nielsen J • ISO 9241-11
학습 용이성	처음 시스템을 접했을 때, 작동법을 습득 하거나 배우기가 쉬운가?	
효율성	시스템을 얼마나 쉽고, 빨리 그리고 간단 하게 수행하는가	
만족도	시스템에 대한 사용자의 주관적 만족도는 어떠한가?	• 송유미 (2021) • 김선희, 조용진 (2021) • 박다솜, 반영환 (2021) • 김소령 외 6인 (2011) • 김지영 외 2인 (2021)
신뢰성	시스템의 요소에 대해 신뢰할 수 있는가?	
기능성	시스템이 명시된 요구와 내재된 요구를 수행하기 위해 적절한 기능을 제공하는가?	
반응성	자연스러운 대화를 제공하는가?	
조작 편의성	조작하는 것이 불편한 사용 자들도 쉽게 이용할 수 있는가?	
오류	오류발생 시 신속하게 대처할 수 있는가?	
접근성	기술 친화적이지 않은 계층도 평등하게 활용할 수 있는가?	
안전성	오작동하지 않는가?	
직관성	지각과 인지가 쉽고 빠른 감각 인터페이스를 제공하는가?	
정확도	객관적이고 공평한 태도를 유지하는가?	
친숙성	의료 경험에 대한 불안과 두려움을 극복할 수 있는가?	

기존의 전통적인 사용성 평가 지표를 활용하되 평가 내용만 일부 연구 목적에 맞춰 수정하여 사용되고 있었다.

2) LLM 챗봇 관련 사용성 평가 지표

2022년 11월 말, LLM GPT3.5를 탑재한 ChatGPT가 등장하면서 다양한 이미지 생성 모형을 통해 스스로 결과물을 생성하는 생성형 AI 시장이 형성되었다. LLM 기반 챗봇이 도입되고 관련 연구가 시작된 2023년 이후 문헌을 조사하였다. LLM의 특징을 포함하여 평가 지표로 사용될 수 있는 요소가 포함된 선행 연구를 중심으로 살펴보았다.

【표 2-7】은 안무정, 강태임(2023)의 '디지털 트랜스포메이션 경영을 위한 챗GPT 사용자 경험(UX) 디자인 평가 -오픈AI 챗GPT와 마이크로소프트 빙 챗GPT 교차 활용을 중심으로-'에서 ChatGPT와 빙(Bing)챗의 사용성 평가를 비교하는 연구이다. 모바일 앱, VR, 화상회의, 챗봇, 인공지능 분야의 사용자 사용자 경험 디자인을 연구한 선행 논문을 분석하여 ChatGPT 사용자 경험 디자인 요소 16가지를 추출하였다. 추출된 16가지 디자인 요소는 '공정성', '사회적 책임', '해석 가능성', '투명성', '성능', '신뢰성', '만족도', '안전성', '사용성', '효율성', '조작성', '의인화', '매력성', '접근성', '가치성', '유용성'이며, 특히 '의인화'는 LLM의 주요한 특징 중 하나로 추후 사용성 평가 지표 체계를 정립하는 데 기여할 수 있다.

【표 2-7】 LLM 기반 챗봇 사용성 평가 지표: 디지털 트랜스포메이션 경영을 위한 챗GPT 사용자 경험(UX) 디자인 평가 -오픈AI 챗GPT와 마이크로소프트 빙 챗GPT 교차 활용을 중심으로-[38]

38) 안무정 and 강태임. (2023). 디지털 트랜스포메이션 경영을 위한 챗GPT 사용자 경험(UX) 디자인 평가 -오픈AI 챗GPT와 마이크로소프트 빙 챗GPT 교차활용을 중심으로-. 한국디자인문화학회지, 29(2), 237-247.

평가 지표	평가 내용	선행 연구
공정성	챗GPT는 젠더, 지역, 나이 등에 편향성 없이 결과를 생성한다.	• Jakob Nielsen (1993) • 피터 모빌 • ISO 9241-11 (1998)
사회적 책임	챗GPT가 사회적 가치의 위협, 유해한 질문을 거부한다.	
해석 가능성	챗GPT가 생성한 결과에 대해 이해하고 설명할 수 있다.	
투명성	챗GPT가 생성한 결과에 대한 타당한 논리적 근거를 제공한다.	• 박선영 외 2 (2021) • 임종수 외 2 (2020)
성능	챗GPT의 처리 속도와 질문에 대한 답변의 정확도가 높다.	
신뢰성	챗GPT가 생성한 결과의 맥락이 일관성을 유지한다.	
만족도	챗GPT는 질문하는 과정과 생성된 결과에 대해서 만족한다.	
안전성	챗GPT에 질문하는 내용에 개인 정보 유출과 허위 정보를 차단한다.	
사용성	챗GPT는 사용하기 쉽고 원하는 정보를 검색하는 데 효과적이다.	
효율성	챗GPT는 원하는 정보를 검색하는 데 투입된 시간과 노력 비해 효율적이다	
조작성	챗GPT는 의도하는 방향에 따라 빠르게 결과를 얻을 수 있다.	
의인화	챗GPT를 사용함에 따라 비서, 조력자 등과 같은 감성적 친밀감을 가진다.	
매력성	챗GPT를 다시 사용하고 싶은 감정적 마음이 있다.	
접근성	챗GPT는 성별, 장애 등 관계없이 쉽게 접근할 수 있다.	
가치성	챗GPT가 생성한 결과가 기대한 내용과 부합되고 가치가 있다.	
유용성	챗GPT는 문제를 해결하거나 목적을 달성하는 데 도움이 된다.	

【표 2-8】은 한정윤 외 3(2023)의 ‘챗GPT를 활용한 맞춤형 피드백 생성 및 효과 분석’에서 챗GPT를 활용한 맞춤형 피드백의 생성 및 적용 방법론을 탐색한 후, 피드백에 대한 사용자의 인식 및 태도와 교육적 효과를 확인하는 것을 목적으로 사용한 사용성 평가 지표이다. 관련 참고 문헌에서 내용을 활용하여 총 5가지 요인 ‘명확성/정확성’, ‘유용성’, ‘신뢰성’, ‘사회적 실재감’, ‘지속 사용 의도’로 진행하였다.

【표 2-8】 LLM 기반 챗봇 사용성 평가 지표: 챗GPT를 활용한 맞춤형 피드백 생성 및 효과 분석[39)]

평가 지표	평가 내용	선행 연구
명확성/정확성	피드백의 내용을 얼마나 명확하게 파악할 수 있는지에 대한 정확도	김미량(2003), 이은원, 이수정(2004)
유용성	피드백이 학습자에게 얼마나 유용하게 활용될 수 있는지에 대한 정도	정보통신산업진흥원(2011), 임정훈(2014)
신뢰성	피드백의 내용에 대한 신뢰도	진굉위,이정(2021)
사회적 실재감	피드백 내용이 마치 실제 사람이 피드백을 주는 것처럼 느끼는 정도	민윤정(2020), 임동수, 이순구(2022), 조지영(2014)
지속 사용 의도	제공된 피드백에 대해 수용하고 지속적으로 제공받기를 희망하는 정도	최미애(2011), 허희진, 김우빈(2022)

【표 2-9】는 LLM 기반 챗봇 품질을 평가 프레임워크를 개발하는 연구로, ‘Readability’, ‘Correctness’, ‘Coherence’, ‘Engagemen’, ‘Overall Quality’ 5가지 평가 가능 요소가 있다. 특히 ‘Coherence’는 LLM의 주요한 특징 중 하나다. 챗봇이 맥락에 적합한 답변을 하고 있는지 평가하는 지표로 추후 사용성 평가 지표 체계를 정립하는 데 기여할 수 있다.

【표 2-9】 LLM 기반 챗봇 사용성 평가 지표: Personalized Response with Generative AIImproving Customer Interaction with Zero-Shot Learning LLM Chatbots[40)]

평가 지표	평가 내용
Readability	텍스트가 읽기 쉬운 정도, 유창성
Correctness	텍스트가 사실과 상식을 정확하게 반영하며, 논리적인지 여부와 문법적으로 적합한지
Coherence	텍스트가 특정 주제나 이야기 라인에 일관성 있는지 여부

39) 한정윤, 구예리, 김수진. (2023). 챗GPT를 활용한 맞춤형 피드백 생성 및 효과 분석. 교육정보미디어연구, 29(4), 1123-1151.

40) Bink, J. Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM Chatbots.

Engagemen	텍스트가 흥미롭고 매료시키는 정도
Overall Quality	독해 자료의 전반적인 텍스트 품질

【표 2-10】은 챗봇 인터페이스 디자인과 서비스를 개선하는 연구로, ‘Knowledge Recency’, ‘Logical Reasoning’, ‘Hallucination’, ‘Transparency’, ‘Bias’, ‘Privacy Risks’, ‘Unfairness’, ‘Academic Misuse’, ‘Over-reliance’, ‘Distribution of Wrong Information’ 10가지 평가 가능 요소가 있다. ‘Hallucination’ LLM의 주요한 특징 중 하나로 챗봇이 잘못된 답변을 하는 경우를 평가하는 지표로 추후 사용성 평가 지표 체계를 정립하는 데 기여할 수 있다.

【표 2-10】 LLM 기반 챗봇 사용성 평가 지표: A Complete Survey on LLM-based AI Chatbots[41)]

평가 지표	평가 내용
Knowledge Recency	최신 지식을 유지하는 데 어려움이 있음
Logical Reasoning	다단계 추론 문항의 성능 격차
Hallucination	부정확하고 신뢰할 수 없는 응답 생성
Transparency	챗봇 추론 과정의 명확성 부족
Bias	챗봇 교육 및 대응에 있어 데이터 편향성
Privacy Risks	개인 정보 보호 문제 및 데이터 보호 문제
Unfairness	챗봇 접근성의 언어적, 경제적 불공정성
Academic Misuse	학문적 청렴성을 지키기 위한 과제
Over-reliance	비판적 사고력에 미치는 영향

전통적인 사용성 평가 지표의 평가 내용을 일부 수정하여 사용하거나, 평가 대상의 특성에 맞춰 평가 지표 자체를 새로 구성하여 평가하였다.

41) Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A complete survey on llm-based ai chatbots.arXiv preprint arXiv:2406.16937.

3) 교육 분야 챗봇 관련 사용성 평가 지표

교육 분야에서 챗봇은 학습의 진행 과정에서 대화를 통해 즉각적인 사용자 피드백이 가능하며, 피드백에 따라 맞춤형 콘텐츠 제공이 가능하다. 이러한 상호작용 특성을 바탕으로 어학을 중심으로 학습 분야에 활용되는 사례가 파악되었다.[42] 이미 많은 외국어 교육용 챗봇 들이 개발 중이거나, 사용 중이며, 이를 사용한 교육적 효과나 활용 방안에 대한 연구들이 활발히 진행되고 있다.[43]

다음은 교육 분야에서 평가 지표로 사용될 수 있는 요소가 포함된 선행 연구를 중심으로 살펴보았다. 【표 2-11】은 평가 지표를 개발하는 연구로, '시스템의 가독성', '시스템과 실세계의 일치', '제어의 자유권', '일관성과 표준', '에러 인식', '진단 및 복구', '도움말 및 문서화', '상호작용 관리 능력', '의인화 수준', '학습 동기', '교수 자료' 총 10개의 평가 가능 요소가 있다.

【표 2-11】 교육 분야 챗봇 사용성 평가 지표: 사용자 친화적인 챗봇 튜터 설계 지침 개발 연구[44]

평가 지표	평가 내용	선행 연구
시스템의 가독성 (Visibility of system status)	사용자가 어떻게 진행되고 있는지 피드백을 강조하는 측면	Jakob Nielsen (1993)
시스템과 실세계의 일치 (Match between system and the real world)	사용자의 언어와 실 세계의 개념에 익숙한 디자인을 강조한 측면	Jakob Nielsen (1993)
제어의 자유권 (User control and freedom)	에러발생시 언제라도 빠져나올 수 있도록 하는 자유권을 강조한 측면	Jakob Nielsen (1993)
일관성과 표준 (Consistency and standards)	산업 표준에 따라 사용자가 언어와 행동을 이해할 수 있도록 하는 측면	Jakob Nielsen (1993)
에러 인식, 진단, 복구 (Help users recognise, diagnose, and recover from errors)	오류가 발생시 사용자에게 명확하게 알려줄 수 있도록 하는 측면	Jakob Nielsen (1993)

42) 조희석(2018). 챗봇(ChatBot)의 활용 사례 및 이러닝 도입 전략
43) 박정아, 이향. (2021). 한국어 교육용 AI 챗봇 개발을 위한 챗봇 빌더 활용 방안.외국어로서의 한국어교육,63, 51-91, https://doi.org/10.21716/TKFL.63.3
44) 차현진. (2023). 사용자 친화적인 챗봇 튜터 설계 지침 개발 연구. 컴퓨터교육학회 논문지, 26(5), 79-92.

도움말 및 문서화 (Help and Documentation)	챗봇이 기본적으로 어떤 역할을 하는지에 대한 설명과 문구를 제공하도록 설계하는 측면	Jakob Nielsen (1993)
상호작용 관리 능력 (Interaction management capabilities)	챗봇의 특성을 반영한 영역으로 대화의 시작과 끝, 절차 등 사용자 대화의 흐름을 챗봇이 이해할 수 있도록 설계하는 측면	Moore, R.J, Arar, R(2019)
의인화 수준 (Personification level)	사용자가 챗봇을 사람처럼 상호 작용할 수 있도록 페르소나를 설계하는 측면	민윤정 외 2인 (2020)
학습 동기 (Learning motivation)	학습자가 지속적으로 학습하도록 동기를 부여하는 측면	Alsumait, A, Al-Osaimi, A (2010)
교수 자료 (Learning materials)	챗봇이 제공하는 교수 자료를 학습자에게 적절한 학습 목표와 난이도를 고려하여 설계하는 측면	Albion, P.R. (1999)

【표 2-12】는 교육을 위한 챗봇 수준을 평가하는 연구로, '인지된 유용성', '인지된 용이성', '태도', '사용 의도', '행동적 몰입', '정서적 몰입', '인지적 몰입', '긍정적 감정', '인지된 어려움', '인지된 스킬', '수행 기대감', '수행 목표', '수행 자기효능감' '부정적 감정' 촌 14개의 평가 가능 요소가 있다.

【표 2-12】 교육 분야 챗봇 사용성 평가 지표: 대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석[45)]

평가 지표	평가 내용	선행 연구
인지된 유용성	가상현실 기반 학습 콘텐츠는 학습활동을 효율적으로 만들어 준다.	Davis(1989)
인지된 용이성	가상현실 기반 학습 콘텐츠를 조작하기 쉽다.	Huang, Liaw (2018)
태도	가상현실 기반 학습 콘텐츠를 사용하는 것은 좋은 생각이다.	
사용 의도	나는 가상현실 기반 학습 콘텐츠를 사용하는 수업에 적극적으로 참여할 의향이 있다.	
행동적 몰입	나는 가상현실 기반 STEAM 수업에 참여할 때 일정한 집중력을 유지할 수 있었다.	Fredrick 외 2인 (2004)
정서적 몰입	나는 가상현실 기반 STEAM 수업 활동을 통해 재미를 느꼈다.	Sun, Rueda (2012)
인지적 몰입	나는 STEAM 수업 자료를 읽을 때, 내용을 이해하기 위해 나에게 질문을 하였다.	

45) 윤헌철. (2023). 대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석. 홀리스틱융합교육연구, 27(4), 47-67.

긍정적 감정	나는 가상현실 기반 STEAM 수업에 참여했을 때, '흥미로움'을 느꼈다.	Moneta, Kekkone –Moneta(2007)
인지된 어려움	가상현실 기반 STEAM 수업에 참여했을 때, 학습활동의 어려움 정도는 어떠하였는가?	
인지된 스킬	가상현실 기반 STEAM 수업에 참여했을 때, 학습활동을 성공적으로 수행할 수 있는가?	
수행 기대감	가상현실 기반 STEAM 수업을 듣고, 수행에 대한 예상 점수는 몇 점인가?	
수행 목표	가상현실 기반 STEAM 수업을 듣고, 수행에 대한 목표 점수는 몇 점인가?	
수행 자기효능감	가상현실 기반 STEAM 수업을 듣고, 기말 시험에서 어려운 문제를 해결할 자신이 있는가?	
부정적 감정	나는 가상현실 기반 STEAM 수업에 참여했을 때, '긴장감'을 느꼈다.	

3. 소결

그동안 사용성 평가 지표는 발전하는 시대와 기술에 맞춰 변형됐으며, LLM 기반 챗봇이 등장한 이후도 관련 연구들이 점차 진행되고 있음을 확인할 수 있었다. 이에 따라 본 연구의 주제인 LLM 기반 교육용 챗봇의 사용성 평가 지표 체계 역시 기존 사용성 평가에 대한 정형화된 요소와 LLM과 교육 분야 특징에 따른 평가 요소가 함께 어우러져 도출되어야 한다.

Ⅲ. 연구 방법

다음 【그림 3-1】은 사용성 평가 지표 정립 단계이다. 사용성 평가 지표 개발부터 검증 적용 활용까지 총 8단계의 과정을 거쳐 진행하였다.

【그림 3-1】 사용성 평가 지표 정립 단계: 개발부터 활용까지

1. 사용성 평가 지표 개발

1단계, 각종 문헌을 기반으로 본 연구의 주제와 연관된 다양한 사용성 평가 지표를 수집했다. 사용성 평가 지표는 LLM이 도입되기 전 기존 챗봇의 사용성 평가 요소, LLM 기반 챗봇 특성에 따른 평가 요소, 그리고 교육용 챗봇의 특성과 이에 기반한 사용성 평가 요소이며 총 29개의 국내외 논문을 통해 총 274개의 평가 지표를 수집하였다. 이는 기존 인공지능 챗봇 사용성 평가 지표와 차별화되는 LLM 교육용 챗봇에 특화된 평가 지표 토대를 마련했다는 데 의의가 있다.

2단계, 수집한 274개의 평가 지표 중 중복되거나 불필요한 지표를 제거하여 총 113개의 유의미한 지표를 추출하였다. 이러한 사용성 평가 지표들은 본 연구의 범위에 따라 단어와 의미를 조정하며 추출하였기 때문에 LLM 기반 교육용 챗봇과 관련성이 깊다. 이후 1차 어피니티 다이어그램을 진행하여 주지표와 이에 포함되는 상세 지표를 구분하여 사용성 평가 지표의 체계를 갖추었다. 이는 2차 지표 정리 FGI를 진행하기 위한 기반을 마련했다는 점에서 의의가 있다.

	A	B	C	D	E	F	G
1	*S = 사용성 *O = 품질						
2	S	정확성	일관성	일관성	Coherence Consistency	- 텍스트가 특정 주제나 이야기 라인에 일관성 있는지 여부 - 전반적인 기능/디자인에 관한 일관성	C_L2_DC1_4 L_L2_OB4_3 G_L2_DB5_2 C_L2_OE1_6
3				일관성과 표준	Consistency and standards	- 산업 표준에 따라 사용자가 언어와 행동을 이해할 수 있도록 하는 측면 - 챗봇은 일관성 있는 언어와 스타일로 상호작용할 수 있도록 설계한다. - 챗봇이 제공하는 버튼, 상호작용 방식 등 일관된 디자인과 표준으로 소통하도록 설계한다.	E_L1_DD1_4
4				후속질문 타당성	Relevance of Follow-Up Questions	- 이전 응답을 기반으로 문맥을 유지하고 논리적이고 적절한 질문을 제기할 수 있는 능력	E_L2_OA5_2
5			직관성	정확도	Accuracy	- 성한 응답의 문법, 구문, 의미론적 해석, 전반적인 구조를 평가하는 자동 및 인간 기반 평가를 포괄함	C_L2_DA3_13 G_L1_OB1_1
6				정확성	Correctness	- 텍스트가 사실과 상식을 정확하게 반영하며, 논리적인지 여부와 문법적으로 적합한지	L_L2_OB4_2
7				명확성 정확성 명료성	Perspicuity	- 서비스 조건이 더 명확하고 이해하기 쉬운 인터페이스를 제공하는지	C_L2_DC1_2 G_L2_DB4_1 L_L2_OB2_2
8				시스템 명확성	System Clarity	- 응답이 명확하고 관련성이 있는지	E_L2_OA5_3
9				직관성	Intuitiveness	- 병원 안내 로봇을 효율적으로 사용할 수 있도록 직관적인 인지가 가능한가? - 각 애플리케이션 테스트가 기대에 부합하였나요?	C_L2_DA3_12 C_L2_DA4_1 C_L2_OE1_8
10				투명성	Transparency	- 챗봇 추론 과정의 명확성 부족	G_L2_DA1_4 L_L2_OA1_4
11				인식보다는 기억	Recognition rather than recall	- 학습자가 직관적으로 인식하고 사용할 수 있도록 명확한 시각적 요소와 명확한 지침을 제공한다. - 상호작용하는 동안 학습자가 선택한 정보를 보여준다.	E_L1_DD1_6
12				시스템과 실세계의 일치	Match between system and the real world	- 사용자의 언어와 실 세계의 개념에 익숙한 디자인을 강조한 측면 - 학습자 수준에 맞추어 친숙한 언어와 방법을 통해 챗봇과 상호작용을 경험할 수 있도록 설계한다. - 메타포나 아이콘, 버튼 등 사용자 인터페이스 요소들이 학습자가 직관적으로 이해할 수 있도록 설계한다.	E_L1_DD1_2
13			적절성	근거	Groundedness	- 생성된 응답의 사실적 타당성을 평가하는 것	G_L2_OB1_7
				감각, 특이성,	Sensibility, Specificity,Interestingness (SSI)	- 생성된 응답의 적절성을 측정하는 것	G_L2_OB1_2

【그림 3-2】 사용성 평가 지표 개발: 1차 정리

3단계, 2차 지표 정리를 위해 HCI/서비스디자인 관련분야 연구원, HCI 분야 교수급 전문가 등 4명과 함께 FGI를 통한 2차 어피니티 다이어그램을 진행하였다. 이를 통해 'LLM'과 '교육 분야' 특화 지표의 특성을 구분하고 특화 지표 체계를 추출했다는 점에서 의의가 있다.

4단계, 3차 지표 정리는 2차 지표 정리 내용을 바탕으로 HCI 분야 교수급 전문가 5명에게 사용성 평가 지표에 대한 피드백을 받아 진행하였다. 각 지표의 의미 및 분류가 연구 목적에 부합하는지 확인하고 이를 바탕으로 세부 내용을 수정함으로써 사용성 평가 지표를 고도화했다는 데 의의가 있다.

【그림 3-3】 FGI 진행 사진

2. 사용성 평가 지표 검증

5단계, 사용성 평가 지표에 따라 각각 대응하는 세부 질문을 도출하고, 설문지를 구성하였다. 이후 본 사용성 평가를 시행하기에 앞서 HCI/서비스디자인 관련분야 연구원 7명을 대상으로 파일럿 테스트를 진행하였다. 사용성 평가 목적 및 대상에 대한 설명을 바탕으로 평가 항목의 난이도, 세부 질문 적합성 등

에 대한 피드백을 받았으며, 이를 통해 본 설문을 위한 최종 설문지를 구성했다는 점에서 의의가 있다.

6단계, 앞서 정립한 사용성 평가 지표를 바탕으로 구성된 설문지를 활용하여 본 설문 진행하였다. 본 설문은 LLM 기반 교육용 챗봇 중 영어 회화 앱 서비스에 한정하여 사용 경험이 있는 사용자를 대상으로 진행하였다. 설문 결과를 바탕으로 통계적 분석을 하여, 항목별 상관관계 및 의미적 연관성이 있는 사용성 평가 지표 체계를 최종적으로 도출했다는 점에서 의의가 있다.

3. 사용성 평가 지표 적용

7단계, 최종적으로 개발된 사용성 평가 지표의 유효성을 확인하기 위해 사용성 태스트를 진행하였다. LLM 기반 교육용 챗봇 중 영어 회화전용 앱 서비스 3가지를 선정하여 서비스디자인 분야 연구원을 10명을 대상으로 사용성 평가를 진행하였다. 연구원은 지정해 준 태스크에 따라 각각 3가지 앱 서비스를 경험해 본 후 각각의 사용성 평가 설문을 진행하였고, 이에 따라 도출된 설문 결과를 분석하는 과정을 거쳐 사용성 지표의 유효성을 확인했다는 점에서 의의가 있다.

4. 사용성 평가 지표 활용

8단계, LLM 기반 교육용 챗봇 사용성 평가 지표를 활용한 챗봇 서비스 디자인 가이드라인을 제안하였다. 개발과 검증, 적용 단계를 거쳐 도출된 사용성 평가 지표가 실질적으로 활용될 수 있는 기반을 마련하기 위해 ‘사용성 평가 지표를 활용한 LLM 기반 교육용 챗봇 서비스디자인 가이드라인’을 출하였다. 디자인 가이드라인은 추후 LLM 기반 교육용 챗봇의 초기 디자인 및 개발 단계에서부터 활용될 수 있다는 점에서 의의가 있다.

Ⅳ. 사용성 평가 지표 개발

1. 연구 방법

LLM이 도입된 교육용 챗봇 사용성 평가 지표 개발을 위해 관련 문헌을 수집하였다. 2024.04.25.~2024.06.18. 총 55일간 Google Scholar, DBpia, RISS, arXiv 등의 학술 DB를 통해 관련성 높은 문헌을 추려내어 최종적으로 국문 논문 14개, 해외 논문 15개 총 29개의 논문을 선정하였다. 기존 AI 기반 챗봇, 교육용 챗봇, ChatGPT 기반 챗봇, LLM 기반 챗봇 총 4개의 수집 키워드를 기준으로 조사했으며, 기존 AI 기반 챗봇 관련 문헌은 (C) 코드, 교육용 챗봇 관련 문헌은 (E) 코드, ChatGPT 기반 챗봇 관련 문헌은 (G) 코드, LLM 기반 챗봇 관련 문헌은 (L) 코드를 부여하여 비교 구분하였다.

앞서 수집된 지표 중 의미가 중복되거나 불필요한 지표를 합치고 제거하여 사용성 평가 지표 체계를 구성하기 위한 기반을 마련하였다. 이후 유사성에 따라 의미 있게 분류하는 방법인 어피니티 다이어그램(Affinity Diagram) 방법론을 활용하여 다량의 복잡한 내용을 1차적으로 구조화하였다.

이때 주지표의 카테고리를 구분하고 위계 구조를 수립하기 위하여 인지심리학 이론인 SOR(Stimulus-Organism-Response) Theory[46]를 도입하였다. SOR 이론은 환경에 의한 자극에 영향을 받는 인지 및 감정 상태를 분석하여 인간의 행동을 설명하는 메커니즘이다. 외부 자극(S)이 사람에게 영향을 미치면, 내적 유기체 변화(O)가 발생하며, 이를 통해 행동반응(R)이 일어난다는 이론이다. 자극(S)은 개인을 흥분시키는 영향력을 의미하며, 광고나 가격, 입소문 등 외부에서 개인의 흥미를 유발하는 외부요인이다. 유기체(O)는 고객의 정서적 또는

46) Mehrabian A., Russell J. A. (1974) An approach to environmental psychology

인지적 상태를 의미하며, 정서적 상태는 자극받은 후 고객이 느끼는 감정이나 기분을 말한다. 반응(R)은 특정 행동을 수행한 결과로 정의된다. SOR 이론은 반전하면서 다양한 분야에 적용되었으며, 특히 웹사이트 경험, 앱 서비스 경험에서 나타나는 다양한 소비자 행동 등을 설명하는 데 유용하게 활용되고 있다.[47] 본 연구에서는 SOR 이론이 사용자가 서비스 여정을 통해 경험하는 과정과 유사하다고 보았다. 사용자는 서비스를 사용하면서 인지적으로 특성을 이해하고, 정서적으로 판단을 내리며, 행동적으로 반응한다. 이러한 단계는 사용성을 평가할 때 단순히 앱의 기능적 사용성만 보는 것이 아니라, 사용자가 느끼는 전반적인 사용자 경험(예: 사용성, 정서적 만족감, 지속적 사용 의도)을 총체적으로 이해하는 데 기여한다. 서비스의 지속적 사용 여부를 평가하기 위해서는 기능적 사용성뿐 아니라, 사용자의 전반적인 경험과 만족도를 종합적으로 고려해야 한다. 이에 본 연구는 사용성 평가 지표를 분류하는 초기 단계에서 SOR 이론을 활용하여, 인지적, 정서적, 행동적 반응을 포괄하는 체계적이고 총체적인 지표 체계를 개발하고자 하였다.

사용성 평가 지표 2차 정리를 위해 HCI/서비스디자인 관련분야 석사과정 연구원 2명, 박사과정 연구원 2명, 교수급 전문가 1명 총 5명이 참여한 FGI(Focus Group Interview)를 진행하였다. FGI는 약 90분간 오프라인으로 진행되었으며, 1차 지표 정리한 내용을 기반으로 2차 어피니티 다이어그램을 진행하여 주지표와 상세 지표의 분류 및 의미 재정의, ‘LLM’, ‘교육 분야’ 특화 지표 도출을 목적으로 진행하였다.

사용성 평가 지표 개발 마지막 단계인 사용성 평가 지표 3차 정리 및 고도화를 위해 5명의 서비스디자인 관련 교수급 전문가에게 주관식 설문을 진행하였다. 24.08.24~24.08.29 총 6일간 진행되었으며, ‘그루핑의 적합성’, ‘지표 각각

47) Kim, M. J., Lee, C.-K., & Jung, T. (2020). Exploring Consumer Behavior in Virtual Reality Tourism Using an Extended Stimulus-Organism-Response Model. Journal of Travel Research, 59(1), 69-89. https://doi.org/10.1177/0047287518818915

의 정의 및 의미 명확성’ ‘추가될 만한 지표’, ‘기타 의견’ 등에 대해 자문하였다. 3차 정리를 마지막으로 사용성 평가 지표 개발 단계의 LLM 기반 교육용 사용성 평가 지표가 도출되었다.

2. 사용성 평가 지표 수집

문헌 연구 단계에서 본 연구 목적에 부합하는 총 29개의 논문을 분석하였다. LLM 기반 교육용 챗봇 서비스의 사용성 평가 지표와 기존 AI기반 챗봇 서비스의 사용성 평가의 차이를 알아보기 위해 29개의 논문에 각각의 코드를 부여하였다. 기존 AI기반 챗봇 관련 문헌(C) 9개, 교육용 챗봇 관련 문헌(E) 7개, ChatGPT 기반 챗봇 관련 문헌(G) 6개, LLM 기반 챗봇 관련 문헌(L) 7개를 분석하였다【표 4-1】. (G)와 (L)은 모두 생성한 AI기반의 연구라는 점에서 본 연구의 목적상 같은 내용이 포함되었다고 볼 수 있기에, 추후 분석 단계에서는 합쳐졌다. 【표 4-2】는 29의 논문에서 각각 기준을 두고 있는 대상 챗봇 기술 수준, 대상 챗봇, 사용성 평가 지표 혹은 요인을 정리한 표다. 총 274개의 사용성 평가 요소가 추출되었으며, 1차 분석을 위해 의미가 중복되거나 불필요한 지표를 합치고 제거하여 총 113개의 요소로 정리되었다.

다음【표 4-3】은 사용성 평가 지표 논문을 수집하여 발견한 총 274개의 사용성 평가 요소 중 의미가 중복되거나 불필요한 지표를 합치고 제거하여 정리한 총 113개의 요소의 결과이다. 앞서 도출된 총 274개의 사용성 평가 요소를 각 문헌에 부여된 코드에 따라 정리해 본 결과 기존 AI기반 챗봇 관련 문헌(C)에서 추출된 요소 20개, ChatGPT 기반 챗봇 관련 문헌(G) 및 LLM 기반 챗봇 관련 문헌(L)에서 추출된 요소 36개, 교육용 챗봇 관련 문헌(E)에서 추출된 요소 32개로 정리되었다.

(G)/(L)+(C) 9개, (E)+(C) 3개에 속하는 요소들과 (C), (G)/(L), (E)에서 모두 사용되고 있는 주요 지표 14개 가시성, 몰입, 사용 의도, 사용성, 실패에 대한 회복력, 유용성, 즐거움, 일관성, 접근성, 정확성, 프라이버시, 효과성은 대체로 앞서 문헌 조사 내용에서 전통적인 사용성 평가 지표로 분류되었던 요소들을 포함하고 있다.

단순히 접근성의 정도, 프라이버시 보호 정도 등 AI 챗봇 초기에 중요하게 여겨졌던 사용성 측면을 강조하고 있는 기존 AI기반 챗봇 관련 문헌(C)에서 추출된 요소에 비해 LLM 기반 챗봇(G), (L)에 관련된 요소는 사회적 실재감, 의인화, 최신성, 환각, 편향 등 LLM은 특성으로 거론된 내용이 더 강조되고 있었다. 또한 교육 분야 챗봇(E)에 관련된 요소는 학습 동기, 학습자 맞춤성, 학문적 피드백 등 교육(학습)에서 중요시하는 요소들이 포함되어 있었다.

【표 4-1】 사용성 평가 지표 논문 수집

국내외	연구자	(C)	(G)	(L)	(E)
국내	곽헌동(2023)	●			
	김민지(2023)	●			
	김형조(2023)	●			
	노지혜, 우승현, 박진영(2023)	●			
	모온가(2023)	●			
	양정아(2023)	●			
	이승희, 손원준(2022)	●			
	박지원 외 4(2024)		●		
	강신천, 허희옥(2023)		●		
	안무정, 강태임(2023)		●		
	한정윤, 구혜리, 김수진(2023)		●		
	차주혜 외 4(2024)				●
	윤헌철(2023)				●
	차현진(2023)				●
해외	Mudrikah Nasyiah, Bayu Kelana, and Anggar Riskinato(2024)	●			
	Simone Borsci외 6인(2022)	●			
	Goran Bubaš외 2인(2024)		●		
	Mahyar Abbasian외 13인(2023)		●		
	Emily Theophilou외 12인(2023)		●		
	Caitlin Silvestri 외 9인(2024)			●	
	Changrong Xiao외 4인(2024)			●	
	Jonathan Dortheimer외 3인(2024)			●	
	Samuel Kernan Freire, Chaofan Wang, Evangelos Niforatos(2024)			●	

	Sumit Kumar Dam외 3인(2024)			●	
	Bink, Joëlle M(2023)			●	
	Henansh Tanwar외 3인(2024)				●
	Maung Thway외 4인(2024)				●
	Yusuke Kajiwara, Kouhei Kawabata(2024)				●
	Galina Ilieva(2023)				●
합계(계)	29	9	7	6	7

【표 4-2】 사용성 평가 지표 논문 수집(2)

제목	저자(발행 연도)	기술	대상	코드	평가 지표 및 요인
컴퓨터 기반 협력적 논증에서 개인 논증을 지원하기 위한 챗봇 개발[48]	곽헌동(2023)	AI	학생 개인 논증 구축을 지원하는 교육용 챗봇	(C)	유용성, 사용성, 매력성, 신뢰성, 검색성
협력학습을 지원하는 인공지능 챗봇 설계원리 개발 연구[49]	김민지(2023)	AI	협력학습을 지원하는 (교육용) 인공지능 챗봇	(C)	도구의 사용성, 접근성, 신뢰성, 매력성, 유용성, 가치성
온라인 토론 성찰을 지원하는 대시보드 기반 챗봇 개발[50]	김형조(2023)	AI	온라인 토론에서 학습자들의 대시보드 기반 성찰을 돕기 위한 챗봇	(C)	유용성, 사용 용이성, 태도, 사용 의도
의료서비스 로봇의 사용성 평가를 위한 사용성 평가 지표 개발 : 병원 안내 로봇, 키즈 로봇 중심으로[51]	노지혜외 2인 (2023)	AI	병원 안내 로봇, 병원 키즈 로봇	(C)	사용성, 기능성, 신뢰성, 만족도, 효과성, 학습 용이성, 효율성, 만족도, 신뢰성, 기능성, 반응성, 조작 편의성, 오류, 접근성, 안전성, 직관성, 정확도, 친숙성, 만족도, 신뢰성, 외적 요소, 사용성, 기능성, 안전성, 효용성, 재미, 몰입
중국 온라인 쇼핑몰의 챗봇 서비스에 대한 지각된 가치 및 지속 이용 의도 연구[52]	모온가(2023)	AI	온라인 쇼핑몰 챗봇	(C)	가시성, 명확성, 자율성, 일관성, 오류 방지성, 효율성, 유연성, 심미성, 조작성, 도움말 제공

대화형 챗봇(Chat-Bot) 서비스디자인 - 패브릭 인테리어 컨설팅을 중심으로 -[53]	양정아(2023)	AI	컨설팅 챗봇	(C)	효과성, 효율성, 만족도, 이해 용이성
시니어를 위한 애플리케이션의 사용성 평가지표에 관한 연구 - 국내 유통기업 사례를 중심으로[54]	이승희, 손원준(2022)	AI	동영상 저작도구 애플리케이션	(C)	직관성, 조작성, 반응성, 오류성, 만족성
생성형 AI 기반의 동화책 제작 서비스 설계 및 구현[55]	박지원 외 4인 (2024)	GAI (GPT-4)	동화책 제작 서비스	(G)	사용 편리성, 시각디자인, 효과성, 효율성, 만족도
생성형 AI 기반 교수설계 지원 플랫폼 개발 및 시범 적용[56]	강신천, 허희옥(2023)	GAI	교수설계 지원 플랫폼	(G)	Accessibility, Consistency, Efficiency, Intimacy, Feedback, Providing Information, Simplicity, Convince of operation, Visibility, Message, Icon, Layout, Denomination, Screen Transition, Size, Structure, Category
디지털 트랜스포메이션 경영을 위한 챗 GPT 사용자 경험(UX) 디자인 평가 -오픈AI 챗GPT와 마이크로소프트 빙 챗 GPT 교차 활용을 중심으로-[57]	안무정, 강태임(2023)	ChatGPT	ChatGPT, 빙(Bing)	(G)	공정성, 사회적 책임, 해석 가능성, 투명성, 성능, 신뢰성, 만족도, 안전성, 사용성, 효율성, 조작성, 의인화, 매력성, 접근성, 가치성, 유용성
챗GPT를 활용한 맞춤형 피드백 생성 및 효과 분석[58]	한정윤 외 2인(2023)	GAI	맞춤 피드백 생성 모델	(G)	명확성/정확성, 유용성, 신뢰성, 사회적 실재감, 지속 사용 의도

48) op.cit
49) 김민지. (2023). 협력학습을 지원하는 인공지능 챗봇 설계원리 개발 연구: The Developmental Study of AI Chatbot Design Principles for Supporting Collaborative Learning.
50) 김형조. "온라인 토론 성찰을 지원하는 대시보드 기반 챗봇 개발." 국내석사학위논문 서울대학교 대학원, 2023. 서울
51) op.cit

개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석[59]	차주혜 외 4인(2024)	AI	교육용 챗봇 (솔미챗, Speak AI 튜터, AI 헬피, 키위챗, 클래스팅 젤로)	(E)	학습 콘텐츠 추천, 피드백, Q&A 제공, 목표 설정 및 학습 진행 상황 모니터링, 주제별 토론 유지 가능, 특정 질문에 대한 답변 가능, 인사말/유쾌한 성격 제공, 자연스러운 대화, 재미있는/매력적인, 의미와 의도를 감지할 수 있습니다, 사회적 신호에 적절하게 대응합니다.
대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석[60]	윤헌철(2023)	AI	대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석	(E)	긍정적 감정, 인지된 어려움, 인지된 스킬, 수행 기대감, 수행 목표, 수행 자기효능감, 부정적 감정, 행동적 몰입, 정서적 몰입, 인지적 몰입, 인지된 유용성, 인지된 용이성, 태도, 사용 의도
사용자 친화적인 챗봇 튜터 설계 지침 개발 연구[61]	차현진(2023)	AI	교육용 챗봇	(E)	시스템의 가독성, 시스템과 실세계의 일치, 제어의 자유권, 일관성과 표준, 오류 방지, 인식보다는 기억, 사용의 유연성과 효율성, 심미적이고 최소한의 디자인, 상황 이해, 사용자가 오류를 인식하고 진단하며 복구할 수 있도록 돕기, 상호작용 관리 능력, 도움말 및 문서화, 의인화 수준, 학습 동기, 학습 절차와 방법, 교수 자료

52) 모온가. "중국 온라인 쇼핑몰의 챗봇 서비스에 대한 지각된 가치 및 지속이용의도 연구." 국내석사학위논문 건국대학교 대학원, 2023. 서울

System Usability Scale for Measuring Usability of Social Network Applications from User Perspectives[62]	Mudrikah Nasyiah 외 2인 (2024)	AI		(C)	Frequency of use, Web/system simplicity, Ease of use, Technical support, Integration with web/system, Consistency, Speed of, learning, Intuitiveness, Confidence, Need to learn
The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents[63]	Simone Borsci외 6인(2022)	AI	CRM 챗봇	(C)	Ease to start a conversation, Access to chatbot, Expectation setting, Flexibility and communication effort, Ability to maintain a themed discussion, Reference to the service, Users' privacy and security, Recognition and facilitation of users' goal and intent, Relevance of information, Maxim of quantity, Resilience to failure, Understandability and politeness, Perceived conversational credibility, Speed of answer

53) op.cit
54) 이승희, 손원준. (2022). 시니어 세대를 위한 모바일 어플리케이션에 관한 사용성 평가 연구 – 국내 유통기업 사례를 중심으로. 상품문화디자인학연구,(68), 1-12.
55) 박지원, 정하성, 임동희, 박주은, 정종진. (2024-01-24). 생성형 AI 기반의 동화책 제작 서비스 설계 및 구현. 한국HCI학회 학술대회, 강원.
56) 강신천, 허희옥. (2023). 생성형 AI 기반 교수설계 지원 플랫폼 개발 및 시범 적용. 컴퓨터교육학회 논문지, 26(6), 143-153.

Development of an Assessment Scale for Measurement of Usability and User Experience Characteristics of Bing Chat Conversational AI[64]	Goran Bubaš외 2인(2024)	GAI	Bing Chat	(G)	Perceived Usefulness, General Usability, Learnability, System Reliability, Visual Design and Navigation, Information Quality, Information Display, Cognitive Involvement, Design Appeal, Trust, Personification, Risk Perception
Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI[65]	Mahyar Abbasian외 13인 (2023)	GAI	healthcare chatbot	(G)	Accuracy, Trustworthiness, Empathy, Performance, Intrinsic Sensibility, Specificity, Interestingness (SSI), Robustness, Generalization, Conciseness, Up-to-dateness, Groundedness, Safety and Security, Privacy, Bias, Interpret ability, Emotional Support, Health Literacy, Fairness, Personalization, Memory Efficiency, Floating point Operations (FLOP), Token Limit, Number of Parameters

57) op.cit
58) op.cit
59) 채주혜, 김민영, 류강, 유영만, 신윤희. (2024). 개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석. 디지털콘텐츠학회논문지, 25(4), 1053-1068. 10.9728/dcs.2024.25.4.1053
60) op.cit

Learning to Prompt in the Classroom to Understand AI Limits: A Pilot Study[66]	Emily Theophilou외 1인 (2023)	ChatGPT	ChatGPT	(G)	Perceived level of identity threat, Functionality, Self-Reported, Emotions after interaction, Interaction quality(UX), Perception of human likeness, Social Presence, Semantic Differential Hedonic dimension, Semantic Differential Pragmatic dimension
Evaluation of a Novel Large Language Model (LLM) Powered Chatbot for Oral-Boards Scenarios[67]	Caitlin Silvestri 외 9인 (2024)	LLM	LLM	(L)	Inappropriate Content, Missing Content, Likelihood of Harm, Extent of Harm, Hallucinations
Evaluating Reading Comprehension Exercises Generated by LLMs: A Showcase of ChatGPT in Education Applications[68]	Changrong Xiao외 4인 (2024)	LLM	ChatGPT	(L)	Readability, Correctness, Coherence, Engagement, Overall, Quality
Evaluating large-language-model chatbots to engage communities in large-scale design projects[69]	Jonathan Dortheimer외 3인 (2024)	LLM	AI 챗봇, GPT4-LLM	(L)	Service Quality, Perceived enjoyment, Perceived usefulness, Perceived ease of use, Satisfaction, Continuance intention
Conversational Assistants in Knowledge-Intensive Contexts: Interactions with LLM- versus Intent-based Systems[70]	Samuel Kernan Freire 외 2인 (2024)	LLM	Intent 챗봇 LLM 챗봇	(L)	Attractiveness, Perspicuity, Efficiency, Dependability

61) op.cit

62) Nasyiah, M., Kelana, B., & Riskinato, A. (2024). System Usability Scale for Measuring Usability of Social Network Applications from User Perspectives. In E3S Web of Conferences (Vol. 483, p. 03010). EDP Sciences.

A Complete Survey on LLM-based AI Chatbots[71]	Sumit Kumar Dam외 3인 (2024)	LLM	LLM	(L)	Knowledge Recency, Logical Reasoning, Hallucination, Transparency, Bias, Privacy Risks, Unfairness, Academic Misuse, Over-reliance, Distribution of Wrong Information
Personalized Response with Generative AIImproving Customer Interaction with Zero-Shot Learning LLM Chatbots[72]	Bink, Joëlle M(2023)	LLM	AI 챗봇, GPT4	(L)	Similarity, Readability, Complexity, Sentiment
OpineBot: Class Feedback Reimagined Using a Conversational LLM[73]	Henansh Tanwar외 3인 (2024)	LLM	OpineBot	(E)	Enhanced Engagement, Relevance of Follow-Up, Questions, System Clarity, Course-Specific Feedback, Cognitive Involvement, Personalized Experience, Usability and Access
Battling Botpoop using GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Learning[74]	Maung Thway외 4인 (2024)	GenAI	Professor Leodar라는 맞춤형 대화 증강 생성(RAG) 챗봇	(E)	Usefulness, Effectiveness, Clear explanations, Personalized
AI literacy for ethical use of chatbot: Will students accept AI ethics?[75]	Yusuke Kajiwara, Kouhei Kawabata(2024)	LLM	ChatGPT	(E)	Usefulness, Fairness, Privacy, Data protection
Effects of Generative Chatbots in Higher Education[76]	Galina Ilieva(2023)	ChatGPT	ChatGPT	(E)	Personalized, Self-directed

63) Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The chatbot usability scale: the

design and pilot of a usability scale for interaction with AI-based conversational agents. Personal and ubiquitous computing, 26, 95-119.

64) Bubaš, G., Čižmešija, A., & Kovačić, A. (2023). Development of an assessment scale for measurement of usability and user experience characteristics of Bing chat conversational AI. Future Internet, 16(1), 4.

65) Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., ⋯ & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Medicine, 7(1), 82.

66) Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G., Telari, A., ⋯ & Ognibene, D. (2023, November). Learning to prompt in the classroom to understand AI limits: a pilot study. In International Conference of the Italian Association for Artificial Intelligence (pp. 481-496). Cham: Springer Nature Switzerland.

67) Silvestri, C., Roshal, J., Shah, M., Widmann, W. D., Townsend, C., Brian, R., ⋯ & Sathe, T. S. (2024). Evaluation of a Novel Large Language Model (LLM) Powered Chatbot for Oral-Boards Scenarios. medRxiv, 2024-05.

68) Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023, July). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 610-625).

69) Dortheimer, J., Martelaro, N., Sprecher, A., & Schubert, G. (2024). Evaluating large-language-model chatbots to engage communities in large-scale design projects. AI EDAM, 38, e4.

70) Freire, S. K., Wang, C., & Niforatos, E. (2024). Chatbots in knowledge-intensive contexts: Comparing intent and llm-based systems. arXiv preprint arXiv:2402.04955.

71) op.cit

72) op.cit

73) Kumar, D. (2024). OpineBot: Class Feedback Reimagined Using a Conversational LLM.

74) Thway, M., Recatala-Gomez, J., Lim, F. S., Hippalgaonkar, K., & Ng, L. W. (2023). Battling Botpoop using GenAI for higher education: A study of a retrieval augmented generation chatbots impact on learning. 2023.

75) Kajiwara, Y., & Kawabata, K. (2024). AI Literacy for Ethical use of Chatbot: Will Students accept AI Ethics?. Computers and Education: Artificial Intelligence, 100251.

76) Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M., & Angelov, D. (2023). Effects of generative chatbots in higher education.

【표 4-3】 사용성 평가 지표 논문 수집 결과

LLM 기반 챗봇(G)/(L)	기존 AI기반 챗봇(C)	교육 분야 챗봇(E)
간결함, 감정, 공감, 과도한 의존, 복잡성, 부적절한 콘텐츠, 불공정성, 인지된 용이성, 유연성, 사회적 실재감, 사회적 책임, 상호작용 관리 능력, 서비스 품질, 조작의 편리성, 위험 인식, 유사성, 의미 차원적 실용성, 의인화, 인식보다는 기억, 일반화, 잘못된 정보의 배포, 전반적인 품질, 주제 중심의 논의를 유지하는 능력, 정보 제공, 정보 품질, 정체성 위협에 대한 인식 수준, 지속 사용 의도, 최신성, 투명성, 편향, 피드백, 피해 가능성, 학문적 오용, 해석 가능성, 환각	검색성, 견고성, 기술 지원, 대화를 시작하기 쉬움, 도움말 제공, 반응성, 사용 빈도, 사용자 목표와 의도의 인식 및 촉진, 사용자 프라이버시와 보안, 유연성, 응답 속도, 인식된 대화, 자신감, 자율성, 직관성, 친숙성, 학습 속도, 학습 필요성, 효용성	가독성, 개인 맞춤형, 공감, 교수 자료, 교육, 목표 설정 및 학습 진행 상황 모니터링, 행동적 몰입, 인지적 몰입, 부정적 감정, 사회적 신호에 적절한 대응, 수행 기대감, 수행 목표, 수행 자기효능감, 시스템 명확성, 시스템과 실세계의 일치, 영향, 정서적 몰입, 의미와 의도 감지, 유쾌한 성격 제공, 인지된 스킬, 인지된 어려움, 인지적 참여, 자기 주도적, 자연스러운 대화, 제어의 자유권, 참여 촉진, 특정 질문에 대한 답변 가능, 학문적 피드백, 학습 동기, 학습 절차와 방법, 학습 콘텐츠 추천, 학습자 (맞춤성)
기능성, 만족도, 매력성, 명확성, 학습 용이성, 신뢰성, 안전성, 조작 편의성, 효율성		–
–	심미성, 오류 방지, 태도	
가시성, 몰입성, 사용 의도, 사용성, 실패에 대한 회복력, 유용성, 즐거움, 이해 용이성, 일관성, 접근성, 정확성, 타당성, 프라이버시, 효과성		

(G)/(L)	36
(C)	20
(E)	31
(G)/(L)+(C)	9
(E)+(C)	3
(G)/(L)+(C)+(L)	14
총합(개)	113

3. 사용성 평가 지표 체계 구성

1) 지표 체계 정립 과정

정리된 113개의 지표 요소를 SOR 이론에 따라 분류한 뒤, 어피니티 다이어그램 방법론을 활용한 1차 분석을 진행하였다. 그 결과 11개의 주지표와 40개의 상세 지표가 도출되었다. 이후 SOR 이론의 자극(S)은 서비스 자체를 평가하는 지표와 관련된다고 판단하여 사용성(Usability)으로, 유기체(O)는 평가 대상의 기능적/감정적 가치와 만족감과 관련 있다고 판단하여 사용자 가치(User Value)로, 반응(R)은 사용자의 서비스를 수용 여부, 지속 사용 여부와 관련있다고 판단하여 사용자 수용도(User Acceptance)로 재정의하였다.

이후 FGI를 통한 2차 분석에서도 어피니티 다이어그램 방법론을 활용했다. 그 결과 그 결과 11개의 주지표와 31개의 상세 지표로 수정 및 정리되었다. 2차 분석에서는 기존 문헌에서 활용된 정의에 기반하여 LLM, 교육 분야 지표도 구분하였다. 추출된 'LLM' 특화 지표는 주지표인 사회적 실재감과 상세 지표인 최신성, 환각, 의인화, 유연성, 공감, 인간다운 자연스러움으로 총 7개다. '교육 분야' 특화 지표는 주지표인 교육적 상호작용성과, 상세 지표인 개인 맞춤성, 학습자 인지성, 학습 정보 제시성, 피드백, 이해 가능성, 몰입성, 자기 주도성, 학습 동기로 총 9개이다.

사용성 평가 지표 개발 단계의 마지막 3차 정리를 위해 약 6일간 수집된 전문가 의견을 종합해 본 결과, 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

① LLM의 주요 특징인 맥락 일치를 강조해야 한다는 의견에 따라, 기존에 유사한 의미를 내포하고 있는 일관성을 ‘맥락 적합성’으로 수정하고 LLM 특화 지표로 변경하였다.

② 챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용 하는 정도라는 정의가 유연성이라는 지표와 어울리지 않는다는 의견에 따라 지표명을 ‘적응성’으로 수정하였다.

③ 기능적 가치의 상세 지표인 효과성, 효율성, 신뢰성이 상세 지표보다 상위 개념의 지표로 느껴진다는 의견에 따라 지표명을 ‘학습 효과성’, ‘학습 효율성’, ‘학습 신뢰성’으로 수정하였다.

④ 다른 상세 지표명과 지표명의 수준에 차이가 느껴진다는 의견에 따라 학습 동기를 ‘학습 동기부여’로 변경하였다.

b. 의미가 모호한 지표 정의 수정

⑤ 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.

c. 상세 지표 그룹 이동

⑥ 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.

d. 지표 추가 및 삭제

⑦ 접근성의 상세 지표인 물리적 접근성이 너무 많은 의미를 내포하고 있다는 의견에 따라, ‘신체적 접근성’과 ‘환경적 접근성’으로 지표를 분리하였다.

⑧ 챗봇 서비스가 생성한 부정확한 정보를 사용자가 잘못 활용하는 것을 방지하는 태도를 의미하는 오용 방지의 평가 내용이 모호하다는 의견에 따라 지표를 삭제하였다.

⑨ 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.

그 결과 주지표 11개, 상세 지표 31개로 최종 수정 및 정리되었다. 3차 분석을 통해 추출된 ‘LLM’ 특화 지표는 주지표인 사회적 실재감과 상세 지표 맥락 적합성, 최신성, 환각 방지, 의인화, 적응성, 공감, 인간다운 자연스러움, 친밀감으로 총 9개이다. ‘교육 분야’ 특화 지표는 주지표인 교육적 상호작용성과 개인 맞춤성, 학습 정보 제시성, 피드백, 몰입성, 학습 동기부여, 자기 주도성, 이해 가능성, 학습 효과성, 학습 효율성, 학습 신뢰성, 자기효능감으로 총 12개이다.

2) 지표 체계 정립 결과

본 연구는 사용성 평가 지표 개발 단계를 거쳐 LLM 기반 교육용 사용성 평가 지표 체계를 구성하였다.

【표 4-4】 사용성 평가 지표 및 정의 1차 도출

카테고리	주지표	주지표 정의	상세 지표	상세 지표 정의
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 서비스가 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)에 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 서비스를 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 절절히 대처하는 정도
			U.4.3. 윤리성 (Ethicality)	챗봇 서비스가 제공하는 정보나 상호작용 과정에서 불공정한 편향과 차별이 없는 정도
	U.5. 사회적 실재감 (Social Presence)	챗봇 서비스가 사용자와 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용 하는 정도	U.5.1. 의인화 (Personification)	챗봇 서비스가 상황과 태스크에 따라 적절한 인격이 투영되어 의인화된 정도
			U.5.2. 적응성 (Adaptiveness)	챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용 하는 정도
			U.5.3. 공감 (Empathy)	챗봇 서비스가 사용자에게 인지적/감정적으로 공감하여 상호작용 하는 정도
			U.5.4. 인간다운 자연스러움 (Human Naturalness)	챗봇 서비스와의 상호작용이 실제 인안같이 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

사용자 가치 (User Value)	기능적 가치 (Functional Value)	챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도	학습 효과성 (Learning Effectiveness)	챗봇 서비스를 통해 교육(학습) 목적 달성의 효과를 얻는 정도
			학습 효율성 (Learning Efficiency)	챗봇 서비스를 통해 교육(학습)을 효율적으로 하는 정도
			학습 신뢰성 (Learning Credibility)	챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있는 정도
	감정적 가치 (Emotional Value)	챗봇 서비스 사용시 기능적으로 얻는 혜택과 가치	즐거움 (Enjoyment)	챗봇 서비스 사용 경험에서 즐거움과 흥미를 느끼는 정도
			심미성 (Aesthetics)	챗봇 서비스 사용 경험에서 심미성을 느끼는 정도
			친밀감 (Intimacy)	챗봇 서비스 사용 경험에서 심리적 친밀감을 느끼는 정도
			자기효능감 (Self Efficacy)	챗봇 서비스 사용 경험에서 본인의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻는 정도
사용자 수용도 (User Acceptance)	만족도 (Satisfaction)	챗봇 서비스의 전반적인 경험에 대해 만족하는 정도		
	태도 (Attitude)	챗봇 서비스 경험을 긍정적으로 생각하는 정도		
	지속 사용 의도 (Continuous Intention)	챗봇 서비스를 지속적으로 사용하고자 하는 정도		

총 세 가지 카테고리 Usability, User Value, User Acceptance로 지표 체계를 구분된다. U. Usability는 서비스 자체를 평가하는 지표가 포함되어 있으며 주지표 6개, 상세 지표 24개로 구성되어 있다. V. User Value는 사용자가 얻게 되는 감성적인 가치를 평가하는 지표가 포함되어 있으며 주지표 2개 상세 지표 7개로 구성되어 있다. A. User Acceptance는 서비스의 지속 사용 여부를 평가하는 지표가 포함되어 있으며 주지표 3개로 구성된다. 다음은 사용성 평가 지표 체계에 대한 정의 및 설명이다.

[주지표 정의 및 설명]

U.1. 정보 전달성 (Information Delivery)

정보 전달성은 LLM 특성을 포함하는 주지표로, 챗봇 서비스가 제공하는 정보가 의미상으로 명확하게 전달되고 있는지를 의미한다. 챗봇이 전달하는 텍스트, 이미지, 음성 등을 사용자가 한 번에 이해 가능하다면 정보 전달성이 높게 평가될 수 있다.

U.2. 시각적 전달성 (Visual Delivery)

시각적 전달성은 챗봇 서비스가 제공하는 정보가 시각적으로 명확하게 전달되고 있는 정도를 의미한다. 챗봇의 진행 상황, 대화 진행 중 발생한 문제 등을 챗봇이 제공하는 시각적 표현을 통해 사용자가 즉각적으로 인지할 수 있다면 시각적 전달성이 높게 평가될 수 있다.

U.3. 접근성 (Accessibility)

접근성은 챗봇 서비스가 사용자의 신체적/환경적/인지적 수준의 제약 없이 시작할 수 있는 정도를 의미한다. 사용자가 챗봇을 시작할 때 특정한 제약으로

인해 시작하기 어렵다고 느끼지 않는다면 접근성이 높게 평가될 수 있다.

U.4. 안전성 (Safety)

안전성은 챗봇 서비스가 제공하는 정보나 상호작용 과정이 안전하다고 지각되는 정도를 의미한다. 인공지능 기술에 의해 초래될 수 있는 개인 정보 침해와 권리 위반 같은 문제가 없다면 안전성이 높게 평가될 수 있다.

U.5. 사회적 실재감 (Social Presence)

사회적 실재감은 LLM 특성을 포함하는 주지표로, 챗봇 서비스가 사용자와 사회적/감정적으로 자연스러운 실재감을 제공하는 상호작용을 하는 정도를 의미한다. 챗봇이 사용자와 상호작용 하는 과정에서 이질감 없이 실제 사람과 대화하는 것과 유사하다고 느낄수록 사회적 실재감이 높게 평가될 수 있다.

U.6. 교육적 상호작용성 (Educational Interaction)

교육적 상호작용은 교육 분야 특성을 포함하는 주지표로, 챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도를 의미한다. 챗봇이 사용자의 교육(학습) 목적을 달성하기 위해 기여하고 있음을 느꼈다면 교육적 상호작용이 높게 평가될 수 있다.

V.1. 기능적 가치 (Functional Value)

기능적 가치는 챗봇 서비스 사용시 기능적으로 얻는 혜택과 가치를 의미한다. 사용자가 챗봇을 사용함으로써 교육(학습)의 목적을 효과적으로 달성했다고 느꼈다면 기능적 가치가 높게 평가될 수 있다. 즉 챗봇이기에 얻을 수 있는 혜택과 가치가 많다고 느낄수록 그 가치는 높다.

V.2. 감정적 가치 (Emotional Value)

감정적 가치는 LLM 특성과 교육 분야 특성을 모두 포함하는 주지표로, 챗봇 서비스 사용시 감정적으로 얻는 혜택과 가치를 의미한다. 사용자가 챗봇의 상호작용 요소로부터 심리적 만족감을 느꼈다면 감정적 가치가 높게 평가될 수 있다.

A.1 만족도 (Satisfaction)

만족도는 챗봇 서비스의 전반적인 경험에 대해 만족하는 정도를 의미한다. 챗봇과의 전반적인 상호작용이 사용자가 기대했던 수준에 충족할수록 만족도는 높게 평가될 수 있다.

A.2. 태도 (Attitude)

태도는 챗봇 서비스 경험을 긍정적으로 생각하는 정도를 의미한다. 전반적인 서비스의 내용과 수준이 적당하다고 느껴졌다면 태도는 높게 평가될 것이다.

A.3 지속 사용 의도 (Continuous Intention)

지속 사용 의도는 챗봇 서비스를 지속적으로 사용하고자 하는 정도를 의미한다. 챗봇 서비스를 통해 원하는 목적에 달성할 수 있을 것이라고 느껴진다면 지속 사용 의도는 높게 평가될 것이다.

[상세 지표 정의 및 설명]

U.1.1. 명료성 (Clarity)

명료성은 챗봇 서비스가 제공하는 정보가 명확하고 간결하게 의미를 전달하는 정도를 의미한다. 의미 없이 긴 문장을 구사하지 않고 문단의 구분을 명확하게 하고 있다면 명료성은 높게 평가될 것이다.

U.1.2. 투명성 (Transparency)

투명성은 챗봇 서비스가 제공하는 정보가 시스템의 의미를 투명하게(사실대로) 전달하는 정도를 의미한다. 챗봇이 전달하는 내용의 출처가 명확하게 제공되고 있다면 투명성이 높게 평가될 수 있다.

U.1.3. 맥락 적합성 (Contextual Conformity)

맥락 적합성은 LLM 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스가 제공하는 정보가 특정 주제나 대화의 맥락에 적합한 내용을 전달하는 정도를 의미한다. 챗봇이 단순히 문장을 나열하지 않고 이전 대화 내용을 고려하여 흐름에 맞는 적절한 응답을 생성할수록 높게 평가될 수 있다. 챗봇이 언어 맥락을 이해하는 것은 LLM의 중요한 특성이므로 맥락 적합성이 높게 평가될수록 사용성이 높은 챗봇이라고 할 수 있다.

U.1.4. 최신성 (Up to Dateness)

최신성은 LLM 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스가 제공하는 정보가 최신 정보를 반영하고 있는 정도를 의미한다. 챗봇이 정치, 경제, 사회, 문화, 라이프, 스포츠 등의 최신 정보를 반영하여 대화를 진행한다면 최신성이 높게 평가될 수 있다.

U.1.5. 환각 방지 (Hallucination Prevention)

환각 방지는 LLM 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스가 부정확하거나 신뢰할 수 없는 정보를 생성하지 않는 정도를 의미한다. '세종대왕 맥북프로 던짐 사건'과 같은 역사적 사실에 근거가 없는 내용을 사실처럼 전달할 때 챗봇이 환각 반응을 하고 있다고 말한다. 챗봇은 잘못된 데이터를 학습(데이터에 의한 환각)하거나 잘못된 추론(학습 및 추론에 의한 환각)할 때 환각 현상이 발생할 수 있다.[77] 챗봇이 사실과 허구를 구분하는 능력이 있고, 사용자를 속이지 않고 있다고 느껴진다면 환각 방지는 높게 평가될 수 있다.

U.2.1. 가시성 (Visibility)

가시성은 챗봇 서비스가 제공하는 정보의 시각적 표현이 명확하게 전달되고 있는 정도를 의미한다. 챗봇의 모든 기능이 눈에 잘 띄게 설계되어 있고, 그 의미를 한 번에 명확히 파악할 수 있다면 가시성이 높게 평가될 수 있다.

U.2.2. 직관성 (Intuitiveness)

직관성은 챗봇 서비스가 제공하는 정보의 시각적 표현이 직관적으로 이해하기 쉬운 정도를 의미한다. 접속이 지연되고 있는 상황, 음성이 인식되고 있는 상황, 챗봇이 음성을 인식하고 대답하고 있는 상황 등 챗봇이 작동하면서 제공하는 모든 시각적 표현이 사고를 거치지 않고 한 번에 무엇을 의미하는지 쉽게 알아차릴 수 있다면 직관성이 높게 평가될 수 있다.

U.3.1 신체적 접근성 (Physical Accessibility)

신체적 접근성은 챗봇 서비스를 신체적 조건 및 수준의 제약 없이 시작할 수 있는 정도를 의미한다. 챗봇이 나이, 성별, 장애 등 신체적 차이에 구애받지 않

77) Kt enterprise (2024). LLM의 환각현상, 어떻게 보완할 수 있을까? https://enterprise.kt.com/bt/dxstory/2521.do

고 챗봇을 사용하는 데 어려움을 느끼지 않는다면 신체적 접근성이 높게 평가될 수 있다.

U.3.2. 환경적 접근성 (Environmental Accessibility)

환경적 접근성은 챗봇 서비스를 시간/장소 등 사용 환경의 제약 없이 시작할 수 있는 정도를 의미한다. 주변 소음으로 인해 음성 인식률이 떨어지는 것을 예로 들 수 있으며, 이처럼 이동 중 혹은 공공 장소와 같은 환경에서 챗봇 사용에 제약이 생기지 않도록 시스템을 구축해 두었다면 환경적 접근성이 높게 평가될 수 있다.

U.3.3. 인지적 접근성 (Cognitive Accessibility)

인지적 접근성은 챗봇 서비스를 인지적 수준(연령/교육 수준 등) 에 제약 없이 시작할 수 있는 정도를 의미한다. 챗봇 시작 전 연령 혹은 교육 수준에 관해 묻거나 테스트하여 이에 맞는 정보나 학습 내용을 제공하는 절차가 있는 경우, 대화 중 사용자가 대화 중 수동으로 난이도를 조절할 수 있는 기능이 있거나 사용자의 수준을 인지하여 챗봇이 직접 난이도를 조절한다고 느껴진다면 인지적 접근성이 높게 평가될 수 있다.

U.4.1. 프라이버시 보호 (Privacy Protection)

프라이버시 보호는 챗봇 서비스가 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하는 정도를 의미한다. 챗봇과의 대화 전중후 과정에서 사용자의 개인 정보가 사전의 동의 없이 수집되거나 악용된다고 느끼지 않는다면 프라이버시 보호는 높게 평가될 수 있다.

U.4.2. 오류 관리 (Error Management)

오류 관리는 챗봇 서비스가 제공하는 정보가 시스템 사용시 발생할 수 있는

오류를 예방하고, 발생 시 적절히 대처하는 정도를 의미한다. 오류가 발생하기 쉬운 상황을 제거하여 사용자의 실수를 예방하는 기능이 있다면 오류 관리가 높게 평가될 수 있다. 예를 들어 사용자가 대화 중이던 방을 나가려고 시도했을 때, 대화 종료 여부를 다시 확인하거나 대화 내용 저장 여부를 묻는 메시지 창이 이에 해당한다.

U.4.3 윤리성 (Ethicality)

윤리성은 챗봇 서비스가 제공하는 정보나 상호작용 과정에서 불공정한 편향과 차별이 없는 정도를 의미한다. 챗봇이 폭력적이거나 차별적인 내용을 전달하지 않는다면 윤리성이 높게 평가될 수 있다.

U.5.1. 의인화 (Personification)

의인화는 LLM 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스가 상황과 태스크에 따라 적절한 인격이 투영되어 의인화된 정도를 의미한다. 챗봇이 제공하는 혹은 대화 상황에 따라 챗봇에 각각의 적합한 페르소나가 투영되어 있다고 느끼면 의인화가 높게 평가될 수 있다.

U.5.2 적응성 (Adaptiveness)

적응성은 LLM 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용 하는 정도를 의미한다. 사용자의 갑작스러운 감정 변화나 환경 변화를 챗봇이 즉각적으로 받아들이고 반응한다면 적응성이 높게 평가될 수 있다.

U.5.3. 공감 (Empathy)

공감은 LLM 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 사용자에게

인지적/감정적으로 공감하여 상호작용하는 정도를 의미한다. 기쁘거나 슬프거나 무서워하는 등의 감정을 챗봇이 즉각적으로 파악하여 공감하는 답변을 제공한다면 공감이 높게 평가될 수 있다. 단, 무분별한 공감은 사용자에게 오히려 반감을 일으킬 수 있기 때문에 적재적소에 공감하는 것이 중요하다.

U.5.4. 인간다운 자연스러움 (Human Naturalness)

인간다운 자연스러움은 LLM 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스와의 상호작용이 실제 인간같이 자연스럽고 이질감 없는 정도를 의미한다. 챗봇이 기계로 인식되지 않고 실제 사람과 대화한다고 느껴져 대화에 완전히 몰입할 수 있다면 인간다운 자연스러움은 높게 평가될 수 있다.

U.6.1. 개인맞춤성 (Personalization)

개인맞춤성은 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 사용자의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 학습 내용을 제공하는 정도를 의미한다. 챗봇이 다양한 수준의 학습내용을 보유하고 제공한다면 개인맞춤성이 높게 평가될 수 있다.

U.6.2. 학습정보 제시성 (Presentation of Learning Information)

학습정보 제시성은 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 교육(학습) 내용 및 진행 상황에 대한 정보를 명확히 제시하는 정도를 의미한다. 챗봇과의 대화 전중후 과정을 사전에 고지하거나 사용자가 진행 상황을 실시간으로 확인 할 수 있다면 학습정보 제시성이 높게 평가될 수 있다.

U.6.3. 피드백 (Feedback)

피드백은 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 교육

(학습)에 필요한 피드백을 적시적소에 제공하는 정도를 의미한다. 대화 중 사용자의 응답을 파악하여 해석 혹은 문법적 오류 등에 대한 적절한 피드백을 준다면 피드백이 높게 평가될 수 있다.

U.6.4. 몰입성 (Immersion)

몰입성은 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있는 정도를 의미한다. 사용자가 챗봇 서비스를 이용함으로서 교육(학습) 내용에 더 깊게 몰입할 수 있었다면 몰입성이 높게 평가될 수 있다.

U.6.5. 학습 동기 부여 (Learning Motivation)

학습 동기 부여는 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하는 정도를 의미한다. 챗봇을 사용함으로서 교육(학습)을 자발적으로 참여하게 되었다면 학습 동기 부여가 높게 평가될 수 있다.

U.6.6 자기 주도성 (Self Directedness)

자기 주도성은 교육 분야 특화 챗봇 사용성 평가 상세지표로, 챗봇 서비스가 자기주도적인 교육(학습) 경험을 제공하는 정도를 의미한다. 교육(학습) 시간 알림 혹은 학습 내용과 학습 시간을 관리할 수 있도록 페이지를 마련하는 등 챗봇이 자기 주도 학습을 유도하고 있다고 느껴진다면 자기 주도성이 높게 평가될 수 있다.

U.6.7. 이해 가능성 (Understandability)

이해 가능성은 교육 분야 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스

가 제동하는 교육(학습) 내용 및 정보를 잘 이해할 수 있는 정도를 의미한다. 교육(학습) 목적에 적합한 내용을 제공한다면 이해 가능성이 높게 평가될 수 있다.

V.1.1. 학습 효과성 (Learning Effectiveness)

학습 효과성은 교육 분야 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스를 통해 교육(학습) 목적 달성의 효과를 얻는 정도를 의미한다. 챗봇 사용으로 학습 수준이 높아졌다고 느껴진다면 학습 효과성이 높게 평가될 수 있다.

V.1.2. 학습 효율성 (Learning Efficiency)

학습 효율성은 교육 분야 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스를 통해 교육(학습)을 효율적으로 하는 정도를 의미한다. 챗봇 사용으로 학습에 불필요한 시간 사용을 줄이고, 학습 효과는 더 늘었다고 느껴진다면 학습 효율성이 높게 평가될 수 있다.

V.1.3. 학습 신뢰성 (Learning Credibility)

학습 신뢰성은 교육 분야 특화 챗봇 사용성 평가 상세 지표로, 챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있는 정도를 의미한다. 챗봇이 제공하는 교육(학습) 내용과 절차가 사용자의 실력을 향상시킬 것이라는 기대하고 있다면 학습 신뢰성이 높게 평가될 수 있다.

V.2.1. 즐거움 (Enjoyment)

즐거움은 챗봇 사용 경험에서 즐거움과 흥미를 느끼는 정도를 의미한다. 챗봇의 전반적인 UX/UI 경험에 만족한다면 즐거움이 높게 평가될 수 있다.

V.2.2. 심미성 (Aesthetics)

심미성은 챗봇 서비스 사용 경험에서 심미성을 느끼는 정도를 의미한다. 챗봇의 전반적인 인터페이스 디자인에 만족한다면 심미성이 높게 평가될 수 있다.

V.2.3. 친밀감 (Intimacy)

친밀감은 챗봇 서비스 사용 경험에서 심리적 친밀감을 느끼는 정도를 의미한다. 챗봇의 전반적인 인터랙션 과정에서 익숙하고 친밀하다고 느꼈다면 친밀감이 높게 평가될 수 있다.

V.2.4. 자기효능감 (Self Efficacy)

자기효능감은 챗봇 서비스 사용 경험에서 본인의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻는 정도를 의미한다. 챗봇 사용으로 교육(학습) 내용에 대한 자신감을 얻었다면 자기효능감이 높게 평가될 수 있다.

4. 소결

본 연구는 LLM기반 교육용 챗봇의 사용성 평가 지표 체계를 도출하기 위해 사용성 평가 지표 개발 단계를 거쳤다. 문헌 분석을 통해 도출된 사용성 평가 요소를 관련 연구원 및 전문가와 총 3차례 분석 및 정리함으로써 사용성 평가 지표 개발을 하였다. 그 결과 3개의 지표 카테고리 안에서 주지표 11개 상세 지표 31개를 도출할 수 있었다. 지표의 카테고리는 크게 3가지, 사용성(Usability), 사용자 가치(User Value), 사용자 수용도(User Acceptance)로 구분된다. 사용성에 속하는 주지표는 '정보 전달성', '시각적 전달성', '접근성', '안전성', '사회적 실재감', '교육적 상호작용성'이 있다. '정보 전달성'의 상세 지표는 총 5개로 '명료성', '투명성'과 LLM 특화 지표인 '맥락 적합성', '최신성', '환각 방지'가 있다. '시각적 전달성'의 상세 지표는 '가시성', '직관성' 총 2개, '접근성'의 상세 지표는 '신체적 접근성', '환경적 접근성', '인지적 접근성' 총 3개, '안전성'의 상세 지표는 '프라이버시 보호', '오류 관리', '윤리성'이 있다. 주지표와 상세 지표가 모두 LLM 특화 지표인 '사회적 실재감'은 '의인화', '적응성', '공감', '인간다운 자연스러움' 총 4개의 상세 지표가 있으며, 주지표와 상세 지표가 모두 교육 분야 특화 지표인 '교육적 상호작용'은 '개인 맞춤성', '학습 정보 제시성', '피드백', '몰입성', '학습 동기부여', '자기 주도성', '이해 가능성' 총 7개의 상세 지표가 있다.

사용자 가치에 속하는 주지표는 '기능적 가치'와 '감정적 가지'가 있다. '기능적 가치'의 상세 지표는 모두 교육 특화 지표로 '학습 효과성', '학습 효율성', '학습 신뢰성' 총 3개이며, '감정적 가치'의 상세 지표는 '즐거움', '심미성'과 LLM 특화 지표인 '친밀감'과 교육 분야 특화 지표인 '자기효능감' 총 4개다. 사용자 수용도에는 주지표만 속해 있으며 '만족도', '태도', '지속 사용 의도'가 있다.

본 연구 단계는 LLM, 교육 분야 특화 지표를 포함하는 LLM기반 교육용 챗봇의 사용성 평가 지표 구성을 도출했다는 점에서 의의가 있다.

Ⅴ. 사용성 평가 지표 검증

1. 연구 방법

본 설문 진행 전 HCI/서비스디자인 관련 연구원 7명을 대상으로 파일럿 테스트를 진행하였다. 설문 응답을 얻기 위해 구글 폼을 활용하여 설문지를 구성하였으며, 사용성 평가 설문의 난이도와 설문지 구성의 적합성을 확인하기 위해 주지표의 세부 질문은 평서형 문장과 7점 리커트 척도로, 상세 지표의 세부 질문은 의문형 문장과 5점 리커트 척도로 구성하였다.

파일럿 테스트를 진행한 결과를 바탕으로 사용성 평가 지표 검증을 위한 설문지를 최종 구성하였다. 사용성 평가 검증은 변수의 상호 관련성에 기초하여 변수의 공통적인 잠재 구조 혹은 차원을 파악하고 설명하는 데 목적이 있다. 이에 따라 본 연구는 앞서 개발 단계에서 도출된 사용성 평가 지표 42개 간의 구조를 파악하여 LLM기반 교육용 사용성 평가 지표를 정립하고자 한다.

본 설문은 2024년 9월 12일 ~ 2024년 10월 4일 총 23일간 진행되었으며 설문 응답을 얻기 위해 구글 폼을 활용하여 설문지를 구성하였다. 연령과 성별에 관계없이 LLM 기반으로 하는 영어 교육용 챗봇 사용자를 대상으로 진행하였으며 설문지는 인구통계학적 특성을 포함한 기본 문항 6개, 주지표와 상세 지표를 포함하는 사용성 평가 문항 42개로 구성하였다. 주지표와 상세 지표 간 위계관계를 이루고 있어 질문 문항이 중복된다고 느낄 수 있으므로 기본 문항 6개, 주지표 문항 11개, 상세 지표 문항 31개 순서로 질문을 분리하여 진행하였다.

본 설문 후 사용성 평가 지표 개발 단계에서 도출한 지표가 유의미하게 그룹지어졌는지 그 구조를 확인하기 위해 탐색적 요인분석(Exploratory Factor

Analysis, EFA)를 진행하였다. EFA는 서로 상관된 많은 변수들 사이의 복잡한 구조를 몇 개의 공통 요인으로 단순화하여 분석하는 통계적 기법으로, 심리학 및 사회과학 분야에서 순서형 리커트 척도로 구성된 설문 항목의 상관 구조를 설명하고자 할 때 주로 사용된다.[78] 본 연구에서는 상세 지표를 변수 공통 요인을 주지표로 설정하여, 주지표와 상세 지표가 적절하게 그룹을 이루고 있는지 확인하였으며, EFA를 위해 통계 소프트웨어인 SPSS(Statistical Package for Social Science)를 사용하였다. SPSS 통계분석 중 사용자 가치와 수용도가 사용성보다 상위 가치로 구분되기 때문에 사용성 카테고리의 상세 지표 문항 24개와 사용자 가치 카테고리에 있는 상제 지표 문항 7개를 구분하여 EFA를 총 2세트 진행하였다.

78) 고광이, 류도향. (2021). 부부의 취업형태에 따른 연령별 가족관계 탐색적 요인분석 - 여성가족패널조사를 중심으로. 한국데이터정보과학회지, 32(1), 169-197, 10.7465/jkdi.2021.32.1.169

2. 사용성 평가 파일럿 테스트 및 설문 구성

【표 5-1】,【표 5-2】는 7명의 HCI/서비스디자인 관련 연구원을 대상으로 진행한 파일럿 테스트 결과이다. 평서형보다는 의문형으로 문항을 구성하는 방식이 답변하는데 더 용이했다는 응답이 57.1%로 더 높았다. 또한 7점 리커트 척도보다 5점 리커트 척도로 문항을 구성하는 방식이 답변하는데 더 용이했다는 응답이 85.7%로 더 높았다. 이를 바탕으로 주지표와 상세 지표의 세부 질문을 최종 수정하고, 5점 리커트 척도 형태로 설문지를 최종 도출하였다.

【표 5-1】 주 지표 세부 질문 추출

주지표	주지표 세부 질문
1. 정보 전달성 (Information Delivery)	챗봇 서비스가 제공하는 정보가 의미상으로 명확하게 전달되고 있나요?
2. 시각적 전달성 (Visual Delivery)	챗봇 서비스가 제공하는 정보가 시각적으로 명확하게 전달되고 있나요?
3. 접근성 (Visual Delivery)	챗봇 서비스는 나의 신체적/환경적/인지적 수준의 제약 없이 시작할 수 있나요?
4. 안전성 (Safety)	챗봇 서비스가 제공하는 정보나 상호작용 과정이 안전한가요?
5. 사회적 실재감 (Social Presence)	챗봇 서비스와의 상호작용이 사회적/감정적으로 자연스러운 실재감을 제공하고 있나요?
6. 교육적 상호작용성 (Educational Interaction)	챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하나요?
7. 기능적 가치 (Functional Value)	챗봇 서비스 사용시 기능적으로 혜택과 가치를 얻나요?
8. 감정적 가치 (Emotional Value)	챗봇 서비스 사용시 감정적으로 혜택과 가치를 얻나요?
9. 만족도 (Satisfaction)	챗봇 서비스의 전반적인 경험에 대해 만족하나요?
10. 태도 (Attitude)	챗봇 서비스 경험을 긍정적으로 생각하나요?
11. 지속 사용 의도 (Continuous Intention)	챗봇 서비스를 지속적으로 사용하고자 하나요?

【표 5-2】 상세 지표 세부 질문 추출

상세 지표	상세 지표 세부 질문
12. 명료성 (Clarity)	챗봇 서비스가 제공하는 정보가 명확하고 간결하게 의미를 전달하고 있나요?
13. 투명성 (Transparency)	챗봇 서비스가 제공하는 정보가 시스템의 상태를 투명하게(사실대로) 전달하고 있나요?
14. 맥락 적합성 (Contextual Conformity)	챗봇 서비스가 제공하는 정보가 특정 주제나 대화의 맥락에 적합한 내용을 전달하고 있나요?
15. 최신성 (Up to Dateness)	챗봇 서비스가 제공하는 정보가 최신 정보를 반영하고 있나요?
16. 환각 방지 (Hallucination Prevention)	챗봇 서비스가 부정확하거나 신뢰할 수 없는 정보를 생성하지 않고 있나요?
17. 가시성 (Visibility)	챗봇 서비스가 제공하는 정보의 시각적 표현이 명확하게 전달되고 있나요?
18. 직관성 (Intuitiveness)	챗봇 서비스가 제공하는 정보의 시각적 표현이 직관적으로 이해하기 쉽나요?
19. 신체적 접근성 (Physical Accessibility)	챗봇 서비스를 신체적 조건 및 수준의 제약 없이 시작할 수 있나요?
20. 환경적 접근성 (Environmental Accessibility)	챗봇 서비스를 시간/장소 등 사용 환경의 제약 없이 시작할 수 있나요?
21. 인지적 접근성 (Cognitive Accessibility)	챗봇 서비스를 인지적 수준(연령/교육 수준 등)에 제약 없이 시작할 수 있나요?
22. 프라이버시 보호 (Privacy Protection)	챗봇 서비스가 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하고 있나요?
23. 오류 관리 (Error Management)	챗봇 서비스가 제공하는 정보가 시스템 사용 시 발생할 수 있는 오류를 예방하고, 발생 시 적절히 대처하고 있나요?
24. 윤리성 (Ethicality)	챗봇 서비스가 제공하는 정보나 상호작용 과정에서 불공정한 편향과 차별이 없나요?
25. 의인화 (Personification)	챗봇 서비스가 상황과 태스크에 따라 적절한 인격이 투영되어 의인화 되었나요?
26. 적응성 (Adaptiveness)	챗봇 서비스가 나의 반응이나 변화를 수용하고 적용하여 상호작용하고 있나요?
27. 공감 (Empathy)	챗봇 서비스가 나에게 인지적/감정적으로 공감하여 상호작용하고 있나요?
28. 인간다운 자연스러움 (Human Naturalness)	챗봇 서비스와의 상호작용이 실제 인간같이 자연스럽고 이질감이 없나요?
29. 개인 맞춤성 (Personalization)	챗봇 서비스가 나의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 교육(학습) 내용을 제공하고 있나요?

30. 학습 정보 제시성 (Presentation of Learning Information)	챗봇 서비스가 교육(학습) 내용 및 진행 상황에 대한 정보를 명확하게 제시하고 있나요?
31. 피드백 (Feedback)	챗봇 서비스가 교육(학습)에 필요한 피드백을 적시 적소에 제공하고 있나요?
32. 몰입성 (Immersion)	챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있나요?
33. 학습 동기부여 (Learning Motivation)	챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하고 있나요?
34. 자기 주도성 (Self Directedness)	챗봇 서비스가 자기 주도적인 교육(학습) 경험을 제공하고 있나요?
35. 이해 가능성 (Understandability)	챗봇 서비스가 제공하는 교육(학습) 과정과 내용을 신뢰할 수 있나요?
36. 학습 효과성 (Learning Effectiveness)	챗봇 서비스를 통해 교육(학습) 목적 달성 효과를 얻고 있나요?
37. 학습 효율성 (Learning Efficiency)	챗봇 서비스를 통해 교육(학습)을 효율적으로 하고 있나요?
38. 학습 신뢰성 (Learning Credibility)	챗봇 서비스의 교육(학습) 과정과 내용을 신뢰할 수 있나요?
39. 즐거움 (Enjoyment)	챗봇 서비스 사용 경험에서 즐거움과 흥미를 느끼나요?
40. 심미성 (Aesthetics)	챗봇 서비스 사용 경험에서 심미성을 느끼나요?
41. 친밀감 (Intimacy)	챗봇 서비스 사용 경험에서 심리적인 친밀감을 느끼나요?
42. 자기효능감 (Self Efficacy)	챗봇 서비스 사용 경험에서 나의 성공적인 교육(학습) 수행 능력에 대한 믿음을 얻고 있다고 느끼나요?

3. 사용성 평가 통계 결과 및 분석

LLM이 도입된 영어 교육용 챗봇 사용자를 대상으로 한 설문 결과 총 204개의 응답을 수집하였다. 다음은 설문 참가자의 인구통계학적 특성이다. 설문 참여자는 남성 101명(49.2%), 여성 103명(50.2%)으로 구성되며, 연령은 10대 2명(0.9%), 20대 128명(62.4%), 30대 59명(28.7%), 40대 13명(6.3%), 50대 이상(0.9%) 2명이다. 이들의 교육 수준의 초등학교 졸업 1명(0.4%), 고등학교 졸업 27명(13.1%), 대학교 졸업 154명(75.1%), 대학원 이상 22명(10.7%)으로 나타났다. 다양한 애플리케이션과 디지털 기기에 대한 친숙도는 전혀 친숙하지 않다 9명(4.3), 친숙하지 않다 16명(7.8%), 보통이다 32명(15.6%), 친숙하다 84명(40.9%), 매우 친숙하다 63명(30.7%)이다. 이처럼 다양한 성별, 연령, 교육 수준, 디지털 기기 친숙도를 가진 대상자를 모집했으며, 접근성 그리고 교육적 상호작용성과 기능성 가치 등 특화 지표를 다방면에서 평가한 내용을 확인할 수 있었다.

영어 교육용 챗봇 서비스 평균 이용 빈도는 지속적으로 사용하지 않음 49명(23.9%), 월 4회 미만 27명(13.1%), 주 1~2회 81명(39.5%), 주 3~4회 31명(15.1%), 주 5~6회 6명(2.9%), 매일 1회 이상 10명(4.8%)이다. 추가로 사용해 본 LLM 모델(ChatGPT 등)이 탑재된 영어 교육용 챗봇 서비스로는 스픽(Speak), 듀오링고 맥스(Duolingo Max), 스피킹맥스-맥스 AI(Speaking Max-Max AI), 멤라이즈(Memrise), 말해보카, ChatGPT 등이 있었으며, 사용안 해 보거나 없다고 응답한 답변이 4개로 나타났다. LLM기반 교육용 챗봇을 사용해 보지 않은 응답자는 평가 대상 기준에서 벗어나기 때문에 이후 진행된 통계분석에서는 사용해 본 서비스가 없다고 응답한 4명을 제외한 200명을 대상으로 진행하였다.

LLM이 도입된 영어 교육용 챗봇 사용성 평가 지표를 정립하기 위해 탐색적

요인분석(EFA)을 실시하였다. 아래 【표 5-3】,【표 5-4】는 EFA를 통해 1차로 추출한 상세 지표 33개의 항목에 대해 항목 구조를 파악한 결과이다. 주성분 분석은 VARIMAX 직교 회전으로 수행되었으며, 요인 추출 기준은 고유값 1 이상으로 설정하였다.

사용성 카테고리의 주지표와 상세 지표 관계를 검증하기 위해 총 6차례의 반복된 EFA 단계를 거쳤다. 우선 사용성 카테고리의 상세 지표인 변수 12 ~ 변수 35를 대상으로 변수 제거 없이, 고유값 기준으로 1차 EFA를 진행하였다. 개발 단계의 상세 지표 그룹과 다르게 공통성을 띠고 있음을 확인한 후, 고정된 요인 수를 고유값 기준으로 추출된 4 이상으로 설정하여 4차례 추가 EFA를 진행하였다【표 부록 5-1】,【표 부록 5-2】. 그 결과 변수 16, 변수 19, 변수 25가 지속적으로 불규칙적으로 구분되었으며, 개발 단계의 상세 지표 그룹과 다른 속성을 띄고 있음을 확인할 수 있었다. 이 점을 고려하여 변수 16, 변수 19, 변수 25를 제외하고 21개의 변수를 대상으로 변수 제거 없이, 교유값 기준으로 EFA를 진행하였다 【표 5-3】.

【표 5-3】은 사용성 카테고리에 있는 상세 지표의 EFA 최종 결과 지표이다. 【표 5-3】의 표본 적합도(MSA)는 0.926으로 나타나 본 자료가 요인분석에 적합하다고 볼 수 있다. 또한 Bartlett의 구형성 검정 결과, 근사 카이제곱이 2667.393, 유의확률은 0.001 미만으로 유의수준 0.005를 기준으로 '만족도 척도' 변수 간의 상관성이 인정되어 전반적으로 요인분석이 가능하다고 할 수 있다. 이에 총 4가지의 요인과 구성요소가 추출되었다. 요인 1은 '가시성', '맥락적합성', '최신성', '직관성', '명료성', '투명성'이 구성요소로 있으며, 요인 2는 '인간다운 자연스러움', '공감', '적응성', '개인 맞춤성', '피드백', '학습 정보 제시성'이, 요인 3은 '인지적 접근성', '프라이버시 보호', '오류 관리', '윤리성', '환경적 접근성'이, 요인 4는 '이해 가능성', '몰입성', '학습 동기부여', '자기 주도성'이 구성요소로 추출되었다.

【표 5-4】는 사용자 가치 카테고리에 있는 상세 지표의 EFA 최종 결과 지표이다. 【표 5-4】의 표본 적합도(MSA)는 0.871로 나타나 본 자료가 요인분석에 적합하다고 볼 수 있다. 또한 Bartlett의 구형성 검정 결과, 근사 카이제곱이 675.952, 유의확률은 0.001 미만으로 유의수준 0.005를 기준으로 '만족도 척도' 변수 간의 상관성이 인정되어 전반적으로 요인분석이 가능하다고 할 수 있다. 이에 총 2가지의 요인과 구성요소가 추출되었다. 요인 5는 '학습 효과성', '학습 효율성', '학습 신뢰성', '즐거움'이, 요인 6은 '심미성', '친밀감', '자기효능감'이 구성요소로 추출되었다.

【표 5-3】 최종 EFA결과: 사용성 카테고리

KMO의 표본 적합도(MSA) m 검정							
Bartlett의 구형성 검정	근사 카이제곱	2667.393					
	자유도	210					
	유의확률	<.001					
요인	요소	성 분				공통성	
		1	2	3	4		
요인 1	17. 가시성 (Visibility)	0.763	0.193	0.175	0.253	0.714	
	14. 맥락 적합성 (Contextual Conformity)	0.753	0.285	0.181	0.222	0.731	
	15. 최신성 (Up to Dateness)	0.746	0.252	0.181	0.149	0.676	
	18. 직관성 (Intuitiveness)	0.702	0.227	0.163	0.245	0.630	
	12. 명료성 (Clarity)	0.702	0.239	0.116	0.392	0.716	
	13. 투명성 (Transparency)	0.686	0.261	0.156	0.184	0.597	
요인 2	28. 인간다운 자연스러움 (Human Naturalness)	0.310	0.776	0.183	0.058	0.735	
	27. 공감 (Empathy)	0.238	0.765	0.162	0.093	0.677	
	26. 적응성 (Adaptiveness)	0.344	0.650	0.214	0.265	0.657	

	29. 개인 맞춤성 (Personalization)	0.185	0.618	0.160	0.411	0.610
	31. 피드백 (Feedback)	0.290	0.608	0.117	0.448	0.668
	30. 학습 정보 제시성 (Presentation of Learning)	0.254	0.568	0.152	0.419	0.606
요인 3	21. 인지적 접근성 (Cognitive Accessibility)	0.044	0.041	0.816	0.210	0.714
	22. 프라이버시 보호 (Privacy Protection)	0.189	0.110	0.815	0.162	0.738
	23. 오류 관리 (Error Management)	0.344	0.234	0.706	0.119	0.686
	24. 윤리성 (Ethicality)	0.258	0.110	0.696	0.163	0.590
	20. 환경적 접근성 (Environmental Accessibility)	0.031	0.251	0.627	0.069	0.462
요인 4	35. 이해 가능성 (Understandability)	0.295	0.009	0.293	0.761	0.753
	32. 몰입성 (Immersion)	0.229	0.280	0.177	0.708	0.664
	33. 학습 동기부여 (Learning Motivation)	0.288	0.315	0.191	0.703	0.714
	34. 자기 주도성 (Self Directedness)	0.328	0.337	0.153	0.687	0.716

【표 5-4】 최종 EFA결과: 사용자 가치 카테고리

KMO의 표본 적합도(MSA) m 검정			
Bartlett의 구형성 검정	근사 카이제곱	675.952	
	자유도	210	
	유의확률	<.001	

요인	요소	성 분		공통성
		1	2	
요인 5	36. 학습 효과성 (Learning Effectiveness)	0.843	0.190	0.746
	37. 학습 효율성 (Learning Efficiency)	0.842	0.219	0.756
	38. 학습 신뢰성 (Learning Credibility)	0.707	0.427	0.600
	39. 즐거움 (Enjoyment)	0.629	0.452	0.682
요인 6	40. 심미성 (Aesthetics)	0.174	0.879	0.716
	41. 친밀감 (Intimacy)	0.289	0.770	0.803
	42. 자기효능감 (Self Efficacy)	0.404	0.744	0.676

4. LLM기반 교육용 챗봇 사용성 평가 지표 최종 도출

본 연구는 통계분석을 마지막으로 LLM기반 교육용 챗봇의 사용성 평가 지표 세트를 최종 도출하였다. 사용성 평가 지표 개발 단계에서 주지표 11개, 상세지표 31개였던 사용성 평가 지표는 사용성 평가 지표 검증 단계를 거쳐 주지표 9개, 상세지표 28개로 축소되어 최종 도출되었다.

기존 문헌 연구를 통한 조사 항목 등을 결합한 내용을 기반으로 주지표 영역인 요인 1~6의 명칭을 재정의 하였다. 요인 1의 이름은 '효과적인 전달성' 요인 2의 이름은 '신뢰가능한 상호작용', 요인 3의 이름은 '사회적 실재감', 요인 4의 이름은 '교육적 상호작용성'으로 정의한다. 요인 5의 이름은 '감정적 가치', 요인 6의 이름은 '기능적 가치'

요인 1. '효과적인 전달성'은 챗봇 서비스가 제공하는 정보가 의미적/시각적으로 명확하게 전달되고 있는 정도를 의미한다. 사용성 평가 지표 개발 단계에서 정보 전달성(주지표)의 상세 지표인 '환각 방지'가 제거되고 '명료성', '투명성', '맥락 적합성', '최신성'과 시각적 전달성(주지표)의 상세 지표인 '가시성', '직관성'이 합쳐져 변경된 결과이다. EFA 분석 과정에서 제거된 '환각 방지'는 가상의 상황을 기반으로 대화를 통해 학습을 유도하는 영어 회화 챗봇의 경우, 해당 지표가 사용성 평가에 적합하지 않은 것으로 판단되어 제외된 것으로 해석할 수 있다. 사용성을 평가하는 데 어울리지 않기 때문에 제거되었다고 해석할 수 있다. 또한 '효과적인 전달성'이라는 명칭은 단순히 정보 제공 여부를 넘어서, 사용자와 챗봇 간의 상호작용에서 정보가 효과적으로 이해되고 수용될 수 있는지를 평가하는 지표로서의 역할을 강조한다. 즉 요인 1은 정보의 전달성과 관련된 세부 지표를 하나로 통합하여, 챗봇의 전반적인 전달 능력을 더욱 종합적으로 반영한다는 특징을 지닌다.

요인 2. '신뢰가능한 상호작용'은 챗봇 서비스가 안전하고 신뢰할 만한 상호

작용을 제공하는 정도를 의미한다. 사용성 평가 지표 개발 단계에서 접근성(주지표)의 상세 지표인 ‘신체적 접근성’이 제거되고 ‘환경적 접근성’, ‘인지적 접근성’과 안전성(주지표)의 상세 지표 ‘프라이버시 보호’, ‘오류 관리’, ‘윤리성’이 합쳐진 결과이다. EFA 분석 과정에서 제거된 ‘신체적 접근성’은 평가 대상이 된 앱 특성상 신체적 제약이 있더라도 음성으로 상호 작용할 수 있으므로 교육용 챗봇의 사용성을 판단하는 데 큰 영향을 미치지 않았다고 해석할 수 있다. 즉 사용자의 신체적 조건은 교육용 챗봇 사용성 평가에 있어 제약이 적다는 것을 의미한다. '신뢰 가능한 상호작용'이라는 명칭은 해당 요인이 챗봇 서비스의 안정성과 신뢰성을 중시하는 평가 항목임을 반영한다. 챗봇과의 상호작용에서 사용자는 정보를 제공받고, 개인적인 데이터를 다루며, 때로는 중요한 결정을 내리게 된다. 따라서 챗봇이 제공하는 상호작용이 사용자가 신뢰할 수 있고, 개인정보보호와 오류 관리가 적절히 이루어지는지를 평가하는 것이 중요하다. 이 명칭은 사용자가 챗봇과의 상호작용에서 경험하는 신뢰성을 핵심 요소로 삼고 있으며, 그 신뢰성은 결국 사용자의 지속적인 이용 여부와 직결된다는 점에서 중요한 의미가 있다.

요인 3. 사회적 실재감은 챗봇 서비스가 사용자와 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용 하는 정도를 의미한다. 사용성 평가 지표 개발 단계에서 사회적 실재감(주지표)의 상세 지표인 ‘의인화’가 제거되고 ‘적응성’, ‘공감’, ‘인간다운 자연스러움’과 교육적 상호작용(주지표)의 일부 상세 지표였던 ‘개인 맞춤성‘, ‘피드백’, 학습 정보 제시성이 합쳐진 결과이다. 이때 ‘학습 정보 제시성’은 LLM 특화 지표인 사회적 실재감(주지표)으로 상위 그룹이 변경되면서 이에 맞춰 지표명을 ‘과정 제시성’으로 수정하였다. ‘과정 제시성’은 챗봇 서비스가 진행 내용 및 상황을 자연스럽게 느끼도록 정보를 명확하게 제시하는 정도를 의미한다. EFA 분석 과정에서 ‘의인화’가 제거된 이유는 교육용 챗봇의 특성상 챗봇이 인간처럼 형상화되는 것보다 오히려 사람처럼 대화하는

것이 중요하기 때문이라고 해석할 수 있다.

요인 4. 교육적 상호작용은 사용성 평가 지표 개발 단계에서 교육적 상호작용성(주지표)의 일부 세부 지표인 ‘몰입성’, ‘학습 동기부여’, ‘자기 주도성’, ‘이해 가능성’이 요소로 있다.

요인 5. 기능적 가치는 챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공하는 정도를 의미한다. 사용성 평가 지표 개발 단계에서 기능적 가치(주지표)의 상세 지표인 ‘학습 효과성’, ‘학습 효율성’, ‘학습 신뢰성’과 감정적 가치(주지표)의 상세 지표인 즐거움이 유사성을 보이며 같은 요인으로 묶였다. 이때 즐거움이 교육 분야 특화 지표인 기능적 가치로 상위 그룹이 변경되면서 이에 맞춰 지표명을 ‘학습 흥미성’으로 수정하였다. ‘학습 흥미성’은 챗봇 서비스의 교육(학습) 과정에서 즐거움과 흥미를 느끼는 정도를 의미한다.

요인 6. 감정적 가치는 챗봇 서비스 사용시 기능적으로 얻는 혜택과 가치를 의미한다. 사용성 평가 지표 개발 단계에서 감정적 가치(주지표)의 상세 지표인 즐거움이 분리되어, ‘심미성’, ‘친밀감’, ‘자기효능감’이 요소로 있다.

【표 5-5】 사용성 평가 최종 지표 및 정의

카테고리	주지표	주지표 정의	상세 지표	상세 지표 정의
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. 신뢰가능한 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)에 제약 없이 시작할 수 있는 정도
			U.2.3. 프라이버시 보호 (Privacy Protection)	챗봇 서비스를 제공하는 정보가 개인 정보 침해를 방지하고 프라이버시를 보호하는 정도
			U.2.4. 오류 관리 (Error Management)	챗봇 서비스가 제공하는 정보가 시스템 사용 시 발생할 수 있는 오류를 예방하고, 오류발생 시 절절히 대처하는 정도
			U.2.5. 윤리성 (Ethicality)	챗봇 서비스가 제공하는 정보나 상호작용 과정에서 불공정한 편향과 차별이 없는 정도

	U.3. 사회적 실재감 (Social Presence)	챗봇 서비스가 사용자와 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용 하는 정도	U.3.4. 개인 맞춤성 (Personalization)	챗봇 서비스가 사용자의 개별 교육(학습) 수준 및 진행 상황에 맞춤화된 학습 내용을 제공하는 정도
			U.3.5. 과정 제시성 (Presentation of Progress)	챗봇 서비스가 진행 내용 및 상황을 자연스럽게 느끼도록 정보를 명확하게 제시하는 정도
			U.3.6. 피드백 (Feedback)	챗봇 서비스가 사용자에게 필요한 피드백을 적시 적소에 제공하는 정도
			U.3.1. 적응성 (Adaptiveness)	챗봇 서비스가 사용자의 반응이나 변화를 수용하고 적용하여 상호작용 하는 정도
			U.3.2. 공감 (Empathy)	챗봇 서비스가 사용자에게 인지적/감정적으로 공감하여 상호작용 하는 정도
			U.3.3. 인간다운 자연스러움 (Human Naturalness)	챗봇 서비스가 상황과 태스크에 따라 적절한 인격이 투영되어 실제 인간같이 자연스럽고 이질감이 없는 정도
	U.4. 교육적 상호작용성 (Educational Interaction)	챗봇 서비스가 학습의 효과적인 교육적 상호작용을 제공하는 정도	U.4.1. 몰입성 (Immersion)	챗봇 서비스가 제공하는 교육(학습) 활동에 몰입할 수 있는 정도
			U.4.2. 학습 동기부여 (Learning Motivation)	챗봇 서비스가 교육(학습) 동기를 효과적으로 부여하는 정도
			U.4.3.자기 주도성 (Self Directedness)	챗봇 서비스가 자기 주도적인 교육(학습) 경험을 제공하는 정도
			U.4.4. 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 사용시 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)	챗봇 서비스 사용시 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		

5. 소결

본 연구는 사용성 평가 지표 검증 과정을 거쳐 LLM기반 교육용 챗봇의 최종 사용성 평가 지표를 최종 도출하였다. 지표 검증 과정은 파일럿 테스트와 본 설문 및 분석 총 두 단계를 거쳤으며, 파일럿 테스트를 통해서 설문지를 완성하고 본 설문과 통계적 분석을 통해 사용성 평가 지표 개발 단계에서 도출된 11개의 주지표와 31개의 상세 지표 그루핑의 적합성을 확인하였다. 그 결과 11개로 도출되었던 주지표는 상세지표 수정 내용에 따라 '효과적 전달성', '신뢰가능한 상호작용', '사회적 실재감', '교육적 상호작용성', '기능적 가치', '감정적 가치', '만족도', '태도', '지속 사용 의도' 총 9개의 주지표로 축소되었다. 더하여 상세 지표 중 '환각 방지', '신체적 접근성', '의인화' 총 3개의 지표가 삭제되어 상세 지표는 31개에서 총 28개로 축소되었다.

'효과적 전달성'의 상세 지표는 '명료성', '투명성', '맥락 적합성', '최신성', '가시성', '직관성' 총 6개, '신뢰가능한 상호작용'의 상세 지표는 '환경적 접근성', '인지적 접근성''프라이버시 보호', '오류 관리', '윤리성' 총 5개, '사회적 실재감'의 상세 지표는 '개인 맞춤성', '과정 제시성', '피드백', '적응성', '공감', '인간다운 자연스러움' 총 6개, '교육적 상호작용성' 상세 지표는 '몰입성', '학습 동기부여', '자기 주도성', '이해 가능성' 총 4개, '기능적 가치' 상세 지표는 '학습 효과성', '학습 효율성', '학습 신뢰성', '학습 흥미성' 총 4개, '감정적 가치' 상세 지표는 '심미성', '친밀감', '자기효능감' 총 3개로 도출되었다.

본 연구 단계는 LLM기반 교육용 챗봇 사용성 평가 지표를 통계 기반으로 분석하고 검증했다는 점에서 의의가 있다.

Ⅵ. 사용성 평가 지표 적용

1. 연구 방법

본 연구는 LLM기반 교육용 사용성 평가 지표 개발이라는 연구 목적을 달성하기 위해 사용성 평가 지표 적용 단계를 거치고 있다. 사용성 평가 지표 적용 단계는 앞서 사용성 평가 지표 개발, 검증 단계에서 최종 도출한 LLM기반 교육용 챗봇 사용성 평가 지표의 유의성을 확인하는 단계이다. 현재 출시된 영어 회화를 위한 교육용 챗봇 앱 서비스 3가지를 대상으로 사용성 평가를 진행하여, 개발된 사용성 평가 지표가 실제 사용자가 느끼는 앱 서비스 기능 및 요소 각각의 만족도와 유사하게 평가되고 있는지 분석하였다.

2024년 11월 4일 ~ 2024년 11월 6일 총 3일간 HCI/서비스디자인 관련 연구원 10명을 대상으로 사용성 평가 지표 적용을 위한 사용성 평가를 진행했다. 대상 앱 서비스 선정 기준은 LLM기반 챗봇 기능이 있고, 마지막 업데이트가 최근 3개월 이내에 이루어졌으며, 영어 교육할 수 있어야 한다. 세 가지 기준을 모두 충족하면서 서로 다른 인터랙션을 적용하고 있는 스픽(Speak), 프랙티카(Praktika), 멤라이즈(Memrise) 3개를 대상 서비스로 선택하였다. 대상 앱서비스 각각의 특징을 분석한 후 유사한 태스크를 지시하여 앱서비스 별 기능과 서비스에 대한 만족도 차이를 비교적 쉽게 느낄 수 있도록 한다.

사용성 평가 지표 검증 단계에서 최종 도출된 지표 체계를 기반으로 설문지를 재구성하였으며, 설문 응답을 얻기 위해 구글 폼을 활용하였다. 추가로 앱서비스 각각의 전반적인 만족도 원인을 확인하기 위해 각각의 설문 마지막에 주관식으로 서비스에 대한 총평 및 의견을 받았다. 이후 3개 그룹 간 응답을 비교하는 것에 대한 통계적 유의성을 확인하기 위해 SPSS 통계분석 프로그램을

활용하여 분석하였다. One-way ANOVA(일원 배치 분산분석)를 실행하고자 하였으나, 정규성 검정 결과 모든 변수의 유의 확률이 0.050을 넘지 않는 다는 점에서 정규성을 띠지 않고, 표본의 수가 30을 넘지 않으므로 비모수 검정인 프리드먼(Friedman) 검정을 진행하였다. 이후 개별 집단 간 통계적 유의성을 검증하기 위해 사후분석으로 맨 휘트니(Mann-Whitney) 검정을 진행하였다.

2. 사용성 평가 지표 적용 연구 설계

사용성 평가를 진행하기 위해 LLM기반 챗봇 기능이 있는 앱 서비스 중 영어 회화 교육을 위한 앱 서비스 3가지를 선정하고 분석하였다. 2024년 11월 4일 기준 ISO의 앱스토어(APP Store) 교육 카테고리 무료 다운로드 수 2위인 '스픽(Speak)', '19위인 프랙티카(Praktika)'와 앞서 진행된 사용성 평가 대상자들이 사용해 본 LLM기반 영어 교육용 챗봇 서비스로 언급된 '멤라이즈(Memrise)'를 대상 서비스로 선정하였다. 세 서비스 모두 최근 1개월 이내에 업데이트가 진행되었으며 3천 개 이상의 리뷰글을 보유하고 있었다는 점에서 활발하게 사용되고 있는 교육용 앱서비스라고 할 수 있다. 또한 대상 앱 세 가지는 모두 사용자가 AI와 상호작용하며 대화를 주고 받는다는 점에서 공통점을 가지고 있지만 '스픽'은 음성 기반으로 GhatGPT와 같이 말풍선을 주고받으며 대화창 내에서 상호작용하며, '프랙티카'는 음성과 문자를 기반으로 상호작용하며 시각적으로 실재감을 줄 수 있는 AI아바타 기능이 있다. 그리고 '멤라이즈'는 음성과 문자를 기반으로 상호작용하며 스픽과 같이 대화창 내에서 말풍선을 주고받으며 상호작용한다는 점에서 서로다른 인터랙션 방식을 가지고 있다. 다음 【표 6-1】은 세 가지 LLM기반 교육용 챗봇 서비스를 상세하세 분석한 표이다.

스픽(Speak)은 AI 영어 튜터와 대화를 기반으로 하는 스피킹 앱 서비스이다.

스픽의 핵심 기능 중 하나인 프리톡은 ChatGPT의 제작사인 OpenAI의 GPT-4 기술을 적용한 대화형 인터페이스 기반의 챗봇으로, 외국인과 실제로 대화하는 듯한 경험을 할 수 있다. 프리톡은 서비스에서 제공하는 시나리오, 다른 사용자가 만든 시나리오 혹은 '나만의 시나리오 만들기' 기능으로 사용자가 직접 본인의 역할, AI의 역할, 대화 주제 및 상황을 구성하는 대로 시뮬레이션 할 수 있다는 것이 주요 특징이며, 설정된 상황이 마무리되면 대화가 자동으로 종료되며 대화에 대한 종합적인 평가와 이어서 복습할 수 있는 서비스를 제공한다. 프리톡은 다른 두 앱의 챗봇과 달리 대화 시 음성으로만 답변할 수 있으며, 대화 중 챗봇이 제공하는 답변 각각에 대한 피드백이 가능하다. 피드백 시 답변에 대한 부정적 의견만 제시할 수 있으며 '터무니없는 답변', '반복적인 답변', '부적절하거나 불쾌함', '기타' 옵션으로 문제를 신고할 수 있다. 또한 대화 중 사용자 답변의 어휘, 문법에 대한 교정 및 피드백도 즉각적으로 확인할 수 있으며, '왕초보 모드', '고급 어휘 사용' 등 챗봇의 난이도를 변경할 수 있다.

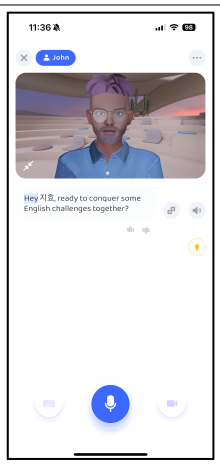
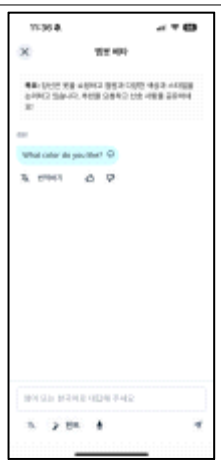
프랙티카(Praktika)는 AI 아바타 튜터와 대화하면서 학습할 수 있는 GPT-4 기반 1:1 영어 과외 앱서비스이다. 약 25명의 국적과 이에 따른 엑센트, 교육 성향 등이 다른 AI 아바타 튜터와 대화하며 영어 회화 학습을 할 수 있다. 마이크 버튼을 눌러 음성으로 대화하거나 키보드 버튼을 누르면 문자로 답변할 수도 있으며, 이때 문법이나 단어, 혹은 발음의 오류를 즉각적으로 수정해 주며 튜터가 오류에 대한 설명을 직접 제공하며 대화를 이끌어나간다. 챗봇 각각의 답변에 대한 피드백은 긍/부정 선택옵션으로 가능하며, 대화 중 난이도 변경은 불가능하다. 프랙티카가 다른 두 앱의 챗봇과 다른 점 두 가지는 AI 아바타와 대화한다는 점과 챗봇의 역할이 튜터로 고정되어 있어, 25개의 아바타 모두 상황에 따라 사용자와 챗봇의 역할이 바뀌는 것이 아닌 튜터와 학생이 대화하는 프로세스로 진행된다는 것이다.

멤라이즈(Memrise) 다양한 언어를 공부할 수 있는 기억 훈련 기반 학습 앱

서비스이다. 비디오 클립, 오디오 발음, 대화형 퀴즈 등 멀티미디어 요소가 통합되어 단어를 학습할 수 있는 서비스로 시작하여, 최근 GPT-3 기술에 기반한 AI 언어 파트너 멤봇(MemBot)이 도입되면서 다양한 주제로 자유롭게 대화할 수 있는 언어학습의 기회를 제공하였다. 주어진 시나리오 중 원하는 주제를 선택하면, 멤봇의 역할과 대화의 구체적인 방향과 목표를 제안해 준다. 멤봇은 문자 인식을 기반으로 하며 마이크 버튼을 눌러 음성 인식도 가능하다. 이때 문법에 대해 즉각적인 교정을 해주지만 이에 대한 설명을 따로 제공해 주지는 않는다. 챗봇 각각의 답변에 대한 피드백은 긍/부정 선택옵션으로 가능하며, 대화 중 난이도 변경은 불가능하다.

이처럼 세 가지 앱서비스는 LLM기반의 대화형 인터페이스를 기반으로 인터랙션 하는 챗봇이지만 서로 다른 인터랙션 방식을 제공하고 있다는 점에서 사용성 평가에 차이가 있을 수 있다. 따라서 분석한 세 가지 앱서비스를 대상으로 앞서 개발한 사용성 평가 지표를 적용하여 사용성 평가 지표의 유의성을 확인하고자 한다.

【표 6-1】 LLM 기반 영어 교육용 챗봇이 도입된 앱 서비스별 특징

		스픽(Speak)	프랙티카(Praktika)	멤라이즈(Memrise)
APP				
소개		개인 맞춤형 AI 영어 튜터와 대화하는 스피킹 앱서비스	3D 아바타 AI 튜터와 말하기, 읽기, 영어단어를 함께 공부할 수 있는 앱서비스	다양한 언어를 공부할 수 있는 기억 훈련 기반 학습 앱 서비스
앱 서비스 출시일		2019년 3월 (AI 챗봇 도입: 2023년 1월)	2022년 5월	2013년 5월 (AI 챗봇 도입: 2023년)
챗봇 명		프리톡	AI 아바타	멤봇(MemBot)
챗봇 기술 수준 (2024.11.29. 기준)		GPT-4	GPT-4	GPT-3
AI 챗봇 매칭 방식		• 서비스에서 생성한 시나리오 선택(Topic) • 사용자가 생성한 시나리오 선택(Community)	• 서비스에서 생성한 시나리오 선택 • 자유주제 대화 선택	• 서비스에서 생성한 시나리오 선택
AI 챗봇 특징		• 사용자가 직접 페르소나를 제작하여 상황 연출 가능 • 대화 종료 후 복습 서비스 자동 제공	• 약 25명의 국적과 성격이 다른 아바타를 튜터로 선택 가능 • 대화 중 AI 아바타가 3D 그래픽으로 등장	• 대화 주제, 목표, 챗봇의 역할을 구체적으로 제공해 줌
챗봇 지원 기능	음성 답변	O	O	O
	문자 답변	X	O	O
	챗봇 답변 피드백	O	O	O
	사용자 답변 피드백	O	O	O
	난이도 변경	O	X	X
실제 대화 화면				

3. 결과 분석 및 비교

본 연구는 SPSS를 이용하여 다음과 같은 프로세스로 통계분석을 진행하였다. 먼저 스픽, 프랙티카, 멤라이즈 세 앱의 평균과 표준편차 결과이다【표 6-2】. 주지표 단독으로 점수를 받았을 때의 평균과 상세 지표를 점수 평균값의 평균을 비교했을 때, 값의 차이가 모두 0.4를 넘지 않고 비슷한 패턴을 갖고 있다는 점에서 개발된 사용성 평가 지표를 활용한 평가가 일관성 있게 잘 되고 있음을 예측해 볼 수 있다.

‘효과적 전달성’은 스픽과 프랙티카가 4.4로 높았고, 멤라이즈가 3.9로 가장 낮았다. 이는 스픽의 화면이 깔끔하고 보기 좋았다는 의견, 멤라이즈에 문자뿐만 아니라 음성 인식 기능이 있지만 음성 인식 버튼이 잘 작동하지 않고 인식도 잘하지 못한다는 의견이 이를 뒷받침할 수 있다.

‘신뢰가능한 상호작용’은 프랙티카가 4.4로 가장 높았고, 스픽이 4.3, 그리고 멤라이즈가 3.9로 가장 낮았다. 이는 다른 앱서비스와 달리 스픽은 문자 기능 없이 음성 인식 기능만 있어 시공간의 제약을 받고, 멤라이즈는 한 번 교육 수준이 설정되면 바꿀 수 없으며 평가된 수준에 비해 대화 내용이 너무 쉽게 제공되어 불만족스러웠다는 의견이 이를 뒷받침할 수 있다.

‘사회적 실재감’은 프랙티카가 4.6으로 가장 높았고, 스픽이 4.0, 그리고 멤라이즈가 3.1로 가장 낮았다. 이는 다른 앱서비스와 달리 프랙티카는 피드백을 AI 튜터가 말하는 방식으로 피드백을 제공하고 있어서 좋았다는 점. 반면에 멤라이즈는 수정 사항을 컬러로 구분하여 제공해 주는 것이 좋았지만 그에 따른 추가적인 설명이 없어서 아쉬웠다는 의견이 이를 뒷받침할 수 있다.

‘교육적 상호작용’은 프랙티카가 4.6으로 가장 높았고, 스픽이 4.0, 그리고 멤라이즈가 3.5로 가장 낮았다.

‘기능적 가치’는 스픽이 4.4로 가장 높았고, 프랙티카가 4.0, 그리고 멤라이즈가 3.4로 가장 낮았다. 이는 다른 앱서비스와 달리 스픽이 대화 종료 후 대화

에서 말한 문장에 대해 한 번에 확인하고 피드백을 받을 수 있다는 점에서 학습에 효과적이라고 느꼈다는 의견이 이를 뒷받침할 수 있다.

‘감정적 가치’는 프랙티카가 4.1로 가장 높았고, 스픽이 3.4, 그리고 멤라이즈가 2.4로 가장 낮았다. 이는 프랙티카가 사용자의 답변 각각에 이모티콘으로 감정을 표현해 줘서 비교적 더 공감해 준다고 느꼈고 몰입할 수 있었다는 의견, 감정적으로 공감해 주는 점이 기억에 남는다는 의견을 통해 결과를 뒷받침할 수 있다.

다음은 SPSS 통계분석 프리드먼(Friedman) 검정 분석 결과이다. LLM 특화 지표 중 ‘과정 제시성’, ‘피드백’을 제외한 ‘맥락 적합성(0.019)’, ‘최신성(0.018)’, ‘개인 맞춤성(p<. 001)’, ‘적응성(0.022)’, ‘공감(0.008)’, ‘인간다운 자연스러움(0.004), ‘친밀감(0.002)’이, 교육 분야 특화 지표 중 ‘학습 신뢰성’을 제외한 ‘몰입성(0.032)’, ‘학습 동기부여(0. 026)’, ‘자기 주도성(0.028)’, ‘이해 가능성(0.022)’, ‘학습 효과정(0.002)’, ‘학습 효율성(0.007)’, ‘학습 흥미성(0.016)’,‘자기 효능감(0.029)’가 유의확률이 0.050이하로 통계적으로 유의미한 결과를 확인할 수 있었다..

이에 반해 LLM/교육 분야 특화 지표가 아닌 지표는 유의확률이 낮았다. 이는 스픽, 프랙티카, 멤라이즈 모두 LLM기반 챗봇 서비스는 최근 나온 서비스이기 때문의 기본적인 기능과 기술이 유사하여 이에 대한 만족도는 유사하기 때문이라고 예측해 볼 수 있다. 모든 분석을 종합해 본 결과, 개발된 LLM기반 교육용 챗봇 사용성 평가 지표가 다양한 앱서비스에 적용하여 평가하는 데 사용될 수 있음을 확인할 수 있었다.

【표 6-2】 앱서비스 기술 통계 및 프리드먼(Friedman) 검정

	평가 지표		대상 애플리케이션		
			스픽 (Speak)	프랙티카 (Praktika)	멤라이즈 (Memrise)
주지표	효과적 전달성 (Effective Delivery)		4.4(0.84)	4.4(0.51)	3.9(0.87)
상세지표	명료성 (Clarity)	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 (Transparency)	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 (Contextual Conformity)	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 (Up to Dateness)	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 (Visibility)	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 (Intuitiveness)	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		
	상세 지표 평균		4.2	4.4	3.8
주지표	신뢰가능한 상호작용 (Reliable Interaction)		4.3(0.67)	4.4(0.51)	3.9(0.99)

상세지표	환경적 접근성 (Environmental Accessibility)	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 (Cognitive Accessibility)	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 (Privacy Protection)	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 (Error Management)	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 (Ethicality)	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		
	상세 지표 평균		3.9	4.0	3.7
주지표	사회적 실재감 (Social Presence)		4.0(0.94)	4.6(0.69)	3.1(0.99)
상세지표	개인 맞춤성 (Personalization)	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		
	과정 제시성 (Presentation of Progress)	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 (Feedback)	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		
	적응성 (Adaptiveness)	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 (Empathy)	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 (Human Naturalness)	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*		
	상세 지표 평균		4.1	4.4	3.3
주지표	교육적 상호작용성 (Educational Interaction)		4.0(0.81)	4.6(0.69)	3.5(1.17)
상세지표	몰입성 (Immersion)	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 (Learning Motivation)	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 (Self Directedness)	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 (Understandability)	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*		
	상세 지표 평균		3.9	4.5	3.3
주지표	기능적 가치 (Functional Value)		4.4(0.69)	4.5(0.52)	3.4(1.17)
상세지표	학습 효과성 (Learning Effectiveness)	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 (Learning Efficiency)	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 (Learning Credibility)	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 (Learning Interest)	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*		
	상세지표 평균		4.2	4.5	3.4
주지표	감정적 가치 (Emotional Value)		3.4(1.26)	4.1(0.73)	2.4(1.07)
상세지표	심미성 (Aesthetics)	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 (Intimacy)	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 (Self Efficacy)	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*		
	상세지표 평균		3.5	4.2	2.7
주지표	만족도 (Satisfaction)	평균 (표준편차)	4.0(0.81)	4.3(0.48)	3.0(1.05)
	태도 (Attitude)	평균 (표준편차)	4.5(0.70)	4.7(0.48)	3.2(1.13)
	지속 사용 의도 (Continuous Intention)	평균 (표준편차)	4.2(1.03)	4.4(0.84)	2.7(1.33)

【표 6-3】 앱서비스 별 맨 휘트니(Mann-Whitney) 검정

지표	스픽-프랙티카		프랙티카-멤라이즈		멤라이즈-스픽	
	M/W U	유의확률	M/W U	유의확률	M/W U	유의확률
효과적 전달성 (Effective Delivery)	46.00	0.737	34.00	0.165	33.00	0.168
명료성 (Clarity)	30.00	0.081	36.00	0.250	24.00	*0.033
투명성 (Transparency)	45.00	0.661	37.50	0.304	34.00	0.185
맥락 적합성 (Contextual Conformity)	47	0.786	34.00	0.182	32.00	0.138
최신성 (Up to Dateness)	30	0.108	22.00	*0.026	42.00	0.526
가시성 (Visibility)	29.50	0.090	31.50	0.122	50.00	1.000
직관성 (Intuitiveness)	36	0.255	43.00	0.559	43.50	0.603
신뢰가능한 상호작용 (Reliable Interaction)	47.00	0.796	36.00	0.246	39.00	0.371
환경적 접근성 (Environmental Accessibility)	39.50	0.404	44.00	0.603	46.50	0.779
인지적 접근성 (Cognitive Accessibility)	39	0.371	40.50	0.431	46.50	0.755
프라이버시 보호 (Privacy Protection)	43	0.573	38.00	0.339	44.00	0.630
오류 관리 (Error Management)	45.50	0.717	31.00	0.131	35.50	0.237
윤리성 (Ethicality)	46.00	0.738	36.00	0.261	35.00	0.230
사회적 실재감 (Social Presence)	30.00	0.098	12.50	*0.003	25.00	*0.050
개인 맞춤성 (Personalization)	18.00	*0.007	10.00	*0.001	27.00	0.065
과정 제시성 (Presentation of Progress)	38.50	0.363	38.50	0.364	47.50	0.846
피드백 (Feedback)	47.00	0.802	29.00	0.091	31.00	0.134
적응성 (Adaptiveness)	48.50	0.888	30.50	0.101	32.50	0.144

공감 (Empathy)	34.00	0.185	15.00	*0.005	27.50	0.078
인간다운 자연스러움 (Human Naturalness)	45.50	0.711	15.50	*0.007	14.00	*0.003
교육적 상호작용성 (Educational Interaction)	27.00	0.055	23.50	*0.030	36.50	0.285
몰입성 (Immersion)	27.00	0.051	21.00	*0.016	34.50	0.223
학습 동기부여 (Learning Motivation)	18.00	*0.011	20.50	*0.020	48.00	0.876
자기 주도성 (Self Directedness)	33.50	0.175	23.50	*0.033	36.50	0.293
이해 가능성 (Understandability)	40.00	0.383	29.00	0.084	36.00	0.246
기능적 가치 (Functional Value)	47.50	0.831	20.00	*0.017	23.50	*0.035
학습 효과성 (Learning Effectiveness)	40.50	0.410	18.50	*0.012	25.00	*0.49
학습 효율성 (Learning Efficiency)	39.00	0.343	20.00	*0.016	29.50	0.107
학습 신뢰성 (Learning Credibility)	43.50	0.584	34.00	0.192	39.00	0.377
학습 000 (Learning Interest)	34.50	0.192	20.00	*0.016	31.00	0.135
감정적 가치 (Emotional Value)	33.50	0.190	11.00	*0.002	27.00	0.075
심미성 (Aesthetics)	33.00	0.179	18.00	*0.012	33.50	0.200
친밀감 (Intimacy)	24.00	*0.042	9.00	*0.002	27.50	0.088
자기효능감 (Self Efficacy)	35.00	0.233	21.00	*0.023	37.50	0.332
만족도 (Satisfaction)	41.50	0.423	14.00	*0.003	22.00	*0.020
태도 (Attitude)	43.50	0.557	11.00	*0.002	16.50	*0.008
지속 사용 의도 (Continuous Intention)	45.00	0.676	15.00	*0.006	19.00	*0.016

4. 소결

본 연구는 사용성 평가 지표 적용 과정을 거쳐 LLM기반 교육용 챗봇의 사용성 평가 지표의 유의성을 확인하였다. 우선 사용성 평가 지표 유의성 확인을 위해 지표를 적용할 대상 앱서비스 세 가지를 선정하고 분석하였다. 영어 회화 교육용 챗봇 기능이 있는 스픽, 프랙티카, 멤라이즈는 서로 다른 인터랙션 방식을 가지고 있어 사용성 평가에 차이를 확인하기에 적합하다고 판단하였다. 이후 HCI/Service 분야 연구원 10명을 대상으로 세 가지 앱의 사용성 평가를 각각 하여 결과값에 대한 평균 비교 및 통계분석을 통해 유의성을 확인하였다. 주지표 평균과 같은 그룹 내 상세지표 평균의 평균값 차이가 모두 0.4를 넘지 않고 비슷한 패턴을 갖고 있다는 점에서 개발된 사용성 평가 지표를 활용한 평가가 일관성 있게 잘 되고 있음을 예측해 볼 수 있다.

더하여 일반 지표의 사용성 평가 평균은 차이가 없는 것으로 나타나지만, LLM 특화 지표인 ‘공감’, ‘인간다운 자연스러움’, ‘개인 맞춤성’, ‘친밀감’, 교육 분야 특화 지표인 ‘학습 동기부여’, ‘학습 효과성’, ‘학습 효율성’, ‘학습 흥미성’이 유의확률이 0.050 이하로 세 개의 서비스가 통계적으로 유의미하다는 것을 확인함으로써 개발된 사용성 평가 지표가 LLM, 교육 분야 사용성이 높은 서비스와 그렇지 않은 서비스를 잘 판별하여 평가할 수 있다는 것을 확인할 수 있었다. 특히 사후 검정을 통해 프랙티카와 멤라이즈가 큰 차이를 보인다는 점에서 이를 다시 한번 확인 할 수 있다.

VII. 사용성 평가 지표 활용

1. 연구 방법

본 연구는 LLM기반 교육용 사용성 평가 지표를 활용하기 위해 앞서 개발된 사용성 평가 지표를 기반으로 LLM기반 교육용 챗봇 개발 단계에서 활용될 수 있는 디자인 가이드라인을 제안한다. 디자인 가이드라인은 추후 교육용 챗봇 개발하는 단계에서 참고 및 활용될 수 있다. 구체적인 디자인 솔루션을 낼 수 있는 부분은 사용성(usability) 부분이기 때문에 사용자 가치(User Value), 사용자 수용도(User Acceptance) 부분을 제외하고 사용성의 평가 지표에서 활용될 수 있는 구체적인 디자인 가이드라인을 제시했다. 이후 LLM 특화, 교육 분야 특화 지표의 디자인 가이드에 따른 상세 예시를 제시하였다.

2. 지표를 활용한 효과적인 사용성 평가 방안 제안

본 연구에서는 일반적인 AI 사용성 평가 지표, LLM 특화 지표, 그리고 교육 분야 특화 지표를 체계적으로 구분하여 개발하였다. 이러한 사용성 평가 지표는 이미 개발된 LLM 기반 교육용 챗봇 서비스 및 영어 회화용 챗봇 서비스의 사용성을 평가하고, 이를 바탕으로 서비스 품질을 향상시키는 데 기여할 수 있다. 다음은 본 연구에서 제안된 지표를 활용하여 효과적으로 사용성을 평가할 수 있는 방안을 설명한 내용이다.

첫째, 개발 단계에서 프로토타입이 완료된 베타버전의 챗봇을 평가하여 초기 설계 및 개발 과정에서 발생할 수 있는 문제점과 개선점을 도출할 수 있다. 이 과정에서 사용성 평가 지표는 실제 사용자와의 상호작용을 기반으로 사용자가

예상하는 사용 수준을 평가하고, 이를 통해 서비스의 시장성을 예측하는 데 도움을 줄 수 있다. 특히, 챗봇의 대화 품질, 사용자 편의성, 피드백 적합성 등을 세부적으로 평가하여 서비스의 완성도를 높일 수 있다.

둘째, 이미 상용화된 여러 챗봇 서비스를 비교 평가하여 자사의 서비스가 경쟁 시장에서 얼마나 우위를 점할 수 있는지 평가할 수 있다. 이를 통해 서비스의 차별화 요소를 강화하고, 경쟁 서비스 대비 사용자 경험이 뛰어난 영역을 분석하여 지속적인 개선 방향을 제시할 수 있다. 이러한 비교 평가는 단순한 기술적 우수성을 넘어, 실제 사용자가 느끼는 만족도를 기반으로 서비스의 품질을 정량적으로 분석할 수 있는 기회를 제공한다.

이와 같이 사용성 평가 지표를 활용할 때, 특히 교육용 챗봇 서비스에서는 평가 시점과 테스트 운영 방법에 대해 몇 가지 유의해야 할 점이 있다. 우선, 챗봇과 사용자의 상호작용이 충분히 이루어질 수 있도록 충분한 대화 시나리오를 구성해야 한다. 사용자와의 대화가 제한적이거나 불충분할 경우, 실제 사용성을 제대로 평가하지 못할 가능성이 있다. 따라서 대화의 자연스러운 흐름을 고려하여 평가가 이루어져야 한다. 또한, 실제 사용자를 대상으로 영어 회화 테스트를 진행할 때는 언어적 배경, 영어 실력, 학습 목적 등 다양한 사용자 특성을 고려해야 한다. 이는 사용성 평가 결과의 신뢰도를 높이고, 지표가 포괄적으로 적용될 수 있도록 하는 데 중요한 역할을 한다. 예를 들어, 초급 사용자와 고급 사용자가 챗봇을 사용하는 방식과 기대치가 다를 수 있으므로, 이를 반영한 평가 체계를 마련하는 것이 필요하다.

3. 사용성 평가 지표 활용: 디자인 가이드라인 수립

본 연구는 앞서 개발된 사용성 평가 지표를 활용한 다양한 교육 분야에서 활용될 수 있는 LLM기반 챗봇의 디자인 가이드라인을 수립하였다. 【표 7-1】은 사용성(usability)에 포함되는 주지표 4개의 디자인 지침과 상세지표 21개의 디자인 솔루션이다.

다음【표 7-2】,【표 7-3】은 LLM, 교육 분야 특화 상세지표의 구체적인 디자인 솔루션과 실제로 확인할 수 있는 예시이다.

‘맥락 적합성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 사용자가 말하는 문맥에 적합한 답변을 제공할 수 있도록 설계하는 것이 있다. 스픽에서 사용자가 갑자기 챗봇의 질문 문맥에 벗어나는 답변을 했을 때, 문맥에서 벗어나고 있음을 인지하고 그 점을 사용자에게 고지하는 것, 문맥에 맞춰 답변을 제공하는 것을 참고할 수 있다. 나아가 사용자와 나눴던 이전의 답변을 기억하고 맥락에 맞는 답변을 제공할 수 있도록 발전할 수 있다.

‘최신성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 정치, 경제, 사회, 문화, 라이프, 스포츠 등의 최신 정보를 반영한 답변을 제공할 수 있도록 기반을 마련해야 한다. 예를 들어 스픽의 프리톡에서 최근 화두에 오른 이슈인 ‘정우성 뉴스 봤냐’를 주제로 대화할 수 있도록 제공된 시나리오가 이에 해당한다. 실제 대화 내용에서도 ‘Do you think it’s normal for celebrities to have many relationships?’라는 질문을 하는 등 해당 이슈와 연관된 질문을 주고받을 수 있었다. 이뿐만 아니라 최신 뉴스 기사에서 반영된 내용이라면 해당 기사의 제목과 기사 발행 날짜를 함께 언급하여 대화를 이어 나가는 것도 솔루션이 될 수 있다.

‘개인 맞춤성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 사용자의 교육 수준에 맞는 과정을 적절하게 제공하거나 사용자가 직

접 자신의 교육 수준에 맞는 과정을 선택할 수 있는 기능을 제공하는 것이 있다. 스픽에서 대화 도중 우측 설정 바를 통해 '왕초보 모드', '고급 어휘 활용' 등 난이도 관련 설정을 변경할 수 있는 것을 참고할 수 있다.

'과정 제시성'이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇과의 대화 전중후에 대한 절차를 사전에 고지하거나 사용자가 대화 진행 상황을 실시간으로 확인할 수 있도록 명시하는 기능을 제공하는 것이 있다. 이는 스픽에서 복습 과정 중 화면 상단에 학습 진도를 바 형태로 제시하는 것을 참고할 수 있다. 또는 챗봇이 어떤 프로세스를 통해서 나의 학습을 분석하고 있는지 고지하는 것도 이에 포함된다.

'피드백'이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 사용자 응답의 의미적 문법적 오류 등을 파악하여 적절한 피드백을 제공하는 것이 있다. 스픽에서 대화 중 사용자 답변 옆에 있는 '*'버튼을 누르면 어휘나 문법에 대한 피드백을 즉각적으로 받을 수 있는 기능이 이에 포함된다.

'적응성'이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 사용자의 대화 흐름에 적절한 대답을 하지 못하는 경우 답변에 대한 힌트를 제공하는 것이 있다. 스픽에서 사용자가 챗봇의 질문에 대해 대답하기 어려운 경우 원활한 대화를 위해 힌트를 제공하고 있다. 오른쪽의 전구 버튼을 누르면 '이렇게 말해보세요'라는 문구와 함께 질문에 적절한 답변을 알려준다.

'공감'이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 사용자의 기분을 파악하여 적시 적소에 공감하는 답변을 제공하는 것이 있다. 프랙티카에서 사용자의 답변에 공감하는 이모티콘을 부여하고 챗봇의 답변으로도 "It is indeed a classic!" 등과 같은 공감의 문장을 포함하여 말하고 있는 것을 참고할 수 있다.

‘인간다운 자연스러움’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 사용자가 원하는 상황에 맞게 적절한 페르소나를 제공하거나 대화 시작 전 사용자가 원하는 페르소나를 직접 설정할 수 있는 기능을 제공하는 것이 있다. 스픽에서 챗봇과 대화 시작 전 사용자가 직접 설정한 사용자 본인, 챗봇의 역할과 상황에 따라 챗봇에 인격이 투영되어 시뮬레이션하는 것을 참고할 수 있다.

‘몰입성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 사용자가 교육(학습) 상황에 몰입할 수 있도록 대화 중 실제 사람 같은 아바타 화면을 제공하는 것이 있다. 프랙티카에서 대화 중 AI 아바타와 마주보고 얘기하는 듯한 인터페이스는 사용자에게 몰입감을 줄 수 있다.

‘학습 동기부여’가 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 사용자가 챗봇을 사용할수록 만족감을 느끼도록 출석 도장 혹은 학습 완료 벳지와 같은 교육(학습) 동기부여 요소를 제공하는 것이 있다. 듀오링고에서 매일매일 학습 시 불꽃을 부여하고, 하루라도 참석하지 않으면 불꽃이 꺼지는 페널티를 제공해 사용자에게 동기부여를 제공한다. 또는 대화 완료 후 채팅창 내에서 학습 완료 알림과 벳지를 보여주는 것도 동기부여 요소로 작용할 수 있다.

‘자기 주도성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 지정된 시간에 교육(학습)할 수 있도록 알림을 제공하는 것이 있다. 말해보카에서 사용자가 학습 일정을 잊지 않도록 학습하지 않은 사용자에게 알림을 보내는 것은 자기주도 학습을 이끄는 발판이 될 수 있다.

‘이해 가능성’이 높게 평가될 수 있도록 설계과정에서 활용될 수 있는 디자인 솔루션은 챗봇이 교육(학습) 목적을 명확하게 인식하고 정보를 제공할 수 있도록 구성해야 하는 것. 즉 교육 콘텐츠 자체 나 방식을 사용자가 잘 이해할 수 있도록 챗봇이 잘 설명해 주는 것이 있다. 멤라이즈에서 대화 시작 전, 그리고

대화 중 대화 목표를 지속적으로 보여주는 것이 예시가 될 수 있다. 또한 매뉴얼을 숙지하지 않아도 잘 이해할 수 있도록 익숙한 인터페이스를 제공하는 것도 솔루션이 될 수 있다.

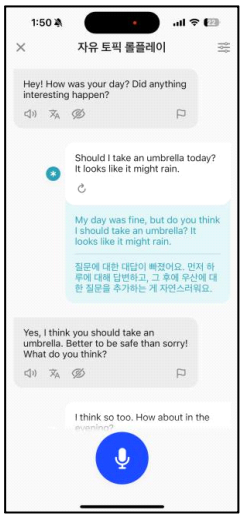
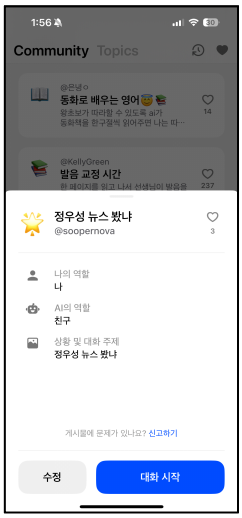
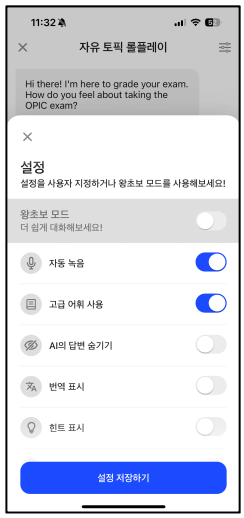
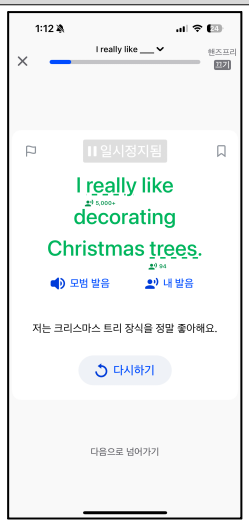
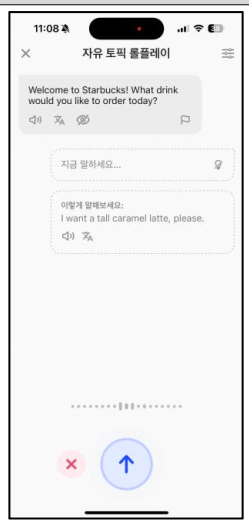
【표 7-1】 LLM 기반 교육용 챗봇 사용성 평가 지표를 활용한 디자인 가이드라인

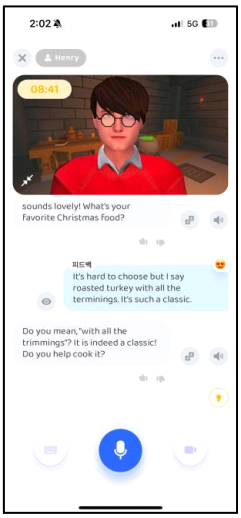
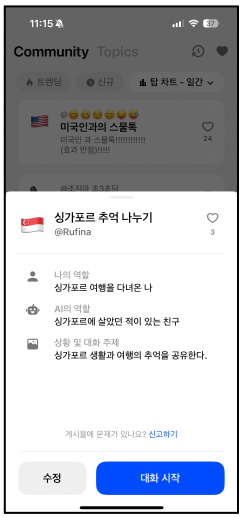
주지표	상세지표	디자인 솔루션
효과적 전달성 (Effective Delivery)	챗봇 서비스가 제공하는 정보가 의미적/시각적으로 명확하게 전달되어야 함	
	명료성 (Clarity)	• 챗봇이 문상을 간결하게 구사하고 문단의 구분을 명확하게 해야 함
	투명성 (Transparency)	• 챗봇이 사실에 기반한 내용을 전달할 때 정확한 출처를 제공함 • 나의 학습 정보가 어떻게 쓰이고 있는지 설명 (내 정보가 어떻게 활용되고 있는지)
	맥락 적합성 (Contextual Conformity)	• 챗봇이 사용자와 나눴던 이전의 답변을 기억하고 맥락에 맞는 답변을 제공함
	최신성 (Up to Dateness)	• 챗봇이 정치, 경제, 사회, 문화, 라이프, 스포츠 등의 최신 정보를 반영한 답변을 제공함
	가시성 (Visibility)	• 챗봇이 음성 제어 버튼, 해석 제공 버튼 등 사용자가 그 의미를 명확하게 알고 사용할 수 있는 디자인을 제공함
	직관성 (Intuitiveness)	• 챗봇이 응답을 생성하는 것을 사용자가 즉각적으로 인지할 수 있도록 모션을 제공함 • 챗봇이 사용자가 음성으로 대답하는 경우 말의 시작과 마침을 인식할 수 있는 모션을 제공함
신뢰가능한 상호작용 Reliable Interaction	챗봇 서비스가 안전하고 신뢰할 만한 상호작용을 제공해야 함	
	프라이버시 보호 (Privacy Protection)	• 챗봇이 사용자와의 대화 내용을 수집하고 있다면 사전에 고지하고 동의를 받아야 함
	오류 관리 (Error Management)	• 대화 중 사용자가 대화 종료를 시도했을 때, 대화 종료 여부를 확인하는 메시지를 보냄 • 대화 중 사용자가 대화 종료를 시도했을 때, 대화 저장 여부를 확인하는 메시지를 보냄

	윤리성 (Ethicality)	• 챗봇이 윤리에 어긋나는 발언을 하는 경우를 대비하여 발화전 자체 검열 시스템을 제공함 • 챗봇이 윤리에 어긋나는 발언을 하는 경우 사용자가 신고할 수 있는 기능을 제공함
	환경적 접근성 (Environmental Accessibility)	• 챗봇이 사용자가 이동 중이거나 공공 장소에 있을 때 주변 소음으로 음성으로 대답하는 데 어려움이 없도록 대화 방식에 다양한 옵션을 제공해야 함
	인지적 접근성 (Cognitive Accessibility)	• 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 (Social Presence)	챗봇 서비스가 사용자와 사회적/감정적으로 자연스러운 실재감을 제공하며 상호작용 해야 함	
	개인 맞춤성 (Personalization)	• 사용자의 교육 수준에 맞는 과정을 적절하게 제공함 • 사용자가 직접 자신의 교육 수준에 맞는 과정을 선택할 수 있는 기능을 제공함
	과정 제시성 (Presentation of Progress)	• 챗봇과의 대화 전중후에 대한 절차를 사전에 고지함 • 사용자가 대화 진행 상황을 실시간으로 확인할 수 있도록 명시하는 기능을 제공함 • 챗봇이 어떤 프로세스를 통해서 나의 학습을 분석하고 있는지
	피드백 (Feedback)	• 챗봇이 사용자 응답의 의미적/문법적 오류 등을 파악하여 적절한 피드백을 제공함
	적응성 (Adaptiveness)	• 챗봇이 사용자의 대화 흐름에 적절한 대답을 하지 못하는 경우 해당 답변을 평가할 수 있는 기능을 제공함
	공감 (Empathy)	• 챗봇이 사용자의 기분을 파악하여 사용자 질문에 대답 전, 적시 적소에 공감하는 답변을 제공함
	인간다운 자연스러움 (Human Naturalness)	• 챗봇이 사용자가 원하는 상황에 맞게 적절한 페르소나를 제공함 • 대화 시작 전 사용자가 원하는 페르소나를 직접 설정할 수 있는 기능을 제공함
교육적 상호작용성 (Educational Interaction)	챗봇 서비스가 학습에 효과적인 교육적 상호작용을 제공해야 함	

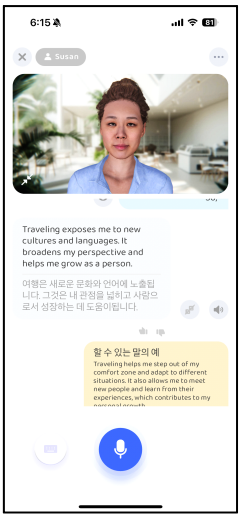
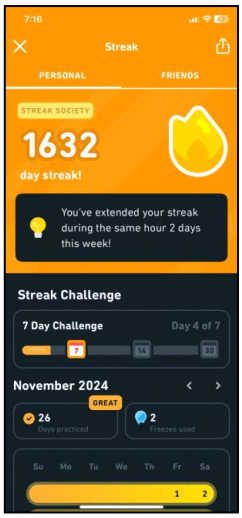
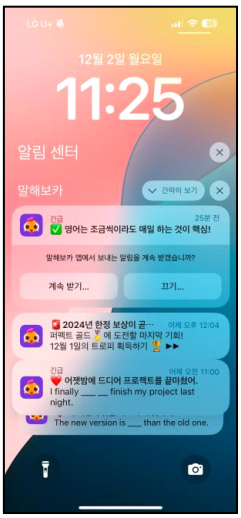
	몰입성 (Immersion)	• 사용자가 교육(학습) 상황에 몰입할 수 있도록 대화 중 실제 사람 같은 아바타 화면을 제공함
	학습 동기부여 (Learning Motivation)	• 사용자가 챗봇을 사용할수록 만족감을 느낄 수 있도록 출석 도장 혹은 학습 완료 벳지와 같은 교육(학습) 동기부여 요소를 제공함
	자기 주도성 (Self Directedness)	• 챗봇이 지정된 시간에 교육(학습)할 수 있도록 알림을 제공함 • 사용자가 교육(학습)을 진행한 내용과 시간을 스스로 관리할 수 있는 페이지를 마련함
	이해 가능성 (Understandability)	• 챗봇이 교육(학습) 목적을 명확하게 인식하고 정보를 제공할 수 있도록 구성해야 함 • 교육 콘텐츠 자체 나 방식을 사용자가 잘 이해할 수 있도록 챗봇이 잘 설명해 주는 것 • 매뉴얼을 숙지하지 않아도 잘 이해할 수 있는 것

【표 7-2】 LLM 특화 지표 디자인 가이드라인 예시

	LLM 특화 지표		
	맥락 적합성	**최신성**	**개인 맞춤성**
화면 예시			
화면 설명	[스픽] 갑자기 다른 주제로 얘기해도 이에 맞는 답변을 제공함	[스픽] 챗봇이 최신 사회 이슈를 다루고 있음	[스픽] 대화 중 챗봇 난이도를 설정할 수 있음
	과정 제시성	**피드백**	**적응성**
화면 예시			
화면 설명	[스픽] 화면 상단에 학습 진도를 상단 바로 제시함	[스픽] '*'을 누르면 내 답변에 대한 피드백을 확인할 수 있음	[스픽] 사용자가 대답할 수 있도록 힌트(전구 아이콘)를 제공함

	공감	인간다운 자연스러움
화면 예시		
화면 설명	[프랙티카] 공감하는 이모티콘과 대답을 제공함	[스픽] 대화 시작 전 사용자 챗봇 각각의 역할과 상황을 구체적으로 명시함

【표 7-3】 교육 분야 특화 지표 디자인 가이드라인 예시

	교육 분야 특화 지표	
	몰입성	학습 동기부여
화면 예시		
화면 설명	[프랙티카] AI 아바타와 영상통화 하는 듯한 몰입감을 제공함	[듀오링고] 매일매일 학습 완료 시 불꽃 아이템을 제공함
	자기 주도성	이해 가능성
화면 예시		
화면 설명	[말해보카] 잊지 않고 학습하도록 알림을 보냄	[멤라이즈] 대화 전 대화의 목표를 구체적으로 사용자에게 알려줌

4. 소결

본 연구는 LLM 기반 교육용 챗봇의 사용성 평가 지표 체계가 실질적으로 활용될 수 있도록 사용성 평가 지표 활용 단계를 체계적으로 수행하였다. 그 결과 본 연구에서 제안된 사용성 평가 지표는 LLM 기반 교육용 챗봇 서비스의 품질 향상 및 경쟁력 강화에 중요한 역할을 할 수 있다. 특히, 평가 지표를 활용한 체계적인 접근은 서비스 설계 단계부터 상용화 이후까지 전반적인 서비스 개선 과정을 지원하며, 사용자 중심의 설계 및 개발을 가능하게 한다. 추후 다양한 사용 사례와 평가 환경에서의 검증을 통해, 본 지표 체계가 더욱 폭넓게 활용될 수 있기를 기대한다.

또한, 본 연구에서는 사용성 평가 요소의 세 가지 주요 카테고리 중 구체적인 솔루션 제안이 가능한 '사용성'에 속하는 지표를 중심으로 디자인 가이드라인을 도출하였다. 이를 통해 서비스 설계자와 개발자가 평가 지표를 직접적으로 활용할 수 있는 지침을 제공하였으며, LLM 및 교육 분야 특화 지표를 참고하여 실제 사례에 적용할 수 있는 예시 화면과 구체적인 설명을 추가로 제안하였다. 이러한 가이드라인은 기존의 사용성 평가에 그치지 않고, 개발 단계부터 상용화 이후의 개선 단계에 이르기까지 평가 지표의 활용 범위를 확장하는 데 기여할 수 있다.

결론적으로, 본 연구의 결과는 개발된 사용성 평가 지표가 이미 출시된 앱 서비스의 사용성 개선뿐만 아니라, 초기 설계 및 개발 단계에서도 실질적으로 활용될 수 있음을 시사한다. 앞으로 본 연구에서 제안된 지표 체계가 다양한 LLM 기반 서비스와 교육 분야뿐 아니라, 비교육 분야의 서비스에서도 폭넓게 적용되어 사용성 평가 및 서비스 개선에 유용한 자료로 활용되기를 기대한다.

VIII. 결 론

1. 연구 요약

본 연구는 LLM 기술 시장이 확대됨에 따라 교육 분야 챗봇의 사용성을 향상시키기 위한 특화된 지표 체계를 개발하는 데 목적을 두고 진행되었다. 이를 위해 다양한 선행 연구를 검토하고, 기존의 AI 평가 지표를 기반으로 평가 요소를 수집하였다. 특히, 일반적인 AI 평가 지표뿐만 아니라, 평가 목적에 적합한 LLM 기반 및 교육 분야 특화 지표를 추출하기 위해 체계적인 개발 및 검증 과정을 거쳤다. 연구 과정에서 평가 지표 개발을 위한 체계적인 절차를 설계하고 이를 검증하여 LLM 챗봇의 사용성 평가에 적합한 최종 지표 체계를 도출하였다. 나아가 실무에서 활용 가능성을 높이기 위해 평가 지표를 실제 설계 및 개발 단계에서 적용할 수 있는 디자인 가이드라인 형태로 제안하였다.

본 연구의 연구 질문별 주요 결과는 다음과 같다.

연구 질문 1: LLM 기반 교육용 챗봇 서비스의 사용성 평가 지표는 기존 AI 기반 챗봇의 사용성 평가 지표와 어떠한 차이가 있는가?

LLM 기반 교육용 챗봇의 사용성 평가 지표는 기존 AI 기반 챗봇의 지표와 비교하여, 사회적 실재감, 신뢰 가능성, 적응성과 같은 LLM의 특성을 반영하는 세부 지표를 포함하고 있다는 점에서 차별성을 보였다. 특히, LLM 기반 챗봇은 자연스러운 언어 생성과 상호작용에서 기존 AI 챗봇보다 우수한 기능을 제공하므로, 사용자가 이를 어떻게 인지하고 경험하는지를 평가하는 지표가 추가로 요구되었다.

연구 질문 2: LLM 기반 챗봇 서비스의 사용성 평가 지표 중 'LLM'과 '교육 분야'에 특화된 사용성 평가 지표 요인은 어떤 것이 있는가?

'LLM' 특화 지표로는 자연스러움, 맥락 이해, 과정 제시성 등이 도출되었으며, 이는 LLM이 가진 대규모 데이터 처리 및 생성 능력에서 기인한 요인이다. 한편, '교육 분야' 특화 지표로는 개인 맞춤형 학습 제공, 학습 피드백의 명확성, 교육적 상호작용의 효과성이 확인되었다. 이들 지표는 학습자에게 적합한 교육 콘텐츠를 제공하고 학습 효과를 극대화하는 데 중요한 역할을 한다는 점에서 교육적 맥락의 특성을 반영한다.

연구 질문 3: 본 사용성 평가 지표 체계가 실제 교육용 챗봇 서비스의 사용성 평가에 효과적인가?

연구 결과, 본 연구에서 개발한 사용성 평가 지표 체계는 실제 교육용 챗봇 서비스의 사용성 평가에서 유효성을 보였다. 실험을 통해 평가 지표를 적용한 결과, 사용자의 만족도와 서비스 효율성을 체계적으로 평가할 수 있었으며, 기존 평가 방식보다 LLM 기반 챗봇의 고유한 특성을 더 잘 반영하였다.

연구 질문 4: 본 사용성 평가 지표 체계가 실제 교육용 챗봇 서비스에서 어떻게 활용될 수 있는가?

본 사용성 평가 지표 체계는 실제 서비스 설계 및 개선 과정에서 유용하게 활용될 수 있다. 구체적으로, 서비스 기획 단계에서는 디자인 가이드라인으로 활용되어 챗봇의 기능 설계와 사용자 경험을 강화할 수 있으며, 서비스 출시 전 검증 단계에서는 사용자 테스트를 통해 잠재적 문제점을 식별하고 수정하는 데 도움을 줄 수 있다. 또한, 서비스 운영 단계에서는 사용성 데이터를 바탕으로 지속적으로 서비스 품질을 개선하는 데 기여할 수 있다.

본 연구는 LLM 기반 교육용 챗봇의 사용성을 평가하고 개선하기 위한 체계적인 지표를 제안함으로써, 학문적 및 실무적 기여를 제공하였다. 추후 연구에서는 본 지표 체계를 다양한 서비스 맥락에 적용하여, 서비스 특성에 따라 정의 및 문항을 변형하는 추가 검증이 필요할 것이다.

2. 연구 가치 및 기여

본 연구의 연구 가치 및 기여는 학문적, 실무적 부분 각각 구분하여 생각해 볼 수 있다. 우선 학문적으로는 첫째, 최신 LLM기반으로 개발된 평가 지표의 학문적 경향을 반영하고 있다. 둘째, 통계적으로 검증된 지표 체계를 수집하였고, 지표 체계가 '일반적인 UX 관점의 지표', '일반적인 AI 지표', 'LLM 특화 지표', '교육 분야 특화 지표' 각각의 특색있는 지표를 추출하고 있다는 점에서 가치가 있다. 마지막으로, SOR 이론을 바탕으로 지표 체계를 구성하여 지표 체계에 위계를 구분했다는 것이다. 사용자의 인지와 행동에 차이가 있음에도 기존 지표 체계는 대체로 위계관계 없이 지표 체계가 구성되어 있어 혼용의 우려가 있다. 이에 반해 본 연구에서 개발된 사용성 평가 지표 체계는 SOR 이론은 도입하여 사용성뿐만 아니라 사용자 가치, 사용자 수용도와 같이 주 지표의 상위 카테고리를 만들었다는 점에서, 사용 이후에 발생하는 사용자 경험을 좀 더 포괄적이고 체계적으로 지표 체계를 구성했다는 점에서 가치가 있다.

실무적으로는 첫째, 실제 사용되고 있는 앱서비스를 대상으로 지표 체계를 검증하고 실제 차이가 보이는 앱서비스의 차이를 확인함으로써 지표 체계가 유의미함을 확인했다. 둘째, 개발된 지표 체계가 이미 시장에 나와 있는 서비스들을 개선하기 위해, 또는 출시 전 앱을 테스트하기 위해서 사용될 수 있다는 점. 셋째, 서비스를 기획하고 개발하는 초기 단계에서, 효과적으로 활용되기 위해 개발 단계의 디자인 가이드라인으로 활용되어 실제 디자이너 둘에게 직접적인 참고 자료를 제공했다는 점에서 실무적 기여가 크다.

3. 연구 한계 및 추후 연구 방안

본 연구를 통해 발견한 연구의 한계와 추후 연구 방안은 다음과 같다. 첫째, 본 연구는 영어 회화 분야에 초점을 맞추어 실험을 진행하였다. 이러한 접근은 영어 회화 분야에서의 LLM 기반 교육용 챗봇의 사용성을 심층적으로 분석하는 데 유용하였으나, 다른 교육 분야나 비언어 학습 영역에 대한 적용 가능성을 충분히 검토하지 못했다는 한계가 있다. 이에 따라 추후 연구에서는 수학, 과학 등 다양한 학문 분야나 비언어 학습 환경에서도 사용성 평가 지표를 적용하여 일반화를 검토하는 것이 필요하다.

둘째, 사용성 평가 지표를 적용한 실험 단계에서 참여 대상자의 수가 10명으로 제한되었다. 이는 사용성 평가 결과에 대한 통계적 유의성을 확보하기에 표본 크기가 충분하지 않다는 점에서 한계가 존재한다. 따라서 추후 연구에서는 다양한 연령대와 배경을 가진 사용자로 표본을 확대하여 보다 신뢰할 수 있는 결과를 도출하고, 평가 결과의 일반화를 보완할 필요가 있다.

셋째, 디자인 가이드라인의 예시로 영어 교육 서비스에 초점을 두고 작성하였다는 점이다. 이는 가이드라인이 영어 교육 외의 다른 교육・비교육 서비스에 적용될 가능성을 충분히 제시하지 못했음을 의미한다. 이에 따라 추후 연구에서는 다양한 교육 서비스와 산업 분야에서 활용할 수 있는 보편적 가이드라인을 제안하고, 실제 현장에서의 적용 가능성을 검증하는 노력이 필요하다.

이러한 한계를 보완하기 위해, 향후 연구에서는 영어 회화를 넘어선 다양한 교육 분야와 사용 사례를 포괄하고, 표본 규모를 확대하며, 보다 폭넓은 산업적 활용을 위한 설계 및 검증 과정을 포함해야 할 것이다. 이를 통해 본 연구에서 제안된 사용성 평가 지표 체계는 교육용 챗봇의 효과적인 설계와 개발에 실질적으로 기여할 수 있을 것으로 기대되며 향후 LLM 기술을 활용한 다양한 분야에서 연구와 실무 적용에 유용한 참고 자료로 활용될 수 있다.

참 고 문 헌

- TTA정보통신용어사전. https://terms.tta.or.kr/dictionary/dictionaryViewhttps: //terms.tta.or.kr/dictionary/dictionaryView.do?word_seq=193122
- 인공지능 챗봇 트렌드 2021: 산업 별 전망 (교육). https://tonyaround.com/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5-%EC%B1%97%EB%B4%87-%ED%8A%B8%EB%A0%8C%EB%93%9C-2021-%EC%82%B0%EC%97%85-%EB%B3%84-%EC%A0%84%EB%A7%9D-%EA%B5%90%EC%9C%A1/
- 동아일보(2020). 비대면 교육서비스 수요 증가 대학가에 AI 챗봇 '현명한 앤써니' 보급
- 교육부(2023). 공교육과 기술이 함께 발전하는 '교육 정보 기술(에듀테크)' 시대 열린다.
- Tian, H., Lu, W., Li, T. O., Tang, X., Cheung, S. C., Klein, J., & Bissyandé, T. F. (2023). Is ChatGPT the Ultimate Programming Assistant-How far is it?. arXiv preprint arXiv:2304.11938.
- 김지현(2024). AI의 시작과 발전 과정, 미래의 전망.
- Li, W., Zhang, X., Li, J., Yang, X., Li, D., & Liu, Y. (2024). An explanatory study of factors influencing engagement in AI education at the K-12 Level: an extension of the classic TAM model.Scientific Reports, 14.
- Oermann, E. K., & Kondziolka, D. (2023). On Chatbots and Generative Artificial Intelligence.Neurosurgery,92(4), 665-666. https://doi.org/10.1227/neu.0000000000002415
- Hornb æk, K. (2006). Current practice in measuring usability: Challenges

to usability studies and research.International journal of human-computer studies,64(2), 79-102.

- 김근용, 윤기하, 김량수, 류지형, & 김성창. (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
- 조영임. (2023). 초거대 AI 와 생성형 인공지능.ICT Standard Weekly,1145, 1-9.
- 이현주, 성창수, & 전병훈. (2023). 빅카인즈를 활용한 GenAI (생성형 인공지능) 기술 동향 분석: ChatGPT 등장과 스타트업 영향 평가.벤처창업연구,18(4), 65-76.
- 김성희, & 이승민. (2024). 생성형 AI 의 기술적 특성과 사서의 개인적 특성이 생성형 AI 사용의도에 미치는 영향.한국비블리아학회지,35(2), 109-133.
- 최수진. (2023). 챗 GPT 따라잡아라… 속도 내는 IT 거인들.한경 BUSINESS, (1420), 30-31.
- Stade, E.C., Stirman, S.W., Ungar, L.H.et al.Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation.npj Mental Health Res3, 12 (2024).
- Yang, J., Wang, Z., Lin, Y., & Zhao, Z. (2024). Global Data Constraints: Ethical and Effectiveness Challenges in Large Language Model.arXiv preprint arXiv:2406.11214.
- 박대민. (2023). 신뢰할 수 있는 인공지능 기반의 저널리즘 인공지능: 언론 신뢰와 인공지능 신뢰성 간 통약가능성을 바탕으로.언론과 사회,31(4), 5-47.
- 이항, & 김준환. (2023). 통합기술수용모델이 챗 GPT 이용자의 디지털리터러시와 수용의도에 미치는 영향.융복합지식학회논문지,11(2), 33-43.

- 황홍섭. (2021). 초등 사회과 마이크로러닝을 위한 챗봇의 개발. 사회과교육, 60(3), 81-104.
- Miller, R.B., 1971. Human ease of use criteria and their tradeoffs. IBM Report TR 00.2185, 12 April. IBM Corporation, Poughkeepsie, NY.
- Shackel, B. (2009). Usability-Context, framework, definition, design and evaluation.Interacting with computers,21(5-6), 339-346.
- Shackel, B. (1981).The concept of usability(pp. 1-30). Poughkeepsie, New York: Proceedings of IBM Software and Information Usability Symposium.
- Nielsen, J. (1993).Usability Engineering.Boston: Morgan Kaufmann.
- 이승희, 손원준. (2022). 시니어 세대를 위한 모바일 어플리케이션에 관한 사용성 평가 연구 - 국내 유통기업 사례를 중심으로. 상품문화디자인학연구,(68), 1-12.
- 서창희. "시니어를 위한 애플리케이션의 사용성 평가지표에 관한 연구." 국내석사학위논문 상명대학교 일반대학원, 2022. 서울
- Kim, N.-H. (2020). User Experience Validation Using the Honeycomb Model in the Requirements Development Stage.International Journal of Advanced Smart Convergence,9(3), 227-231. https://doi.org/10.7236/IJASC.2020.9.3.227
- Morville, P. (2004). Ambient Findability, Educational technology research and development, 54(6), 623-626.
- Nielsen, J. (1994). 10 Usability Heuristics for User Interface Design [On-Line]. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/ten-usability-heuristics/
- https://www.iso.org/standard/16722.html
- Lopez, C.M., Lopez, J.E., Buchely, A.B., & Lopez, D.F. (1998). Ergonomic

requirements for office work with visual display terminals (VDTs).

- Joseph Weizenbaum. 1966. ELIZA-a computer program for the study of natural language communication between man and machine. Commun. ACM 9, 1 (Jan. 1966), 36-45. https://doi.org/10.1145/365153.365168
- ELIZA Archaeology. ELIZA Archaeology Project. https://sites.google.com/view/elizaarchaeology/
- Berry, DM (2023). 계산의 한계: Joseph Weizenbaum과 ELIZA 챗봇.Weizenbaum Journal of the Digital Society,3(3). https://doi.org/10.34669/WI.WJDS/3.3.2
- SK hynix newsroom (2024). [All Around AI 1편] AI의 시작과 발전 과정, 미래 전망. https://news.skhynix.co.kr/post/all-around-ai-1
- Electronic book review (2024). Reading ELIZA: Critical Code Studies in Action. https://electronicbookreview.com/essay/reading-eliza-critical-code-studies-in-action/
- 곽헌동. "컴퓨터 기반 협력적 논증에서 개인 논증을 지원하기 위한 챗봇 개발." 국내석사학위논문 서울대학교 대학원, 2023. 서울
- 강신천, 허희옥. (2023). 생성형 AI 기반 교수설계 지원 플랫폼 개발 및 시범 적용. 컴퓨터교육학회 논문지, 26(6), 143-153.
- 양정아. (2023). 대화형 챗봇(Chat-Bot) 서비스디자인 : 패브릭 인테리어 컨설팅을 중심으로 [석사학위논문, 홍익대학교].
- 안무정 and 강태임. (2023). 디지털 트랜스포메이션 경영을 위한 챗GPT 사용자 경험(UX) 디자인 평가 -오픈AI 챗GPT와 마이크로소프트 빙 챗GPT 교차 활용을 중심으로-. 한국디자인문화학회지, 29(2), 237-247.
- 한정윤, 구예리, 김수진. (2023). 챗GPT를 활용한 맞춤형 피드백 생성 및 효과 분석. 교육정보미디어연구, 29(4), 1123-1151.

- 채주혜, 김민영, 류강, 유영만, 신윤희. (2024). 개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석. 디지털콘텐츠학회논문지, 25(4), 1053-1068. 10.9728/dcs.2024.25.4.1053
- Bink, J. Personalized Response with Generative AI: Improving Customer Interaction with Zero-Shot Learning LLM Chatbots.
- Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A complete survey on llm-based ai chatbots.arXiv preprint arXiv:2406.16937.
- 조희석(2018). 챗봇(ChatBot)의 활용 사례 및 이러닝 도입 전략
- 박정아, 이향. (2021). 한국어 교육용 AI 챗봇 개발을 위한 챗봇 빌더 활용 방안.외국어로서의 한국어교육,63, 51-91, https://doi.org/10.21716/TKFL.63.3
- 차현진. (2023). 사용자 친화적인 챗봇 튜터 설계 지침 개발 연구. 컴퓨터교육학회 논문지, 26(5), 79-92.
- 윤헌철. (2023). 대학 STEAM 교육에서 가상현실 활용에 대한 학습자의 인식 분석. 홀리스틱융합교육연구, 27(4), 47-67.
- Kim, M. J., Lee, C.-K., & Jung, T. (2020). Exploring Consumer Behavior in Virtual Reality Tourism Using an Extended Stimulus-Organism-Response Model. Journal of Travel Research, 59(1), 69-89. https://doi.org/10.1177/0047287518818915
- 김민지. (2023). 협력학습을 지원하는 인공지능 챗봇 설계원리 개발 연구: The Developmental Study of AI Chatbot Design Principles for Supporting Collaborative Learning.
- 김형조. "온라인 토론 성찰을 지원하는 대시보드 기반 챗봇 개발." 국내석사학위논문 서울대학교 대학원, 2023. 서울
- 모온가. "중국 온라인 쇼핑몰의 챗봇 서비스에 대한 지각된 가치 및 지속이용의도 연구." 국내석사학위논문 건국대학교 대학원, 2023. 서울

- 이승희, 손원준. (2022). 시니어 세대를 위한 모바일 어플리케이션에 관한 사용성 평가 연구 - 국내 유통기업 사례를 중심으로. 상품문화디자인학연구,(68), 1-12.
- 박지원, 정하성, 임동희, 박주은, 정종진. (2024-01-24). 생성형 AI 기반의 동화책 제작 서비스 설계 및 구현. 한국HCI학회 학술대회, 강원.
- 강신천, 허희옥. (2023). 생성형 AI 기반 교수설계 지원 플랫폼 개발 및 시범 적용. 컴퓨터교육학회 논문지, 26(6), 143-153.
- 채주혜, 김민영, 류강, 유영만, 신윤희. (2024). 개별화 맞춤형 학습을 지원하는 AI 챗봇 기반 플랫폼 분석. 디지털콘텐츠학회논문지, 25(4), 1053-1068.
- Nasyiah, M., Kelana, B., & Riskinato, A. (2024). System Usability Scale for Measuring Usability of Social Network Applications from User Perspectives. In E3S Web of Conferences (Vol. 483, p. 03010). EDP Sciences.
- Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The chatbot usability scale: the design and pilot of a usability scale for interaction with AI-based conversational agents. Personal and ubiquitous computing, 26, 95-119.
- Buba□, G., Či엖e쉏ja, A., & Kovačić, A. (2023). Development of an assessment scale for measurement of usability and user experience characteristics of Bing chat conversational AI. Future Internet, 16(1), 4.
- Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., ... & Rahmani, A. M. (2024). Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Medicine, 7(1), 82.
- Theophilou, E., Koyutürk, C., Yavari, M., Bursic, S., Donabauer, G.,

Telari, A., ... & Ognibene, D. (2023, November). Learning to prompt in the classroom to understand AI limits: a pilot study. In International Conference of the Italian Association for Artificial Intelligence (pp. 481-496). Cham: Springer Nature Switzerland.

- Silvestri, C., Roshal, J., Shah, M., Widmann, W. D., Townsend, C., Brian, R., ... & Sathe, T. S. (2024). Evaluation of a Novel Large Language Model (LLM) Powered Chatbot for Oral-Boards Scenarios. medRxiv, 2024-05.
- Xiao, C., Xu, S. X., Zhang, K., Wang, Y., & Xia, L. (2023, July). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 610-625).
- Dortheimer, J., Martelaro, N., Sprecher, A., & Schubert, G. (2024). Evaluating large-language-model chatbots to engage communities in large-scale design projects. AI EDAM, 38, e4.
- Freire, S. K., Wang, C., & Niforatos, E. (2024). Chatbots in knowledge-intensive contexts: Comparing intent and llm-based systems. arXiv preprint arXiv:2402.04955.
- Kumar, D. (2024). OpineBot: Class Feedback Reimagined Using a Conversational LLM.
- Thway, M., Recatala-Gomez, J., Lim, F. S., Hippalgaonkar, K., & Ng, L. W. (2023). Battling Botpoop using GenAI for higher education: A study of a retrieval augmented generation chatbots impact on learning. 2023.
- Kajiwara, Y., & Kawabata, K. (2024). AI Literacy for Ethical use of

Chatbot: Will Students accept AI Ethics?. Computers and Education: Artificial Intelligence, 100251.

- Ilieva, G., Yankova, T., Klisarova-Belcheva, S., Dimitrov, A., Bratkov, M., & Angelov, D. (2023). Effects of generative chatbots in higher education. Information, 14(9), 492.
- 고광이, 류도향. (2021). 부부의 취업형태에 따른 연령별 가족관계 탐색적 요인분석 - 여성가족패널조사를 중심으로. 한국데이터정보과학회지, 32(1), 169-197, 10.7465/jkdi.2021.32.1.169
- Kt enterprise (2024). LLM의 환각현상, 어떻게 보완할 수 있을까? https://enterprise.kt.com/bt/dxstory/2521.do

ABSTRACT

Development and Application of Usability Evaluation scale for Educational Chatbots based on LLM : Focusing on an English Conversation App Services

Kim JiHyo
Department of Future Convergence Technology Engineering
Graduate School of
Sungshin University

This study aims to develop usability evaluation metrics for educational chatbot services based on large language models (LLMs) and propose design guidelines that can be applied in practice through their validation. The EdTech market is growing rapidly, and LLM-based chatbots provide personalized learning experiences for learners, bringing innovation to traditional educational methods. However, research on reliable usability evaluation metrics that reflect the characteristics of LLM-based educational chatbots is still lacking.

In this context, the study analyzes existing usability evaluation metrics for AI-based chatbots and proposes a new usability evaluation metric system that integrates LLM-specific and education-specific factors based on four major research questions. The research was conducted through a literature

review, affinity diagramming, expert group reviews, pilot testing, main survey, and exploratory factor analysis. As a result, the usability evaluation metrics for LLM-based educational chatbots included elements that differentiate them from traditional AI-based evaluation metrics.

The proposed usability evaluation metric system is presented as design guidelines that can be utilized in the design and development phases of LLM-based educational chatbot services and provides practical tools to improve education service quality and enhance user experience. The findings of this study suggest the potential for expanding LLM-based technology to various industries and will serve as a valuable reference for future research and practical applications.

부 록

【표 부록-1】 4차 EFA결과: 사용성 카테고리

KMO의 표본 적합도(MSA) m 검정	[redacted]							
Bartlett의 구형성 검정	근사 카이제곱			3066.328				
	자유도			276				
	유의확률			<.001				
요소	성 분							공통성
	1	2	3	4	5	6	7	
32. 몰입성 (Immersion)	0.781	0.225	0.211	-0.024	0.100	0.262	0.049	0.787
33. 학습 동기부여 (Learning Motivation)	0.730	0.197	0.191	0.167	0.232	0.190	0.077	0.732
34. 자기 주도성 (Self Directedness)	0.707	0.175	0.086	0.226	0.330	0.155	0.139	0.741
35. 이해 가능성 (Understandability)	0.656	0.219	0.248	0.209	0.323	-0.272	0.163	0.788
13. 투명성 (Transparency)	0.182	0.817	0.183	0.064	0.222	0.125	0.041	0.805
12. 명료성 (Clarity)	0.495	0.670	0.063	0.198	0.057	0.226	0.091	0.800
19. 신체적 접근성 (Physical Accessibility)	0.059	0.641	0.227	0.321	0.295	0.003	0.135	0.675
14. 맥락 적합성 (Contextual Conformity)	0.314	0.604	0.100	0.339	0.117	0.236	0.149	0.725
15. 최신성 (Up to Dateness)	0.238	0.574	0.035	0.472	0.076	0.222	0.204	0.708
24. 윤리성 (Ethicality)	0.121	0.236	0.812	0.099	0.187	-0.053	0.156	0.802
25. 의인화 (Personification)	0.180	0.074	0.716	-0.039	0.142	0.406	0.083	0.744

23. 오류 관리 (Error Management)	0.199	0.151	0.686	0.299	0.060	0.208	0.292	0.755
22. 프라이버시 보호 (Privacy Protection)	0.244	0.056	0.584	0.231	- 0.079	0.157	0.540	0.779
16. 환각 방지 (Hallucination Prevention)	0.013	0.166	0.111	0.829	0.197	0.180	- 0.040	0.800
17. 가시성 (Visibility)	0.320	0.480	0.135	0.609	0.132	0.085	0.078	0.752
18. 직관성 (Intuitiveness)	0.301	0.384	0.194	0.581	0.209	0.084	0.026	0.664
29. 개인 맞춤성 (Personalization)	0.265	0.161	0.098	0.187	0.755	0.186	0.158	0.769
31. 피드백 (Feedback)	0.374	0.232	0.132	0.194	0.643	0.250	0.051	0.727
26. 적응성 (Adaptiveness)	0.262	0.235	0.159	0.285	0.508	0.397	0.154	0.669
30. 학습 정보 제시성 (Presentation of Learning)	0.414	0.235	0.096	0.092	0.483	0.352	0.135	0.620
27. 공감 (Empathy)	0.228	0.126	0.185	0.228	0.217	0.763	0.035	0.784
28. 인간다운 자연스러움 (Human Naturalness)	0.115	0.305	0.198	0.158	0.362	0.673	0.071	0.760
20. 환경적 접근성 (Environmental Accessibility)	- 0.004	0.215	0.091	- 0.077	0.305	0.054	0.812	0.815
21. 인지적 접근성 (Cognitive Accessibility)	0.213	0.036	0.401	0.079	0.013	0.017	0.762	0.795

【표 부록-2】 5차 EFA결과: 사용성 카테고리

KMO의 표본 적합도(MSA) m 검정			
Bartlett의 구형성 검정	근사 카이제곱	3066.328	
	자유도	276	
	유의확률	<.001	

요소	성 분								공통성
	1	2	3	4	5	6	7	8	
13. 투명성 (Transparency)	0.775	0.140	0.162	0.227	0.016	0.141	0.056	0.293	0.806
12. 명료성 (Clarity)	0.745	0.405	0.113	0.159	0.120	0.175	0.059	-0.029	0.806
14. 맥락 적합성 (Contextual Conformity)	0.689	0.218	0.168	0.225	0.329	0.164	0.106	-0.050	0.750
15. 최신성 (Up to Dateness)	0.656	0.160	0.105	0.157	0.410	0.165	0.171	-0.056	0.719
33. 학습 동기부여 (Learning Motivation)	0.211	0.761	0.170	0.182	0.145	0.275	0.078	0.083	0.795
32. 몰입성 (Immersion)	0.297	0.734	0.238	0.170	-0.078	0.252	0.004	-0.092	0.791
35. 이해 가능성 (Understandability)	0.195	0.702	0.213	0.264	0.205	-0.187	0.167	0.228	0.802
34. 자기 주도성 (Self Directedness)	0.228	0.688	0.105	0.351	0.197	0.175	0.126	-0.024	0.746
23. 오류 관리 (Error Management)	0.216	0.137	0.759	0.156	0.255	0.131	0.180	-0.051	0.783
24. 윤리성 (Ethicality)	0.147	0.175	0.759	0.105	0.099	0.023	0.118	0.383	0.810
25. 의인화 (Personification)	0.090	0.131	0.738	0.207	-0.072	0.364	-0.003	0.023	0.751
22. 프라이버시 보호 (Privacy Protection)	0.124	0.215	0.682	-0.021	0.195	0.103	0.435	-0.122	0.780
29. 개인 맞춤성 (Personalization)	0.131	0.241	0.088	0.743	0.196	0.205	0.186	0.167	0.778

30. 학습 정보 제시성 (Presentation of Learning)	0.347	0.246	0.193	0.695	0.034	0.196	0.071	-0.212	0.792
31. 피드백 (Feedback)	0.246	0.313	0.138	0.692	0.178	0.231	0.051	0.082	0.750
26. 적응성 (Adaptiveness)	0.219	0.264	0.151	0.460	0.280	0.448	0.174	0.134	0.678
16. 환각 방지 (Hallucination Prevention)	0.197	0.038	0.103	0.145	0.822	0.224	-0.039	0.077	0.806
17. 가시성 (Visibility)	0.524	0.307	0.148	0.132	0.568	0.105	0.064	0.089	0.754
18. 직관성 (Intuitiveness)	0.453	0.245	0.225	0.277	0.536	0.048	-0.014	0.018	0.682
27. 공감 (Empathy)	0.164	0.210	0.202	0.206	0.201	0.785	0.024	-0.068	0.816
28. 인간다운 자연스러움 (Human Naturalness)	0.275	0.120	0.179	0.295	0.144	0.736	0.096	0.135	0.799
20. 환경적 접근성 (Environmental Accessibility)	0.135	0.047	0.133	0.191	-0.054	0.117	0.848	0.169	0.839
21. 인지적 접근성 (Cognitive Accessibility)	0.082	0.189	0.517	0.058	0.060	-0.039	0.683	-0.122	0.800
19. 신체적 접근성 (Physical Accessibility)	0.514	0.153	0.139	0.108	0.327	0.166	0.205	0.515	0.760